

APPLIED MATHEMATICS

Liew How Hui

Draft 2025-12-21

Contents

Preface	xviii
Notations	xix
I Logic for Proving and Computing	xxi
1 Propositional Logic	1
1.1 Syntax of Compound Statements	1
1.2 Semantics of Statements and Truth Table	5
1.3 Logical Equivalences and Laws of Logical Equivalences	10
1.4 Applications: Logic Circuits, Solving Logic Puzzles, etc.	19
1.5 Logical Implication and Argument for Statements	22
1.6 Tableaux: Using Diagram to Check Validity	30
1.7 Rules of Inference for Statements	34
2 Predicate Logic	38
2.1 Syntax of Quantified Statements	38
2.2 Formal versus Informal Language	40
2.3 Semantics of Quantified Statements	45
2.4 Logical Implication and Arguments for Quantified Statements (Topic 2b)	51
2.5 Satisfiability Modulo Theories (SMT)	55
2.6 Rules of Inference for Quantified Statements	57
3 Numbers and Methods of Proofs	67
3.1 Notations	67
3.1.1 Natural Numbers	67
3.1.2 Integers: Even, Odd, Prime, Composite	69
3.1.3 Rational Numbers and Real Numbers	72
3.2 Methods of Proof	74
3.2.1 Direct Proof	74
3.2.2 Proof by Contraposition/Contrapositive Proof	78
3.2.3 Proof by Contradiction	79
3.2.4 Classical Theorems: Irrationality of $\sqrt{2}$; Infinitude of Primes	81
3.3 Disproving Statements	84
3.3.1 Disproving Universal Statements with Counterexamples	84
3.3.2 Disproving Existential Statements	86
3.4 Mathematical Induction	87
3.5 Divisibility	90
3.5.1 Properties of Divisibility	91
3.5.2 Greatest Common Divisor	92
3.5.3 Unique Factorisation Theorem & Standard Factor Form	95
3.5.4 Quotient-Remainder Theorem, div and mod	96
3.5.5 Floor and Ceiling	97

3.6	Modular Arithmetic	99
3.6.1	Properties of Congruence	99
3.6.2	Euclidean Algorithm	103
3.6.3	Linear Congruences	105
3.7	Application: SMT For Number System	107
3.8	Application: Cryptography	109
3.8.1	Shift Cipher	110
3.8.2	Affine Cipher	112
3.8.3	Substitution Cipher	112
3.8.4	RSA Cryptosystem	113
3.9	Application: Random Number Generator and Simulation	117
4	Sets and Binary Relations	121
4.1	Sets and Relations	121
4.1.1	Definitions	121
4.1.2	Properties of Relations	124
4.1.3	Application of Sets and Relations	126
4.2	Representing Relations: Boolean Matrix and Digraph	126
4.2.1	(Boolean/Logical) Matrix	126
4.2.2	Digraphs	130
4.3	Closures of (Binary) Relations	134
4.3.1	Reflexive Closure	134
4.3.2	Symmetric Closure	135
4.3.3	Transitive Closure and Warshall's Algorithm	135
4.4	Equivalence Relation	137
4.5	Partial Order Relation	141
4.5.1	Topological Sorting	148
II	Real and Complex Analysis	150
5	Real Number System	151
5.1	Topological Properties of Real Line	151
5.2	Real-Valued Functions	151
5.3	Sequences and Series	153
6	Limits and Continuity	154
6.1	Limit of Functions	154
6.1.1	The limit of a function	154
6.1.2	Some limit laws	157
6.2	Continuity	159
7	Sequences and Series of Functions	164
8	Measures and Integration	165
8.1	Riemann Sum and Definite Integral	165
9	Differentiation	167
9.1	Derivative and Differentiation Rules	167
9.2	Derivative of Some Analytic Functions	167
9.3	Mean Value Theorem?	167
9.4	Antiderivative and Methods of Integration	168
9.4.1	Integration by Parts	170
9.4.2	Integration by Substitution	171
9.4.3	Hyperbolic Functions	174

9.4.4	Integration using Partial Fractions	175
9.5	Definite Integral and Applications	177
9.5.1	Areas, Volumes, Arc Lengths and Area of Surface of Revolution	178
9.5.2	Work Done	181
9.5.3	Centre of Mass	182
9.6	Improper Integrals????	183
9.7	Calculus with Maxima	183
9.8	Summary of Results	184
10	Theory of Interest	186
10.1	Measurement of Interest	186
10.1.1	Accumulation and Amount Functions	186
10.1.2	Effective Rate of Interest	186
10.1.3	Simple Interest	187
10.1.4	Compound Interest	187
10.1.5	Present Value	188
10.1.6	Effective Rate of Discount	188
10.1.7	Nominal Rates of Interest and Discount	189
10.1.8	Force of Interest and Discount	190
10.1.9	Varying Interest	191
10.2	Solutions to Problems in Interest	192
10.2.1	Equations of Value	192
10.2.2	Unknown Time	193
10.2.3	Unknown Rate of Interest	194
10.2.4	Determining Time Periods	194
10.3	Basic Annuities	195
10.3.1	Annuity-immediate, Annuity-due	195
10.3.2	Annuity Values on Any Date	198
10.3.3	Perpetuities	198
10.3.4	Unknown Time	199
10.3.5	Unknown Rate of Interest	201
10.3.6	Varying Interest	202
10.3.7	Extra: Annuities Not Involving Compound Interest	203
10.4	More General Annuities	203
10.4.1	Differing Payment and Interest Conversion Period	204
10.4.2	Annuities Payable More Frequently Than Interest is Convertible	205
10.4.3	Optional: Annuities Payable Less Frequently Than Interest is Convertible	207
10.4.4	Continuous Annuities	208
10.4.5	Increasing and Decreasing Annuities	209
10.4.6	More General Varying Annuities	212
10.4.7	Continuous Varying Annuities	213
10.4.8	Summary	214
10.5	Computations with Calculators and Excel	214
10.5.1	BA II Plus	214
10.5.2	Introduction to Excel Financial Functions	221
10.5.3	Setting loan amortization model on Excel	221
10.5.4	Calculate the value of stock, given the pattern of dividends and yield rates	221
10.5.5	Calculate net present value, given the investment contributions and investment returns	221

11 Complex Numbers	223
11.1 Algebraic Structure of $\mathbb{C}$	223
11.1.1 Algebraic Operations of $\mathbb{C}$ and Their Properties	224
11.1.2 The i Notation	226
11.1.3 Matrix Representation	229
11.2 Geometric Interpretation and Triangle Inequality	229
11.3 Polar Form and Roots	231
11.4 Topological Structure of $\mathbb{C}$	237
11.5 Basic Geometrical Objects in Complex Plane	239
11.6 Effective Periods	241
12 Complex Functions	242
12.1 Complex-Valued Functions	242
12.2 Limits and Continuity	245
12.2.1 Limit at $z_0 \in \mathbb{C}$	245
12.2.2 Limit at ∞	247
12.2.3 Continuity	248
12.3 Holomorphic Functions and C-R Equations	249
12.3.1 Derivatives	250
12.3.2 Holomorphic Functions	252
12.3.3 Cauchy-Riemann (C-R) Equations	254
12.3.4 Harmonic Functions	258
12.3.5 Cauchy-Riemann Equations in Polar Coordinates	259
12.4 Elementary Functions	260
12.4.1 Complex Exponential Function	260
12.4.2 Complex Trigonometric Functions	263
12.4.3 Hyperbolic Functions	265
12.5 Inverse Mapping and Multi-Valued Functions	267
12.5.1 n th Root Function	267
12.5.2 Single-Valued vs Multi-Valued Functions	269
12.5.3 Complex Logarithmic Function	270
12.5.4 Inverse Trigonometric and Hyperbolic Functions	273
12.5.5 Arbitrary Power Function	273
12.6 Extended Complex Plane and Riemann Sphere	274
13 Further Topics in Geometry	275
13.1 Translations	275
13.2 Magnification and Rotation	276
13.3 Linear Mapping	277
13.4 $w = \frac{1}{z}$ Transformation	279
13.5 Möbius Transformation on $\hat{\mathbb{C}}$	282
14 Complex Integration	287
14.1 Arcs, Curves and Contours	287
14.2 Integrals and Contour Integrals	289
14.3 Anti-Derivatives	293
14.4 Cauchy Theorems	295
14.4.1 Simply and Multiply Connected Domains	295
14.4.2 Cauchy Theorem	296
14.4.3 Cauchy-Goursat Theorem	296
14.5 General Cauchy Integral Formula	297
14.5.1 Cauchy Integral Formula on a Disk	297
14.5.2 Cauchy Integral Formula for Simply Connected Domain	298
14.5.3 Cauchy Integral Formula for Multiply Connected Domain	300

14.6	Maximum Modulus Principle	304
14.7	Liouville's Theorem	307
15	Analyticity	309
15.1	Number and Function Sequence	309
15.2	Number and Function Series	311
15.3	Power Series and Taylor Series	314
15.4	Holomorphicity and Analyticity	320
15.5	Isolated Zeroes Principle	322
15.6	Analytic Continuation	323
16	Residues and Poles	326
16.1	Singularities of a Function	326
16.2	Laurent Series	328
16.3	Characterisation of Zero and Pole	334
16.4	Residue Theorem and Its Computation	338
16.5	Examples of Applications	341
16.5.1	Evaluation of Definite Integrals Involving Sine and Cosine	344
16.5.2	Evaluation of Improper Integral of Rational Function on $\mathbb{R}$	347
16.5.3	Evaluation of Improper Integrals from Fourier Analysis	353
16.5.4	Evaluating Integrals with "Multi-Valued" Functions	357
16.6	Argument Principle and Rouché Theorem	359
17	Harmonic Function	361
17.1	$\frac{\partial}{\partial z}$ and $\frac{\partial}{\partial \bar{z}}$	361
17.2	Harmonic Conjugate	362
17.2.1	Relation with Holomorphic Functions	362
17.2.2	Finding Harmonic Conjugate	363
17.3	Mean Value Property	365
17.4	Application to Simple Dirichlet Problem	368
18	Conformal Mapping	371
18.1	Local Properties of Holomorphic Maps	371
18.1.1	Preservation of Angles	371
18.1.2	Local Amplification	373
18.1.3	Local Inverse	374
18.1.4	Local Property at Critical Points	374
18.2	Elementary Conformal Mapping	375
18.2.1	Conformal Mapping	375
18.2.2	Conformal Mappings between $\mathbb{H}$ and Δ	376
18.2.3	Conformal Mappings for Generalised Rectangles and Circular Shape Objects	379
18.3	Riemann Mapping Theorem	381
18.4	Numerical Conformal Mapping	381
19	Applications	384
19.1	Integral Transforms	384
19.1.1	Fourier Transform	384
19.1.2	Laplace Transform	385
19.1.3	Complex Numbers in AC Circuit Theory	387
19.1.4	Classical Control Theory	388
19.2	Application to Potential Theory	389
19.2.1	General Dirichlet Problem	389
19.2.2	Transformations of Boundary Conditions	390
19.2.3	Poisson's Integral Formula	392
19.2.4	Heat Problems	393

19.2.5	Electrostatic Field	394
19.2.6	Fluid Flow	396
III	Linear Algebra	398
20	Vectors	399
20.1	Two-Dimensional Vectors	399
20.2	Three-Dimensional Vectors	402
20.3	Length and Angle	405
20.4	Lines, Cross Product and Planes	407
20.5	Projections	411
20.6	Euclidean n -Space	413
21	Matrices and Linear Equations	417
21.1	Algebra of Matrices	417
21.2	The Inverse of a Matrix	422
21.3	Systems of Linear Equations and Linear Independence	424
21.4	Special Matrices and Row Echelon Form	428
21.5	LU Factorisation and Its Variants	432
21.6	Applications of Gaussian Elimination	438
21.7	Working with Maxima	442
22	Determinants, Eigenvalues, Eigenvectors	443
22.1	The Determinant of a Matrix	443
22.2	Properties of Determinants	447
22.3	Geometric Applications of Determinants	449
22.4	Eigenvalues and Eigenvectors	452
22.5	Similarity and Diagonalisation	460
22.6	Jordan Canonical Form	465
22.7	Working with Maxima	466
23	Vector Spaces	467
23.1	The Theory of Linear Equations	467
23.2	Vector Spaces and Subspaces	471
23.3	Linear Independence	477
23.4	Basis and Dimension, Change of Basis	480
24	Linear Transformations	486
24.1	Linear Transformations	486
24.2	Kernel and Range	489
24.3	Matrix Representation of Linear Transformation	492
24.4	Changing Coordinates	498
25	Inner Product Spaces and Orthogonality	500
25.1	Inner Product and Norm	500
25.2	Least Square Approximations and Projection	500
25.3	Orthogonality in $\mathbb{R}^n$	504
25.4	The Gram-Schmidt Process and the QR Factorisation	506
25.5	Orthogonal Diagonalisation of Symmetric Matrices	510
IV	Multivariable Calculus	513
26	Vector Calculus	514
26.1	Euclidean Space	514

26.1.1	Special Case $n = 3$: $\mathbb{R}^3$	515
26.2	Parametric Curves	516
26.3	Multivariable Functions	517
26.3.1	Graphing two-variable functions in $\mathbb{R}^3$	518
26.3.2	Linear function	519
26.3.3	Visualizing and sketching graphs of two-variable (nonlinear) functions	521
26.3.4	Graphs of some common non-linear functions	523
26.4	Differentiation of multivariable functions	526
26.4.1	Partial derivative	526
26.4.2	Total Differential	528
26.4.3	Chain rule and tree diagram	529
26.4.4	Implicit partial differentiation	531
27	Vector Field	532
27.1	Gradient, Divergence and Curl	532
27.2	Formulae involving ∇	533
27.3	Grad, Div, Laplacian, and Curl in Cartesian and Curvilinear Coordinate Systems	535
27.4	Vector Fields in Cylindrical and Spherical Coordinate Systems	535
27.4.1	Transformation of Coordinates	535
27.4.2	Orthogonal Curvilinear Coordinates	536
27.4.3	Unit Vectors in Curvilinear Systems	536
27.4.4	Arc Length and Volume Elements	536
27.4.5	Special Orthogonal Coordinate Systems	536
27.4.6	The gradient vector	537
27.4.7	Directional derivative	539
27.5	Unconstrained optimization:Local extrema and saddle point	540
27.5.1	The second derivative test for two-variable functions	541
27.6	Constrained optimization:The Lagrange multipliers method	542
28	Multiple Integral	546
28.1	Double Integrals	546
28.1.1	Double Integrals over Rectangular Regions	546
28.1.2	Double Integrals over Non-rectangular Regions	547
28.1.3	Double Integrals in Polar Coordinates	548
28.2	Triple Integrals and Volume Integrals	548
28.2.1	Triple Integrals in A Rectangular Coordinates	549
28.2.2	Triple Integrals in Cylindrical and Spherical Coordinates	549
28.2.3	Volume (Triple) Integrals	551
28.3	Line Integrals	555
28.3.1	Line Integrals of Vector Fields	558
28.4	Surface Integrals	560
28.4.1	Parametric Surfaces	560
28.4.2	Surface Area and Surface Integrals	561
28.5	Green's Theorem, Stokes' Theorem and Divergence Theorem	564
28.5.1	Green's Theorem in the Plane	565
28.5.2	Stokes' Theorem	566
28.5.3	Gauss' Divergence Theorem	567
28.6	Potential Functions, Path Independence, Conservative Fields	568
28.6.1	Path Independence and Conservative Fields	568

V	Differential Equations	572
29	First Order Equations	573
29.1	Definitions	573
29.2	Separable Equations	575
29.3	Linear Equations and Integrating Factors	577
29.4	Exact Equations and Integrating Factors	579
29.5	Change of Variables (or Symmetry)	582
29.6	Picard's Existence and Uniqueness Theorem	585
29.7	Formulation and Solution of Application Specific Problems	587
30	Second Order Equations	591
30.1	Linear Equations and Fundamental Solutions	591
30.2	Homogeneous Equations with Constant Coefficients	596
30.3	Non-Homogeneous Equations with Constant Coefficients & Method of Undetermined Coefficients	599
30.4	Homogeneous Linear Equation & Reduction of Order	601
30.5	Non-Homogeneous Linear Equation and Variation of Parameters	605
30.6	Mechanical and Electrical Vibrations	608
30.6.1	Free Vibration	609
30.6.2	Damped Free Vibration	610
30.6.3	Forced Free Vibration	610
30.6.4	Damped Forced Vibration	611
30.7	Working with Maxima	611
31	Series Solutions of Linear Equations	612
31.1	Review of Power Series	613
31.2	Ordinary Points and Singular Points	616
31.3	Series Solutions Near an Ordinary Point	618
31.4	Series Solutions Near a Regular Singular Point	624
31.5	Comments on Convergence & Irregular Singular Points	632
32	Laplace Transform	634
32.1	Definitions	634
32.2	Rules of Laplace Transform	636
32.3	Inverse Laplace Transform	638
32.4	Applications in IVP of Solving Ordinary (Integro-)Differential Equation	641
32.5	Transfer Function and Convolution	644
32.6	System Responses	648
32.6.1	Impulse Function and Impulse Response	649
32.6.2	Step Function and Step Response	650
32.6.3	Frequency Response	652
32.7	Formulation and Solution of Application-Specific Problems	653
32.8	Working with Maxima	658
33	Systems of ODEs	659
33.1	Development of Solution Methods	659
33.2	System of First Order ODEs and Phase Space	660
33.2.1	Critical Point	665
33.2.2	(Lyapunov) Stability of Critical Point	666
33.3	Homogeneous Linear System with Constant Coefficients	667
33.3.1	Different Real Eigenvalues	667
33.3.2	Some Complex Eigenvalues	670
33.3.3	Repeated Eigenvalues	671
33.4	Non-Homogeneous Linear System with Constant Coefficients	676

33.5	Fundamental Matrix and Transition Matrix	679
33.5.1	Matrix Functions	680
33.5.2	Fundamental Matrices	680
33.5.3	Transition Matrix	683
33.6	Non-Homogeneous Linear Systems	684
33.6.1	Theory of Linear Systems	685
33.6.2	Variation Of Parameters	686
33.6.3	Matrix Exponential Solutions	688
33.6.4	Undetermined Coefficients	690
33.7	Application-Specific Problems	691
33.8	Introduction to Nonlinear Systems	694
33.9	One-Dimensional Phase Space Analysis	694
33.10	Two-Dimensional Phase Space Analysis	696
33.10.1	Classification of Integral Curve Portrait of Autonomous Constant Linear System	696
33.10.2	Classification of Phase Portrait of Autonomous Constant Linear System	699
33.10.3	Classification of Phase Portrait of Almost Linear Systems	701
33.11	Three- and Higher-Dimensional Phase Space Analysis	708
33.12	Competing Species and Predator-Prey Equation	709
34	Introduction to PDE and Fourier Series	711
34.1	Introduction to PDE	711
34.1.1	Definitions	711
34.1.2	Notations and Examples	712
34.1.3	How to Solve PDEs?	714
34.2	Fourier Series	714
34.2.1	(Trigonometric) Fourier Series	714
34.2.2	Fourier Series On the Half Interval	721
34.2.3	Pointwise Convergence of Fourier Series	723
34.2.4	Uniform Convergence and Gibbs Phenomena	724
34.2.5	Parseval's Theorem and Mean Square Error	725
34.2.6	Complex Form of Fourier Series	728
34.2.7	Application: Isoperimetric Theorem	735
34.3	Multiple Fourier Series	736
34.4	Sturm-Liouville (Eigenvalue) Problems	738
34.5	Hilbert Space and Generalised Fourier Series	742
35	Boundary-Value Problems in Rectangular Coordinates	747
35.1	First Order Quasi-Linear PDEs	747
35.1.1	Change of Variables and Method of Characteristic	748
35.1.2	Constant Coefficient and Homogeneous Examples	749
35.1.3	Variable Coefficient or Non-Homogeneous Examples	750
35.2	Second Order Linear PDEs	752
35.2.1	Hadamard's Classification	752
35.2.2	Heat or Diffusion Equation	756
35.2.3	Laplace Equation and Poisson Equation	759
35.2.4	Wave Equation	759
35.2.5	Domains, Boundary Conditions and Initial Conditions	760
35.3	1-D Heat Equation	762
35.3.1	(Homogeneous) Type 1 BCs	762
35.3.2	(Homogeneous) Type 2 BCs	764
35.3.3	(Homogeneous) Type 3 BCs	767
35.3.4	(Homogeneous) Mixed BCs	768
35.3.5	Nonhomogeneous PDE and Homogeneous BCs	771

35.3.6	Homogeneous PDE and Nonhomogeneous BC	773
35.3.7	Nonhomogeneous PDE and BCs	776
35.3.8	“General” Heat Equation Problem	777
35.4	2-D and 3-D Heat Equations	778
35.5	2-D and 3-D Laplace Equation	781
35.6	Poisson Equation	785
35.7	Vibrating String — 1-D Wave Equation	786
35.7.1	(Homogeneous) Type 1 BCs	787
35.7.2	(Homogeneous) Type 2 BCs	790
35.7.3	(Homogeneous) Type 3 BCs	790
35.7.4	(Homogeneous) Mixed BCs	791
35.7.5	Nonhomogeneous PDE and Homogeneous BCs	792
35.7.6	Homogeneous PDE and Nonhomogeneous BCs	793
35.7.7	Nonhomogeneous PDE and Nonhomogeneous BCs	795
35.8	2-D and 3-D Wave Equations	797
36	Boundary-Value Problems in Cylindrical Coordinates	804
36.1	Second Order Linear PDE in Cylindrical Coordinates	804
36.1.1	Domains, Boundary Conditions and Initial Conditions	804
36.1.2	Laplacian in Cylindrical Coordinates	805
36.1.3	Heat Equations	806
36.1.4	Laplace and Poisson Equations	806
36.1.5	Wave Equations — Vibrating Drumhead	807
36.2	Special Functions	807
36.2.1	Bessel Equations and Bessel Functions	807
36.2.2	Orthogonal Property of Bessel Functions	810
36.2.3	Modified Bessel Functions	813
36.3	1-D Heat Flow (in Infinite Cylinder)	814
36.3.1	(Homogeneous) Type 1 BC	814
36.3.2	(Homogeneous) Type 2 BC	818
36.3.3	(Homogeneous) Type 3 BC	819
36.4	2-D and 3-D Heat Equation	821
36.4.1	In Finite Cylinder	821
36.4.2	In Infinite Cylinder	823
36.5	2-D and 3-D Laplace Equations	826
36.5.1	“Inner” Disk, “Outer” Disk and “Ring”	827
36.5.2	“Wedge” and “Sector”	832
36.5.3	Inner “Tube”	833
36.6	Poisson Equation	836
36.7	1-D Wave Equation	837
36.8	2-D and 3-D Wave Equations	839
37	Boundary-Value Problems in Spherical Coordinates	843
37.1	Second Order Linear PDEs in Spherical Coordinates	843
37.1.1	Domains, Boundary Conditions and Initial Conditions	843
37.1.2	Laplacian in Spherical Coordinates	844
37.1.3	Heat or Diffusion Equation	844
37.1.4	Laplace Equation and Poisson Equation	845
37.1.5	Wave Equation	845
37.2	Special Functions	845
37.2.1	Spherical Bessel Equation	845
37.2.2	Legendre Polynomials and Associated Legendre Functions	847
37.2.3	Orthogonal Property of Legendre Polynomials	851
37.2.4	Orthogonal Property of Associated Legendre Functions	853

37.2.5	Spherical Harmonics	854
37.3	1-D Heat Equation (on a Sphere with Homogeneous Distribution)	856
37.3.1	(Homogeneous) Type 1 BC	857
37.3.2	(Homogeneous) Type 3 BC	860
37.3.3	Nonhomogeneous PDE and Homogeneous BC	861
37.4	2-D and 3-D Heat Equations (on a Sphere with Non-Homogeneous Distribution)	862
37.5	2-D and 3-D Laplace Equations	864
37.5.1	BVP in a Sphere	864
37.5.2	BVP exterior to a Sphere	868
37.5.3	BVP on a Hemisphere	872
37.6	1-D Wave Equation	873
37.7	2-D and 3-D Wave Equations	874
38	Fourier Transforms and Applications	876
38.1	Fourier Transform Basics	876
38.2	Properties of Fourier Transform when $n = 1$	879
38.3	Fourier Transform for Generalised Functions	884
38.4	Heat Equation	887
38.5	Laplace Equation	893
38.6	Wave Equation	898
VI	Numerical Methods	905
39	Computer Algebra with Reduce	906
40	Introduction to Numerical Methods	911
40.1	Overview	911
40.2	What is a Numerical Algorithm?	912
40.3	Numeric Data Structures	913
40.3.1	IEEE Floating Point Representation	914
40.3.2	Floating Point Arithmetic	917
40.3.3	Floating-Point Mathematical Functions	919
40.4	Programming Environment	922
40.4.1	Assembly Language	922
40.4.2	C11 and C++11	925
40.5	Numerical Functions in Software	927
41	Numerical Linear Algebra	930
41.1	C++ Armadillo Library and OpenBLAS	931
41.2	Array and I/O	931
41.3	Array Arithmetic and Norms	935
41.3.1	Vector Norms	937
41.3.2	Matrix Norms	938
41.3.3	BLAS	940
41.4	Direct Solution of Linear Systems	943
41.4.1	LAPACK	943
41.4.2	Naming Scheme	943
41.4.3	Naïve Gaussian Elimination	946
41.4.4	Pivoting Strategies	948
41.4.5	Matrix Inversion	954
41.4.6	Application in Statistics	954
41.4.7	Applications in Electrical Engineering	956
41.5	Matrix Factorisation	957
41.5.1	LU Decomposition	958

41.5.2	PLU Decomposition	959
41.5.3	Cholesky Decomposition	962
41.5.4	QR Decomposition	965
41.5.5	SVD Decomposition	973
41.6	Iteration Methods	977
41.6.1	Iteration Process and Convergence Rate	977
41.6.2	Convergence Acceleration when $n = 1$	978
41.6.3	Sparse Matrix	981
41.6.4	Jacobi Method	982
41.6.5	Gauss-Seidel Method	986
41.6.6	Successive Over-Relaxation (SOR) Method	988
41.6.7	Conjugate Gradient (CG) Methods	992
41.6.8	Fortran Module and ITPACK	995
41.7	Eigenvalue Problems	998
41.7.1	Power Methods	999
41.7.2	Google PageRank and Markov Chain	1006
41.7.3	QR Algorithm	1011
42	Interpolation & Curve Fitting	1021
42.1	Working with Polynomials	1022
42.2	(Lagrange) Interpolation	1024
42.2.1	Natural Form	1024
42.2.2	Lagrange Form	1026
42.2.3	Application of Lagrange Form	1029
42.2.4	Newton Form and Newton Forward Divided-Differences	1031
42.2.5	Newton Backward Divided Differences	1035
42.2.6	Runge Phenomenon	1036
42.3	Hermite Interpolation	1038
42.4	Piecewise Polynomial Interpolation & Univariate Spline	1041
42.4.1	Piecewise Linear Interpolation & Linear Spline	1042
42.4.2	Scientific Data Visualisation	1043
42.4.3	Piecewise Cubic Interpolation & Cubic Spline	1045
42.5	Application: Static Term Structure Models	1055
42.6	Parametric Curve & Bézier Curve	1056
42.7	Application: Computer Graphics and CAD	1063
43	Nonlinear Systems	1065
43.1	The Bisection Method	1066
43.1.1	Properties of Continuous Functions	1066
43.1.2	Implementation	1066
43.2	Newton(-Raphson) Method & Secant Method	1070
43.2.1	Properties of Differentiable Functions & Taylor's Series	1070
43.2.2	Newton's Method and Implementation	1072
43.2.3	Secant Method and Implementation	1078
43.3	Finite Difference Approximation	1079
43.3.1	Finite Difference	1080
43.3.2	Richardson's Extrapolation	1082
43.4	System of Equations	1085
43.4.1	Bisection Methods	1086
43.4.2	Newton's Method	1087
43.4.3	Approximation of Jacobian and Newton's Method of n -Dimension	1091
43.4.4	Variations of Newton's Method and SUNDIALS' KINSOL	1092
43.5	Fixed-Point Iteration	1097
43.5.1	Existence Theorem	1097

43.5.2	Convergence Acceleration	1099
43.5.3	Fortran Module	1099
43.6	Optimisation and Machine Learning	1104
43.6.1	Machine Learning	1105
43.6.2	ML Software Libraries and A Little Bit About Training	1106
44	Numerical Integration	1111
44.1	Evenly-Spaced Quadrature	1112
44.1.1	Basic Quadrature Rules	1112
44.1.2	Composite Rules	1114
44.1.3	Romberg Method	1121
44.2	Gaussian Quadrature	1127
44.3	Adaptive Quadrature	1137
44.4	Quadrature Software	1140
44.5	Hard Integration Problems	1142
45	Non-stiff ODE-IVPs	1144
45.1	IVP Theory and Examples	1145
45.2	Taylor Methods	1149
45.2.1	Euler Method	1149
45.2.2	Higher-Order Taylor Method	1155
45.3	Runge-Kutta Methods	1158
45.3.1	Runge-Kutta Methods of Order Two	1159
45.3.2	Runge-Kutta Method of Order Three	1164
45.3.3	Runge-Kutta Methods of Order Four	1165
45.4	Error Control and Runge-Kutta-Fehlberg Method	1172
45.5	Multi-Step Formulae	1176
45.6	Predictor-Corrector Methods	1178
45.7	Convergence, Stability and Stiffness	1180
46	ODE-BVPs	1185
46.1	Existence and Uniqueness Theorem	1185
46.2	Linear Shooting Method	1187
46.3	Nonlinear (Single) Shooting Method	1192
46.4	Linear Finite Difference Method	1195
46.5	Nonlinear Finite Difference Method	1200
46.6	Finite Element Method (FEM)	1202
47	Linear Partial Differential Equations	1206
47.1	Overview	1207
47.2	Elliptic Problems	1213
47.2.1	Finite Difference Approximation	1214
47.2.2	Finite Difference Method on Rectangular Grid	1215
47.3	Parabolic Problems	1222
47.3.1	FTCS (Explicit) Scheme	1223
47.3.2	BTCS (Implicit) Scheme	1228
47.3.3	Crank-Nicolson (Implicit) Method	1230
47.3.4	Method of Lines	1234
47.3.5	Application: Financial Mathematics	1235
47.4	Hyperbolic Problems	1238
47.4.1	Finite Difference Method on Rectangular Grid	1239
47.4.2	Method of Lines	1243
47.5	Open Source PDE Software	1245

VII	Signal Analysis and Control	1252
48	Signal and Filter	1253
48.1	Systems and Linear Transforms	1253
48.2	Fourier Integrals	1255
48.3	Fourier Transforms	1256
48.4	Step Function and Piecewise Continuous Function	1257
48.5	Properties of Fourier Transform	1258
48.6	Impulse Function and Generalised Fourier Transforms	1260
48.7	Plancherel's Theorem	1261
48.8	Application in Solving Ordinary Differential Equations	1262
48.9	DFT	1263
48.10	Laplace Transform	1264
49	Linear Difference Equations & Z-Transform	1265
49.1	Definition and Examples	1266
49.2	Properties of Z-Transform	1267
49.3	Inverse z -Transform	1268
49.4	Applications in Solving Difference Equation	1269
49.5	Transfer Function and State-Space Model	1271
50	Linear Difference System & State Space Method	1274
VIII	Probability and Statistics	1275
51	Probability Theory and Statistical Modelling	1276
51.1	Population Space, Random Variables and Data	1276
51.2	General Probability Theory	1277
51.3	Discrete Distributions	1277
51.4	Continuous Distributions	1277
51.5	Joint Distributions	1277
51.6	Distributions Functions of Random Variables	1278
51.7	Inferences on the Mean of a Distribution	1278
51.8	Inference on Proportions	1278
51.9	Comparing Two Means and Two Proportions	1278
51.10	The Analysis of Variance	1278
51.11	Chi-Square Tests	1279
51.12	Probability Models	1279
51.13	Convergence	1279
51.14	Statistical Models and Inferences	1279
51.15	Criteria for Evaluating Point Estimators	1279
51.16	Tests of Hypotheses	1279
51.17	Interval Estimation	1279
51.18	Information and Conditioning	1280
51.19	General Conditional Expectation	1280
51.20	Single period multi-state model	1280
52	Statistical Decisions	1281
52.1	Chapter 1 Probability — Preparations	1281
52.1.1	1.1 Betting Problems	1281
52.1.2	1.2 Probability models	1282
52.1.3	1.3 Conditional Probability and Independent Events	1285
52.1.4	1.4 Partition Theorem and Bayes's Formula	1287
52.1.5	1.5 Random Variables	1287

52.1.6	1.7 Continuous Random Variables	1290
52.1.7	1.8 Expected Values of Continuous Random Variables	1290
52.1.8	1.9 Some Special Distribution	1291
52.1.9	1.11 Joint and Marginal Distributions	1295
52.1.10	1.13 Independence of Random Variables	1295
52.1.11	1.14 Conditional distributions	1296
52.1.12	1.15 Properties of Expected Values	1296
52.2	Chapter 2 Decision Problems	1298
52.2.1	2.1 Introduction	1298
52.2.2	2.2 Some Examples	1300
52.2.3	2.3 Decision	1302
52.2.4	2.4 Formulating Decision Problems	1302
52.3	Chapter 3 Utility	1304
52.3.1	3.2 Preference	1305
52.3.2	3.3 Preference Axioms	1305
52.3.3	3.4 Coding Intermediate Prospects	1306
52.3.4	3.5 Properties of Utility Function	1307
52.3.5	3.6 Utility for Arbitrary Prospects	1308
52.4	No-Data Problem	1309
52.5	Topic 3: Using Data in Decisions	1309
53	Applied Statistical Models	1310
53.1	Multiple Linear Regression	1310
53.2	Model Adequacy Checking	1310
53.3	Transformations and Weighting to Correct Model Inadequacies	1311
53.4	Exponential Smoothing Methods	1311
53.5	Nonlinear Regression	1311
53.6	MA, AR, ARMA, the Box-Jenkins Methodology for ARIMA Models and the Autocorrelation Function	1311
54	Design and Analysis of Experiment	1313
54.1	Introduction	1313
54.2	Experiments with single factor	1313
54.3	Randomized blocks, Latin squares and related designs	1313
54.4	Introduction to factorial designs	1313
54.5	The 2^k factorial designs	1314
55	Statistical Simulation	1315
55.1	Monte Carlo Approach	1315
55.2	Random Number Generation	1315
55.3	Variance Reduction Techniques	1315
55.4	Simulation of Stochastic Processes	1316
55.5	Discrete Event Simulation	1316
55.6	Applications in Finance and Insurance	1316
56	Applications in Statistics and Econometrics	1317
56.1	Non-negative Matrices	1317
56.1.1	Theory	1317
56.1.2	Leontief Models in Economics	1321
56.1.3	Markov Chains	1323
56.1.4	Genetic Models	1324
56.1.5	Population Growth Models (Demography)	1327
56.2	Linear Models	1328
56.2.1	Some Useful Lemmas	1329
56.2.2	Estimation in a Linear Model	1329

56.2.3	Estimation of Variance	1330
56.2.4	Method of MINQUE	1330
56.3	Econometrics	1332
56.3.1	Classical Approach: Linear Regression with OLS & GLS	1333
56.3.2	Time-Series Analysis	1335
56.3.3	Qualitative Response (QR) Models	1335
56.3.4	Software: gretl (GUI) and R (CLI)	1336
56.4	Real-Life Statistics	1336
57	Methods for Quality Improvements	1337
57.1	Introduction	1338
57.1.1	Dimensions of quality	1338
57.1.2	Three aspects of quality	1339
57.1.3	Quality characteristics	1339
57.1.4	Types of variation in processes	1340
57.2	Statistical Process Control	1341
57.2.1	The Magnificent Seven	1341
57.2.2	Some Other Methods	1344
57.3	Introduction to ISO	1348
58	Applied Nonparametric Statistics	1353
58.1	Prerequisites	1353
58.2	Procedures that Utilize Data from a Single Sample	1353
58.3	Procedures that Utilize Data from Two Independent Samples	1353
58.4	Procedures that Utilize Data from Two Related Samples	1353
58.5	Procedures that Utilize Data from Three or More Independent Samples	1354
58.6	Procedures that Utilize Data from Three or More Related Samples	1354
58.7	Goodness-of-Fit Tests	1354
58.8	Rank Correlation And Other Measures of Association	1354
59	Stochastic Processes	1355
59.1	Introduction	1355
59.2	Introduction to Markov Chain	1355
59.3	Poisson Processes	1355
59.4	Continuous-Time Markov Chain	1356
59.5	Queuing Theory	1356
IX	Operations Research & Optimisation	1357
60	Optimisation	1358
60.1	Numerical Optimisation Problems and Direct Search Methods	1358
60.2	Golden Section Search	1358
60.3	Simple Gradient Methods	1359
60.4	Other topics	1360
61	Operations Research	1361
61.1	Linear Programming	1361
61.1.1	The Simplex Algorithm and Goal Programming	1361
61.1.2	Sensitivity Analysis and Duality	1362
61.2	Transportation, Assignment and Transshipment Problems	1362
61.3	Integer Programming	1362
61.4	Game Theory	1362
61.5	Inventory Models	1362
61.6	Nonlinear Programming	1362

61.7 Queuing Theory	1363
61.8 Decision Analysis	1363
61.9 Dynamic Programming	1363
61.10Network Models	1363

Preface

This is a compilation of the lecture notes the author used in his undergraduate teaching. The combined lecture notes are posted in the hope that those who are really fond of mathematics, whether they have the chance to study in a university or not, can use it to learn what a part of applied mathematics is all about.

Various programming languages are used in the past such as C, C++, Fortran 90, Python, Julia. In future, these notes will update all the programming examples to Chez Scheme (r6rs), Common Lisp and C++ since they are portable and works on various POSIX compliant operating systems such as GNU/Linux, NetBSD, Haiku and OpenBSD. Common Lisp based computer algebra systems (MAXIMA and REDUCE) will also be used in linear algebra and calculus.

Note that Lisp is a class of programming languages that fits well with mathematical thinking based on lambda calculus and is used in the development of older generation of computer algebra systems such as MAXIMA, REDUCE and AXIOM while C++ is still one of the most efficient compiled programming language.

Some of the notes are arranged by based on the numerical methods and ordinary differential equations by Dr Goh Yong Kheng, the complex analysis and statistical decision by Dr Wong Wing Yue. the notes on multivariate calculus by Dr Lem Kok Hong is yet to be updated. So if there are any mistakes, the mistakes are mine and if you spot any mistakes, please let me know.

Notations

Logic Notations

$\therefore$	denotes	“since”, “because of”.
$\therefore$	denotes	“therefore”, “thus”, “hence”.
$\exists$	denotes	“there exists”, “there is a/an”, “there are some”.
$\nexists$	denotes	“there does not exists”.
$\forall$	denotes	“for all”, “for every”, “for each”.
$\in$	denotes	“belongs to”.
$\notin$	denotes	“does not belongs to”.
iff	denotes	“if and only if”.
s.t.	denotes	“such that”.

Set Notations

Set is a collection of objects, called the elements of the set. We write $a \in X$ to mean that a is an element of the set X . A set without any element is called *null set* or *empty set*. It is denoted by $\{\}$ or $\emptyset$. *Nonempty* set has at least one element.

Sets can be categorised into *finite sets* and *infinite sets*. Finite sets has finite number of elements and can be abstractly written as

$$X = \{a_1, \dots, a_n\}.$$

This is a set with n elements. For example, $X = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. It is a finite set with $n = 10$ elements, and $a_i = i - 1$, $i = 1, \dots, 10$.

Infinite sets are sets with infinite elements, important examples are

The set of natural numbers	$\mathbb{N} = \{0, 1, 2, 3, \dots\}$.
The set of positive natural numbers	$\mathbb{N}^* = \{1, 2, 3, \dots\}$.
The set of integers	$\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$.
The set of rational numbers	$\mathbb{Q} = \{\frac{p}{q} \mid p, q \in \mathbb{Z}\}$.
The set of all real numbers	$\mathbb{R} = \{n_1 \cdots n_r \cdot d_1 d_2 d_3 \cdots \mid n_i, d_j \in X\}$.
The set of all non-zero real numbers	$\mathbb{R}^* = \mathbb{R} \setminus \{0\}$.
The set of all complex numbers	$\mathbb{C} = \{a + bi \mid a, b \in \mathbb{R}, i = \sqrt{-1}\}$.
The set of all non-zero complex numbers	$\mathbb{C}^* = \mathbb{C} \setminus \{0\}$.
The set of n -tuples of real numbers	$\mathbb{R}^n = \{(x_1, x_2, \dots, x_n) \mid x_i \in \mathbb{R}, i = 1, 2, \dots, n\}$.

Various of the above number sets have relation as follows:

$$\mathbb{N}^* \subset \mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}.$$

For real number set, it's subset are commonly used in calculus, the following is the notation that will be used:

$[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$	denotes	a closed interval.
$(a, b) = \{x \in \mathbb{R} \mid a < x < b\}$	denotes	an open interval.
$[a, b) = \{x \in \mathbb{R} \mid a \leq x < b\}$	denotes	an open-closed interval.
$(a, b] = \{x \in \mathbb{R} \mid a < x \leq b\}$	denotes	an closed-open interval.

Note that in some french books $]a, b[$ is used to denote open interval. This makes it possible to distinguish a point on a plane (x, y) from an interval $]a, b[$. However, this is not commonly used in Malaysia, we will stick to (a, b) to represent an open interval.

Vector Notations

1. (a) $\mathbf{a}, \mathbf{b}, \mathbf{c}, \dots, \mathbf{u}, \mathbf{v}, \mathbf{w}, \dots$ denote plane/space vectors.
 (b) $a, b, c, \dots, u, v, w, \dots$ denote scalars
2. $\mathbf{i}, \mathbf{j}, \mathbf{k}$ denote the coordinates direction unit vectors. So

$$\mathbf{i} = (1, 0, 0), \quad \mathbf{j} = (0, 1, 0), \quad \mathbf{k} = (0, 0, 1).$$

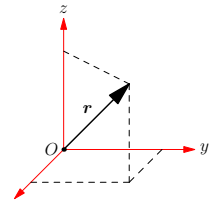
3. A vector $\mathbf{a}$ with components a_1, a_2, a_3 is written as

$$\mathbf{a} = (a_1, a_2, a_3) \quad \text{or} \quad \mathbf{a} = a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}.$$

4. $\mathbf{a}, \mathbf{b}, \mathbf{c}, \dots, \mathbf{x}, \mathbf{y}, \mathbf{z}, \dots$ denotes vectzors in space
5. The position vector of a point $P(x, y, z)$ is denoted by

$$\mathbf{r} = \overrightarrow{OP} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}.$$

6. $f, g, h, \dots$ denote scalar functions.
7. $\mathbf{F}, \mathbf{G}, \mathbf{H}, \dots$ denote vector fields.



Part I

Logic for Proving and Computing

Chapter 1

Propositional Logic

Discrete mathematics is a subject to prepare students to learn the ideas behind mathematical proving, interpreter/compiler construction and formal specification and verification. Mathematical proving is difficult because we are dealing with the characterisation of infinite objects.

Interpreter/compiler construction involves formal syntax, operational semantics, denotational semantics, graph algorithms which are beyond year 1 subject. Formal specification and verification formalises informal problems and uses an interpreter or a compiler to check if the problem has a solution or to verify the proposed solution formally.

The following are suggestions for improvement by students:

- Tableaux method need more examples — to be implemented in future. The reason why there are not much examples is that the method is not very unified across textbooks.
- Make number theory more systematic: Definitions first, follow by applications in proof.
- Have a summary table of definitions to be used in tests and final.

They will only be included in the lecture notes but not in this text since the plan is to make this text a reference to mathematical concepts used in applications rather than to have it as lecture notes.

Logic is concerned with representing statements in formal symbols. We will explore propositional logic, predicate logic, methods of proofs, induction, recursion, recurrence relations and we will explore the applications of logic to mathematical proofs, computer systems and program verification.

By the end of this chapter, it is hope that the reader has a basic understanding regarding the two domains of mathematical logic:

- semantics : https://en.wikipedia.org/wiki/Model_theory; and
- syntactic : https://en.wikipedia.org/wiki/Proof_theory.

1.1 Syntax of Compound Statements

Informally, **statements** or **propositions**, are **sentences** that are either **true** or **false**, but not both. Normally they will be denoted as p , q , r , etc. or the indexed letters p_1 , p_2 , etc. These letters are called statement variables, that is, variables that can be replaced by statements.

General mathematical statements are combinations of simpler statements formed through some choice of the words *not*, *and*, *or*, *if ... then ...*, and *if and only if*. These are called **(logical) connectives** (Ref: https://en.wikipedia.org/wiki/Logical_connective) and are denoted by the following symbols:

Connectives	Examples	Meaning
$\sim, \neg, !$	$\sim P, \neg P, \overline{P}, -P, !P$	Not
$\wedge, \cdot, \&, \&\&$	$P \wedge Q, P \cdot Q, PQ, P\&Q, P\&\&Q$	And
$\vee, +, , $	$P \vee Q, P + Q, P Q, P Q$	Or
$\rightarrow, \supset$	$P \rightarrow Q, P \supset Q$	If ..., then ...
$\leftrightarrow$	$P \leftrightarrow Q$	If and only if

The following logical connectives are used more often in electronics (e.g. MOS integrated circuit design) rather than mathematics:

Connectives	Examples	Meaning
$\overline{}, \uparrow, \bar{}$	$P\overline{}Q, P \uparrow Q, \overline{P \cdot Q}$	NAND
$\overline{}, \downarrow, \overline{+}$	$P\overline{}Q, P \downarrow Q, \overline{P + Q}$	NOR
$\underline{}, \oplus$	$P\underline{}Q, P \oplus Q$	XOR

Formally, the rules to form a proper proposition (or statement) is stated in the following definition.

def:formula

Definition 1.1 (Well-Formed Formula). *Statements* (or *propositions*) are either atomic or compound.

1. Constants $T, \top, \perp, F$ and single statement variables $p, q, r, s, t, p_i, i = 1, 2, 3, \dots$ are **atomic** (or *simple* or *primitive*) **statements or formulas**.
2. If ϕ and ψ are statements (abbreviated notations), then the expressions

$$(\sim \phi), (\phi \wedge \psi), (\phi \vee \psi), (\phi \rightarrow \psi), (\phi \leftrightarrow \psi)$$

are **(compound) statements** or **(compound) formulas**.

Too many parentheses in a formula (Definition 1.1) can be annoying. Parentheses are usually “simplified” in writing based on the *precedence*:

1. Evaluate parentheses first;
2. Then evaluate negations;
3. Then evaluate $\wedge$;
4. Then evaluate $\vee$;
5. Then evaluate $\rightarrow$;
6. Then evaluate $\leftrightarrow$.

This means that the “simplified” expression $p \wedge q \vee \sim r$ means $(p \wedge q) \vee (\sim r)$. If you intend to mean $p \wedge (q \vee \sim r)$, you *must* use the parentheses. The expression $p \rightarrow q \rightarrow r$ means $p \rightarrow (q \rightarrow r)$ [Huth and Ryan, 2004, Convention 1.3]. When $\rightarrow$ and $\leftrightarrow$ appear together, it is advisable to write the parentheses.

Representation of Proposition in Scheme (topic1prop.scm)

```
(define (prop->string prop)
  (define (left l) (car (cdr l)))
  (define (rght l) (car (cdr (cdr l))))
  (cond
    ((equal? prop #t) "T")
    ((equal? prop #f) "F")
    ((eq? (car prop) 'Atom)
     (let ((v (left prop))) (if (symbol? v) (symbol->string v) v)))
    ((eq? (car prop) 'Neg) (string-append "~(" (prop->string (left prop)) ")"))))
```

```

((eq? (car prop) 'Conj) (string-append " (" (prop->string (left prop)) ")\\\\" (
                                     (prop->string (right prop)) ") "))
((eq? (car prop) 'Disj) (string-append " (" (prop->string (left prop)) ")\\\\" (
                                     (prop->string (right prop)) ") "))
((eq? (car prop) 'Impl) (string-append " (" (prop->string (left prop)) ")->("
                                     (prop->string (right prop)) ") "))
((eq? (car prop) 'Iff ) (string-append " (" (prop->string (left prop)) ")<->("
                                     (prop->string (right prop)) ") "))
(else (string-append "Error Invalid operation: " prop))))

(define prop1
; (list 'Iff
;   (list 'Impl (list 'Atom "p") (list 'Atom "q"))
;   (list 'Impl
;     (list 'Disj (list 'Atom "q") (list 'Conj (list 'Atom "p") (list 'Atom "q"))))
;   #t)))
(define prop1 '(Iff
                 (Impl (Atom "p") (Atom "q"))
                 (Impl (Disj (Atom "q") (Conj (Atom "p") (Atom "q")))) #t)))

(display (prop->string prop1))
(newline)

```

eg:1.3

Example 1.1. (a) $5 < 3$ atomic statement

(b) A dog is not an animal. compound statement

(c) If the earth is flat, then $3 + 4 = 7$ compound statement

Reading formulas:

- (a) $\sim \phi$ or $\neg \phi$ is read as “not ϕ ”. This statement is called the “negation of ϕ ”.
- (b) $\phi \wedge \psi$ is read as the *conjunction* of ϕ and ψ or just “ ϕ and ψ ”.
- (c) $\phi \vee \psi$ is read as the *disjunction* of ϕ and ψ , or just “ ϕ or ψ ”.
- (d) $\phi \rightarrow \psi$ is read as “ ϕ implies ψ ” or “if ϕ then ψ ”. This form of statement is called a *conditional statement* or *implication*. In $\phi \rightarrow \psi$, the statement ϕ is called the *hypothesis* and the statement ψ is called the *conclusion* or *consequent*. The statement ϕ is called the *sufficient condition* for ψ and ψ is called the *necessary condition* for ϕ .
- (e) $p \leftrightarrow q$ is read as “ ϕ if and only if ψ ” and is normally written as “ p iff q ” (popularised by the mathematician Paul Halmos). This form of statement is called the *biconditional* of ϕ and ψ and ϕ is called the *necessary and sufficient condition* for ψ .

Remark 1.2.

1. An English sentence “ ϕ but ψ ” usually means “ ϕ and ψ ”.
2. An English sentence “neither ϕ nor ψ ” usually means “ $\sim \phi$ and $\sim \psi$ ”.
3. The notation for inequalities involves “and” and “or” statements. Let a , b and c be real numbers. Then
 - $a \leq b$ means “ $a < b$ ” or “ $a = b$ ”
 - $a < b < c$ means “ $a < b$ ” and “ $b < c$ ”.

rem:conditional

Remark 1.3. A variety of informal sentences are used to express $p \rightarrow q$:

- If p then q
- p implies q
- q follows from p
- p is sufficient for q

q is necessary for p
 q when p
 q if p
 p only if q

Note that to say “ p only if q ” means that p can take place only if q takes place also. That is, if q does not take place, then p cannot take place (symbolically “ $\sim q \rightarrow \sim p$ ”). Another way to say this is that if p occurs, then q must also occur.

def:1.10

Definition 1.2. Let p and q be statement variables:

1. The **negation** of $p \rightarrow q$ is $p \wedge \sim q$.
2. The **contrapositive** of $p \rightarrow q$ is $\sim q \rightarrow \sim p$.
3. The **converse** of $p \rightarrow q$ is $q \rightarrow p$.
4. The **inverse** of $p \rightarrow q$ is $\sim p \rightarrow \sim q$.

eg:1.14

Example 1.4. Write the negation, contrapositive, converse and inverse of the following conditional statements:

1. If 3 is positive then 3 is nonnegative.

Solution:

Negation: 3 is positive and 3 is not nonnegative.
 Contrapositive: If 3 is not nonnegative then 3 is not positive.
 Converse: If 3 is nonnegative then 3 is positive.
 Inverse: If 3 is not positive then 3 is not non-negative.

2. If you study hard, then you will not fail UECM1303.

Solution.

Negation: You study hard and you will fail UECM1303
 Contrapositive: If you fail UECM1303, then you didn't study hard.
 Converse: If you will not fail UECM1303, then you study hard.
 Inverse: If you didn't study hard, then you will fail UECM1303. □

t1:q11

Example 1.5 (Tutorial 1, Q4). Give the negation, converse, inverse and contrapositive of each the following statements.

- (a) I will pass the course if I work hard.
- (b) If $A = B \cap C$, then $A \subset C$.
- (c) If $-2 < 4$ and $3 + 8 = 11$, then $\sin(\pi/2) = 1$.

Solution. (a) Negation: I work hard and I will not pass the course.

Contrapositive: If I do not pass the course, then I do not work hard.

Inverse: If I do not work hard, then I will not pass the course.

Converse: If I pass the course, then I work hard.

(b) Negation: $A = B \cap C$ and $A \not\subset C$

Contrapositive: If $A \not\subset C$, then $A \neq B \cap C$.

Inverse: If $A \neq B \cap C$, then $A \not\subset C$.

Converse: If $A \subset C$, then $A = B \cap C$.

(c) Negation: $-2 < 4$ and $3 + 8 = 11$ and $\sin(\pi/2) \neq 1$.

Contrapositive: If $\sin(\pi/2) \neq 1$ then $2 \geq 4$ or $3 + 8 \neq 11$

Inverse: If $2 \geq 4$ or $3 + 8 \neq 11$ then $\sin(\pi/2) \neq 1$.

Converse: If $\sin(\pi/2) = 1$ then $2 < 4$ and $3 + 8 = 11$.

□

1.2 Semantics of Statements and Truth Table

Every language has two aspects: **syntax** and **semantics**. The syntax in Section 1.1 deals with the form or structure of the language of propositional logic. The semantics in this section adds “meaning” to the form. In the mathematical set theory language, suppose all the compound statements form a set Prop , the “semantics” or **(truth) assignment** is a mapping

$$v : \text{Prop} \rightarrow \{T, F\}.$$

Definition 1.3. The *truth value* of a statement/proposition is true (T or 1), if it is a true statement/proposition and false (F or 0), if it is a false statement/proposition.

Definition 1.4. [Hinman, 2005]

1. An *atomic truth assignment* is a function v that maps an atomic symbol (Definition 1.1) to a value in $\{T, F\}$ (or $\{\top, \perp\}$).
2. A **truth assignment** (or a *model* or a *valuation*) is a function v such that for any sentences ϕ and ψ ,
 - (a) $v(\sim \phi) = \text{if } v(\phi) = F \text{ then } T \text{ else } F$
 - (b) $v(\phi \vee \psi) = \text{if } v(\phi) = T \text{ or } v(\psi) = T \text{ then } T \text{ else } F$
 - (c) $v(\phi \wedge \psi) = \text{if } v(\phi) = F \text{ or } v(\psi) = F \text{ then } F \text{ else } T$
 - (d) $v(\phi \rightarrow \psi) = \text{if } v(\phi) = F \text{ then } T \text{ else } v(\psi)$
 - (e) $v(\phi \leftrightarrow \psi) = \text{if } v(\phi) = v(\psi) \text{ then } T \text{ else } F$

Implementation of the Evaluation of Proposition (topic1teval.scm)

```
(load "topic1prop.scm")

(define (prop-eval prop tvals)
  (define (left l) (car (cdr l)))
  (define (right l) (car (cdr (cdr l))))
  (define (get-tval s tvals)
    (let ((cur-pair (car tvals)))
      (let ((vv (car cur-pair)))
        (let ((v (if (symbol? vv) (symbol->string vv) vv)))
          (if (string=? v s) (left cur-pair) (get-tval s (cdr tvals)))))))
  (cond
    ((equal? prop #t) #t)
    ((equal? prop #f) #f)
    ((eq? (car prop) 'Atom) (get-tval (if (symbol? (left prop))
                                           (symbol->string (left prop))
                                           (left prop)) tvals))
    ((eq? (car prop) 'Neg) (not (prop-eval (left prop) tvals)))
    ((eq? (car prop) 'Conj) (and (prop-eval (left prop) tvals)
                                (prop-eval (right prop) tvals)))
    ((eq? (car prop) 'Disj) (or (prop-eval (left prop) tvals)
                               (prop-eval (right prop) tvals)))
    ((eq? (car prop) 'Impl) (or (not (prop-eval (left prop) tvals))
                                (prop-eval (right prop) tvals))))
```

```

    ((eq? (car prop) 'Iff ) (eq? (prop-eval (left prop) tvals)
                                   (prop-eval (right prop) tvals)))
    (else (string-append "Error Invalid operation: " prop))))

;;; Example 1.2.3
(define prop2
  '(Impl (Atom "q")
        (Impl (Neg (Atom "r"))
              (Impl (Atom "r")
                    (Disj (Atom "p") (Atom "s"))))))
(display "\nThe proposition\n")
(display (prop->string prop2))
(newline)

;(trace prop-eval)
(define tvals '(("p" #t) ("q" #t) ("r" #f) ("s" #f) ("t" #f)))
(display "with the truth assignment ")
(display tvals)
(display " is ")
(display (prop-eval prop2 tvals))
(newline)

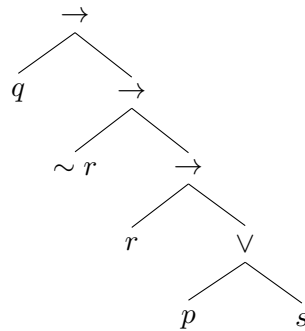
```

Example 1.6. Suppose $v(p)=T$, $v(q)=T$, $v(r)=F$ and $v(s)=F$, find $v(q \rightarrow (\sim r \rightarrow (r \rightarrow (p \vee s))))$.

Solution: Given $v(p) = v(q) = T$, $v(r) = v(s) = F$.

$$\begin{aligned}
 & v(q \rightarrow (\sim r \rightarrow (r \rightarrow (p \vee s)))) \\
 &= T \text{ if } v(q) = F \text{ else } v(\sim r \rightarrow (r \rightarrow (p \vee s))) = v(\sim r \rightarrow (r \rightarrow (p \vee s))) \\
 &= T \text{ if } v(\sim r) = F \text{ else } v(r \rightarrow (p \vee s)) \\
 &= T \text{ if } (T \text{ if } v(r) = F \text{ else } F) = F \text{ else } v(r \rightarrow (p \vee s)) \\
 &= T \text{ if } T = F \text{ else } v(r \rightarrow (p \vee s)) = v(r \rightarrow (p \vee s)) = T \text{ if } v(r) = F \text{ else } v(p \vee s) = T.
 \end{aligned}$$

It is easier to understand this if we draw the parse tree of the statement $q \rightarrow (\sim r \rightarrow (r \rightarrow (p \vee s)))$.



One can also run the Scheme program by turning on the trace to see how computer encodes the truth assignment.

t1:q5

Example 1.7 (Tutorial 1, Q2). Given that p and q are true and r , s and t are false, find the truth value of each statement below.

- (a) $(p \vee \sim q) \rightarrow (r \wedge s \wedge t)$
- (b) $(q \rightarrow (r \rightarrow s)) \wedge ((p \rightarrow s) \rightarrow (\sim t))$

Solution. Given $v(p) = v(q) = T$, $v(r) = v(s) = v(t) = F$.

(a) $v((p \vee \sim q) \rightarrow (r \wedge s \wedge t))$

$$\begin{aligned}
&= \text{T if } v(p \vee \sim q) = \text{F else } v(r \wedge s \wedge t) \\
&= \text{T if } \text{T}=\text{F else } v(r \wedge s \wedge t) \\
&= v(r \wedge s \wedge t) = \text{F}
\end{aligned}$$

where $v(p \vee \sim q) = \text{T}$ since $v(p) = \text{F}$

$$v(r \wedge s \wedge t) = \text{F} \text{ since } v(r) = v(s) = v(t) = \text{F}.$$

$$\begin{aligned}
\text{(b) } &v(q \rightarrow (r \rightarrow s)) \wedge ((p \rightarrow s) \rightarrow (\sim t)) \\
&= \text{F if any one of } v(q \rightarrow (r \rightarrow s)), v((p \rightarrow s) \rightarrow (\sim t)) \text{ is F else } \text{T} = \text{T}
\end{aligned}$$

where

$$\begin{aligned}
v(q \rightarrow (r \rightarrow s)) &= \text{T if } v(q) = \text{F else } v(r \rightarrow s) \\
&= v(r \rightarrow s) = \text{T if } v(r) = \text{F else } v(s) = \text{T}.
\end{aligned}$$

$$\begin{aligned}
v((p \rightarrow s) \rightarrow (\sim t)) \\
&= \text{T if } v(p \rightarrow s) = \text{F else } v(\sim t) = \text{T if } \text{F} = \text{F else } v(\sim t) = \text{T}
\end{aligned}$$

$$v(p \rightarrow s) = \text{T if } v(p) = \text{F else } v(s) = v(s) = \text{F}$$

□

eg:1.13

Example 1.8. Determine the hypothesis and conclusion for each of the following conditional statements. Then determine the truth value.

1. The moon is square only if the sun rises in the East.

Solution: p : The moon is square, q : the sun rises in the East.

$$v(p) = \text{F}, v(q) = \text{T}.$$

According to remark [1.3](#), the informal statement could be written as $p \rightarrow q$.

$$\text{Therefore, } v(p \rightarrow q) = \text{T}.$$

2. 1 and 3 are prime if 1 multiply 3 is prime.

Solution: p : 1 multiply 3 is prime

q : 1 and 3 are prime

$$v(p) = \text{T}, v(q) = \text{F}, \text{ so } v(p \rightarrow q) = \text{F}.$$

3. $(\sin \pi)(\cos \pi) = 0$ when $\sin \pi = 0$ or $\cos \pi = 0$.

Solution: p : $\sin \pi = 0$ or $\cos \pi = 0$

q : $(\sin \pi)(\cos \pi) = 0$

$$v(p) = \text{T}, v(q) = \text{T}, \text{ so } v(p \rightarrow q) = \text{T}.$$

4. If $1 + 1 = 3$, then cats can fly.

Solution: p : $1 + 1 = 3$

q : cats can fly

$$v(p) = \text{F}, v(q) = \text{F}, \text{ so } v(p \rightarrow q) = \text{T}.$$

t1:q6

Example 1.9 (Tutorial 1, Q3). If statement q is true, determine all truth values assignments for the statements p , r and s for which the truth value of the following statement is true:

$$(q \rightarrow [(\sim p \vee r) \wedge \sim s]) \wedge [\sim s \rightarrow (\sim r \wedge q)].$$

Solution. Given $v(q) = \text{T}$ and

$$v(q \rightarrow [(\sim p \vee r) \wedge \sim s]) \wedge [\sim s \rightarrow (\sim r \wedge q)] = \text{T}.$$

By definition, we have

$$v((\sim p \vee r) \wedge \sim s) = \text{T} \text{ and } v(\sim s \rightarrow (\sim r \wedge q)) = \text{T}.$$

i.e. $v(\sim p \vee r) = T$ and $v(\sim s) = T$ and $v(\sim r \wedge q) = T$
 which implies $v(s) = F$, $v(\sim r) = T$ and $v(q) = T$ (consistent).
 From here, $v(r) = F$ and $v(\sim p) = T$, i.e. $v(p) = F$. □

Some statements in propositional logic is special, i.e. for all possible assignments, it will be true. Going through all possible assignments leads to the notion of truth table.

Definition 1.5. The *truth table* for a given statement displays the truth values that correspond to all possible combinations of truth values for its atomic statement variables.

Remark 1.10. If a statement s has n **atomic statements**, there will need to be 2^n **rows in the truth table for s** . So truth table is only viable for statements with less than 6 atomic statement variables for manual calculation.

Inefficient Implementation of Truth Values Generation (topic1tttbl.scm)

```
;;; https://stackoverflow.com/questions/20333184/easier-way-to-generate-a-truth-table
;;; This implementation is very inefficient (and does not work well)!!!
;;; Need further improvement!!!!!!!!!!
(define (tt alist)
  (if (null? alist)
      '()
      (let ((x (car alist)) (xs (cdr alist)))
        (append (map (lambda l (cons (list x #t) l)) (tt xs))
                  (map (lambda l (cons (list x #f) l)) (tt xs))))))

;;; A function to print a truth table generated by the function tt
(define (ttstr tab)
  (define sep " | ")
  (define (myjoin l)
    (if (null? l) ""
        (string-append (car l) sep (myjoin (cdr l)))))
  ;; Get propositional name
  (define (get-prop l)
    (if (null? l) '() (cons (car (car l)) (get-prop (car (cdr l))))))
  ;; Get values
  (define (get-val l)
    (if (null? l) '() (cons (car (cdr (car l))) (get-val (car (cdr l))))))
  (define (syml->string l)
    (cond
      ((null? l) "")
      ((= (length l) 1) (symbol->string (car l)))
      (else (string-append (symbol->string (car l)) sep (syml->string (cdr l)))))
  (define h1 (syml->string (get-prop (car tab))))
  (define h2 (make-string (string-length h1) #\~))
  (define (booll->string l)
    (cond
      ((null? l) "")
      ((= (length l) 1) (if (car l) "T" "F"))
      (else (string-append (if (car l) "T" "F") sep (booll->string (cdr l)))))
  (define (st l)
    (if (null? l) ""
        (string-append (booll->string (get-val (car l))) "\n" (st (cdr l)))))
  (string-append h1 "\n" h2 "\n" (st tab)))

(define tab (tt '(p q r)))
(display (ttstr tab))
(newline)
```

Output of topic1tttbl.scm Program

p | q | r

T		T		T
T		T		F
T		F		T
T		F		F
F		T		T
F		T		F
F		F		T
F		F		F

def:1.5

Theorem 1.11 (Logical Unary Operation). *Let p be an atomic statement. If $v(p) = T$, then $v(\sim p) = F$; if $v(p) = F$, then $v(\sim p) = T$. This can be summarised as the truth table below by ignoring the cumbersome evaluation symbol $v(\cdot)$.*

Truth Table for $\sim p$

p	$\sim p$
T	F
F	T

def:1.6

Theorem 1.12 (Logical Binary Operations). *Let p and q be atomic statement variables. The evaluation of the logical binary operations $\wedge$, $\vee$, $\rightarrow$, $\leftrightarrow$ are given by the following truth table.*

$v(p)$	$v(q)$	$v(p \wedge q)$	$v(p \vee q)$	$v(p \rightarrow q)$	$v(p \leftrightarrow q)$
T	T	T	T	T	T
T	F	F	T	F	F
F	T	F	T	T	F
F	F	F	F	T	T

The truth assignment $v(\cdot)$ will be ignored in the construction of a truth table.

eg:1.10

Example 1.13. Construct a truth table for the statement $(p \wedge q) \vee \sim r$.

Solution: Steps to construct a truth table for the statement:

1. This statements involves p , q and r 3 statement variables.
2. Set up columns labeled p , q , r , $\sim r$, $(p \wedge q)$ and $(p \wedge q) \vee \sim r$.
3. Fill in the p , q and r columns with all the logically possible combinations of T's and F's. There are altogether $2^3 = 8$ possible combinations of truth values for p , q and r .
4. Use the truth tables for $\wedge$ and $\vee$ to fill in the $\sim r$ and $(p \wedge q)$ columns with the appropriate truth values.
5. Finally, fill in the $(p \wedge q) \vee \sim r$ column by considering truth values for $(p \wedge q)$ and $\sim r$.

The truth table for $(p \wedge q) \vee \sim r$ is shown below:

p	q	r	$\sim r$	$p \wedge q$	$(p \wedge q) \vee \sim r$
T	T	T	F	T	T
T	T	F	T	T	T
T	F	T	F	F	F
T	F	F	T	F	T
F	T	T	F	F	F
F	T	F	T	F	T
F	F	T	F	F	F
F	F	F	T	F	T

eg:1.11c

Example 1.14. Construct a truth table for each statement in (a) $q \wedge \sim (\sim p \rightarrow r)$; (b) $(\sim p \leftrightarrow \sim r) \vee (r \leftrightarrow q)$.

Solution.

(a)

p	q	r	$q \wedge \sim (\sim p \rightarrow r)$
T	T	T	F
T	T	F	F
T	F	T	F
T	F	F	F
F	T	T	F
F	T	F	T
F	F	T	F
F	F	F	F

(b)

p	q	r	$(\sim p \leftrightarrow \sim r) \vee (r \leftrightarrow q)$
T	T	T	T
T	T	F	F
T	F	T	T
T	F	F	T
F	T	T	T
F	T	F	T
F	F	T	F
F	F	F	T

□

eg:4var

Example 1.15. Construct the truth table for ϕ : $(p_1 \wedge p_2) \vee (p_3 \wedge p_4)$.

Solution. This statement involves 4 statement variables and there are $2^4 = 16$ different rows in the truth table below:

p_1	p_2	p_3	p_4	ϕ
T	T	T	T	T
T	T	T	F	T
T	T	F	T	T
T	T	F	F	T
T	F	T	T	T
T	F	T	F	F
T	F	F	T	F
T	F	F	F	F
F	T	T	T	T
F	T	T	F	F
F	T	F	T	F
F	T	F	F	F
F	F	T	T	T
F	F	T	F	F
F	F	F	T	F
F	F	F	F	F

□

1.3 Logical Equivalences and Laws of Logical Equivalences

The truth table is used to define “logical equivalence”. First, we introduce the concept of “tautology” (and related concepts).

def:tautology

Definition 1.6.

1. A statement ϕ is said to be a **tautology** or a *tautologous statement* if **for all truth assignments** for the atomic statement variables in ϕ , $v(\phi) = T$, i.e. its truth values in the truth table are **all true**. If ϕ is a tautology, it is denoted as $\models \phi$.
2. A statement ϕ is said to be a **contradiction** or a **contradictory statement** if its truth values in all rows in the truth table are all false.
3. A statement ϕ is said to be a **contingency** if it is neither a tautology nor a contradiction.

Remark 1.16. The constant T or $\top$ is a tautology and the constant F or $\perp$ is a contradiction.

eg:1.12

Example 1.17. Let p , q and r be statement variables. Show that the statement form

1. $\sim p \vee p$ is a tautology.

Proof: First, we construct the truth table for $\sim p \vee p$:

p	$\sim p$	$\sim p \vee p$
T	F	T
F	T	T

The truth values of $\sim p \vee p$ are **all true** in the truth table, so it is a tautology by Definition 1.6.

2. $\sim p \wedge p$ is a contradiction.

Proof: We construct the truth table for $\sim p \wedge p$:

p	$\sim p$	$\sim p \wedge p$
T	F	F
F	T	F

The truth values of $\sim p \wedge p$ are **all false** in the truth table, so it is a contradiction by Definition 1.6.

3. $(p \wedge q) \vee \sim r$ is a contingency.

Proof: From the truth table in Example 1.13, some of the truth values of $(p \wedge q) \vee \sim r$ are true and some are false and so it is a contingency.

t1:q2

Example 1.18 (Tutorial 1, Q6). Construct truth tables for the following statements:

- (a) $(p \leftrightarrow q) \leftrightarrow (\sim p \leftrightarrow \sim q)$
 (b) $\sim p \rightarrow (p \vee q)$
 (c) $(p \rightarrow q) \leftrightarrow (\sim q \rightarrow \sim p)$

Then determine whether each statement above is a tautology, a contingency or a contradiction.

Solution. The truth tables for the statements are

p	q	$(p \leftrightarrow q) \leftrightarrow (\sim p \leftrightarrow \sim q)$	$\sim p \rightarrow (p \vee q)$	$(p \rightarrow q) \leftrightarrow (\sim q \rightarrow \sim p)$
T	T	T	T	T
T	F	T	T	T
F	T	T	T	T
F	F	T	F	T

From the truth table, (a) is a tautology; (b) is a contingency; (c) is a tautology. $\square$

t1:q8

Example 1.19 (Tutorial 1, Q7). Determine whether the following statements are tautologies.

- (a) $p \rightarrow [q \rightarrow (p \wedge q)]$
 (b) $(p \vee q) \wedge \sim (q \rightarrow q)$
 (c) $(p \vee q) \rightarrow [q \rightarrow (p \wedge q)]$

Solution. The truth tables for the statements are

p	q	$p \rightarrow [q \rightarrow (p \wedge q)]$	$(p \vee q) \wedge \sim (q \rightarrow q)$	$(p \vee q) \rightarrow [q \rightarrow (p \wedge q)]$
T	T	T	F	T
T	F	T	F	T
F	T	T	F	F
F	F	T	F	T

From the truth table, (a) is a tautology, (b) is a contradiction and (c) is a contingency. $\square$

Applying all truth assignments to the proposition from Example 1.19(c)

```
(load "topic1tttbl.scm")
(load "topic1teval.scm")

(define (flatten x)
```

```

(cond ((null? x) '())
      ((pair? x) (append (flatten (car x)) (flatten (cdr x))))
      (else (list x))))

(define (pairs x)
  (if (null? x) '() (cons (list (car x) (car (cdr x))) (pairs (cdr (cdr x))))))

(define (merge prop-string tassigns tvals)
  (if (= (length tassigns) (length tvals))
      (begin
        (let loop ((row (car tassigns)))
          (if (= (length row) 1)
              (begin
                (display " ") (display (car (car (car row)))) (display " | ")
                (display prop-string) (newline))
              (begin
                (let ((rc (car row)))
                  (display " ") (display (car rc)) (display " ")
                  (loop (cdr row))))))
        (display (make-string
                  (+ (* 4 (length (car tassigns))) (string-length prop-string)) #\-)
                  (newline))
        (let loop ((r1 tassigns) (r2 tvals))
          (if (null? r1)
              '()
              (let ((row (car r1)) (v (car r2)))
                (begin
                  (let loop2 ((row row))
                    (if (= (length row) 1)
                        (begin (display (car (cdr (car (car row))))))
                        (begin (display (car (cdr (car row))) (display " ")
                                      (loop2 (cdr row))))
                    (display " | ") (display v) (newline)
                    (loop (cdr r1) (cdr r2))))))
                (display "Different table lengths"))))
      ;; Example 1.3.5(c): merge all possible truth assignments into the proposition
      (define prop3
        '(Impl (Disj (Atom p) (Atom q))
              (Impl (Atom q)
                    (Conj (Atom p) (Atom q)))))

      (display (make-string 70 #\-)
      (newline)

      (define ttbl (tt '(p q)))
      (display ttbl)
      (newline)
      (display (make-string 70 #\-)
      (newline)
      (define tvals (map (lambda (tv) (prop-eval prop3 (pairs (flatten tv)))) ttbl))

      (merge (prop->string prop3) ttbl tvals)

```

The following example illustrates the use of the notation $\models$ for tautology.

Example 1.20 (Final Exam May 2013, Q1(a)). Use truth table to show that $\models (p \rightarrow q) \wedge (r \rightarrow s) \rightarrow ((p \wedge r) \rightarrow (q \wedge s))$. The truth table must contain columns for both $(p \rightarrow q) \wedge (r \rightarrow s)$ and $(p \wedge r) \rightarrow (q \wedge s)$.

(10 marks)

Solution: Let ϕ denote $(p \rightarrow q) \wedge (r \rightarrow s)$ and ψ denote $((p \wedge r) \rightarrow (q \wedge s))$. The truth table is

..... [9 marks]

p	q	r	s	$(p \rightarrow q) \wedge (r \rightarrow s)$	$(p \wedge r) \rightarrow (q \wedge s)$	$\phi \rightarrow \psi$
T	T	T	T	T	T	T
T	T	T	F	F	F	T
T	T	F	T	T	T	T
T	T	F	F	T	T	T
T	F	T	T	F	F	T
T	F	T	F	F	F	T
T	F	F	T	F	T	T
T	F	F	F	F	T	T
F	T	T	T	T	T	T
F	T	T	F	F	T	T
F	T	F	T	T	T	T
F	T	F	F	T	T	T
F	F	T	T	T	T	T
F	F	T	F	F	T	T
F	F	F	T	T	T	T
F	F	F	F	T	T	T

Since $\phi \rightarrow \psi$ is always T, it is a tautology. [1 mark]

Two compound statements that have the same truth values in all possible truth assignments are called *logically equivalent*. We can also define this notion as follows.

def:logical-eq

Definition 1.7. [Hinman, 2005, Definition 1.3.2] Two statements ϕ and ψ are called **logically equivalent** or **tautologically equivalent** if $\phi \leftrightarrow \psi$ is a tautology. We use notations $\phi \equiv \psi$ and $\phi \Leftrightarrow \psi$ to denote that ϕ and ψ are logically equivalent.

In the following examples, we will show that we can either use the truth table of $\phi \leftrightarrow \psi$ or comparing the truth tables of ϕ and ψ to determine if $\phi \equiv \psi$.

eg:1.16

Example 1.21. Let p and q be two statement variables. Determine whether the statements $\sim p \vee \sim q$ and $\sim (p \vee q)$ are logically equivalent or not.

Solution: By Definition 1.7, we can construct a truth table for $(\sim p \vee \sim q) \leftrightarrow (\sim (p \vee q))$ to determine if it is logically equivalent.

p	q	$\sim p$	$\sim q$	$p \vee q$	$\sim (p \vee q)$	$\sim p \vee \sim q$	$(\sim p \vee \sim q) \leftrightarrow (\sim (p \vee q))$
T	T	F	F	T	F	F	T
T	F	F	T	T	F	T	F
F	T	T	F	T	F	T	F
F	F	T	T	F	T	T	T

Since $(\sim p \vee \sim q) \leftrightarrow (\sim (p \vee q))$ is not a tautology, $\sim p \vee \sim q$ and $\sim (p \vee q)$ are not logically equivalent.

t1:q14

Example 1.22 (Tutorial 1, Q8). Determine whether the 2 statements forms are equivalent.

(a) $p \leftrightarrow q$ and $(p \wedge q) \vee (\sim p \wedge \sim q)$.

(b) $p \rightarrow (q \rightarrow r)$ and $(p \rightarrow q) \rightarrow r$.

(c) $(p \vee q) \rightarrow r$ and $(p \rightarrow r) \wedge (q \rightarrow r)$.

Solution. (a) They are logically equivalent since they have same truth values as shown below:

p	q	$p \leftrightarrow q$	$(p \wedge q) \vee (\sim p \wedge \sim q)$
T	T	T	T
T	F	F	F
F	T	F	F
F	F	T	T

- (b) $p \rightarrow (q \rightarrow r)$ and $(p \rightarrow q) \rightarrow r$ are not logically equivalent since they have different truth values (refer to Table in part (c)).
- (c) $(p \vee q) \rightarrow r$ and $(p \rightarrow r) \wedge (q \rightarrow r)$ are logically equivalent since they have the same truth values as shown below:

p	q	r	$p \rightarrow (q \rightarrow r)$	$(p \rightarrow q) \rightarrow r$	$(p \vee q) \rightarrow r$	$(p \rightarrow r) \wedge (q \rightarrow r)$
T	T	T	T	T	T	T
T	T	F	F	F	F	F
T	F	T	T	T	T	T
T	F	F	T	T	F	F
F	T	T	T	T	T	T
F	T	F	T	F	F	F
F	F	T	T	T	T	T
F	F	F	T	F	T	T

□

Mathematicians have identified the useful laws for simplifying statements based on the concept of logical equivalence. They are summarised in the following theorem.

thm:equiv-laws

Theorem 1.23 (Laws of Logical Equivalences). *Given any atomic statements p , q and r , the following logical equivalences hold [Epp, 2020].*

1. *Double negative law:*

- $\sim(\sim p) \equiv p$.

2. *Idempotent laws:*

- $p \wedge p \equiv p$;
- $p \vee p \equiv p$.

3. *Universal bound laws:*

- $p \vee \text{T} \equiv \text{T}$;
- $p \wedge \text{F} \equiv \text{F}$.

4. *Identity laws:*

- $p \wedge \text{T} \equiv p$;
- $p \vee \text{F} \equiv p$.

5. *Negation laws:*

- $p \vee \sim p \equiv \text{T}$;
- $p \wedge \sim p \equiv \text{F}$.

6. *Commutative laws:*

- $p \wedge q \equiv q \wedge p$;
- $p \vee q \equiv q \vee p$.

7. *Absorption laws:*

- $p \vee (p \wedge q) \equiv p$;
- $p \wedge (p \vee q) \equiv p$.

8. *Associative laws:*

- $(p \wedge q) \wedge r \equiv p \wedge (q \wedge r)$;
- $(p \vee q) \vee r \equiv p \vee (q \vee r)$.

9. *Distributive laws:*

- $p \wedge (q \vee r) \equiv (p \wedge q) \vee (p \wedge r)$;
- $p \vee (q \wedge r) \equiv (p \vee q) \wedge (p \vee r)$.

10. *De Morgan's laws:*

- $\sim(p \wedge q) \equiv \sim p \vee \sim q$;
- $\sim(p \vee q) \equiv \sim p \wedge \sim q$.

11. *Implication law:*

- $p \rightarrow q \equiv \sim p \vee q$

12. *Contrapositive law:*

- $p \rightarrow q \equiv \sim q \vee \sim \sim p$

13. *Biconditional law:*

- $p \leftrightarrow q \equiv (p \rightarrow q) \wedge (q \rightarrow p)$.

page:equivlaws

Proof: Since any statements can only be T or F, truth table can be used to prove 1. to 13. The following are just two demonstrations and the rest are left as exercises. Note that Example 1.18(c) actually proves the contrapositive law.

7. One of the absorption law

p	q	$p \wedge q$	$p \vee (p \wedge q)$	$(p \vee (p \wedge q)) \leftrightarrow p$
T	T	T	T	T
T	F	F	T	T
F	T	F	F	T
F	F	F	F	T

Since $\models (p \vee (p \wedge q)) \leftrightarrow p$, by definition, $p \vee (p \wedge q) \equiv p$.

11. Implication law

p	q	$\sim p$	$p \rightarrow q$	$\sim p \vee q$
T	T	F	T	T
T	F	F	F	F
F	T	T	T	T
F	F	T	T	T

The last two columns have the same truth values, hence the two statements are logical equivalent.

An important application of the logical equivalence laws is to write sentences in more readable form as demonstrated in the following example.

eg:1.17

Example 1.24. Write the negation of the given statements:

- John is smart but lazy.

In $\sim (p \wedge q)$ form: John is *not* “smart and lazy”.

In equivalent form $\sim p \vee \sim q$: John is not smart or he is not lazy.

- $2 \leq \sqrt{2}$

In $\sim (p \vee q)$ form: $2 \not\leq \sqrt{2}$

In equivalent form $\sim p \wedge \sim q$: $2 \not\leq \sqrt{2}$ and $2 \neq \sqrt{2}$.

Note that the statement $\sim (p \wedge q)$ is called the *negation of conjunction of p and q* while the statement $\sim (p \vee q)$ is called the *negation of disjunction of p and q* .

t1:q7

Example 1.25 (Tutorial 1, Q9). Rewrite the following statements in (the logically equivalent) if-then form.

- Fix my ceiling or I won't pay my rent.
- Study hard or I won't pass Discrete Mathematics.
- Catching the 7am bus is a sufficient condition for my being on time for school.
- Doing homework regularly is a necessary condition to pass the course.
- Ali studies calculus only if he is a math major.
- P is a square only if P is a rectangle.
- n is divisible by 6 is a sufficient condition for n to be divisible by 2 and n is divisible by 3.

Solution. (a) If the ceiling is not fixed, then I won't pay the rent. $c \vee \sim p \equiv \sim c \rightarrow \sim p$
 (b) If I don't study hard, then I won't pass Discrete Mathematics. ... $s \vee \sim p \equiv \sim s \rightarrow \sim p$
 (c) If I catch the 7am bus, then I will be on time for school. $c \rightarrow t$ (Remark 1.3)
 (d) If I pass the course, then I do homework regularly. $p \rightarrow d$ (Remark 1.3)
 (e) If Ali study calculus, then he is a math major. $s \rightarrow m$ (Remark 1.3)
 (f) If P is a square, then P is a rectangle. $s \rightarrow r$ (Remark 1.3)
 (g) If n is divisible by 6, then n is divisible by 2 and n is divisible by 3 $p_6 \rightarrow p_2$
 (Remark 1.3)

□

To make the laws of logical equivalences useful, mathematicians have shown that the logical equivalence “ $\equiv$ ” is an equivalence relation (see Set Relations in later topic). We also require logically equivalent statements to be logically equivalent under substitution [Chiswell and Hodges, 2007, Herrmann, 29 Jan 2006].

def:subst

Definition 1.8 (Uniform Substitution). If ϕ, ψ, ψ_i are statements/formulas, and p_i are (atomic) propositional variables, then $\phi[\psi/p_i]$ denotes the result of replacing each occurrence of p_i by an occurrence of ψ in ϕ ; similarly, the *simultaneous substitution* S of $p_1, \dots, p_n$ by formulas $\psi_1, \dots, \psi_n$ is denoted by $\psi_1/p_1, \dots, \psi_n/p_n$.

thm:substitution

Theorem 1.26 (Substitution Theorem). Let S be a simultaneous substitution (Definition 1.8) and $\phi \equiv \psi$. Then $\phi[S] \equiv \psi[S]$. More generally, let ϕ_i and ψ_i be statements such that $\phi_i \equiv \psi_i, 1 \leq i \leq n$. Then $\phi[\phi_1/p_1, \dots, \phi_n/p_n] \equiv \psi[\psi_1/p_1, \dots, \psi_n/p_n]$.

Proof. See Chiswell and Hodges [2007, Theorem 3.7.6]. $\square$

eg:1.19

Example 1.27. Show each logical equivalence below using laws of logical equivalences (Theorem 1.23).

$$1. (p \vee q) \wedge ((p \vee q) \vee r) \wedge \sim (\sim (p \vee q)) \equiv p \vee q$$

Proof: To illustrate the power of the substitution, we note that a substitution of $p \vee q$ for p and r for q in absorption law gives

$$(p \vee q) \wedge ((p \vee q) \vee r) \equiv p \vee q.$$

This leads to the first step in the following logical equivalence, the rest of the steps are of similar reasoning.

$$\begin{aligned} (p \vee q) \wedge ((p \vee q) \vee r) \wedge \sim (\sim (p \vee q)) &\equiv (p \vee q) \wedge \sim (\sim (p \vee q)) && \text{(Absorption law)} \\ &\equiv (p \vee q) \wedge (p \vee q) && \text{(Double negative law)} \\ &\equiv p \vee q && \text{(Idempotent law)} \end{aligned}$$

$$2. (p \vee q) \wedge \sim (\sim p \wedge q) \equiv p$$

Solution.

$$\begin{aligned} (p \vee q) \wedge \sim (\sim p \wedge q) &\equiv (p \vee q) \wedge (\sim \sim p \vee \sim q) && \text{(De Morgan)} \\ &\equiv (p \vee q) \wedge (p \vee \sim q) && \text{(Double negative)} \\ &\equiv p \vee (q \wedge \sim q) && \text{(Distributive)} \\ &\equiv p \vee F && \text{(Negation)} \\ &\equiv p && \text{(Identity)} \end{aligned}$$

 $\square$

$$3. \sim [\sim ((p \vee q) \wedge r) \vee \sim q] \equiv q \wedge r$$

Solution.

$$\begin{aligned} \sim [\sim ((p \vee q) \wedge r) \vee \sim q] &\equiv \sim \sim ((p \vee q) \wedge r) \wedge \sim \sim q && \text{(De Morgan)} \\ &\equiv ((p \vee q) \wedge r) \wedge q && \text{(Double Negative)} \\ &\equiv ((p \vee q) \wedge q) \wedge r && \text{(Associative)} \\ &\equiv q \wedge r && \text{(Absorption)} \end{aligned}$$

 $\square$

t1:q18

Example 1.28 (Tutorial 1, Q10). Explain why the statement “If today is not cold, then today is cold” is logically equivalent to the statement “Today is cold”.

Solution. Let p be the statement “today is cold”. Formally, we have $\sim p \rightarrow p \equiv \sim (\sim p) \vee p \equiv p \vee p \equiv p$. $\square$

t1:q15

Example 1.29 (Tutorial 1, Q11). (a) Show that the following statements are all logically equivalent.

$$p \rightarrow (q \vee r), \quad (p \wedge \sim q) \rightarrow r \quad \text{and} \quad (p \wedge \sim r) \rightarrow q.$$

- (b) Use the logical equivalences in (a) to rewrite the sentence “If n is prime, then n is odd or n is 2.” in 2 different ways.

Solution. (a) $p \rightarrow (q \vee r) \equiv \sim p \vee (q \vee r) \equiv (\sim p \vee q) \vee r \equiv \sim (p \wedge \sim q) \vee r \equiv (p \wedge \sim q) \rightarrow r$
 $p \rightarrow (q \vee r) \equiv \sim p \vee (q \vee r) \equiv \sim p \vee (r \vee q) \equiv \sim (p \wedge \sim r) \vee q \equiv (p \wedge \sim r) \rightarrow q$

- (b) If n is prime and not odd, then n is 2.

If n is prime and n is not 2, then n is odd.

□

The laws of logical equivalence can be used to simplify some long statements to logically equivalent shorter statements as demonstrated below.

eg:1.20

Example 1.30. Simplify the following statement to a statement with no more than 3 logical connectives by stating the law used in each step of the simplification:

$$(p \vee q) \wedge p \wedge (r \vee q) \wedge (p \vee \sim p \vee r) \wedge (r \wedge \sim q).$$

Solution:

$$\begin{aligned} & (p \vee q) \wedge p \wedge (r \vee q) \wedge (p \vee \sim p \vee r) \wedge (r \wedge \sim q) \\ & \equiv (p \vee q) \wedge p \wedge (r \vee q) \wedge (T \vee r) \wedge (r \wedge \sim q) && \text{(Negation)} \\ & \equiv (p \vee q) \wedge p \wedge (r \vee q) \wedge T \wedge (r \wedge \sim q) && \text{(Commutative and Universal Bound)} \\ & \equiv p \wedge [(r \vee q) \wedge (r \wedge \sim q)] && \text{(Absorption, Identity)} \\ & \equiv p \wedge [(r \wedge (r \wedge \sim q)) \vee (q \wedge (r \wedge \sim q))] && \text{(Distributive)} \\ & \equiv p \wedge [(r \wedge r) \wedge \sim q \vee (q \wedge (r \wedge \sim q))] && \text{(Associative, Commutative)} \\ & \equiv p \wedge [(r \wedge \sim q) \vee (F \wedge r)] && \text{(Idempotent, Associative, Negation)} \\ & \equiv p \wedge [(r \wedge \sim q) \vee F] && \text{(Universal bound)} \\ & \equiv p \wedge (r \wedge \sim q) && \text{(Identity)} \end{aligned}$$

t1:q16

Example 1.31 (Tutorial 1, Q12). Use the laws of logical equivalence to show the following:

- (a) $(p \wedge (\sim (\sim p \vee q))) \vee (p \wedge q) \equiv p$.

Solution. $(p \wedge (\sim (\sim p \vee q))) \vee (p \wedge q) \equiv (p \wedge (p \wedge \sim q)) \vee (p \wedge q)$
 $\equiv (p \wedge \sim q) \vee (p \wedge q)$
 $\equiv p \wedge (\sim q \vee q) \equiv p \wedge T \equiv p$

□

- (b) $\sim (p \vee \sim q) \vee (\sim p \wedge \sim q) \equiv \sim p$.

Solution. $\sim (p \vee \sim q) \vee (\sim p \wedge \sim q) \equiv (\sim p \wedge q) \vee (\sim p \wedge \sim q)$
 $\equiv \sim p \wedge (q \vee \sim q)$
 $\equiv \sim p \wedge T \equiv \sim p$

□

- (c) $\sim ((\sim p \wedge q) \vee (\sim p \wedge \sim q)) \vee (p \wedge q) \equiv p$.

Solution. $\sim ((\sim p \wedge q) \vee (\sim p \wedge \sim q)) \vee (p \wedge q) \equiv \sim (\sim p \wedge (q \vee \sim q)) \vee (p \wedge q)$
 $\equiv \sim (\sim p \wedge T) \vee (p \wedge q)$
 $\equiv p \vee (p \wedge q) \equiv p$

□

- (d) $(\sim p \vee \sim q) \rightarrow (p \wedge q \wedge r) \equiv p \wedge q$.

Solution. $(\sim p \vee \sim q) \rightarrow (p \wedge q \wedge r) \equiv \sim (\sim p \vee \sim q) \vee (p \wedge q \wedge r)$
 $\equiv (p \wedge q) \vee (p \wedge q \wedge r) \equiv p \wedge q$

□

t1:q17

Example 1.32 (Tutorial 1, Q13). Simplify the following compound statement to a statement with no more than 3 logical connectives involving $\sim$, $\vee$ and $\wedge$:

$$[[[(p \wedge q) \wedge r] \vee [(p \wedge q) \wedge \sim r]] \vee \sim q] \rightarrow s.$$

Solution.
$$\begin{aligned} & [[[(p \wedge q) \wedge r] \vee [(p \wedge q) \wedge \sim r]] \vee \sim q] \rightarrow s \equiv [[(p \wedge q) \wedge (r \vee \sim r)] \vee \sim q] \rightarrow s \\ & \equiv [(p \wedge q) \wedge T] \vee \sim q \rightarrow s \\ & \equiv (p \wedge q) \vee \sim q \rightarrow s \\ & \equiv [(p \vee \sim q) \wedge (q \vee \sim q)] \rightarrow s \\ & \equiv [(p \vee \sim q) \wedge T] \rightarrow s \\ & \equiv (p \vee \sim q) \rightarrow s \\ & \equiv \sim p \wedge q \vee s \end{aligned}$$

□

t1:q20

Example 1.33 (Tutorial 1, Q14). Verify that

(a) $(p \rightarrow q) \wedge [\sim q \wedge (r \vee \sim q)] \equiv \sim (p \vee q)$

Solution.
$$\begin{aligned} (p \rightarrow q) \wedge [\sim q \wedge (r \vee \sim q)] & \equiv (\sim p \vee q) \wedge \sim q \\ & \equiv (\sim p \wedge \sim q) \vee (q \wedge \sim q) \\ & \equiv \sim (p \vee q) \vee F \equiv \sim (p \vee q) \end{aligned}$$

□

(b) $p \vee q \vee (\sim p \wedge \sim q \wedge r) \equiv p \vee q \vee r$

Solution.
$$\begin{aligned} p \vee q \vee (\sim p \wedge \sim q \wedge r) & \equiv (p \vee q) \vee (\sim (p \vee q) \wedge r) \\ & \equiv ((p \vee q) \vee \sim (p \vee q)) \wedge ((p \vee q) \vee r) \\ & \equiv T \wedge (p \vee q \vee r) \equiv p \vee q \vee r \end{aligned}$$

□

(c) $(p \leftrightarrow q) \wedge (q \leftrightarrow r) \wedge (r \leftrightarrow p) \equiv (p \rightarrow q) \wedge (q \rightarrow r) \wedge (r \rightarrow p)$

Solution: One can use truth table to prove that they are logically equivalent, this is simpler compare to using laws of logical equivalence as follows (let the right hand side be ψ and the associative law and commutative law are sometimes used without explicitly mentioned).

First, we reduce the right-hand-side to an equivalent form involving only negation, conjunction and disjunction.

$$\begin{aligned} \psi & \equiv (\sim p \vee q) \wedge (\sim q \vee r) \wedge (\sim r \vee p) && [\text{implication law}] \\ & \equiv (\sim p \wedge \sim q \vee \sim p \wedge r \vee q \wedge \sim q \vee q \wedge r) \wedge (\sim r \vee p) && [\text{distributive law}] \\ & \equiv (\sim p \wedge \sim q \vee \sim p \wedge r \vee q \wedge r) \wedge (\sim r \vee p) && [\text{negation \& identity laws}] \\ & \equiv (\sim p \wedge \sim q \wedge r \vee \sim p \wedge \sim q \wedge p \vee \sim p \wedge r \wedge \sim r \vee \sim p \wedge r \wedge p \\ & \quad \vee q \wedge r \wedge \sim r \vee q \wedge r \wedge p) && [\text{distributive law}] \\ & \equiv (\sim p \wedge \sim q \wedge r) \vee (p \wedge q \wedge r) && [\text{negation \& identity laws}] \end{aligned}$$

$$\begin{aligned} & (p \leftrightarrow q) \wedge (q \leftrightarrow r) \wedge (r \leftrightarrow p) \\ & \equiv (p \rightarrow q) \wedge (q \rightarrow p) \wedge (q \rightarrow r) \wedge (r \rightarrow q) \wedge (r \rightarrow p) \wedge (p \rightarrow r) && [\text{biconditional law}] \\ & \equiv (p \rightarrow q) \wedge (q \rightarrow r) \wedge (r \rightarrow p) \wedge (q \rightarrow p) \wedge (r \rightarrow q) \wedge (p \rightarrow r) && [\text{associative \& commutative laws}] \\ & \equiv \psi \wedge (\sim q \vee p) \wedge (\sim r \vee q) \wedge (\sim p \vee r) && [\text{implication law}] \\ & \equiv \psi \wedge (\sim q \wedge \sim r \vee \sim q \wedge q \vee p \wedge \sim r \vee p \wedge q) \wedge (\sim p \vee r) && [\text{distributive law}] \\ & \equiv \psi \wedge (\sim q \wedge \sim r \vee F \vee p \wedge \sim r \vee p \wedge q) \wedge (\sim p \vee r) && [\text{negation law}] \\ & \equiv \psi \wedge (\sim q \wedge \sim r \vee p \wedge \sim r \vee p \wedge q) \wedge (\sim p \vee r) && [\text{identity law}] \\ & \equiv \psi \wedge (\sim q \wedge \sim r \wedge p \vee \sim q \wedge \sim r \wedge r \\ & \quad \vee p \wedge \sim r \wedge \sim p \vee p \wedge \sim r \wedge r \vee p \wedge q \wedge \sim p \vee p \wedge q \wedge r) && [\text{distributive law}] \\ & \equiv \psi \wedge (\sim q \wedge \sim r \wedge \sim p \vee F \vee F \vee F \vee p \wedge q \wedge r) && [\text{negation law}] \\ & \equiv \psi \wedge (\sim p \wedge \sim q \wedge \sim r \vee p \wedge q \wedge r) \equiv \psi \wedge \psi = \psi && [\text{identity law}] \end{aligned}$$

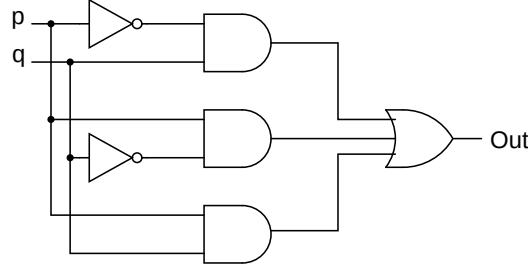
Hence the two statements are logically equivalent.

1.4 Applications: Logic Circuits, Solving Logic Puzzles, etc.

sec:apps1

Logical equivalence is used in logic circuit design [Nisan and Schocken, 2005] and in digital signal processing systems [Woods et al., 2008] to identify logical equivalent circuits which are minimal to save power (larger circuits use more power).

Example 1.34. Simplify the logic diagram below.



For the names of the symbols, see Nisan and Schocken [2005, Chapter 1].

Solution: Step 1: Obtain the formula of the circuit: $Out = (\sim p \wedge q) \vee (p \wedge \sim q) \vee (p \wedge q)$

Step 2: Simplify the above formula:

$$\begin{aligned}
 Out &= (\sim p \wedge q) \vee (p \wedge (\sim q \vee q)) && \text{(Distributive Law)} \\
 &\equiv (\sim p \wedge q) \vee (p \wedge T) && \text{(Negation Law)} \\
 &\equiv (\sim p \wedge q) \vee p && \text{(Identity Law)} \\
 &\equiv (\sim p \vee p) \wedge (q \vee p) && \text{(Distributive Law)} \\
 &\equiv T \wedge (p \vee q) && \text{(Negation Law \& Commutative Law)} \\
 &\equiv p \vee q && \text{(Identity Law)}
 \end{aligned}$$

Step 3: The above circuit is equivalent to an or-gate

Example 1.35 (Tutorial 1, Q15). In logic circuit design, one of the basic logic gate is the NAND gate. It is logically equivalent to $\sim (p \wedge q)$ and denoted by $(p \uparrow q)$ for any statements p and q .

t1:q9

(a) Represent the logic gates (i) $\sim p$ and (ii) $p \rightarrow q$ using the NAND gate.

Solution. (i) $\sim p \equiv \sim (p \wedge p) \equiv p \uparrow p$

(ii) $p \rightarrow q \equiv \sim p \vee q \equiv \sim (p \wedge \sim q) \equiv p \uparrow \sim q \equiv p \uparrow (q \uparrow q)$

□

t1:q21

(b) Are $p \uparrow (q \uparrow r)$ and $(p \uparrow q) \uparrow r$ logically equivalent?

Solution. $p \uparrow (q \uparrow r) \equiv \sim (p \wedge (q \uparrow r)) \equiv \sim (p \wedge \sim (q \wedge r)) \equiv (\sim p) \vee (q \wedge r)$

$(p \uparrow q) \uparrow r \equiv \sim ((p \uparrow q) \wedge r) \equiv \sim (\sim (p \wedge q) \wedge r) \equiv (p \wedge q) \vee (\sim r)$

When $p \equiv F, q \equiv F, r \equiv T$, $(p \uparrow q) \uparrow r \equiv T$ but $p \uparrow (q \uparrow r) \equiv F$. Hence, they are not logically equivalent.

□

We will introduce generalised concepts related to the “truth assignment” from Definition [1.4](#). Note that we will use the notion of **formula** (which includes propositions and predicates) rather than just proposition.

def:SAT

Definition 1.9. [Schöning and Torán, 2013] A formula ϕ is called **satisfiable** if there exists an assignment v such that $v(\phi) = T$. Otherwise, ϕ is called **unsatisfiable** (or contradiction as defined above).

The set of all satisfiable formulas is denoted by SAT. A satisfying assignment v for a formula ϕ is often also called a **model** for ϕ , in which case we write: $v \models \phi$. (Generation of [Definition 1.4](#)).

A formula ϕ is called **valid** or **tautology** if, for all assignments v , $v(\phi) = T$. The set of tautologies is denoted by TAUT. The notation $\models \phi$ expresses that ϕ is a tautology.

A formula ψ **follows** from ϕ (or, ψ is a **consequence** from ϕ , denoted by $\phi \models \psi$, if $\phi \rightarrow \psi$ (or $\sim \phi \vee \psi$) is a tautology.

Based on the definition, ϕ is a **tautology** iff $\sim \phi$ is **unsatisfiable**; ϕ is **satisfiable** iff $\sim \phi$ is not a tautology.

Apart from the truth table method, another method to check for satisfiability is the **tableaux method**, which checks a proposition ϕ is satisfiable by expanding ϕ to conjunction and disjunction forms using laws of logical equivalences and identify “open” branches.

eg:tableau1

Example 1.36. Use tableaux method to check the satisfiability of the proposition $\sim ((p \wedge \sim r) \vee (\sim q \wedge r))$.

Solution:

1.	$\sim ((p \wedge \sim r) \vee (\sim q \wedge r))$	Hypothesis
2.	$(\sim (p \wedge \sim r)) \wedge (\sim (\sim q \wedge r))$	De Morgan
3.	$\sim p \vee r$	2 Conjunction & De Morgan
4.	$q \vee \sim r$	2 Conjunction & De Morgan
5.		3 disjunction
6.		4 disjunction

From the right tree, we can see that a valuation $v(q) = T$ and $v(r) = T$ to make $v(\sim p \vee r) = T$ and $v(q \vee \sim r) = T$. Hence the hypothesis is satisfiable.

From the example above, we can see that **encoding** propositions into propositions involving only negation, conjunction and disjunction is the way to solve **satisfiability problem** (it is important because it is at the crossroads of logic, graph theory, computer science, computer engineering, and operations research). CNF is used in the computerisation of **SAT solver** as well as the generalisation to **SMT (Satisfiability Modulo Theories)** solvers.

def:cnf

Definition 1.10. A **literal** is an atom p or a negated atom $\sim p$.

A **clause** is a disjunction of **literals**.

A formula ϕ is in **conjunctive normal form (CNF)** if it is a conjunction of clauses:

$$F = C_1 \wedge C_2 \wedge \cdots \wedge C_m$$

where $C_i = u_{i,1} \vee u_{i,2} \vee \cdots \vee u_{i,k_i}$, $i = 1, \dots, m$ are clauses, $u_{i,j}$ are literals. If $|C_i| = k_i \leq k$ for some constant k for all i , ϕ is said to be in k -CNF.

A formula ϕ is in **disjunctive normal form (DNF)** if it is a disjunction of conjunctions of literals.

Example 1.37. $(p \vee \sim q) \wedge (\sim p \vee r) \wedge (q \vee r)$ is a CNF.

Example 1.38. The proposition $\sim ((p \wedge \sim r) \vee (\sim q \wedge r))$ from Example [1.36](#) has a logically equivalent CNF:

$$(\sim p \vee q) \wedge (q \vee \sim r).$$

Exercise: Feel free to check if the lecturer is correct by using truth table or comparison table.

Theorem 1.39. For every proposition ϕ , there exists a CNF formula which is equivalent to ϕ . However, finding the equivalent CNF usually requires exponential computation time.

Theorem 1.40 (Tseitin). For every proposition ϕ , there exists a sat-equivalent formula ψ in 3-CNF. The transformation from ϕ to ψ can be done efficiently in polynomial time.

When a formula ϕ is transformed into an equivalent conjunctive normal form (CNF), the **DPLL algorithm**, named for its developers Davis, Putnam, Logemann, and Loveland, is an efficient approach to solving **boolean satisfiability (SAT) problems**.

```

1: function DPLL( $\Phi$ )
2:   if  $\phi$ =true then return true
3:   if  $\phi$  contains a false clause then return false
4:   for all unit clauses  $\ell$  in  $\phi$  do
5:      $\phi \leftarrow \text{UNIT-PROPAGATE}(\ell, \phi)$ 
6:   for all literals  $\ell$  occurring pure in  $\phi$  do
7:      $\phi \rightarrow \text{PURE-LITERAL-ASSIGN}(\ell, \phi)$ 
8:    $\ell \leftarrow \text{CHOOSE-LITERAL}(\phi)$  return DPLL( $\phi \wedge \ell$ )  $\vee$  DPLL( $\phi \wedge \sim \ell$ )

```

There are many SMT solvers. https://en.wikipedia.org/wiki/Z3_Theorem_Prover, developed by Microsoft, is one of the most popular and will be used in this course.

Example 1.41. Use the Z3 Theorem Prover to solve the SAT problem for the proposition $\sim((p \wedge \sim r) \vee (\sim \wedge r))$ from Example [1.36](#) ^{eg:tableau1}

Solution: A program using Z3 SMT solver to solve SAT problem from Example [1.36](#) ^{eg:tableau1} is shown below.

```

;;; Example 1.3.26 and 1.3.30:  $\sim((p \ \& \ \sim r) \ \vee \ (\sim q \ \& \ r))$ 
(declare-const p Bool)
(declare-const q Bool)
(declare-const r Bool)
(define-fun proposition () Bool
  (not (or (and p (not r)) (and (not q) r))))
(assert proposition)
(check-sat)
(get-model)
(exit)

```

The truth assignment found by Z3 SMT solver is shown below.

```

sat
(
  (define-fun p () Bool
    false)
  (define-fun q () Bool
    false)
  (define-fun proposition () Bool
    (not (or (and p (not r)) (and (not q) r))))
  (define-fun r () Bool
    false)
)

```

Although SAT/SMT solvers use CNF in the algorithm, many logical puzzles are expressed in DNF as illustrated in the following example.

Example 1.42. Four candidates A, B, C and D are to be selected to participate in a chess competition. However, only two candidates will be chosen in the final list and the following constraints must be obeyed:

- Only one from A and B will participate.
- If C participates, then D will also need to participate.
- At most one of B and D will participate.
- If D is not participating, then A is also not participating.

Solution: Let a, b, c, d denote the statements A, B, C, D participates in the competition respectively. Then the constraints can be written as

$$[(a \wedge \sim b) \vee (\sim a \wedge b)] \wedge [c \rightarrow d] \wedge \sim (b \wedge d) \wedge [\sim d \rightarrow \sim a]$$

Let us denote it by p and we can either construct a truth table or use the law of logical equivalence to simplify the expressions to involving negations and conjunctions.

$$\begin{aligned} p &\equiv [(a \wedge \sim b) \vee (\sim a \wedge b)] \wedge [\sim c \vee d] \wedge [\sim b \vee \sim d] \wedge [d \vee \sim a] \\ &\equiv [(a \wedge \sim b) \vee (\sim a \wedge b)] \wedge [(\sim c \wedge \sim a) \vee d] \wedge [\sim b \vee \sim d] \\ &\equiv [(a \wedge \sim b) \vee (\sim a \wedge b)] \wedge [(\sim c \wedge \sim a) \wedge [\sim b \vee \sim d] \vee d \wedge [\sim b \vee \sim d]] \\ &\equiv [(a \wedge \sim b) \vee (\sim a \wedge b)] \wedge [\sim a \wedge \sim c \wedge \sim b \vee \sim a \wedge \sim c \wedge \sim d \vee \sim b \wedge d \vee \sim d \wedge d] \\ &\equiv [(a \wedge \sim b) \vee (\sim a \wedge b)] \wedge [\sim a \wedge \sim b \wedge \sim c \vee \sim a \wedge \sim c \wedge \sim d \vee \sim b \wedge d] \\ &\equiv \underline{a \wedge \sim b \wedge \sim a \wedge \sim b \wedge \sim c} \vee \underline{a \wedge \sim b \wedge \sim a \wedge \sim c \wedge \sim d} \vee \underline{a \wedge \sim b \wedge \sim b \wedge d} \\ &\quad \vee \underline{\sim a \wedge b \wedge \sim a \wedge \sim b \wedge \sim c} \vee \underline{\sim a \wedge b \wedge \sim a \wedge \sim c \wedge \sim d} \vee \underline{\sim a \wedge b \wedge \sim b \wedge d} \\ &\equiv F \vee F \vee \underline{a \wedge \sim b \wedge d} \vee F \vee \underline{\sim a \wedge b \wedge \sim c \wedge \sim d} \vee F \end{aligned}$$

Since we need two participants (so $\sim a \wedge b \wedge \sim c \wedge \sim d$ is not useful because only B goes), they are A and D.

```
;; Example 1.3.32: participate = true; not participate = false
(declare-const A Bool)
(declare-const B Bool)
(declare-const C Bool)
(declare-const D Bool)
(assert (or (and A (not B)) (and (not A) B))) ; condition 1
(assert (implies C D)) ; condition 2
(assert (not (and B D)))
(assert (implies (not D) (not A)))
(check-sat)
(get-model)
(exit)
```

The output matches our mathematical analysis above.

```
sat
(
  (define-fun A () Bool
    false)
  (define-fun D () Bool
    false)
  (define-fun B () Bool
    true)
  (define-fun C () Bool
    false)
)
```

1.5 Logical Implication and Argument for Statements

sec:impl-laws Logical equivalence (Definition def:logical-eg 1.7) is defined in terms of the connective of “ordinary” equivalence $\leftrightarrow$. Analogously, the notion of *logical implication* can be defined in terms of “ordinary” implication “ $\rightarrow$ ”.

Definition 1.11. We say that a statement ϕ *logically implies* a statement ψ or ψ is a *tautological consequence* of ϕ [Hinman, 2005, Definition 1.3.2], denoted by $\phi \Rightarrow \psi$, when the statement $\phi \rightarrow \psi$ is a tautology, i.e. $\models \phi \rightarrow \psi$. When this happens, we sometimes say that the statement ϕ is *stronger* than ψ .

Remark 1.43. The distinction between implication and logical implication is fundamental. Whereas $p \rightarrow q$ is a compound statement, the statement $p \Rightarrow q$ describes a relationship between two compound statement p and q , it is a **meta language**!!!

The link between logical equivalence and logical implication is as follows.

Theorem 1.44. Let ϕ and ψ be two statements. Then $\phi \equiv \psi$ iff $\phi \Rightarrow \psi$ and $\psi \Rightarrow \phi$.

By definition, to show that “ $\phi \Rightarrow \psi$ ”, i.e. a statement ϕ logically implies another statement ψ , we can create the truth table for the statement $\phi \rightarrow \psi$ and examine its last column.

Example 1.45. Show that $p \wedge q \Rightarrow p \vee q$.

Proof: The corresponding truth table is

p	q	$p \wedge q$	$p \vee q$	$p \wedge q \rightarrow p \vee q$
T	T	T	T	T
T	F	F	T	T
F	T	F	T	T
F	F	F	F	T

Since the last column consists entirely of T's, the implication $p \wedge q \rightarrow p \vee q$ is a *tautology*, i.e. $\models p \wedge q \rightarrow p \vee q$ so that $p \wedge q \Rightarrow p \vee q$.

Remark 1.46. For statements related to logical implication, the final step of the truth table is always to evaluate the implication. The *comparison table*, which is a striped version of a truth table, lacks the last column since for implication, can be used. We know that as long as there is no $T \rightarrow F$, we know that the implication will be true.

Example 1.47. Show that $\sim p \Rightarrow p \rightarrow q$ using *comparison table*.

Solution. The corresponding comparison table is

p	q	$\sim p$	$p \rightarrow q$
T	T	F	T
T	F	F	F
F	T	T	T
F	F	T	T

Since for all different assignment of truth values, there is no $T \rightarrow F$, $\models \sim p \rightarrow (p \rightarrow q)$, by definition, $\sim p \Rightarrow (p \rightarrow q)$. $\square$

An argument is an assertion that the conjunction of several statements implies another statement as defined below.

Definition 1.12. For any statements $\phi_1, \phi_2, \dots, \phi_n$ and ψ ,

$$\begin{array}{ccc} & & \phi_1 \\ & & \phi_2 \\ \{ \phi_1, \phi_2, \dots, \phi_n \} / \therefore \psi & \text{or} & \vdots \\ & & \phi_n \\ & & \hline & & \therefore \psi \end{array}$$

are called **arguments** (or *deductive argument*, see https://en.wikipedia.org/wiki/Deductive_reasoning). We call $\{\phi_1, \phi_2, \dots, \phi_n\}$ the set of **hypotheses**, **premises** or **assumptions** and ψ the **conclusion** of the argument. It is customary to ignore the “bracket” for convenience [Herrmann, 29 Jan 2006].

When there is no hypothesis, an argument is written as $/ \therefore \psi$.

Definition 1.13. When the premises $\phi, \phi_1, \phi_2, \dots$ and ϕ_n are true, ψ is true, then we say the argument is **valid**. ψ is said to be *deduced* or *inferred* from the premises, and that ψ *follows logically from* or is a *logical consequence* of the premises or that ψ . An argument that is not valid is said to be **invalid** or *fallacious*.

Three methods for determining the validity of an argument are

Method 1 : using “truth table” or “comparison table”;

Method 2 : using laws of equivalences and implications;

Method 3 : https://en.wikipedia.org/wiki/Method_of_analytic_tableaux or “truth tree”.

Method 1: In a “truth table”, an argument is valid if for all truth assignments, the truth value of

$$\phi_1 \wedge \phi_2 \wedge \dots \wedge \phi_n \rightarrow \psi$$

are all true, i.e. the above proposition is a tautology. Otherwise, the argument is invalid. In a “comparison table”, an argument is valid when there is no truth assignment such that ψ is false but $\phi_1, \phi_2, \dots$ and ϕ_n are true. If there is **one case** where $\phi_1, \phi_2, \dots$ and ϕ_n are true but ψ is false, the argument is invalid. A mathematical formalism of method 1 is given by the deduction theorem below.

Theorem 1.48 (Deduction Theorem). *Let Γ be any finite (possibly empty) set of formulas, $\phi, \phi_1, \dots, \phi_n$ and ψ be formulas [Herrmann, 29 Jan 2006, Section 2.8].*

1. $\Gamma, \phi \models \psi$ iff $\Gamma \models \phi \rightarrow \psi$.
2. $\phi_1, \dots, \phi_n \models \psi$ iff $\phi_1 \wedge \dots \wedge \phi_i, \dots, \phi_n \models \psi$ where $1 < i \leq n$.

Based on this theorem, an argument $\phi_1, \dots, \phi_n / \therefore \psi$ is valid iff $\models \phi_1 \wedge \dots \wedge \phi_n \rightarrow \psi$; otherwise, the argument is invalid.

Example 1.49. Determine if $p \vee q / \therefore p \wedge q$ is valid.

Solution 1: The truth table for the argument is listed below.

p	q	$p \vee q$	$p \wedge q$	$(p \vee q) \rightarrow (p \wedge q)$
T	T	T	T	T
T	F	T	F	F
F	T	T	F	F
F	F	F	F	T

According to the last column the statement $(p \vee q) \rightarrow (p \wedge q)$ is not a tautology. By definition, the argument is invalid.

Solution 2: We don’t need to write down the whole comparison table to confirm that the argument is invalid, we only need to find **one case** where the conclusion $p \wedge q$ is false but the premise $p \vee q$ is true.

Suppose $v(p \wedge q) = F$, then we can have $v(p) = T$ and $v(q) = F$. However, $v(p \vee q) = T$.

We found one case where the premise is true but the conclusion is false. Hence the argument is invalid!

eg:3.1b

Example 1.50. Determine the validity of the argument $p \vee (q \vee r), \sim r / \therefore p \vee q$.

Solution: We will use the comparison table to check the validity of the argument.

p	q	r	$p \vee (q \vee r)$	$\sim r$	$p \vee q$
T	T	T	T	F	T
T	T	F	T	T	T
T	F	T	T	F	T
T	F	F	T	T	T
F	T	T	T	F	T
F	T	F	T	T	T
F	F	T	T	F	F
F	F	F	F	T	F

In each situation where the *premises are both true*, the *conclusion is also true*, so the argument is valid.

eg:3.1a

Example 1.51. Determine the validity of the argument $p \rightarrow q \vee \sim r, q \rightarrow p \wedge r / \therefore p \rightarrow r$ by using the comparison table.

Solution.

p	q	r	$p \rightarrow q \vee \sim r$	$q \rightarrow p \wedge r$	$p \rightarrow r$
T	T	T	T	T	T
T	T	F	T	F	F
T	F	T	F	T	T
T	F	F	T	T	F
F	T	T	T	F	T
F	T	F	T	F	T
F	F	T	T	T	T
F	F	F	T	T	T

The 4th row shows that it is possible for an argument of this form to have true premises and a false conclusion. Hence this form of argument is invalid.

Note: It is not necessary to show the whole table, it is sufficient if able to show the 4th row. □

t3:q2

Example 1.52 (Tutorial 1, Q16). Use comparison tables to determine whether the arguments are valid.

(a) $p \rightarrow (q \vee r), \sim q \vee \sim r / \therefore \sim p \vee \sim r$

Solution. The third row of the comparison table shows that it is possible for an argument to have true premises and false conclusion, this indicates that the argument is invalid.

p	q	r	$p \rightarrow (q \vee r)$	$\sim q \vee \sim r$	$\sim p \vee \sim r$
T	T	T	T	F	F
T	T	F	T	T	T
T	F	T	T	T	F
T	F	F	F	T	T
F	T	T	T	F	T
F	T	F	T	T	T
F	F	T	T	T	T
F	F	F	T	T	T

□

(b)

$$\begin{array}{l}
 (p \wedge q) \rightarrow \sim r \\
 p \vee \sim q \\
 \sim q \rightarrow p \\
 \hline
 \therefore \sim r
 \end{array}$$

Solution. In row 3 of the comparison table, when the conclusion is false, all the premises are true, this implies that the argument is invalid.

p	q	r	$(p \wedge q) \rightarrow \sim r$	$p \vee \sim q$	$\sim q \rightarrow p$	$\sim r$
T	T	T	F	T	T	F
T	T	F	T	T	T	T
T	F	T	T	T	T	F
T	F	F	T	T	T	T
F	T	T	T	F	T	F
F	T	F	T	F	T	T
F	F	T	T	T	F	F
F	F	F	T	T	F	T

□

Some sequence of English sentences with logical meaning and behaves like statements can be formalised and be analysed as “arguments” as illustrated in the examples below.

eg:3.5b

Example 1.53. Is the following argument valid?

If interest rates are going up, stock market prices will go down.

Interest rates are not going up.

Therefore stock market prices will not go down.

Solution: Let p : “Interest rates are going up”; q : “stock market prices will go down”. The argument can be written formally as $p \rightarrow q, \sim p / \therefore \sim q$.

It is not valid since the **third row** of the comparison table indicates that we have a case where the premises are true but the conclusion is false:

p	q	$p \rightarrow q$	$\sim p$	$\sim q$
T	T	T	F	F
T	F	F	F	T
F	T	T	T	F
F	F	T	T	T

t3:q3

Example 1.54 (Tutorial 1, Q17(a)). Write the symbolic form of each argument below and then determine its validity.

If Tom is not on team A, then Hua is on team B.

If Hua is not on team B, then Tom is on team A.

Therefore Tom is not on Team A or Hua is not on Team B.

Solution. Let p : “Tom is on team A”, q : “Hua is on team B”. The argument in formal form is

$$\sim p \rightarrow q, \sim q \rightarrow p / \therefore \sim p \vee \sim q.$$

p	q	$\sim p \rightarrow \sim q$	$\sim q \rightarrow \sim p$	$\sim p \vee \sim q$
T	T	T	T	F
T	F	T	T	T
F	T	T	T	T
F	F	F	F	T

The first row of the comparison table shows that it is possible for an argument to have true premises and false conclusion, this indicates that the argument is invalid. □

eg:3.7

Example 1.55. Show that the argument $\sim p \rightarrow F / \therefore p$ is valid using truth table.

Solution.

p	$(\sim p \rightarrow F) \rightarrow p$
T	T
F	T

Since $\models (\sim p \rightarrow F) \rightarrow p$, the argument is valid. □

Similar to the laws of logical equivalences (Theorem 1.23), the laws of logical implications can be obtained using the truth table method. thm:equiv-laws

Theorem 1.56 (Laws of Logical Implications). *Let F be contradiction. Given any atomic statements p, q and r , the following arguments are valid [Epp, 2020].*

1. *Contradiction Rule:*

$$\sim p \rightarrow F \models p$$

2. *Conjunction:*

$$p, q \models p \wedge q$$

3. *Specialisation:*

$$p \wedge q \models p$$

$$p \wedge q \models q$$

4. *Generalisation:*

$$p \models p \vee q$$

$$q \models p \vee q$$

5. *Elimination:*

$$p \vee q, \sim q \models p$$

$$p \vee q, \sim p \models q$$

6. *Modus Ponens (MP in short):*

$$p \rightarrow q, p \models q$$

7. *Modus Tollens (MT in short):*

$$p \rightarrow q, \sim q \models \sim p$$

8. *Transitivity:*

$$p \rightarrow q, q \rightarrow r \models p \rightarrow r$$

9. *Resolution:*

$$p \vee r, q \vee \sim r \models p \vee q$$

Proof: Use truth table methods. Only the resolution will be proved since it is the most complex (refer to [https://en.wikipedia.org/wiki/Resolution_\(logic\)](https://en.wikipedia.org/wiki/Resolution_(logic))):

p	q	r	$\underbrace{p \vee r}_{\phi_1}$	$\underbrace{q \vee \sim r}_{\phi_2}$	$\underbrace{p \vee q}_{\psi}$	$\phi_1 \wedge \phi_2 \rightarrow \psi$
T	T	T	T	T	T	T
T	T	F	T	T	T	T
T	F	T	F	T	T	T
T	F	F	T	T	T	T
F	T	T	T	T	T	T
F	T	F	T	T	T	T
F	F	T	F	F	F	T
F	F	F	T	F	F	T

Since $\models \phi_1 \wedge \phi_2 \rightarrow \psi$, by definition, $\phi_1, \phi_2 \models \psi$.

To apply the laws of logical implications to general propositions, we need the following theorems [Herrmann, 29 Jan 2006, Section 2.8].

Theorem 1.57 (Substitution of Equivalence). *If $\xi \equiv \phi_n$ and $\phi_1, \dots, \phi_{n-1}, \phi_n \models \psi$, then $\phi_1, \dots, \phi_{n-1}, \xi \models \psi$.*

Theorem 1.58 (“Partial Ordering” Theorem). *Let ϕ_i, ψ_j, ξ be formulas.*

1. $\phi_1, \dots, \phi_n \models \phi_i$ for each $i = 1, \dots, n$.

2. If $\phi_1, \dots, \phi_i \models \psi_j$, where $j = 1, \dots, p$, and $\psi_1, \dots, \psi_p \models \xi$, then $\phi_1, \dots, \phi_n \models \xi$.

Method 2: When an argument is valid, most of the time, it is possible to use on the laws of logical equivalences, laws of logical implication and the various substitution theorems, to show that the conclusion follows from the premises using the laws. However, this method does not directly allow one to check for invalidity. One can use laws of logical equivalences on $\phi_1 \wedge \dots \wedge \phi_n \rightarrow \psi$ and perform forward checking to find possible assignments to make it false.

Example 1.59. Show that the hypotheses

- ϕ_1 : “If John takes the computer course, then John stays in the hostel”
- ϕ_2 : “John does not stay in the hostel”
- ϕ_3 : “If John does not take the computer course, then John takes the language course or stay at home”
- ϕ_4 : “If John takes language course then John buys a motorcycle”
- ϕ_5 : “If John buys a car, then John does not buy motorcycle”
- ϕ_6 : “John buys a car”

lead to the conclusion ξ : “John stays at home”.

Solution: Let

- r_1 : John takes the computer course
- r_2 : John stays in the hostel
- r_3 : John takes the language course
- r_4 : John stays at home
- r_5 : John buys a motorcycle
- r_6 : John buys a car

The above argument can be formally written as

$$r_1 \rightarrow r_2, \sim r_2, \sim r_1 \rightarrow r_3 \vee r_4, r_3 \rightarrow r_5, r_6 \rightarrow \sim r_5, r_6 / \therefore r_4.$$

Instead of using a comparison table, we use the logical equivalence and logical implication rules together with the “Partial Ordering” Theorem to show that the above argument is valid.

$\phi_1 :$	$r_1 \rightarrow r_2$	premise
$\phi_2 :$	$\sim r_2$	premise
$\phi_3 :$	$\sim r_1 \rightarrow r_3 \vee r_4$	premise
$\phi_4 :$	$r_3 \rightarrow r_5$	premise
$\phi_5 :$	$r_6 \rightarrow \sim r_5$	premise
$\phi_6 :$	r_6	premise
$\psi_1 :$	$\sim r_1$	ϕ_1, ϕ_2 , Modus Tollens
$\psi_2 :$	$r_3 \vee r_4$	ϕ_3, ψ_1 , Modus Ponens
$\psi_3 :$	$\sim r_5$	ϕ_5, ϕ_6 , Modus Ponens
$\psi_4 :$	$\sim r_3$	ϕ_4, ψ_3 , Modus Tollens
$\xi :$	r_4	ψ_2, ψ_4 , Elimination

Example 1.60 (Tutorial 1, Q17(b)). Write the symbolic form of the argument below and then show that it is valid using laws of logical implications.

- If I graduate this semester, then I will have passed Calculus.
- If I do not study Calculus for 5 hours a week, then I will not pass Calculus.
- If I study Calculus for 5 hours a week, then I cannot play basketball.
- Therefore, if I play basketball, I will not graduate this semester.

Solution. Let p : “I graduate this semester”, q : “I will have passed Calculus”, r : “I study

Calculus for 5 hours a week”, s : “I play basketball”. The conclusion is $s \rightarrow \sim p$.

$\phi_1 :$	$p \rightarrow q$	premise
$\phi_2 :$	$\sim r \rightarrow \sim q$	premise
$\phi_3 :$	$r \rightarrow \sim s$	premise
$\psi_1 :$	$\sim r \vee \sim s$	ϕ_3 , implication law
$\psi_2 :$	$s \rightarrow \sim r$	ψ_1 , commutative law and implication law
$\psi_3 :$	$s \rightarrow \sim q$	ψ_2, ϕ_2 , transitivity
$\psi_4 :$	$\sim p \vee \sim \sim q$	ϕ_1 , implication law and double negative law
$\psi_5 :$	$\sim q \rightarrow \sim p$	ψ_4 , commutative law and implication law
$\therefore$	$s \rightarrow \sim p$	ψ_3, ψ_5 , implication law

□

1.6 Tableaux: Using Diagram to Check Validity

Method 1 (truth table or comparison table) is not applicable to propositions with too many atomic variables while method 2 (logical laws) cannot be used to show that an argument is invalid. Method 3 (tableaux) can be used to check relatively large arguments.

Method 3: Tableau (the plural form is **tableaux**) or **truth tree** can be used to check the validity of an argument by examining **the premises and the negation of the conclusion**. If they are **satisfiable**, the argument is invalid; If they are **unsatisfiable** or a *contradiction*, the argument is valid (recall Definition 1.9).

A **tableaux** proof is done by constructing a **tableau** or **truth tree** of which the nodes are labelled by formulas of the logic. In particular, every proof is begun with a tree consisting of a single node and then expanded by applying the procedure (https://en.wikipedia.org/wiki/Method_of_analytic_tableaux):

1. Pick an open leaf node. The leaf node in the initial chain is marked open.
2. Pick an applicable node on the branch above the selected node.
3. Apply the applicable node, which corresponds to expanding the tree below the selected leaf node based on expansion rule.

1. $P \wedge Q$ 2. P 1. 3. Q 1.	1. $P \vee Q$ 2. $P \quad Q$ 1.	1. E 2. logical equivalent form 3. of E from page 14

where E can be $\sim (P \wedge Q)$ (De Morgan), $\sim (P \vee Q)$ (De Morgan), $\sim (\sim P)$ (double negative), $P \rightarrow Q$ (implication), $\sim (P \rightarrow Q)$ (implication + De Morgan + double negative), $P \leftrightarrow Q$ (biconditional), $\sim (P \leftrightarrow Q)$ (biconditional + De Morgan).

4. For every newly created node that is both a literal/negated literal, and whose complement appears in a prior node on the same branch, mark the branch as closed. Mark all other newly created nodes as open.

As mentioned in the Section 1.4, tableaux method is tightly link to SAT/SMT solvers. So we will illustrate the use of tableaux method with Z3 SMT solver in the following examples.

Example 1.61. Determine the validity of the argument $p \vee (q \vee r), \sim r / \therefore p \vee q$ from Example 1.50 using truth tree/tableaux method and a SMT (Satisfiability Modulo Theories) solver.

Solution:	1. $p \vee (q \vee r)$	Premise
	2. $\sim r$	Premise
	3. $\sim (p \vee q)$	Negated conclusion
	4. $\sim p$	3 De Morgan
	5. $\sim q$	3 De Morgan
	6. $p \quad q \vee r$	1 disjunction
	7. $\otimes \quad q \quad r$	6 disjunction
		4, 6
		5, 7
		2, 7

```

1 (set-logic QF_UF) ; QF_UF supports Bool (propositional variables)
2 (declare-const p Bool)
3 (declare-const q Bool)
4 (declare-const r Bool)
5 (define-fun argument () Bool
6   (=> (and (or p (or q r)) (not r))
7     (or p q)))
8 (assert (not argument))
9 (check-sat) ; unsat means that there is no (p q r) making the argument invalid
10 (exit)

```


Example 1.62. Use truth tree/tableaux method and an SMT prover to show that the arguments below from Example 1.52 is invalid.

(a) $p \rightarrow (q \vee r), \sim q \vee \sim r / \therefore \sim p \vee \sim r$;

Solution: 1. 2. 3. 4. 5. 6. 7. 8.	$p \rightarrow (q \vee r) \checkmark$ $\sim q \vee \sim r \checkmark$ $\sim (\sim p \vee \sim r) \checkmark$ p r $\swarrow \quad \searrow$ $\sim q \quad \quad \sim r$ $\swarrow \quad \searrow$ $\sim p \quad q \vee r$ $\swarrow \quad \searrow$ $\quad q \quad r$ $\quad \swarrow \quad \searrow$ $\quad \quad \quad$	Premise Premise Negated conclusion 3 De Morgan+Double Negative 3 De Morgan+Double Negative 2 disjunction 1 disjunction 7 disjunction

$\otimes$ 5,6
 $\otimes$ 4,7
 $\otimes$ 6,8

The leaf r is open, so the argument is invalid.

```

1 (declare-const p Bool)
2 (declare-const q Bool)
3 (declare-const r Bool)
4 (define-fun argument () Bool
5   (=> (and (=> p (or q r)) (or (not q) (not r)))
6     (or (not p) (not r))))
7 (assert (not argument))
8 (check-sat)
9 (get-model) ; p=T, q=F, r=T only
10 (exit)
  
```

(b) $(p \wedge q) \rightarrow \sim r, p \vee \sim q, \sim q \rightarrow p / \therefore \sim r$.

Solution.

```

1 (declare-const p Bool)
2 (declare-const q Bool)
3 (declare-const r Bool)
4 (define-fun argument () Bool
5   (=> (and (=> (and p q) (not r)) (or p (not q)) (=> (not q) p))
6     (not r)))
7 (assert (not argument))
8 (check-sat)
9 (get-model)
10 (exit)
  
```

1.	$(p \wedge q) \rightarrow \sim r \checkmark$	Premise
2.	$p \vee \sim q \checkmark$	Premise
3.	$\sim q \rightarrow p \checkmark$	Premise
4.	$\sim(\sim r)$	Negate Conclusion
5.	$\begin{array}{c} \swarrow \quad \searrow \\ \sim(p \wedge q) \quad \sim r \end{array}$	1 implication
6.	$\begin{array}{c} \swarrow \quad \searrow \\ \sim p \quad \sim q \end{array}$	5 De Morgan
7.	$\begin{array}{cc} \swarrow \quad \searrow & \swarrow \quad \searrow \\ p \quad \sim q & q \quad p \end{array}$	2 disjunction; 3 implication
8.	$\begin{array}{ccc} \otimes & & \otimes \\ 6, 7 & & 6, 7 \\ & \swarrow \quad \searrow & \\ & q \quad p & \\ & \otimes \quad \otimes & \\ & 7, 8 \quad 6, 8 & \end{array}$	3 implication
There is an open leaf p , so the argument is invalid. □		

t3:q6

Example 1.63 (Tutorial 1, Q17(c)). Write the symbolic form of the argument below and then determine its validity.

If f is integrable, then g or h is differentiable.

If g is not differentiable, then f is not integrable but it is bounded.

If f is bounded, then g or h is differentiable.

Therefore, g is differentiable.

Solution: Let p : “ f is integrable”, q : “ g is differentiable”, r : “ h is differentiable”, s : “ f is bounded”. Then the argument can be written formally as

$$p \rightarrow (q \vee r), \sim q \rightarrow (\sim p \wedge s), s \rightarrow (q \vee r) / \therefore q.$$

Using Method 2, we assume conclusion $v(q) = F$ and try to find if it is possible to make

1. $v(p \rightarrow (q \vee r)) = T$

2. $v(\sim q \rightarrow (\sim p \wedge s)) = T$ requires $v(\sim p \wedge s) = T$ since $v(\sim q) = T$.

3. $v(s \rightarrow (q \vee r)) = T$

Item 2. suggests that $v(p) = F$, $v(s) = T$ and then Item 3. suggests $v(q \vee r) = v(r) = T$.

<i>Solution.</i>	1.	$p \rightarrow (q \vee r) \checkmark$	Premise
	2.	$\sim q \rightarrow (\sim p \wedge s) \checkmark$	Premise
	3.	$s \rightarrow (q \vee r) \checkmark$	Premise
	4.	$\sim q$	Negate Conclusion
		$\begin{array}{cc} & \swarrow \quad \searrow \\ q & \quad \quad \quad \sim p \wedge s \end{array}$	2 implication
	5.	$\otimes$	5 conjunction
	6.	$\otimes$	5 conjunction
	7.	$\begin{array}{cc} & \swarrow \quad \searrow \\ \sim s & \quad \quad \quad s \end{array}$	
	8.	$\otimes$	3 implication
	9.	$\otimes$	8 disjunction; 8 conjunction
	10.	$\begin{array}{cc} & \swarrow \quad \searrow \\ \sim p & \quad \quad \quad q \vee r \end{array}$	1 implication
	11.	$\begin{array}{cc} & \swarrow \quad \searrow \\ q & \quad \quad \quad r \end{array}$	8 disjunction
		$\otimes$	
		4, 11	

```

(declare-const p Bool)
(declare-const q Bool)
(declare-const r Bool)
(declare-const s Bool)
(define-fun argument () Bool
  (=> (and (=> p (or q r) )
        (=> (not q) (and (not p) s))
        (=> s (or q r)))
    q))
(assert (not argument))
(check-sat)
(get-model)
(exit)

```

□

1.7 Rules of Inference for Statements

Previous laws for “logical inference” are based on **semantics of logic**. In this section, we will venture into **syntactic implication** or https://en.wikipedia.org/wiki/Natural_deduction) which is going to be fundamental tools for mathematicians and computer scientists to conduct proving in future.

There are many syntactic proving software such as Lean 4 prover, Rocq prover, Isabelle prover, etc. We will be using Lean 4 prover, developed by Microsoft, and is becoming very popular among mathematicians since 2020(?) because it supports Unicode on VS Code editor and mathematicians are trying to integrate AI into it to help mathematicians perform proving.

Most of the syntactic proving software use **typed functional programming (FP)** (e.g. Standard ML, Haskell, Ocaml) [Harrison, 2009] to implement https://en.wikipedia.org/wiki/Curry%E2%80%93Howard_correspondence:

Logic	Proposition	Proof	Implication	Conjunction	Disjunction	T	F
FP	Type	Term	function type	Product Type	Sum Type	identity?	()?

Suggestions from students:

- More examples and clearer explanation on Fitch style proof. E.g. provide pointers to some YouTuber’s explanations like Attic Philosophy, William Rose, DG, etc.

There are many notations for natural deduction/syntactic implication:

- Gentzen (sequent calculus) syntax: Gallier [2015], Boolos et al. [2007, Chapter 14]).
- *box-and-line* syntax or *Jaśkowski nested box* syntax: Huth and Ryan [2004], Bornat [2005] (used in Jape, see <http://japeforall.org.uk>).
- https://en.wikipedia.org/wiki/Fitch_notation: Barwise and Etchemendy [1999], <http://forallx.openlogicproject.org/>.

We will be using the Fitch notation which is relatively easy to understand compare to Gentzen syntax used in computer science.

In natural deduction, we have such a collection of **rules of inference**. They allow us to infer a **conclusion** (a **formula**) from a set of **premises** by applying inference rules in succession (called a **proof**) [Huth and Ryan, 2004].

Let ϕ, ψ, ξ be any statements. Mathematicians have identified the following essential rules of inference for natural deduction [Bornat, 2005]:

1. $\wedge$ -introduction:	$\phi, \psi \vdash \phi \wedge \psi$
2. $\wedge$ -elimination:	$\phi \wedge \psi \vdash \phi$ or $\phi \wedge \psi \vdash \psi$
3. $\rightarrow$ -introduction or Conditional Proof (CP in short):	$\boxed{\phi, \dots, \psi} \vdash (\phi \rightarrow \psi)$
4. $\rightarrow$ -elimination:	$\phi \rightarrow \psi, \phi \vdash \psi$
5. $\vee$ -introduction:	$\phi \vdash \phi \vee \psi$ or $\psi \vdash \phi \vee \psi$
6. $\vee$ -elimination:	$\phi \vee \psi, \boxed{\phi, \dots, \xi}, \boxed{\psi, \dots, \xi} \vdash \xi$
7. $\neg$ -introduction or $\sim$ -introduction:	$\boxed{\phi, \dots, \perp} \vdash \sim \phi$
8. $\neg$ -elimination or $\sim$ -elimination:	$\phi, \sim \phi \vdash \perp$
9. $\perp$ -elimination:	$\perp \vdash \phi$

Note that 1. and 2. are called the *rules of conjunction*, 3. and 4. are called the *rules of implication*, 5. and 6. are called the *rules of disjunction*, 7. and 8. are called the *rules of negation* [Bornat, 2005].

In the following examples, we will show how to apply the 8 rules to perform inference.

eg:cp1

Example 1.64. Show that $p \rightarrow q \vdash p \rightarrow (p \wedge q)$.

Proof:

1	$p \rightarrow q$	premise
2	p	assumption
3	q	1,2 $\rightarrow$ -elimination
4	$p \wedge q$	2,3 $\wedge$ -introduction
5	$p \rightarrow p \wedge q$	2,4 $\rightarrow$ -introduction

Proof in Lean 4:

```

theorem eg1_syn01 (P Q : Prop):
  (P -> Q) -> (P -> (P /\ Q)) :=
  fun h1 : P -> Q =>
    fun h2 : P =>
      And.intro h2 (h1 h2)
/- Print types -/
#print eg1_syn01

```

Lines 2–4 serve to justify line 5, but they cannot be used in any subsequent line of the proof, they are closed off from the rest of the proof, but you are free to use line 5 as you need it.

eg:MT

Example 1.65 (Modus Tollens). Show that $p \rightarrow q, \sim q \vdash \sim p$.**Proof:**

1	$p \rightarrow q$	premise
2	$\sim q$	premise
3	p	assumption
4	q	1, 3 $\rightarrow$ -introduction
5	$\perp$	2, 4 $\neg$ -elimination
6	$\sim p$	$\neg$ -introduction

Proof in Lean 4:

```

/- Not X = X -> False -/
theorem modus_tollens:
  (P -> Q) -> (Not Q) -> (Not P) :=
  fun h1 : P -> Q =>
    fun h2 : Q -> False =>
      fun hp : P =>
        show False from h2 (h1 hp)

```

Reference: Huth and Ryan [2004, Section 1.2.2].

eg:transitivity

Example 1.66 (Hypothetical Syllogism (Transitivity)). Show that $p \rightarrow q, q \rightarrow r \vdash p \rightarrow r$.

Hypothetical Syllogism is very easy because $\rightarrow$ corresponds to function mapping of types in type functional programming.

Proof:

1	$p \rightarrow q$	premise
2	$q \rightarrow r$	premise
3	p	assumption
4	q	1, 3 $\rightarrow$ -elimination
5	r	2, 4 $\rightarrow$ -elimination
6	$p \rightarrow r$	3,5 $\rightarrow$ -introduction

Proof in Lean 4:

```

-- Hypothetical Syllogism
theorem transitivity (P Q R : Prop):
  (P -> Q) -> (Q -> R) -> (P -> R) :=
  fun h1: P -> Q =>
    fun h2: Q -> R =>
      fun hp: P =>
        show R from h2 (h1 hp)

```

tut1q20-2013

Example 1.67 (Tutorial 1, Q18). Show that $p \rightarrow q \vdash p \rightarrow (r \rightarrow (s \rightarrow q))$.*Solution.*

1	$p \rightarrow q$	premise
2	p	assumption
3	r	assumption
4	s	assumption
5	q	1,2, $\rightarrow$ -E
6	$s \rightarrow q$	4–5, $\rightarrow$ -I
7	$r \rightarrow (s \rightarrow q)$	3–6, $\rightarrow$ -I
8	$p \rightarrow (r \rightarrow (s \rightarrow q))$	2–7, $\rightarrow$ -I

Proof in Lean 4:

```

set_option linter.unusedVariables false
theorem tut1q18 (P Q R S : Prop):
  (P -> Q) -> (P -> R -> S -> Q) :=
  fun h1: P -> Q =>
    fun hp: P =>
      fun hr: R =>
        fun hs: S =>
          show Q from h1 hp

```

□

Example 1.68. Show that $p \wedge \sim s, q \rightarrow (r \rightarrow s) \vdash (p \rightarrow q) \rightarrow \sim r$.

Proof: Using Lean 4, we can prove using direct Curry-Howard correspondence on the left or the Lean 4 tactics on the right.

```

-- term mode
theorem syn02 (P Q R S : Prop):
  (P /\ Not S) -> (Q -> (R -> S)) ->
  (P -> Q) -> Not R :=
  fun h1 : (P /\ Not S) =>
  fun h2 : (Q -> (R -> S)) =>
  fun h3 : (P -> Q) =>
  fun hr : R =>
    (And.right h1) ((h2 (h3
      (And.left h1))) hr)
#print syn02

```

```

-- tactic mode (keyword: by)
theorem syn02b (P Q R S : Prop):
  (P /\ Not S) -> (Q -> (R -> S)) ->
  (P -> Q) -> Not R := by
  intro h1
  intro h2
  intro h3
  intro hr
  exact (And.right h1)
    ((h2 (h3 (And.left h1))) hr)
#print syn02b

```

<i>Solution.</i>	1	$p \wedge \sim s$	premise	$\square$
	2	$q \rightarrow (r \rightarrow s)$	premise	
	3	$p \rightarrow q$	assumption	
	4	p	1, $\wedge$ -elimination1	
	5	q	3,4 $\rightarrow$ -elimination	
	6	$r \rightarrow s$	2,5 $\rightarrow$ -elimination	
	7	r	assumption	
	8	s	6,7 $\rightarrow$ -elimination	
	9	$\sim s$	1, $\wedge$ -elimination2	
	10	$\perp$	8,9 $\neg$ -elimination	
	11	$\sim r$	10 $\neg$ -introduction	
	12	$(p \rightarrow q) \rightarrow \sim r$	3,11 $\rightarrow$ -introduction	

Example 1.69 (Disjunctive Syllogism). Show that $p \vee q, \sim p \vdash q$; $p \vee q, \sim q \vdash p$.

Proof:

1	$p \vee q$	premise
2	$\sim p$	premise
3	p	assumption
4	$\perp$	2, 3 $\neg$ -elimination
5	q	$\perp$ -elimination
6	q	assumption
7	q	
8	q	1,3-5,6-7 $\vee$ -elimination

Proof in Lean 4:

```

theorem disjunctive_syllogism:
  (P \/ Q) -> (Not P) -> Q :=
  fun h1: P \/ Q =>
  fun h2: P -> False =>
  Or.elim h1
    (fun hp: P =>
      (show Q from
        False.elim (h2 hp)))
    (fun hq: Q =>
      (show Q from hq))
#print disjunctive_syllogism

```

eg:3.3

Example 1.70. By using the rules of inference, show that $\sim p \vee q, q \rightarrow r, q \rightarrow s, p \vdash r \wedge s$.

	1	$\sim p \vee q$	premise	
	2	$q \rightarrow r$	premise	
	3	$q \rightarrow s$	premise	
	4	p	premise	
Solution.	5	$\sim p$	assumption	
	6	$\perp$	4,5 $\neg$ -elimination	
	7	$r \wedge s$	6 $\neg$ -introduction	
	8	q	assumption	
	9	r	2,8 $\rightarrow$ -elimination	
	10	s	3,8 $\rightarrow$ -elimination	
	11	$r \wedge s$	9,10 $\wedge$ -introduction	
	12	$r \wedge s$	1,5–7,8–11 $\vee$ -elimination	
				□

Mathematically, **semantic argument** (using truth table) and **natural deduction** (using type matching) are different. The former is defined through the concept of **valuation** and the later is defined through a **proof** (solely using rules). They are linked through the following theorems [Johnstone, 1987, Barnes and Mack, 1975].

Definition 1.14. A logic system is said to be *sound* if $\phi_1, \dots, \phi_n \vdash \psi$ implies $\phi_1, \dots, \phi_n \models \psi$.

Definition 1.15. A logic system is said to be *adequate* if $\phi_1, \dots, \phi_n \models \psi$ implies $\phi_1, \dots, \phi_n \vdash \psi$.

Theorem 1.71 (The Soundness Theorem). *If $\phi_1, \dots, \phi_n \vdash \psi$, then $\phi_1, \dots, \phi_n \models \psi$.*

Theorem 1.72 (The Completeness Theorem). *If $\phi_1, \dots, \phi_n \models \psi$, then $\phi_1, \dots, \phi_n \vdash \psi$.*

Proof. Refer to Johnstone [1987] or Forster [2003, p. 94] Theorem 20. □

Theorem 1.73 (The Decidability Theorem). *Given propositions/statements $\phi_1, \dots, \phi_n, \psi$. There is an **algorithm** that determines whether or not $\phi_1, \dots, \phi_n \vdash \psi$.*

Chapter 2

Predicate Logic

ch:pred-logic

Propositional logic is **too simple when we need to describe infinitely many statements!**

For example, when we want to compare two things, e.g. $x = y$, $2 - 1 = 1$, $1 = 3 - 2$, etc. It is **infeasible to list all of them**.

We need **predicates**, a function that takes in data and return true or false, as well as **quantifiers** to turn predicates into propositions (which summarised “similar” propositions into one or a few propositions).

The simplest logic system to include propositional logic and handle **predicates** and **quantified statements** is the **first order logic**. Similar to propositional logic, there are two key parts in first order logic, i.e. syntax and semantics. The *syntax* determines which collections of symbols are legal expressions in first-order logic, while the *semantics* determine the meanings behind these expressions.

2.1 Syntax of Quantified Statements

sec:syntax

The *(well-formed) formula* or *quantified statement* for first order logic are the generalisation of Definition 1.1.

def:formula

def:wff

Definition 2.1. [Manin, 1977, Chapter 1] Terms and well-formed formulae are defined recursively as follows.

- **terms**: they represent objects, i.e.
 - Variables (normally denoted $x, y, x_1, x_2, \dots, y_1, \dots$) and constants (normally denoted as $a, b, a_1, a_2, \dots$) are **atomic terms**.
 - If f is a function (which returns a value) of *degree* (or *arity*) r and $t_1, \dots, t_r$ are terms, then $f(t_1, \dots, t_r)$ is a **term**.
- **formulae**: they are relations or functions, i.e.
 - If P is a *relation* of *degree* (or *arity*) r and $t_1, \dots, t_r$ are terms, then $P(t_1, \dots, t_r)$ is an **(atomic) formula**.
 - If ϕ and ψ are formulae (abbreviated notation!), and x is a variable, then the expressions

$$\sim (\phi), (\phi) \wedge (\psi), (\phi) \vee (\psi), (\phi) \rightarrow (\psi), (\phi) \leftrightarrow (\psi), \forall x(\phi), \exists x(\phi)$$

are **formulae**.

def:2.1

Remark 2.1. A predicate with no arguments is an atomic statement, as in propositional logic (Definition 1.1).

def:formula


```
(* https://www.rbjones.com/rbjpub/logic/log021.htm *)
infix 1 Iff
infixr 2 Impl
infix 3 Disj
infix 4 Conj
datatype term = V of string
and formula =
  P of string * term list
| Neg of formula
| Conj of formula * formula
| Disj of formula * formula
| Impl of formula * formula
| Iff of formula * formula
| FAll of string * formula;
```

Similar to propositional calculus, conventions have been developed about the precedence of the logical operators, to avoid the need to write parentheses in some cases:

- $\sim$ is evaluated first
- $\wedge$ and $\vee$ are evaluated next
- Quantifiers $\forall$ and $\exists$ are evaluated next
- $\rightarrow$ and $\leftrightarrow$ are evaluated last.

Remark 2.2. In English, the “predicate” is the part of the sentence that tells us something about the subject. For example,

$\underbrace{\text{James}}_{\text{subject}} \underbrace{\text{is a student at UTAR.}}_{\text{predicate}}$

Example 2.3.

She lives in the city. neither
 live(x, y): x lives in y predicate (of degree / arity 2)
 live(Mary, Austin): “Mary lives in Austin.” .. proposition (or predicate of degree / arity 0)

Example 2.4. The expression $\forall x \forall y (P(f(x)) \rightarrow \sim (P(x) \rightarrow Q(f(y), x, z)))$ is a formula where f is a unary function symbol, P a unary predicate symbol, and Q a ternary predicate symbol.

Example 2.5. Consider the following logic formula

$$\underbrace{(\forall x. \underbrace{P(f(c, y))}_{\phi_2} \wedge \underbrace{Q(g(g(x)), y))}_{\phi_1}}_{\psi_1} \rightarrow (\underbrace{\exists z. \forall w. \sim R(z, w)}_{\psi_1})$$

we can study the formula formally by applying the rules mentioned in Definition [2.1](#). We can also identify def:wff

- function symbols: f, c, g .
- predicate symbols: P, Q, R . ■

Example 2.6. The expression $\forall x x \rightarrow$ is *not* a formula. It is just a string of symbols!

In a formula, a *variable* may occur *free* or *bound*. Intuitively, a variable is free in a formula if it is not quantified: in $\forall y P(x, y)$, variable x is free while y is bound.

def:freevar

Definition 2.2. The free and bound variables of a formula are defined inductively as follows.

- Atomic formulas. If ϕ is an atomic formula then x is free in ϕ if and only if x occurs in ϕ . Moreover, there are no bound variables in any atomic formula.
- Negation. x is free in $\sim \phi$ if and only if x is free in ϕ . x is bound in $\sim \phi$ if and only if x is bound in ϕ .

- Binary connectives. x is free in $(\phi \rightarrow \psi)$ if and only if x is free in either ϕ or ψ . x is bound in $(\phi \rightarrow \psi)$ if and only if x is bound in either ϕ or ψ . The same rule applies to any other binary connective $\wedge, \vee, \leftrightarrow$ in place of $\rightarrow$.
- Quantifiers. x is free in $\forall y\phi$ if and only if x is free in ϕ and x is a different symbol from y . Also, x is bound in $\forall y\phi$ if and only if x is y or x is bound in ϕ . The same rule holds with $\exists$ in place of $\forall$.

Example 2.7. Determine if the variables x, y, z, w are free, bound or neither in the formula $\forall x\forall y(P(x) \rightarrow Q(x, f(x), z))$.

Solution:

- x and y are bound variables
- z is a free variable
- w is neither because it does not occur in the formula.

Freeness and boundness can be also specialised to specific occurrences of variables in a formula.

Example 2.8. For the formula $P(x) \rightarrow \forall x Q(x)$, the first occurrence of x is free while the second is bound, i.e. the x in $P(x)$ is free while the x in $\forall x Q(x)$ is bound.

2.2 Formal versus Informal Language

sec:informal

We use informal languages such as English, Chinese, etc. in real life. Informal languages are *difficult for computer to process*. To ask computer to help us solve problems, we need to express human knowledge in a formal language. The results are represented formally by computer and they will normally be converted to informal languages for human comprehension.

First, let us define the classes of formal statements that we constantly encounter in mathematics.

def:2.3

Definition 2.3. Let $P(x)$ and $Q(x)$ be any predicates with free variable x .

1. A *universal statement* is a formula of the form

$$\forall x P(x)$$

It corresponds to these English sentences: “For all x $P(x)$ ”, “For every x $P(x)$ ” and “For any x $P(x)$ ”.

2. An *existential statement* is a formula of the form

$$\exists x Q(x).$$

It corresponds to these English sentences: “There is/exists an x such that $P(x)$ ”, “There is at least one x such that $P(x)$ ”, “Some x satisfies P ”, “ $P(x)$ for some x ”, etc.

3. A *universal conditional statement* is a formula of the form

$$\forall x (P(x) \rightarrow Q(x)).$$

It corresponds to the English sentence “For every x , if $P(x)$ then $Q(x)$ ”.

Formal to Informal Translation

It is generally straightforward to translate from formal language into English, since you can just turn each logical operator directly into an English word or phrase. We will be introducing some notations which are defined in later sections: $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{R}$ which are the set of natural numbers, integers and real numbers, respectively.

ingAComputer2e

Example 2.9. Let

- $A(x) \quad := x \text{ is an animal.}$
- $C(x) \quad := x \text{ is a cat.}$
- $S(x) \quad := x \text{ is small.}$
- $GP(x) \quad := x \text{ is a good pet.}$

Translate the universal conditional statement

$$\forall x(S(x) \wedge A(x) \rightarrow GP(x)) \quad (*)$$

into English literally.

Solution: We could write the formal sentence $(*)$ into any of the following English sentences.

1. For every thing, if that thing is small and that thing is an animal, then that thing is a good pet.

This is graceless English, but at least it's comprehensible and correct. The style can be improved:

2. Everything which is small and which is an animal is a good pet.

Even better would be:

3. Small animals make good pets.

Such stylistic improvements in the English are optional. It is important to be sure that the effort to improve the literary style doesn't affect the meaning, but this is a question of proper usage of natural language, not of formal logic.

eg:2.6

Example 2.10. Rewrite the following formal statements in a variety of equivalent but more informal ways without using the formal symbols $\forall$ and $\exists$.

1. $\forall x(x \in \mathbb{R} \rightarrow x^2 \geq 0)$.

Solution: Any of the following English sentences are acceptable.

- All real numbers have *nonnegative* squares.
- Every real number has a nonnegative square.
- Any real number has a nonnegative square.
- x has a nonnegative square, for each real number x .
- The square of any real number is nonnegative.

2. $\forall x(x \in \mathbb{R} \rightarrow x^2 \neq -1)$.

Solution:

- All real numbers have square not equal to -1 .
- No real numbers have square equal to -1 .

3. $\exists m(m \in \mathbb{Z} \wedge m^2 = m)$.

Solution. • There is an integer whose square is equal to itself.

- We can find at least one integer equal to its own square.
- $m^2 = m$, for some integer m .
- Some integer equals its own square.
- Some integers equal their own squares.

□

eg:2.9

Example 2.11. Rewrite the following formal statement in a variety of informal ways. Do not use quantifiers or variables.

$$\forall x((x \in \mathbb{R}) \wedge (x > 2) \rightarrow (x^2 > 4)).$$

Solution.

- If a real number is greater than 2 then its square is greater than 4.
- Whenever a real number is greater than 2, its square is greater than 4.
- The square of any real number that is greater than 2 is greater than 4.
- The squares of all real numbers greater than 2 are greater than 4.

□

Informal to Formal Translation

It is sometimes tricky to translate from an informal sentence into a formal sentence, precisely because the informal sentence may not correspond obviously to the logical quantifiers and operators.

eg:bird

Example 2.12. Translate the sentence ‘Some birds can fly’ into logic. Let $B(x)$ mean ‘ x is a bird’ and $F(x)$ mean ‘ x can fly’.

Solution: ‘Some birds can fly’ is translated as

$$\exists x(B(x) \wedge F(x)).$$

If we translate this back to ‘unpolished’ English, it would be ‘there is something which is a bird and this thing can fly.’

Remark 2.13. Warning! A common pitfall is to translate ‘Some birds can fly’ as

$$\exists x(B(x) \rightarrow F(x)) \quad \textbf{Wrong translation!}$$

To see why this is wrong, let p be a frog that somehow got into the universe. Now $B(p)$ is false, so $B(p) \rightarrow F(p)$ is true (remember $F \rightarrow F \equiv T$). This is just saying ‘If that frog were a bird then it would be able to fly’, which is true; it doesn’t mean the frog actually is a bird, or that it actually can fly. However, we have now found a value of x — namely the frog p — for which $B(x) \rightarrow F(x)$ is true, and that is enough to satisfy $\exists x(B(x) \rightarrow F(x))$, which is different from $\exists x(B(x) \wedge F(x))$ which requires us to find a flying bird from birds.

Example 2.14. Using the notations from Example 2.9, the English sentences in on the left side of the table below can be translated into logic formulae on the right side of the table:

Informal	Formal
Cats are animals.	$\forall x(C(x) \rightarrow A(x))$
Cats are small.	$\forall x(C(x) \rightarrow S(x))$
Cats are small animals.	$\forall x(C(x) \rightarrow S(x) \wedge A(x))$
Small animals are good pets.	$\forall x(S(x) \wedge A(x) \rightarrow GP(x))$

Many mathematical quantified statements can be written in one of the form in Definition 2.3. However, mathematical writings normally have many implicitly quantified statements, i.e. statements that do not contain the “keywords” *all* or *every* or *any* or *each* or *exist*, which often tell us what class of quantified statements they belong. Hence, we have to look at the sentence and understand the context.

Example 2.15. Translate the sentence “Every integer is rational.” into a quantified statement.

Solution. $\forall x(x \in \mathbb{Z} \rightarrow x \in \mathbb{Q})$. □

eg:2.12

Example 2.16. Consider the statement “There is an integer that is both prime and even.” Let $P(n)$ be “ n is prime” and $E(n)$ be “ n is even.” Use the notation $P(n)$ and $E(n)$ to rewrite this statement formally.

Solution. $\exists n(n \in \mathbb{Z} \wedge P(n) \wedge E(n))$. □

Example 2.17. In calculus courses, the letter x is mostly always used to indicate a real number, hence the sentence below

“If $x > 2$ then $x^2 > 4$.”

is interpreted to mean the same as the statement

“For all real numbers x , if $x > 2$ then $x^2 > 4$.”

or formally

$$\forall x((x \in \mathbb{R} \wedge x > 2) \rightarrow (x^2 > 4)). \text{ or } \forall x(x \in \mathbb{R} \rightarrow (x > 2) \rightarrow (x^2 > 4)).$$

Example 2.18. Translate the following statement formally.

If a number is an integer, then it is a real number.

Solution: The clue that indicates its universal quantification comes from the presence of the indefinite article “a”. Hence, this statement is a universal statement: $\forall x(x \in \mathbb{Z} \rightarrow x \in \mathbb{R})$.

Example 2.19. Let $E(x)$ be the predicate “ x is even”. Translate the following statement

The number 24 can be written as a sum of two even integers.

into a formal logic formula.

Solution. The statement has an implicit existential quantification:
 $\exists m \exists n(E(m) \wedge E(n) \wedge 24 = m + n)$. □

eg:2.24

Example 2.20. Rewrite the following statements formally using quantifiers and variables.

1. Somebody trusts everybody.

Solution. Let $H(x)$ be the predicate “ x is a human” and $T(x, y)$ be “ x trusts y ”. Then using the quantified conjunction from Theorem 2.47, we have

$$\exists x \underbrace{(H(x))}_q \wedge \underbrace{\forall y(H(y) \rightarrow T(x, y))}_{Q(x)} \equiv \exists x \forall y \underbrace{(H(x))}_q \wedge \underbrace{(H(y) \rightarrow T(x, y))}_{Q(x)}.$$

□

2. Given any positive number, there is another positive number that is smaller than the given number.

Solution. $\forall x \exists y((x \in \mathbb{R}) \wedge (x > 0) \rightarrow (y \in \mathbb{R}) \wedge (y > 0) \wedge (y < x))$. □

3. Any even integer equal twice some other integer.

Solution. $\forall x \exists y ((x \in 2\mathbb{Z}) \rightarrow (y \in \mathbb{Z}) \wedge (x = 2y))$. □

4. For every real number x and every real number y , if x is positive and y is negative, then the product of x and y is negative.

Solution. $\forall x \forall y ((x \in \mathbb{R}) \wedge (y \in \mathbb{R}) \rightarrow ((x > 0) \wedge (y < 0) \rightarrow (xy < 0)))$. □

eg:2.25

Example 2.21. Let S be the set of students in UTAR, M be the set of movies that have ever been released, and let $V(s, m)$ denote “student s has seen movie m ”. Translate each of the following statements into English.

1. $\forall s \exists m ((s \in S) \rightarrow ((m \in M) \wedge V(s, m)))$.

Solution. Every student has seen a movie. □

2. $\exists s \exists t \exists m ((s \in S) \wedge (t \in S) \wedge (m \in M) \wedge (s \neq t) \wedge V(s, m) \wedge V(t, m))$.

Solution. There are 2 different students who have seen the same movie. □

3. $\exists s \exists t ((s \in S) \wedge (t \in S) \wedge (s \neq t) \wedge \forall m ((m \in M) \rightarrow (V(s, m) \rightarrow V(t, m))))$.

Solution. There is a student who has seen all the movie that the other student has seen. □

Often the real difficulty in translating English into logic is in figuring out what the English says, or what the speaker meant to say. For example, many people make statements like ‘All people are not rich’. What this statement actually says is

$$\forall x \sim R(x),$$

where the universe is the set of people and $R(x)$ means ‘ x is rich’. What is usually meant, however, by such a statement is

$$\sim \forall x R(x),$$

that is, it is not true that all people are rich (alternatively, not all people are rich). The intended meaning is equivalent to

$$\exists x \sim R(x).$$

Such problems of *ambiguity* or *incorrect grammar* in English cannot be solved mathematically and they will not be considered in this subject. However, they do illustrate one of the benefits of mathematics: simply translating a problem from English into formal logic may expose confusion or misunderstanding.

2.3 Semantics of Quantified Statements

Semantics of logic is a theory to determine if a formula is true or false as an extension to propositional logic.

In the semantics of propositional logic, we assigned a truth value to each atom.

In the semantics of predicate logic, the truth values of the predicate $P(t_1, \dots, t_n)$ depends on the terms t_i . We need a **model** to **interpret** the terms.

Definition 2.4. A *model* (or *structure*) M consists of a nonempty set D that forms the *domain of discourse* (or *universe of discourse*) and an *interpretation* $(\cdot)^M$, which is a mapping such that

- Each function symbol f of degree n is assigned a function f^M from D^n to D . In particular, each constant symbol is assigned an individual in the domain of discourse.
- Each predicate symbol P of degree n is assigned a relation P^M over D^n , i.e. $P^M \subset D^n$.

The evaluation/interpretation of predicate logic formulas on models is formalised below.

Definition 2.5. Let M be a model and t be a term without variables. Then t^M , the **interpretation of t in M** , is given as follows:

- if t is a constant c , then $t^M = c^M$, $c^M \in D$;
- if t is a variable x , then $t^M = \sigma(x)$, where σ is a variable assignment;
- if t is an n -ary function $f(t_1, \dots, t_n)$, then $t^M = f^M(t_1^M, \dots, t_n^M)$.

For predicates and logical connectives, the interpretation goes through the process **T-schema**:

- Atomic formulae.
 1. A formula $P(t_1, \dots, t_n)$ is associated the value T or F depending on whether $(t_1^M, \dots, t_n^M) \in P^M \subset D^n$.
 2. A formula $t_1 = t_2$ is assigned T if $t_1^M = t_2^M$, i.e. they evaluate to the same object of the domain of discourse.
- Logical connectives. A formula in the form $\sim \phi$, $\phi \rightarrow \psi$, etc. is evaluated according to the truth table for the connective in question, as in propositional logic.
- Existential quantifiers. A formula $\exists x \phi(x)$ is true according to M and σ if there exists an evaluation σ' of the variables that only differs from σ regarding the evaluation of x and such that ϕ is true according to the interpretation M and the variable assignment σ' . This formal definition captures the idea that $\exists x \phi(x)$ is true if and only if there is a way to choose a value for x such that $\phi(x)$ is satisfied.
- Universal quantifiers. A formula $\forall x \phi(x)$ is true according to M and σ if $\phi(x)$ is true for every pair composed by the interpretation M and some variable assignment σ' that differs from σ only on the value of x . This captures the idea that $\forall x \phi(x)$ is true if every possible choice of a value for x causes $\phi(x)$ to be true.

If a formula does not contain free variables, and so is a sentence, then the initial variable assignment does not affect its truth value. In other words, a sentence is true according to M and σ if and only if is true according to M and every other variable assignment σ' .

Example 2.22. [Huth and Ryan, 2004, Example 2.19] Consider the sentence “None of Alma’s lovers’ lovers love her.” Let “Alma” can be a constant a , and the concept “ x loves y ” can be a binary predicate $L(x, y)$.

1. Formalise this sentence.
2. Consider the model M defined by $D = \{p, q, r\}$, $a^M = p$, $L^M = \{(p, p), (q, p), (r, p)\}$. Determine the truth value of the sentence under this model M .

Solution:

1. We can formalise the sentence as

$$\forall x \forall y (L(x, a) \wedge L(y, x) \rightarrow \sim L(y, a)).$$

Intuitively, $L(x, a)$ says that x is Alma's lover, and $L(y, x)$ says that “ y loves x ”, so $L(x, a) \wedge L(y, x)$ says that y is one of Alma's lovers' lovers. The formula $\sim L(y, a)$ says that y does not love Alma, and the quantifiers make sure this is true for any x, y .

2. In order to interpret the formula, we need a variable assignment σ . Consider $\sigma = \{(x, p), (y, q)\}$. Then

$$\begin{aligned} & (\forall x \forall y (L(x, a) \wedge L(y, x) \rightarrow \sim L(y, a)))^M(\sigma) \\ & \equiv L^M(\sigma(x), a^M) \wedge L^M(\sigma(y), \sigma(x)) \rightarrow \sim L^M(\sigma(y), a^M) \\ & \equiv L^M(p, p) \wedge L^M(q, p) \rightarrow \sim L^M(q, p) \equiv T \wedge T \rightarrow \sim T \equiv T \rightarrow F \equiv F. \end{aligned}$$

The formula is false under this model.

To show that a universal formula is true, we need to make sure that it is true for all possible variable assignments.

t2:q11

Example 2.23 (Tutorial 2, Q10). Suppose the model M of the predicate $P(x, y)$ has a universe of discourse $D = \{1, 2, 3\}$ and corresponding Boolean function (an equivalent way to describe a relation) P^M . Write the propositions $\exists y \sim P(1, y)$ and $\forall x P(x, 2)$ using disjunctions and conjunctions.

Solution. With the assignment $\sigma = \{(y, 1), (y, 2), (y, 3)\}$, we have

$$(\exists y \sim P(1, y))^M(\sigma) = \sim P^M(1, 1) \vee \sim P^M(1, 2) \vee \sim P^M(1, 3)$$

With the assignment $\sigma = \{(x, 1), (x, 2), (x, 3)\}$, we have

$$(\forall x P(x, 2))^M(\sigma) = P^M(1, 2) \wedge P^M(2, 2) \wedge P^M(3, 2). \quad \square$$

Remark 2.24. The inspirations of the above examples are

- universal quantifier is like the generalisation of conjunction: $\forall x \phi(x) \equiv \phi(x_0) \wedge \phi(x_1) \wedge \dots$;
- existential quantifier is like the generalisation of disjunction: $\exists x \phi(x) \equiv \phi(x_0) \vee \phi(x_1) \vee \dots$.

Logical Equivalence

The logical equivalence in propositional logic is simple, we only need to verify it to be true for all assignments (using truth table or comparison table).

The logical equivalence in predicate logic needs to be verified to be true for all models!

The laws of logical equivalences (Theorem [I.23](#) and “substitution principle” (Theorem [I.26](#)) are also valid for predicate calculus [Aho and Ullman, 1992]. However, we need to replace the term “atomic statement” with “formula” and introduce the extended notion of logical equivalence.

Definition 2.6. [Chiswell and Hodges, 2007] Let ϕ and ψ be two formulae with free variables $x_1, \dots, x_n$, they are *logically equivalent*, i.e. $\phi \equiv \psi$ if $\models \forall x_1 \dots \forall x_n (\phi \leftrightarrow \psi)$.

In addition to laws of logical equivalences from propositional logic, the following laws for quantified statements.

ntified:de-morgan

Theorem 2.25. [Chiswell and Hodges, 2007, Chapter 7] Let ϕ and ψ be any formulas **with** free variable x (and y). Let ξ be any formula **without** free variable x . Then

- (a) $\sim \forall x \phi \equiv \exists x \sim \phi$; (Generalised de Morgan law)
- (b) $\sim \exists x \phi \equiv \forall x \sim \phi$; (Generalised de Morgan law)
- (c) $\forall x(\phi \wedge \psi) \equiv (\forall x \phi) \wedge (\forall x \psi)$; (Quantified conjunctive law)
- (d) $\exists x(\phi \vee \psi) \equiv (\exists x \phi) \vee (\exists x \psi)$; (Quantified disjunctive law)
- (e) $\forall x \forall y \phi \equiv \forall y \forall x \phi$; (Universal quantifiers swapping law)
- (f) $\exists x \exists y \phi \equiv \exists y \exists x \phi$; (Existential quantifiers swapping law)
- (g) Let $\phi[y/x]$ be the replacement of all occurrence of x with y .
 $\forall x \phi \equiv \forall y \phi[y/x]$; $\exists x \phi \equiv \exists y \phi[y/x]$; (Variable renaming laws)
- (h) $\xi \equiv \forall x \xi \equiv \exists x \xi$; (Independent quantifier law)
- (i) $\forall x(\xi \wedge \psi) \equiv \xi \wedge (\forall x \psi)$; $\exists x(\xi \wedge \psi) \equiv \xi \wedge (\exists x \psi)$; $\forall x(\xi \vee \psi) \equiv \xi \vee (\forall x \psi)$; $\exists x(\xi \vee \psi) \equiv \xi \vee (\exists x \psi)$.
..... (Free variable laws)

Proof. Refer to Chiswell and Hodges [2007, Chapter 7] and Gallier [2015, Chapter 5]. □

t2:q12

Example 2.26 (Tutorial 2, Q5). Determine whether $\exists x(P(x) \vee Q(x))$ and $\exists x P(x) \vee \exists x Q(x)$ have the same truth value. Explain.

Solution. Under any interpretation/model with domain D , $\exists x(P(x) \vee Q(x))$ is true means that $P(s) \vee Q(s) \equiv T$ for some $s \in D$. By definition, either $P(s) \equiv T$ or $Q(s) \equiv T$. This means $\exists x P(x) \equiv T$ or $\exists x Q(x) \equiv T$. Hence, $\exists x P(x) \vee \exists x Q(x)$ is true.

Conversely, if $\exists x P(x) \vee \exists x Q(x) \equiv T$, then $\exists x P(x) \equiv T$ or $\exists x Q(x) \equiv T$. This means, there is an s_1 such that $P(s_1) \equiv T$, which implies $P(s_1) \vee Q(s_1) \equiv T \vee Q(s_1) \equiv T$. Similarly, there is an s_2 such that $Q(s_2) \equiv T$, which implies $Q(s_2) \vee P(s_2) \equiv T$. In either case, $\exists x(P(x) \vee Q(x)) \equiv T$.

Under any interpretation/model with domain D , if $\exists x(P(x) \vee Q(x))$ is false, then $P(a) \vee Q(a)$ is false for any a in the model which means that $P(a) \equiv F$ and $Q(a) \equiv F$. This implies $\exists x P(x) \equiv F$ and $\exists x Q(x) \equiv F$, i.e. $\exists x P(x) \equiv F \vee \exists x Q(x) \equiv F$.

Conversely, if $\exists x P(x) \vee \exists x Q(x) \equiv F$, then $\exists x P(x) \equiv F$ and $\exists x Q(x) \equiv F$. This means, there is nothing in the model such that $P(a)$ and $Q(a)$ is true, i.e. $P(a) \vee Q(a) \equiv F$, i.e. $\exists x(P(x) \vee Q(x)) \equiv F$. □

Corollary 2.27. Let ϕ and ψ be any predicates with free variable x . Let ξ be any formula **without** free variable x . Then

1. $\sim (\forall x(\phi \rightarrow \psi)) \equiv \exists x(\phi \wedge \sim \psi)$;
2. $\exists x(\phi \rightarrow \psi) \equiv (\forall x \phi) \rightarrow (\exists x \psi)$.
3. $\forall x(\xi \rightarrow \psi) \equiv \xi \rightarrow (\forall x \psi)$

Proof. Proof:

1. $LHS \equiv \exists x \sim (\phi \rightarrow \psi) \equiv \exists x \sim (\sim \phi \vee \psi) \equiv \exists x(\sim \sim \phi \wedge \sim \psi) \equiv RHS$
2. $LHS \equiv \exists x(\sim \phi \vee \psi) \equiv (\exists x \sim \phi) \vee (\exists x \psi) \equiv \sim (\forall x \phi) \vee (\exists x \psi) \equiv RHS$
3. $LHS \equiv \forall x(\sim \xi \vee \psi) \equiv (\forall x \sim \xi) \vee (\forall x \psi) \equiv \sim (\exists x \xi) \vee (\forall x \psi) \equiv \sim \xi \vee (\forall x \psi) \equiv RHS$

□

t2:q27

Example 2.28 (Tutorial 2, Q3). Show that the statements $\exists x P(x) \wedge \exists x Q(x)$ and $\exists x(P(x) \wedge Q(x))$ are not logically equivalent in general.

Solution. Consider the model M with the universe of discourse $\mathbb{Z}$ and $P(x)$ be “ $x > 0$ ” and $Q(x)$ be “ $x < 0$ ”.

$\exists x P(x) \wedge \exists x Q(x)$ is true whereas $\exists x P(x) \wedge Q(x)$ is false. $\square$

t2:q28

Example 2.29 (Tutorial 2, Q4). Determine whether the statements $\forall x P(x) \wedge \exists x Q(x)$ and $\forall x \exists y (P(x) \wedge Q(y))$ are logically equivalent.

Solution. Under any interpretation/model with domain D , if $\forall x P(x) \wedge \exists x Q(x)$ is true, $P(x)$ is true for all x and $Q(x)$ is true from some $x = y$. So for all x , there is a y such that $P(x)$ is true and $Q(y)$ is true. Hence, $\forall x \exists y (P(x) \wedge Q(y))$ is true.

On the other hand, if $\forall x \exists y (P(x) \wedge Q(y))$ is true, then for all x , there is a y such that $P(x)$ is true and $Q(y)$ is true, so $\forall x, P(x)$ and $\exists y Q(y)$ are true. Hence, $(\forall x P(x)) \wedge (\exists y Q(y)) \equiv T$. Changing the symbol y to x leads a logical equivalence between the two statements. $\square$

A variation of Definition 1.2 is stated below.

def:2.5

Definition 2.7. Let ϕ and ψ be formulae with free variable x . Consider a statement of the form

$$\forall x (\phi \rightarrow \psi). \quad (2.1)$$

e:forall-imply

1. Its negation is	$\exists x (\phi \wedge \sim \psi).$
2. Its contrapositive form is	$\forall x (\sim \psi \rightarrow \sim \phi).$
3. Its converse form is	$\forall x (\psi \rightarrow \phi).$
4. Its inverse form is	$\forall x (\sim \phi \rightarrow \sim \psi).$

Remark 2.30. A universal conditional statement (2.1) is logically equivalent to its contrapositive form but not logically equivalent to its converse form and inverse form.

eg:2.14a

Example 2.31. Write formal negations for the statement “Every prime number is odd.” using the predicate $\text{prime}(n)$ for “ n is prime” and $\text{odd}(n)$ for “ n is odd”.

Solution: $\sim \forall n (\text{prime}(n) \rightarrow \text{odd}(n)) \equiv \exists n (\text{prime}(n) \wedge \sim \text{odd}(n))$

We need to exercise special care to avoid mistakes when writing negations of statements that are given informally. One way to avoid error is to rewrite the statement formally and take the negation using the formal rule as demonstrated in the example below.

eg:2.16b

Example 2.32. Write an informal negation for the statement

If a computer program has more than 100,000 lines, then it contains a bug.

Solution. Let $\text{program}(x)$ be “ x is a computer program”, $\text{many\_lines}(x)$ be “ x has more than 100,000 lines” and $\text{bug}(x)$ be “ x has a bug.”

- Formal version: $\forall x (\text{program}(x) \wedge \text{many\_lines}(x) \rightarrow \text{bug}(x)).$
- Formal negation: $\exists x (\text{program}(x) \wedge \text{many\_lines}(x) \wedge \sim \text{bug}(x)).$
- Informal negation: There is at least one computer program that has more than 100,000 lines and does not contain a bug.

$\square$

eg:2.18

Example 2.33. Write the contrapositive, converse and inverse for the following statement:

If a real number is greater than 2, then its square is greater than 4.

Solution. Let the model be real number system. The formal version of this statement is

$$\forall x ((x > 2) \rightarrow (x^2 > 4)).$$

1. Contrapositive: $\forall x ((x^2 \leq 4) \rightarrow (x \leq 2)).$

If the square of a real number is less than or equal to 4, then the number is less than or equal to 2.

2. Converse: $\forall x((x^2 > 4) \rightarrow (x > 2))$.

If the square of a real number is greater than 4, then the number is greater than 2.

3. Inverse: $\forall x((x \leq 2) \rightarrow (x^2 \leq 4))$.

If a real number is less than or equal to 2, then the square of the number is less than or equal to 4.

□

Remark 2.34. If we want to specify the domain D of x , then (2.1) will be of the form $\forall x(x \in D \rightarrow (\phi \rightarrow \psi))$ and the contrapositive, converse form and inverse form will be of the forms $\forall x(x \in D \rightarrow (\sim \psi \rightarrow \sim \phi))$, $\forall x(x \in D \rightarrow (\psi \rightarrow \phi))$ and $\forall x(x \in D \rightarrow (\sim \phi \rightarrow \sim \psi))$ respectively.

t2:q23

Example 2.35 (Tutorial 2, Q30). Write a negation for the following statement:

For all real numbers $y > 0$, there exists a real number $z > 0$ such that if $a - z < x < a + z$ then $L - y < f(x) < L + y$.

Solution. Formally

$$\forall y((y > 0) \rightarrow (\exists z((z > 0) \wedge (a - z < x) \wedge (x < a + z) \rightarrow (L - y < f(x)) \wedge (f(x) < L + y))))).$$

It's negation is

$$\exists y((y > 0) \wedge (\forall z((z > 0) \rightarrow ((a - z < x) \wedge (x < a + z) \rightarrow ((L - y \geq f(x)) \vee (f(x) \geq L + y)))))).$$

Informally, the negation is “There is a real numbers $y > 0$ such that for all real numbers $z > 0$, $a - z < x < a + z$ and either $L - y \geq f(x)$ or $f(x) \geq L + y$.” □

The CNF (Definition 1.10) plays an important role in SAT solver for propositional logic.

The predicate logic CNF for SMT solver requires one to transform a formula to prenex normal form (PNF) and then to CNF.

Definition 2.8. A formula of the predicate calculus is in *prenex¹ normal form* if it is written as a string of quantifiers (referred to as the *prefix*) followed by a quantifier-free part (referred to as the *matrix*).

Theorem 2.36. [Rautenberg, 2010, Theorem 4.2] Every formula in predicate logic is logically equivalent to a formula in prenex normal form.

Example 2.37. Given quantifier-free formulae $\phi(y)$, $\psi(z)$, and $\rho(x)$ then

$$\forall x \exists y \forall z (\phi(y) \vee (\psi(z) \rightarrow \rho(x)))$$

is in prenex normal form with matrix $\phi(y) \vee (\psi(z) \rightarrow \rho(x))$, while

$$\forall x ((\exists y \phi(y)) \vee ((\exists z \psi(z)) \rightarrow \rho(x)))$$

is logically equivalent but not in prenex normal form.

Example 2.38. Write logical equivalent form of $\sim \forall x \exists y \exists z P(x, y, z)$ in prenex normal form.

Solution: $\sim \forall x \exists y \exists z P(x, y, z) \equiv \exists x \sim \exists y \exists z P(x, y, z)$ [Generalised de Morgan law (GDM)]
 $\equiv \exists x \forall y \sim \exists z P(x, y, z)$ [GDM]
 $\equiv \exists x \forall y \forall z \sim P(x, y, z)$ [GDM]

eg:2.20

Example 2.39 (Tutorial 2, Q28). Derive the following rule using laws of equivalence:

$$\sim (\forall x(x \in D \rightarrow (\forall y(y \in E \rightarrow P(x, y)))) \equiv \exists x \exists y(x \in D \wedge (y \in E \wedge \sim P(x, y)))$$

Solution. $\sim (\forall x(x \in D \rightarrow (\forall y(y \in E \rightarrow P(x, y))))$
 $\equiv \exists x \sim (x \in D \rightarrow (\forall y(y \in E \rightarrow P(x, y))))$ [Generalised de Morgan law]
 $\equiv \exists x \sim (\sim (x \in D) \vee (\forall y(y \in E \rightarrow P(x, y))))$ [Implication law]
 $\equiv \exists x((x \in D) \wedge \sim (\forall y(y \in E \rightarrow P(x, y))))$ [de Morgan & Double Negation]
 $\equiv \exists x((x \in D) \wedge (\exists y \sim (\sim (y \in E) \vee P(x, y))))$ [de Morgan & Implication Law]
 $\equiv \exists x((x \in D) \wedge (\exists y((y \in E) \wedge \sim P(x, y))))$ [de Morgan & Double Negation]
 $\equiv \exists x \exists y((x \in D) \wedge ((y \in E) \wedge \sim P(x, y)))$ [Quantified conjunction]
 $\square$

Example 2.40 (Tutorial 2, Q29). Show that $\forall x[(C(x) \wedge \exists y(T(y) \wedge L(x, y))) \rightarrow \exists y(D(y) \wedge B(x, y))] \equiv \forall x \forall y \exists z[(C(x) \wedge T(y) \wedge L(x, y)) \rightarrow (D(z) \wedge B(x, z))]$. [Barwise and Etchemendy, 1999, p324]

Solution.
 $\forall x[(C(x) \wedge \exists y(T(y) \wedge L(x, y))) \rightarrow \exists y(D(y) \wedge B(x, y))]$
 $\equiv \forall x[(C(x) \wedge \exists y(T(y) \wedge L(x, y))) \rightarrow \exists z(D(z) \wedge B(x, z))]$ [Variable renaming law]
 $\equiv \forall x [\exists y(C(x) \wedge (T(y) \wedge L(x, y))) \rightarrow \exists z(D(z) \wedge B(x, z))]$ [Quantified conjunction]
 $\equiv \forall x [\sim \exists y(C(x) \wedge (T(y) \wedge L(x, y))) \vee \exists z(D(z) \wedge B(x, z))]$ [implication law]
 $\equiv \forall x [\forall y \sim (C(x) \wedge (T(y) \wedge L(x, y))) \vee \exists z(D(z) \wedge B(x, z))]$ [Generalised de Morgan law]
 $\equiv \forall x [\forall y [\sim (C(x) \wedge (T(y) \wedge L(x, y))) \vee \exists z(D(z) \wedge B(x, z))]]$ [Quantified disjunction]
 $\equiv \forall x [\forall y [\exists z [\sim (C(x) \wedge (T(y) \wedge L(x, y))) \vee (D(z) \wedge B(x, z))]]]$ [Quantified disjunction]
 $\equiv \forall x \forall y \exists z [(C(x) \wedge T(y) \wedge L(x, y)) \rightarrow (D(z) \wedge B(x, z))]$ [implication law]
 $\square$

The following are some semantic examples of two-quantifier statements.

eg:2.20

Example 2.41. Let $P(x, y)$ denote “ $x + y = 0$ ”. What are the truth values of the quantified statements $\exists y \forall x P(x, y)$ and $\forall x \exists y P(x, y)$?

- $\exists y \forall x P(x, y)$ F

Solution. Informally: “Can we find a number y such that for every number x , $x + y = 0$?”

Answer: There is no number y such that $x + y = 0$ for all numbers x . $\square$

- $\forall x \exists y P(x, y)$ T

Solution. Informally: “For every number x , can we find a number y such that $x + y = 0$?”

For every number x , “there is” (i.e. “we can find”) a number $y = -x$ such that $x + y = 0$. $\square$

Example 2.41 demonstrates that $\exists y \forall x P(x, y) \neq \forall x \exists y P(x, y)$ and this shows that the order of quantifiers is IMPORTANT!

eg:2.22

Example 2.42. Let $Q(x, y)$ be the statement “ $x + y = x - y$ ”. If model M of this formula consist the universe of discourse $D = \mathbb{Z}$ and the operations are the normal operations associated with $\mathbb{Z}$. Determine the truth values of the following formula.

1. $\forall y Q(1, y)$ is false since there are many values of y for which $1 + y \neq 1 - y$.

2. $\exists x \exists y Q(x, y)$ is true since we can take $x = y = 0$.
3. $\forall x \exists y Q(x, y)$ is true since we can take $y = 0$ and for each x .
4. $\exists y \forall x Q(x, y)$ is true since there is a $y = 0$.
5. $\forall y \exists x Q(x, y)$ is false since $x + 2 = x - 2$ has no solution when $y = 2$.

t2:q26

Example 2.43 (Tutorial 2, Q9). What are the truth values of $\exists y \forall x (y \geq x)$ and $\forall x \exists y (y \geq x)$ if they are interpreted over the model with the domain of nonnegative integers?

Solution. $\exists y \forall x (y \geq x)$ is false because for every non-negative integer y , there is an integer x such that $y < x$. For example, take $x = y + 1$.

$\forall x \exists y (y \geq x)$ is true because for every non-negative integer x , there is an integer y such that $y < x$. For example, take $y = x + 1$. □

2.4 Logical Implication and Arguments for Quantified Statements (Topic 2b)

sec:argument

To extend the notion of argument from propositional logic to predicate logic, we need very complicated mathematics as Open Logic Project [2019] illustrates. In addition to the laws of logical equivalences (Theorem 1.23), “substitution principle” (Theorem 1.26), we have the following theorems which imply the laws of logical implications (Theorem 1.56).

Theorem 2.44. [Open Logic Project, 2019, Chapter 12] Let $\Gamma = \{\phi_1, \dots, \phi_n\}$. If $\Gamma \subset \Gamma'$ and $\Gamma \models \psi$, then $\Gamma' \models \psi$.

It basically says that if ϕ is true for “preconditions” Γ , it is also true in a larger system Γ' containing Γ .

Theorem 2.45 (Semantic Deduction Theorem). Let $\Gamma = \{\phi_1, \dots, \phi_n\}$. $\Gamma \cup \{\phi\} \models \psi$ iff $\Gamma \models \phi \rightarrow \psi$.

Theorem 2.46. [Open Logic Project, 2019, Chapter 12] If $P(x)$ is a formula with one free variable x and s and a are closed terms. Then

$$P(s) \models \exists x P(x); \quad \forall x P(x) \models P(a).$$

For statements related to quantifiers, we have the following laws.

Proposition 2.1. Let ϕ be a formula with free variable x and a and s are terms free with respect to x in ϕ . Then we have the following laws.

1. *Universal instantiation (or specialisation):* $\forall x \phi \Rightarrow \phi[a/x]$, here $[a/x]$ means “ a replaces x ”. In particular, when ϕ is $P(x)$, we have $\forall x P(x) \Rightarrow P(a)$.
2. *Universal generalisation:* If the term a in $\phi[a/x]$ is arbitrary, then $\phi[a/x] \Rightarrow \forall x \phi$.
3. *Existential instantiation (or specialisation):* $\exists x \phi \Rightarrow \phi[s/x]$, here $[s/x]$ means “ s replaces x ”. In particular, when ϕ is $P(x)$, we have $\exists x P(x) \Rightarrow P(s)$.
4. *Existential generalisation:* For a particular term s , $\phi[s/x] \Rightarrow \exists x \phi$.

The following theorem is a few logical implication involving quantified statements with single variable which can be shown using the laws mentioned above.

Theorem 2.47. Let p and q be two formula without free variable x , $P(x)$ and $Q(x)$ be formula with free variable x . Then

1. $\exists x(P(x) \wedge Q(x)) \Rightarrow (\exists x P(x)) \wedge (\exists x Q(x))$
2. $(\forall x P(x)) \vee (\forall x Q(x)) \Rightarrow \forall x (P(x) \vee Q(x))$
3. $(\exists x P(x)) \rightarrow (\forall x Q(x)) \Rightarrow \forall x (P(x) \rightarrow Q(x))$
4. $\forall x (P(x) \rightarrow Q(x)) \Rightarrow (\forall x P(x)) \rightarrow (\forall x Q(x))$

Proof:

	$\phi_1 : \exists x(P(x) \wedge Q(x))$	premise
	$\psi_1 : P(s) \wedge Q(s)$	ϕ_1 , existential instantiation
	$\psi_2 : P(s)$	ψ_1 , specialisation
1.	$\psi_3 : Q(s)$	ψ_1 , specialisation
	$\psi_4 : \exists x P(x)$	ψ_2 , existential generalisation
	$\psi_5 : \exists x Q(x)$	ψ_3 , existential generalisation
	$\therefore \exists x P(x) \wedge \exists x Q(x)$	ψ_4, ψ_5 , conjunction

Remark: $\Leftarrow$ is not true because we have $(\exists x(x < 0)) \wedge (\exists x(x > 0)) \equiv T \wedge T \equiv T$ but $\exists x((x < 0) \wedge (x > 0)) \equiv F$.

	$\phi_1 : (\forall x P(x)) \vee (\forall x Q(x))$	premise
	$\psi_1 : (\forall x P(x)) \vee (\forall y Q(y))$	ϕ_1 , variable renaming law
	$\psi_2 : \forall y(\forall x P(x) \vee Q(y))$	ψ_1 , free variable equivalence
2.	$\psi_3 : \forall x P(x) \vee Q(t)$	ψ_2 , universal instantiation
	$\psi_4 : \forall x(P(x) \vee Q(t))$	ψ_3 , free variable law
	$\psi_5 : P(t) \vee Q(t)$	ψ_4 , universal instantiation
	$\psi_6 : \forall x(P(x) \vee Q(x))$	ψ_5 , universal generalisation
	$\phi_1 : (\exists x P(x)) \rightarrow (\forall x Q(x))$	premise
	$\psi_1 : \sim \exists x P(x) \vee (\forall x Q(x))$	ϕ_1 , implication law
	$\psi_2 : (\forall x \sim P(x)) \vee (\forall x Q(x))$	ψ_1 , generalised de Morgan
	$\psi_3 : (\forall x \sim P(x)) \vee (\forall y Q(y))$	ψ_2 , change of variable
3.	$\psi_4 : \forall y((\forall x \sim P(x)) \vee Q(y))$	ψ_3 , free variable equivalence
	$\psi_5 : (\forall x \sim P(x)) \vee Q(t)$	ψ_4 , universal instantiation
	$\psi_6 : \sim P(t) \vee Q(t)$	ψ_5 , universal instantiation
	$\psi_7 : P(t) \rightarrow Q(t)$	ψ_6 , implication law
	$\therefore \forall x(P(x) \rightarrow Q(x))$	ψ_7 , universal generalisation

4. Inspire by the syntactic proof in https://en.wikipedia.org/wiki/Universal_generalization, the deduction should be something like:

	$\phi_1 : \forall x(P(x) \rightarrow Q(x))$	premise
	$\phi_2 : \forall x P(x)$	premise
	$\psi_1 : P(t) \rightarrow Q(t)$	ϕ_1 UI
	$\psi_2 : P(t)$	ϕ_2 UI
	$\psi_3 : Q(t)$	ψ_1, ψ_2 MP
	$\psi_4 : \forall x Q(x)$	ψ_3 UG

↓

$$\frac{\phi_1 : \forall x(P(x) \rightarrow Q(x)) \quad \text{premise}}{\psi_5 : \forall xP(x) \rightarrow \forall xQ(x) \quad \phi_2, \psi_4 \text{ Semantic Deduction Theorem}}$$

Example 2.48 (Tutorial 2, Q21). Show that $\sim (\forall x(P(x) \rightarrow Q(x))) \Rightarrow \exists x(P(x) \wedge \sim Q(x))$.

Solution.

$$\begin{array}{ll} \phi_1 : \sim (\forall x(P(x) \rightarrow Q(x))) & \text{premise} \\ \hline \psi_1 : \sim (\forall x(\sim P(x) \vee Q(x))) & \phi_1, \text{ implication law} \\ \psi_2 : \exists x(\sim \sim P(x) \wedge \sim Q(x)) & \psi_1, \text{ Generalised de Morgan} \\ \therefore \exists x(P(x) \wedge \sim Q(x)) & \psi_2, \text{ double negation} \end{array}$$

□

book:p93:q2-30

Example 2.49 (Tutorial 2, Q26). Show that the following argument is valid.

$$\frac{\begin{array}{l} \exists x(F(x) \wedge S(x)) \rightarrow (\forall y(M(y) \rightarrow W(y))) \\ \exists y(M(y) \wedge \sim W(y)) \end{array}}{\therefore \forall x(F(x) \rightarrow \sim S(x))}$$

Solution.

$$\begin{array}{ll} \phi_1 : \exists x(F(x) \wedge S(x)) \rightarrow (\forall y(M(y) \rightarrow W(y))) & \text{premise} \\ \phi_2 : \exists y(M(y) \wedge \sim W(y)) & \text{premise} \\ \hline \psi_1 : \sim \forall y \sim (M(y) \wedge \sim W(y)) & \phi_2, \text{ double negation \& generalised de Morgan} \\ \psi_2 : \sim \forall y(\sim M(y) \vee W(y)) & \psi_1, \text{ de Morgan \& double negation} \\ \psi_3 : \sim \forall y(M(y) \rightarrow W(y)) & \psi_2, \text{ de Morgan \& implication law} \\ \psi_4 : \sim (\exists x(F(x) \wedge S(x))) & \phi_1, \psi_3, \text{ Modus Tollens} \\ \psi_5 : \forall x(\sim F(x) \vee \sim S(x)) & \psi_4, \text{ generalised de Morgan \& de Morgan} \\ \therefore \forall x(F(x) \rightarrow \sim S(x)) & \psi_5, \text{ implication law} \end{array}$$

□

t4:q3

Example 2.50 (Tutorial 2, Q20). For the following arguments, state which are valid and which are invalid. Justify your answers.

1. All healthy people eat an apple a day. John is not a healthy person. Therefore John does not eat an apple a day.

Solution. Let $H(x)$ be “ x is a healthy person” and $A(x)$ be “ x eats an apple a day.” Formally, we have

$$\forall x(H(x) \rightarrow A(x)), \sim H(\text{John}) / \therefore \sim A(\text{John})$$

Consider a model M with $D = \{j, a\}$, where j means John and a means Adam, $H^M = \{a\}$ and $A^M = \{a, j\}$. Hence, the interpretation is

$$\begin{array}{ll} (\forall x(H(x) \rightarrow A(x)))^M & \equiv (H^M(a) \rightarrow A^M(a)) \wedge (H^M(j) \rightarrow A^M(j)) \\ & \equiv (T \rightarrow T) \wedge (F \rightarrow T) \equiv T \\ (\sim H(\text{John}))^M & \equiv \sim H^M(j) \equiv \sim F \equiv T \\ \hline (\sim A(\text{John}))^M & \equiv \sim A^M(j) \equiv \sim T \equiv F. \end{array}$$

This indicates that we have $T \rightarrow F$ in an argument, hence, the argument is invalid. □

2. Every student who studies discrete mathematics is good at logic. John studies discrete mathematics. Therefore John is good at logic.

Solution. Let $D(x)$ be “student x studies discrete mathematics” and $L(x)$ be “ x is good at logic”. Formally, we have

$$\forall x(D(x) \rightarrow L(x)), D(\text{John}) / \therefore L(\text{John})$$

The logical implications:

ϕ_1	$\forall x(D(x) \rightarrow L(x))$	premise
ϕ_2	$D(\text{John})$	premise
ψ_1	$D(\text{John}) \rightarrow L(\text{John})$	ϕ_1 , universal instantiation
$\therefore$	$L(\text{John})$	ϕ_2, ψ_1 , MP

Hence, the argument is valid. □

t4:q7b

3. No heavy object is cheap. XYZ is not a heavy object. Therefore XYZ is cheap.

Solution. Let $H(x)$ be “ x is a heavy object” and $C(x)$ be “ x is cheap.” Formally, we have

$$\forall x(H(x) \rightarrow \sim C(x)), \sim H(\text{XYZ}) / \therefore C(\text{XYZ})$$

Consider a model M with $D = \{h, d\}$, where h means a house and d means a diamond, $H^M = \{h\}$ and $C^M = \{d\}$ and $(\text{XYZ})^M = d$. Hence, the interpretation is

$(\forall x(H(x) \rightarrow \sim C(x)))^M$	$\equiv (H^M(h) \rightarrow \sim C^M(h)) \wedge (H^M(d) \rightarrow \sim C^M(d))$
	$\equiv (T \rightarrow (\sim F)) \wedge (F \rightarrow (\sim F)) \equiv T$
$(\sim H(\text{XYZ}))^M$	$\equiv \sim H^M(d) \equiv \sim F \equiv T$
$(C(\text{XYZ}))^M$	$\equiv C^M(d) \equiv F$

This indicates that we have $T \rightarrow F$ in an argument, hence, the argument is invalid. □

eg:3.14

Example 2.51. Rewrite the following argument using quantifiers, variables, and predicate symbols:

If a number is even, then its square is even.

k is a particular number that is even.

$\therefore k^2$ is even.

Is the argument valid? Why?

Solution. Let $E(x)$ be the predicate “ x is even”. Then

$\phi_1 :$	$\forall x(E(x) \rightarrow E(x^2))$	premise
$\phi_2 :$	$E(k)$	premise
$\psi_1 :$	$E(k) \rightarrow E(k^2)$	ϕ_1 , universal instantiation
$\therefore$	$E(k^2)$	ϕ_2, ψ_1 , modus ponens

The argument is valid. □

2.5 Satisfiability Modulo Theories (SMT)

sec:smt

Satisfiability (Definition ^{def:SAT}1.9), a problem of determining whether a formula expressing a constraint has a solution, is a fundamental problem in theoretical computer science [Björner and de Moura, 2009].

Many-sorted first-order logic is a commonly used formalism and framework for formulating SMT problems. A **(many-sorted) signature** is composed of a set of *sorts* (data types), a set of *function symbols*, and a set of *predicate symbols*. Each function symbol f has associated with it an arity of the form $\sigma_1 \times \cdots \times \sigma_n \rightarrow \sigma$, where $\sigma_1, \dots, \sigma_n, \sigma$ are sorts. If $n = 0$, we say f is a *constant symbol*. Similarly, each predicate symbol p has associated with it an arity of the form $\sigma_1 \times \cdots \times \sigma_n$. If $n = 0$, we say p is a propositional symbol. [de Moura and Björner, 2009]

We assume a set of *variables* X , where each variable is associated with a sort. A *term* t with sort σ has the form x or $f(t_1, \dots, t_n)$, where x is a variable with sort σ , and f is a function symbol with arity $\sigma_1 \times \cdots \times \sigma_n \rightarrow \sigma$, where for each $i \in \{1, \dots, n\}$, t_i has sort σ_i . An *atom* is of the form $p(t_1, \dots, t_n)$ where p is a predicate symbol with arity $\sigma_1, \dots, \sigma_n$, and for each $i \in \{1, \dots, n\}$, t_i is a term with sort σ_i . A *formula*

$$\phi = \text{atom} \mid \sim \phi_0 \mid \phi_0 \wedge \phi_1 \mid \phi_0 \vee \phi_1 \mid \phi_0 \rightarrow \phi_1 \mid \phi_0 \leftrightarrow \phi_1 \mid (\forall x : \sigma. \phi_0) \mid (\exists x : \sigma. \phi_0)$$

where ϕ_0, ϕ_1 are smaller formulas. Note that it is similar to Definition ^{def:wff}2.1 where the variable x has no type while here the variable x has a type σ . A Σ -*formula* ϕ is a formula where each symbol in ϕ is in Σ . We say a variable x is *free* in formula ϕ if it is not bound by any quantifier $\exists, \forall$. A **sentence** is a formula **without free variables**. We use $\text{vars}(\phi)$ to denote the set of free variables in ϕ . A *quantifier-free formula* is a formula not containing $\exists$ or $\forall$. [de Moura and Björner, 2009]

A **structure** M for a signature Σ and variables X consists of non-empty domains $|M|_\sigma$ for each sort in Σ , for each $x \in X$ with sort σ , $M(x) \in |M|_\sigma$, for each function symbol f with arity $\sigma_1 \times \cdots \times \sigma_n \rightarrow \sigma$, $M(f)$ is a total map from $|M|_{\sigma_1} \times \cdots \times |M|_{\sigma_n} \rightarrow |M|_\sigma$, and for each predicate symbol p with arity $\sigma_1 \times \cdots \times \sigma_n$, $M(p)$ is a subset of $|M|_{\sigma_1} \times \cdots \times |M|_{\sigma_n}$. The **interpretation** of a term t is given by $M[[x]] = M(x)$ and $M[[f(t_1, \dots, t_n)]] = M(f)(M[[t_1]], \dots, M[[t_n]])$. We assume that, for each sort σ , the equality $=_\sigma$ is a builtin predicate symbol with arity $\sigma \times \sigma$ that does not occur in any signature and for every structure M , $M(=_\sigma)$ is the identity relation over $|M|_\sigma \times |M|_\sigma$. As a notational convention, we will always omit the subscript. We use $M\{x \mapsto \nu\}$ to denote a structure where the variable symbol x with sort σ is interpreted as $\nu \in |M|_\sigma$, and all other variables, function and predicate symbols have the same interpretation as in M . Given a formula ϕ and a structure M , satisfaction $M \models \phi$ is defined as:

- $M \models p(t_1, \dots, t_n)$ iff $\langle M[[t_1]], \dots, M[[t_n]] \rangle \in M(p)$
- $M \models \sim \phi$ iff $M \not\models \phi$
- $M \models \phi_0 \vee \phi_1$ iff $M \models \phi_0$ or $M \models \phi_1$
- $M \models \phi_0 \wedge \phi_1$ iff $M \models \phi_0$ and $M \models \phi_1$
- $M \models (\exists x : \sigma. \phi)$ iff $M\{x \mapsto \nu\} \models \phi$ for some $\nu \in |M|_\sigma$
- $M \models (\forall x : \sigma. \phi)$ iff $M\{x \mapsto \nu\} \models \phi$ for all $\nu \in |M|_\sigma$

A formula ϕ is **satisfiable** if there is a structure M s.t. $M \models \phi$, and is **valid** if for all structures M , $M \models \phi$. A structure M satisfies a *set of formulas* S ($M \models S$) if $M \models \phi$ for every $\phi \in S$. [de Moura and Björner, 2009]

Related: <https://ocw.mit.edu/courses/6-001-structure-and-interpretation-of-computer-programs-spring-2005/pages/readings/>

A theory is a set of sentences. More formally, a Σ -**theory** is a collection of sentences over a signature Σ . Given a theory T , we say ϕ is satisfiable modulo T if $T \cup \{\phi\}$ is satisfiable. We use $M \models_T \phi$ to denote $M \models \{\phi\} \cup T$.

Example 2.52. Let Σ be the signature containing the symbols $0, 1, +, -$ and $<$, and $\mathbb{Z}$ be the structure that interprets these symbols in the usual way over the integers, then the theory of linear arithmetic is the set of first-order sentences that are true in $\mathbb{Z}$.

Let Ω be a class of structures over a signature Σ , then we use $Th(\Omega)$ to denote the set of all sentences ϕ over Σ such that $M \models \phi$ for every M in Ω . We say the satisfiability problem for theory T is **decidable** if there is a procedure $\mathfrak{S}$ that checks whether any quantifier-free formula is satisfiable or not. In this case, we say $\mathfrak{S}$ is a decision procedure for T . [de Moura and Bjørner, 2009]

Theorem 2.53. *The satisfiability problem for predicate logic is undecidable.*

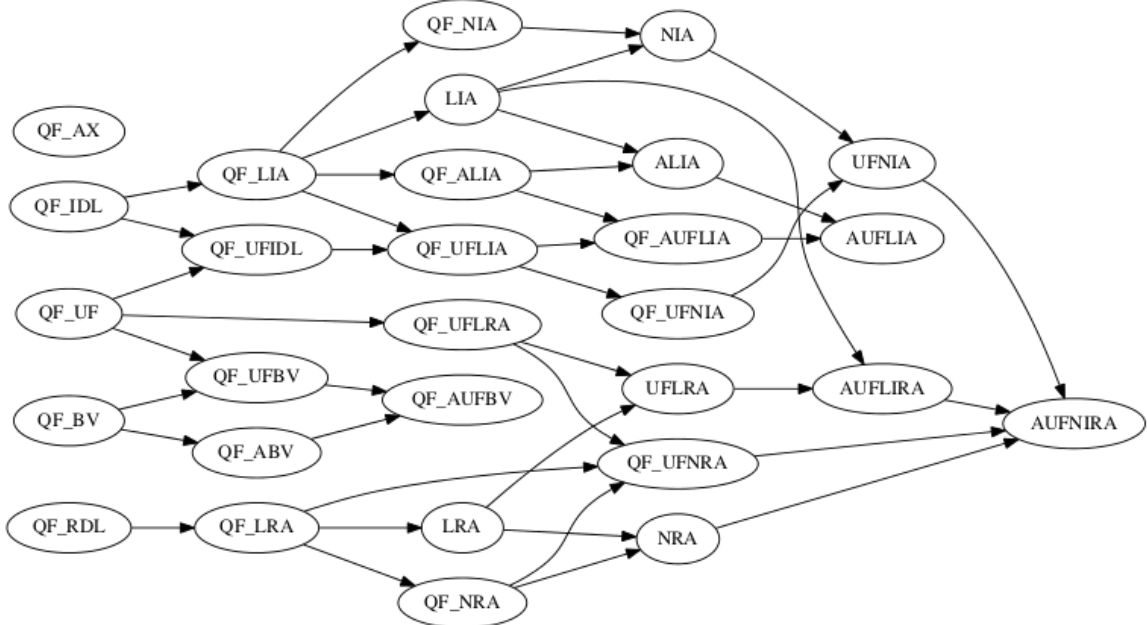
Theorem 2.54. *For any predicate logic formula ϕ , $\vdash \phi$ iff $\models \phi$.*

Satisfiability Modulo Theories (SMT) is an area of automated deduction that studies methods for checking the satisfiability of first-order formulas with respect to some logical theory T of interest.

SMT-LIB (<https://smt-lib.org/index.shtml>) is an international initiative aimed at facilitating research and development in SMT by promoting common specification for SMT solvers such as Z3 (Section 1.4) as well as benchmarks (<https://zenodo.org/communities/smt-lib>). It specifies for the following logical theories:

- Core: Core theory, defining the basic Boolean operators (which we used in Section 1.4)
- Ints: Integer numbers (to use in next topic?)
- Reals: Real numbers
- FloatingPoint: Floating point numbers
- FixedSizeBitVectors: Bit vectors with arbitrary size
- ArraysEx: Functional arrays with extensionality
- Strings: Unicode character strings and regular expressions

SMT-LIB supports the following sub-logics based on the stated theories.



Overview of the sub-logics, ordered by inclusion ($L_1 \rightarrow L_2$ means every formula of L_1 is also a formula of L_2)

Source: <https://smt-lib.org/logics.shtml>

The conventions of the abbreviations are shown below:

- QF for the restriction to quantifier free formulas

- A or AX for the theory ArraysEx
- BV for the theory FixedSizeBitVectors
- FP (forthcoming) for the theory FloatingPoint
- IA for the theory Ints (Integer Arithmetic)
- RA for the theory Reals (Real Arithmetic)
- IRA for the theory Reals_Ints (mixed Integer Real Arithmetic)
- IDL for Integer Difference Logic
- RDL for Rational Difference Logic
- L before IA, RA, or IRA for the linear fragment of those arithmetics
- N before IA, RA, or IRA for the non-linear fragment of those arithmetics
- UF for the extension allowing free sort and function symbols

SMT solvers are used in many formal method based applications such as the mCRL2 is a formal specification language (<https://mcrl2.org/web/index.html>).

Example 2.55. The SMT-LIB script to check the formulas in Example 2.42 in Ints theory is shown below. ^{leg:2.22}

```
(set-logic LIA)
;;; Q(x,y) : x + y = x - y
; 1. forall y Q(1, y)
(push)
(assert (forall ((y Int)) (= (+ 1 y) (- 1 y))))
(check-sat)
(pop)
; 2. E x E y Q(x, y)
(push)
(assert (exists ((x Int) (y Int)) (= (+ x y) (- x y))))
(check-sat) ; Not returning any model
(pop)
(push)
(declare-fun x () Int)
(declare-fun y () Int)
(assert (= (+ x y) (- x y)))
(check-sat)
(get-model)
(pop)
; 3. A x E y Q(x, y)
(push)
(assert (forall ((x Int)) (exists ((y Int)) (= (+ x y) (- x y)))))
(check-sat)
(pop)
; 4. E y A x Q(x, y)
(push)
(assert (exists ((y Int)) (forall ((x Int)) (= (+ x y) (- x y)))))
(check-sat)
(pop)
; 5. A y E x Q(x, y)
(push)
(assert (forall ((y Int)) (exists ((x Int)) (= (+ x y) (- x y)))))
(check-sat)
(pop)
```

2.6 Rules of Inference for Quantified Statements

es-of-inference

The “syntactic implication” or “natural deduction” of quantified statements and the Curry-Howard correspondence is the foundation to modern formal proving software such as Lean 4 prover, Rocq prover, etc.

The Curry-Howard correspondence for quantifiers is

Logic	Proposition	Conjunction	Disjunction	Universal Quantification	Existential Quantification
FP	Type	Product Type	Sum Type	Dependent Product Type	Dependent Sum Type
Logic	Hilbert-style deduction system	natural deduction	hypotheses	implication elimination (modus ponens)	implication introduction
FP	type system for combinatory logic	type system for lambda calculus	free variables	application	abstraction

In predicate logic, we have the following rules of inference related to quantifiers and equality in addition to the rules of inference from Section 1.7:

10. $\forall$ -introduction: $t\phi(t) \vdash \forall x\phi(x)$
11. $\forall$ -elimination: $\forall x\phi(x) \vdash \phi(t)$
12. $\exists$ -introduction: $\phi(s) \vdash \exists x\phi(x)$
13. $\exists$ -elimination: $\exists x\phi(x), \boxed{\phi(s) \cdots \xi} \vdash \xi$
14. $=$ -introduction: $\vdash t = t$
15. $=$ -elimination: $t_1 = t_2, \phi(t_1) \vdash \phi(t_2)$

Here,

- s is a **fresh variable** which does not occur in the premise $\exists x\phi(x)$, conclusion ξ or any assumptions in the syntactic derivation;
- t, t_1, t_2 are arbitrary terms.

Note: $\vdash$ indicates logical entailment based on the rules of inference instead of the laws of logical equivalences and implications based on the semantic theory.

Example 2.56. Consider the formula “ $\forall n(\text{prime}(n) \wedge (n > 2) \rightarrow \text{odd}(n))$. ” By applying universal instantiation with n “instantiate” to the term “ $m + 4$ ” we obtain

$$\text{prime}(m + 4) \wedge (m + 4 > 2) \rightarrow \text{odd}(m + 4).$$

Here $m + 4$ is a **term** and m is an **arbitrary constant**.

Example 2.57 (<https://github.com/OpenLogicProject/fitch>). Show that $\forall x \sim P(x) \vdash \sim \exists x P(x)$.

Proof:	1	$\forall x \sim P(x)$	premise
	2	$\exists x P(x)$	assumption
	3	$\begin{array}{ l} s \\ \hline P(s) \end{array}$	fresh variable
	4	$\begin{array}{ l} \hline \sim P(s) \end{array}$	1, $\forall I$
	5	$\perp$	3,4, $\sim E$
	6	$\perp$	2,3–5, $\exists E$
	7	$\sim \exists x P(x)$	2–6, $\sim I$

eg:3.18a

Example 2.58. Show $\exists x(P(x) \wedge Q(x)) \vdash (\exists x P(x)) \wedge (\exists x Q(x))$ (Theorem 2.47(a)). [thm:quantified:single](#)

Proof:	1	$\exists x(P(x) \wedge Q(x))$	premise
	2	$\begin{array}{ l} s \\ \hline P(s) \wedge Q(s) \end{array}$	assumption with fresh variable s
	3	$P(s)$	2, $\wedge E_1$
	4	$Q(s)$	2, $\wedge E_2$
	5	$\exists x P(x)$	3, $\exists I$
	6	$\exists x Q(x)$	4, $\exists I$
	7	$\exists x P(x) \wedge \exists x Q(x)$	5,6, $\wedge I$
	8	$\exists x P(x) \wedge \exists x Q(x)$	1,2-7, $\exists E$

Formal proving using Lean 4 Prover

```
theorem eg_1_13_3 (a : Type) (p q : a -> Prop) :
  (exists x : a, p x /\ q x) -> (exists x : a, p x) /\ (exists x : a, q x) :=
  fun premise : (exists x : a, p x /\ q x) =>
    Exists.elim premise
      (fun s =>
        fun hs : p s /\ q s =>
          And.intro ⟨s, And.left hs⟩ ⟨s, And.right hs⟩)
```

Note: the $\langle$ and $\rangle$ are in Unicode and needs to be typeset using Unicode editor. E.g. in Emacs, C-x 8 RET 27e8 C-x 8 RET 27e9.

eg:3.18a

Example 2.59. Show that $\sim R(c), \forall t(P(t) \rightarrow Q(t)), \forall t(Q(t) \rightarrow R(t)) \vdash \sim P(c)$.

Proof:	1	$\sim R(c)$	premise
	2	$\forall t(P(t) \rightarrow Q(t))$	premise
	3	$\forall t(Q(t) \rightarrow R(t))$	premise
	4	$P(c) \rightarrow Q(c)$	2, $\forall E$
	5	$Q(c) \rightarrow R(c)$	3, $\forall E$
	6	$\begin{array}{ l} P(c) \\ \hline \end{array}$	assumption
	7	$Q(c)$	6,4 $\rightarrow E$
	8	$R(c)$	7,5 $\rightarrow E$
	9	$\perp$	1,8 $\sim E$
	10	$\sim P(c)$	6,9 $\sim I$

t4:q10

Example 2.60 (Tutorial 2, Q22). Use ONLY the rules of inference to show that

$$\forall x(P(x) \rightarrow (Q(x) \wedge R(x))), \forall x(P(x) \wedge S(x)) \vdash \exists x(R(x) \wedge S(x))$$

<i>Solution.</i>	1	$\forall x(P(x) \rightarrow Q(x) \wedge R(x))$	premise	$\square$
	2	$\forall x(P(x) \wedge S(x))$	premise	
	3	$P(c) \wedge S(c)$	2, $\forall E$	
	4	$P(c) \rightarrow Q(c) \wedge R(c)$	1, $\forall E$	
	5	$P(c)$	3, $\wedge E_1$	
	6	$Q(c) \wedge R(c)$	4,5, $\rightarrow E$	
	7	$R(c)$	6, $\wedge E_2$	
	8	$S(c)$	3, $\wedge E_2$	
	9	$R(c) \wedge S(c)$	7,8, $\wedge I$	
	10	$\exists x(R(x) \wedge S(x))$	9, $\exists I$	

Example 2.61 (Tutorial 2, Q23). Use rules of inference to show that

$$\exists x P(x) \rightarrow \forall x(P(x) \vee Q(x) \rightarrow R(x)), \exists x(P(x) \wedge Q(x)) \vdash \exists y R(y)$$

<i>Solution.</i>	1	$\exists x P(x) \rightarrow \forall x(P(x) \vee Q(x) \rightarrow R(x))$	premise	$\square$
	2	$\exists x(P(x) \wedge Q(x))$	premise	
	3	$s \quad P(s) \wedge Q(s)$	2, assumption	
	4	$P(s)$	3, $\wedge E_1$	
	5	$\exists x P(x)$	4, $\exists I$	
	6	$\forall x(P(x) \vee Q(x) \rightarrow R(x))$	1,5, $\rightarrow E$	
	7	$P(s)$	3, $\wedge E_1$	
	8	$P(s) \vee Q(s)$	7, $\vee I_1$	
	9	$P(s) \vee Q(s) \rightarrow R(s)$	6, $\forall E$	
	10	$R(s)$	8,9 $\rightarrow E$	
	11	$\exists y R(y)$	10, $\exists I$	
	12	$\exists y R(y)$	2,3-11, $\exists E$	

Example 2.62 (Tutorial 2, Q27). What is wrong with the following proof?

1	$\forall x \exists y(x > y)$	premise
2	$\exists y(c > y)$	$\forall$ -elimination, c arbitrary
3	$(c > s)$	$\exists$ -elimination, s specific
4	$\forall x(x > s)$	$\forall$ -introduction
5	$\exists y \forall x(x > y)$	$\exists$ -introduction

Solution. Step 3 is wrong, s is a function of c , i.e. $s = f(c)$. Step 4 is also wrong since $s = f(c)$ and the universal generalisation will affect s as well. $\square$

Example 2.63 (Tutorial 2, Q24). Show that $\forall x(Q(x) \rightarrow R(x)) \wedge (\exists x(Q(x) \wedge I(x))) \vdash \exists x(R(x) \wedge I(x))$.

	1	$\forall x(Q(x) \rightarrow R(x)) \wedge (\exists x(Q(x) \wedge I(x)))$	premise	
	2	$\forall x(Q(x) \rightarrow R(x))$	1, $\wedge E_1$	
	3	$\exists x(Q(x) \wedge I(x))$	1, $\wedge E_2$	
	4	$s \mid Q(s) \wedge I(s)$	3, assumption	
	5	$Q(s) \rightarrow R(s)$	2, $\forall E$	
<i>Solution.</i>	6	$Q(s)$	4, $\wedge E_1$	$\square$
	7	$R(s)$	5,6, $\rightarrow E$	
	8	$I(s)$	4, $\wedge E_2$	
	9	$R(s) \wedge I(s)$	7,8, $\wedge I$	
	10	$\exists x(R(x) \wedge I(x))$	9, $\exists I$	
	11	$\exists x(R(x) \wedge I(x))$	3,4–10 $\exists E$	

Now we look at examples of arguments written in English sentences and validate the argument using the laws and rules mentioned above.

eg:3.18b

Example 2.64. Show that the following arguments is valid with a formal proof using the laws of inference.

No junior or senior is enrolled in calculus class.

Ali is enrolled in calculus class. Therefore Ali is not a senior.

Let $J(x)$: “ x is a junior”; $S(x)$: “ x is a senior”; and $C(x)$: “ x is enrolled in calculus class.”				
	1	$\forall x((J(x) \vee S(x)) \rightarrow \sim C(x))$	premise	
	2	$C(\text{Ali})$	premise	
	3	$(J(\text{Ali}) \vee S(\text{Ali})) \rightarrow \sim C(\text{Ali})$	1, $\forall$ -elimination	
Proof:	4	$S(\text{Ali})$	assumption	
	5	$J(\text{Ali}) \vee S(\text{Ali})$	$\vee$ -intro2	
	6	$\sim C(\text{Ali})$	3,4 $\rightarrow$ -elim	
	7	$\perp$	2,5 $\neg$ -elim	
	8	$\sim S(\text{Ali})$	6, $\neg$ -intro	

eg:3.12

Example 2.65. Show that the premises “A student in this class has not read the book,” and “Everyone in this class passed the first exam” imply the conclusion “someone who passed the first exam has not read the book.”

Solution. Let $C(x)$: “ x is in this class”; $B(x)$: “ x has read the book”; $P(x)$: “ x passed the first exam”. The above argument can be formalised as

$$\exists x(C(x) \wedge \sim B(x)), \forall x(C(x) \rightarrow P(x)) / \therefore \exists x(P(x) \wedge \sim B(x)).$$

We prove the argument as follows:

1	$\exists x(C(x) \wedge \sim B(x))$	premise
2	$\forall x(C(x) \rightarrow P(x))$	premise
3	$\begin{array}{ l} s \\ C(s) \wedge \sim B(s) \end{array}$	1, fresh variable
4	$C(s)$	3, $\wedge E_1$
5	$C(s) \rightarrow P(s)$	2, $\forall E$
6	$P(s)$	4,5 $\rightarrow E$
7	$\sim B(s)$	3, $\wedge E_2$
8	$P(s) \wedge \sim B(s)$	6,7 $\wedge I$
9	$\exists x(P(x) \wedge \sim B(x))$	8, $\exists I$
10	$\exists x(P(x) \wedge \sim B(x))$	1,3-9, $\exists E$

□

book:p92:q2-29c

Example 2.66 (Tutorial 2, Q25). Prove that the following argument is valid: “Every undergraduate is either an arts student or a science student. Some undergraduates are top students. James is not a science student, but he is a top student. Therefore if James is an undergraduate, he is an arts student.”

Solution. Let $U(x)$ be the predicate “ x is an undergraduate”, $A(x)$ be the predicate “ x is an arts student”, $S(x)$ be the predicate “ x is a science student” and $T(x)$ be the predicate “ x is a top student.” Then

1	$\forall x(U(x) \rightarrow A(x) \vee S(x))$	premise
2	$\exists x(U(x) \wedge T(x))$	premise
3	$\sim S(\text{James}) \wedge T(\text{James})$	premise
4	$\begin{array}{ l} U(\text{James}) \end{array}$	assumption
5	$U(\text{James}) \rightarrow A(\text{James}) \vee S(\text{James})$	1, $\forall E$
6	$A(\text{James}) \vee S(\text{James})$	4,5, $\rightarrow E$
7	$\begin{array}{ l} A(\text{James}) \end{array}$	assumption
8	$A(\text{James})$	7
9	$\begin{array}{ l} S(\text{James}) \end{array}$	6, assumption
10	$\sim S(\text{James})$	3, $\wedge E_1$
11	$\perp$	9,10 $\sim E$
12	$A(\text{James})$	11, $\perp E$
13	$A(\text{James})$	6,7-8,9-12, $\vee E$
14	$U(\text{James}) \rightarrow A(\text{James})$	4-13, $\rightarrow I$

□

Applications of Predicate Logic

The formal logic system can be applied in

- logic programming
- *hardware and software specification*: Bjørner [2006], Lamport [2003]
- *hardware verification*: Kropf [1999]
- compiler verification: Appel et al. [2014]

In this section, we will only explore logic programming, the relatively less complex subject.

According to https://en.wikipedia.org/wiki/Logic_programming, **logic programming** is a type of programming paradigm which is largely based on predicate logic. Any program written in a logic programming language is a set of sentences, expressing **facts** and **rules** about some problem domain.

Major logic programming language families include *Prolog* (which stands for Programming in Logic), ECLiPSe, *Datalog* (a subset of Prolog for database and serves as an alternative query system to SQL, refer to Abiteboul et al. [1995]) and *answer set programming* (ASP).

Prolog [Sterling and Shapiro, 1999], which dates back to 1972, was once believed to be the artificial intelligence (AI) language. However, the limitations of the first order logic and the difficulty to get experts to encode domain knowledge have prevented Prolog from getting popular. Despite these, Prolog is still useful in knowledge-based (or ontology) systems [Merritt, 2000].

Definition 2.9. [van Le, 1993] In Prolog, a **term** is either a **constant**, a **variable** (starts with a capital letter or underscore) or a **compound term** of the form $f(t_1, \dots, t_n)$ where f is a function symbol of arity n and t_i ($i = 1, \dots, n$ and $n > 0$) are terms.

An **atomic formula** or **atomic term** is an expression of the form $p(t_1, \dots, t_n)$ where p is a predicate symbol and t_i are terms. If $n = 0$, there is no argument and the parentheses are omitted.

A **constant** can either be a number or an **atomic term** (start from small letter, a string in single quotes or symbols such as $+$, $=$, $!$, etc.).

The rules in Prolog are written in the form of **Horn clauses**, which can be regarded as “arguments” in Definition 1.12.

$$\underbrace{\underbrace{H}_{\text{head}} : \underbrace{-B_1, \dots, B_n}_{\text{body}}}_{\text{rule}} \quad (2.2) \quad \text{e:horn-clause}$$

where $\wedge$ is the same as $\wedge$, $: -$ is the same as $\rightarrow$, H and B_i are usually atomic formulas. So we read (2.2) as a logical implication “ H if B_1 and $\dots$ and B_n .” **Facts** are rules that have **no body**, and are written in the simplified form (which is comparable to a “tautology”):

$$H. \quad (2.3) \quad \text{e:facts}$$

Remarks on Prolog syntax:

- A string starts and ends with double quote "
- Don't use single quotes, they are for ‘atom’ (or symbol)
- Anything that starts with **small letter** is used as “predicate” or atom.
- Anything that starts with **capital letter** is used as “variables”.

Running commands sequentially in Prolog:

```

1 % swipl -q hellolines.pl
2 :- format("This is first line\n").
3 :- format("This is second line: ~d ~d ~d\n", [1, 2, 3]).
4 :- halt.

```

Note that rules without head are called **directives** and are immediately executed by Prolog.

Prolog goes through all predicates and we sometimes need to tell Prolog to stop trying out all possibilities by **cutting** using the “exclamation mark”:

```

1 thegrade(Marks, Grade) :- Marks < 0, Grade="???", !.
2 thegrade(Marks, Grade) :- Marks > 100, Grade="???", !.
3 thegrade(Marks, Grade) :- Marks >= 90, Grade="A+", !.
4 thegrade(Marks, Grade) :- Marks >= 80, Grade="A", !.
5 thegrade(Marks, Grade) :- Marks >= 75, Grade="A-", !.
6 thegrade(Marks, Grade) :- Marks >= 70, Grade="B+", !.
7 thegrade(Marks, Grade) :- Marks >= 65, Grade="B", !.
8 thegrade(Marks, Grade) :- Marks >= 60, Grade="B-", !.
9 thegrade(Marks, Grade) :- Marks >= 55, Grade="C+", !.
10 thegrade(Marks, Grade) :- Marks >= 50, Grade="C", !.
11 thegrade(Marks, Grade) :- Grade="F".

```

Prolog uses recursion on predicates to perform looping. Suppose we want to find the sum of the squares of a list numbers of any size:

$$x_1^2 + x_2^2 + \dots + x_n^2$$

this is how it is done in Standard ML:

```

1 (define (sum_sq alist)
2   (cond
3     ((null? alist) 0)
4     (else (+ (* (car alist) (car alist)) (sum_sq (cdr alist))))))
5
6 (display (sum_sq '(1 3 5 7)))
7 (newline)

```

The implementation in Prolog is similar:

```

1 % gprolog --consult-file sum_sq.pl
2 sum_sq([], 0).
3 % Be careful with X and Xs must be capitalised.
4 % Otherwise, they are treated as atoms.
5 sum_sq([X|Xs], Sum) :- sum_sq(Xs, Sum_rest), Sum is X*X + Sum_rest.
6
7 %sum_sq([1,3,5,7],X).

```

Expression (2.2) is not sufficient for efficient computing. Prolog allows the body of a clause to be negations of atomic formulas to allow non-monotonic logic. In Datalog and ASP, logic programs have only a declarative reading, and their execution is performed by means of a proof procedure or model generator whose behaviour is not meant to be under the control of the programmer.

Prolog could be used to solve some of the logic problems discussed in propositional logic which is illustrated below.

Example 2.67. Use Prolog to determine all truth values assignments in Example [I.9](#).

Solution:

```

1 % Tutorial 1, Q3 (Example 1.2.7)
2 proposition(P,Q,R,S,Val) :- not(P,NotP), or(NotP,R,Expr1),
3   not(S,NotS), and(Expr1,NotS,Expr2), imply(Q,Expr2,Expr3),
4   not(R,NotR), and(NotR,Q,Expr4), imply(NotS,Expr4,Expr5),
5   and(Expr3,Expr5,Val).
6
7 %write_list([]).
8 %write_list([H|T]) :- write(H), write(', '), write_list(T).

```

```

9
10 solv:- format("Determine all truth values assignment [P,R,S]:~n",[]),
11      findall([P,R,S], proposition(P,true,R,S,true), Results),
12      write(Results), nl. % Raw output: write_canonical

```

Remark: More examples are given in Clocksin [1997, Chapter 7].

To illustrate the use of Prolog in real-world problem, I will use illustrate it with the following “to be developed education query system”:

```

                                utar_dmas.pl
1 course('UECM1024', 'Calculus I', core, 4, []).
2 course('UECM1034', 'Calculus II', core, 4, ['UECM1024']).
3 course('UECM1204', 'Probability and Statistics I', core, 4, []).
4 course('UECM1224', 'Probability and Statistics II', core, 4, ['UECM1204']).
5 course('UECM1314', 'Fundamentals of Linear Algebra', core, 4, []).
6 course('UECM1304', 'Discrete Mathematics with Applications', core, 4, []).
7 course('UECM1703', 'Introduction to Scientific Computing', core, 3, []).
8 course('UECM2353', 'Linear Algebra', core, 3, ['UECM1314']).
9 course('UECS1004', 'Programming and Problem Solving', core, 4, []).
10 course('UECS1044', 'Object-Oriented Application Development', core, 4, ['UECS1004']).
11 course('UECS2083', 'Problem Solving with Data Structures and Algorithms', core, 3, ['UECS1044']).
12 course('UECS2094', 'Web Application Development', core, 4, ['UECS1044']).
13 course('UECM2023', 'Ordinary Differential Equations', core, 3, ['UECM1034']).
14 course('UECM3034', 'Numerical Methods', core, 4, []).
15
16 courses_without_prerequisites(X,Y) :- course(X,Y,_,_,[]).
17 courses_with_prerequisites(X,Y) :- course(X,Y,_,_,Z), length(Z, N), N >= 0.

```

1. To **load** the Prolog database, one can use either the commands below:

```

1 ?- [utar_dmas].
2 ?- consult('utar_dmas.pl').

```

2. To **query** the courses with only the code without prerequisite, we can type the following query. Note that underscore is used to indicate a value we are not interested in.

```

1 ?- courses_without_prerequisites(X,_).

```

3. To **query** the courses (showing both code and name) with prerequisite, we can type the following query.

```

1 ?- courses_with_prerequisites(X,Y).

```

4. To **exit** Prolog, type halt.

One interesting use of Prolog is probably in puzzle solving.

Real-world business software consist a lot of SQL statements. A dictionary of SQL relational algebra and Datalog is given below [Ullman and Widom, 2008, Chapter 5].

SQL Concept	Relational Algebra	Datalog
Intersection	$R(x, y) \cap T(x, y)$	$I(x, y) :- R(x, y), T(x, y).$
Union	$R(x, y) \cup T(x, y)$	$U(x, y) :- R(x, y).$ $U(x, y) :- T(x, y).$
Difference	$R(x, y) \setminus T(x, y)$	$D(x, y) :- R(x, y), \text{not } T(x, y).$
Projection	$\pi_x(R)$	$P(x) :- R(x, y)$
Selection	$\sigma_{x>10}(R)$	$S(x, y) :- R(x, y), x > 10$
Product	$R \times T$	$P(x, y, z, w) :- R(x, y), T(z, w)$
Natural Join	$R \bowtie T$	$J(x, y, z) :- R(x, y), T(y, z)$
Theta Join	$R \bowtie_{R.x>T.z} T$	$\theta(x, y, z, w) :- R(x, y), T(z, w), x > z$

The website <http://perugini.cps.udayton.edu/teaching/courses/Spring2017/cps499/Languages/notes/PROLOG.html> did not provide any solution for this, but did provide an example to detect loops in the path as follows (using “not” operator \+ as breaking point to infinite loop?):

```
1 cycle(Start, Visited) :- cycle(Start, Start, [Start], Visited).
2
3 cycle(Orig, Start, Path, Visited) :-
4     edge(Start, Orig),
5     reverse([Orig|Path], Visited).
6
7 cycle(Orig, Start, Path, Visited) :-
8     edge(Start, Next),
9     \+ member(Next, Path),
10    cycle(Orig, Next, [Next|Path], Visited).
```

Chapter 3

Numbers and Methods of Proofs

3.1 Notations

sec:numbers

Using set theory and the notion of equivalence relation (Topic 4), the set of natural numbers, the set of integers defined below can be constructed.

Symbol	Set
$\mathbb{N}$	Set of natural numbers, i.e. $\{0, 1, 2, \dots\}$
$\mathbb{Z}$	Set of integers, i.e. $\{\dots, -2, -1, 0, 1, 2, \dots\}$
$\mathbb{Q}$	Set of rational numbers, i.e. $\{\frac{p}{q} : p, q \in \mathbb{Z}, q \neq 0\}$
$\mathbb{R}$	Set of real numbers, i.e. $\{\pm i_n \dots i_2 i_1 . d_1 d_2 \dots : i_j, d_i \in \mathbb{N}\}$

They are related by $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}$ “semantically”. To differentiate positive, negative and nonzero numbers, we introduce the following notations:

- $\mathbb{R}^+ := \{x \in \mathbb{R} : x > 0\}$, $\mathbb{R}^- := \{x \in \mathbb{R} : x < 0\}$, $\mathbb{R}^* := \{x \in \mathbb{R} : x \neq 0\}$.
- $\mathbb{Q}^+ := \{x \in \mathbb{Q} : x > 0\}$, $\mathbb{Q}^- := \{x \in \mathbb{Q} : x < 0\}$, $\mathbb{Q}^* := \{x \in \mathbb{Q} : x \neq 0\}$.
- $\mathbb{Z}^+ := \{1, 2, 3, \dots\}$, $\mathbb{Z}^- := \{-1, -2, -3, \dots\}$, $\mathbb{Z}^* := \{\pm 1, \pm 2, \dots\}$.

A set of formal propositions, called **axioms**, are used to characterise the set of numbers mentioned above.

3.1.1 Natural Numbers

sec:nat

Peano axioms define the arithmetical properties of natural numbers, usually represented as $\mathbb{N}$.

The symbols for the axioms includes a constant symbol 0 and a unary function symbol S and an equal sign “=”.

1. $0 \in \mathbb{N}$
2. $\forall x \exists y (S(x) = y)$ every natural number has a “next” natural number.
3. $\exists 0 \forall x (\sim (S(x) = 0))$ there is a 0 which has no “previous” natural number.
4. $\forall x (x = x)$ reflexive
5. $\forall x \forall y (x = y \rightarrow y = x)$ symmetric
6. $\forall x \forall y \forall z (x = y \wedge y = z \rightarrow x = z)$ transitive
7. $\forall x \forall y (S(m) = S(n) \rightarrow m = n)$
8. $\forall y (\phi(0, y) \wedge \forall x (\phi(x, y) \Rightarrow \phi(S(x), y)) \Rightarrow \forall x \phi(x, y))$ mathematical induction

def:induction

$\mathbb{N}$ is implemented as `Nat` type in Lean 4's `Init/Prelude.lean`

```
inductive Nat where
| zero : Nat          -- same as 0
| succ (n : Nat) : Nat -- same as n+1
```

With the above definitions, it is possible to obtain the “ordered semiring” properties of $\mathbb{N}$ below.

- A1. $\forall x \forall y (x + y = y + x)$ `Nat.add_comm`
- A2. $\forall x \forall y \forall z ((x + y) + z = x + (y + z))$ associative law
- A3. $\forall x (0 + x = x)$; identity
- M1. $\forall x \forall y (x \cdot y = y \cdot x)$ commutative law
- M2. $\forall x \forall y \forall z ((x \cdot y) \cdot z = x \cdot (y \cdot z))$ associative law
- M3. $\forall x (1 \cdot x = x) \wedge (0 \cdot x = 0)$ identity
- D. $\forall x \forall y \forall z (x \cdot (y + z) = x \cdot y + x \cdot z)$ distributive law
- L1. $\forall x \sim (x < x)$
- L2. $\forall x \forall y ((x < y) \vee (x = y) \vee (y < x))$ total ordering
- L3. $\forall x \forall y \forall z (((x < y) \wedge (y < z)) \rightarrow x < z)$ transitivity
- L4. $\forall x \forall y \forall z ((x < y) \rightarrow (z + x < z + y))$
- L5. $\forall x \forall y \forall z (((x < y) \wedge (0 < z)) \rightarrow z \cdot x < z \cdot y)$.
- L6. $\forall x \forall y ((x < y) \rightarrow \exists z (x + z = y))$
- L7. $(0 < 1) \wedge (\forall x ((x > 0) \rightarrow (x = 1 \vee x > 1)))$.
- L8. $\forall x (x = 0 \vee 0 < x)$.
- I. $\phi(0) \wedge (\forall k (\phi(k) \rightarrow \phi(k + 1))) \rightarrow \forall n \phi(n)$ Induction Principle

Here ϕ is any well-form formula.

A1–D in Lean 4's `Init/Data/Nat/Basic.lean`

```
variable (x y z : Nat)
example : x + y = y + x := Nat.add_comm x y
example : (x + y) + z = x + (y + z) := Nat.add_assoc x y z
example : 0 + x = x := Nat.zero_add x
example : x + 0 = x := Nat.add_zero x
example : x * y = y * x := Nat.mul_comm x y
example : (x * y) * z = x * (y * z) := Nat.mul_assoc x y z
example : x * 1 = x := Nat.mul_one x
example : 1 * x = x := Nat.one_mul x
example : 0 * x = 0 := Nat.zero_mul x
example : x * (y + z) = x * y + x * z := Nat.mul_add x y z
example : x * (y + z) = x * y + x * z := Nat.left_distrib x y z
example : (x + y) * z = x * z + y * z := Nat.add_mul x y z
example : (x + y) * z = x * z + y * z := Nat.right_distrib x y z
theorem L3 : (x < y) /\ (y < z) -> x < z :=
  fun h => Nat.lt_trans h.left h.right
theorem L4 : (x < y) -> (z + x < z + y) :=
  fun h => Nat.add_lt_add_left h _
#check 0 + 1 = 1 -- Print the type of the expression
#eval 0 + 1 = 1 -- Perform calculation to decide the value
```

From the Induction Principle, mathematicians have obtained the well-ordering principle stated below.

Theorem 3.1 (Well-Ordering Property of $\mathbb{N}$). *If $S \subset \mathbb{N}$ and $S \neq \emptyset$, then there is an element $m \in S$ such that $m \leq k$ for all $k \in S$.*

3.1.2 Integers: Even, Odd, Prime, Composite

A set of integers $\mathbb{Z}$ together with the arithmetic operations and relational operations turn $\mathbb{Z}$ into a **commutative ordered ring with unit** whose positive elements are well-ordered.

ordered_comm_ring

Definition 3.1. The *commutative ordered ring with unit* F is a set equipped with constants 0, 1, addition $+$, multiplication $\cdot$ and less than relation $<$ such that

- A1. $\forall x \forall y (x + y = y + x)$; commutativity
- A2. $\forall x \forall y \forall z ((x + y) + z = x + (y + z))$; associativity
- A3. $\forall x (0 + x = x)$; identity
- A4. $\forall x \exists y (x + y = 0)$; additive inverse
- M1. $\forall x \forall y (x \cdot y = y \cdot x)$; commutativity
- M2. $\forall x \forall y \forall z ((x \cdot y) \cdot z = x \cdot (y \cdot z))$; associativity
- M3. $(1 \neq 0) \wedge \forall x (1 \cdot x = x)$; identity
- D. $\forall x \forall y \forall z (x \cdot (y + z) = x \cdot y + x \cdot z)$; distributivity
- O1. $\forall x \forall y ((x = y) \wedge \sim (y < x) \wedge \sim (x < y) \vee \sim (x = y) \wedge (y < x) \wedge \sim (x < y) \vee \sim (x = y) \wedge \sim (y < x) \wedge (x < y))$; total order
- O2. $\forall x \forall y \forall z ((x < y) \wedge (y < z) \rightarrow (x < z))$; transitivity
- O3. $\forall x \forall y \forall z (x < y \rightarrow (z + x < z + y))$;
- O4. $\forall x \forall y \forall z (((x < y) \wedge (0 < z)) \rightarrow z \cdot x < z \cdot y)$

def:integer

Definition 3.2. The *integer set* $\mathbb{Z}$ is a commutative ordered ring with unit such that

- I. $\phi(0) \wedge (\forall k (\phi(k) \rightarrow \phi(k + 1))) \rightarrow \forall n \phi(n)$ Induction Principle

$\mathbb{Z}$ is implemented as Int type in Lean 4's Init/Data/Int/Basic.lean

```
inductive Int : Type where
| ofNat    : Nat -> Int -- n = 0, 1, 2, ...
| negSucc  : Nat -> Int -- negatives are -[1+n]
def negOfNat : Nat -> Int
| 0      => 0
| succ m => negSucc m
```

A1–D in Lean 4's Init/Data/Int/

```
variable (x y z : Int)
theorem A1 : x + y = y + x := Int.add_comm x y
theorem A2 : (x + y) + z = x + (y + z) := Int.add_assoc x y z
theorem A3 : 0 + x = x := Int.zero_add x
theorem A3r : x + 0 = x := Int.add_zero x
theorem M1 : x * y = y * x := Int.mul_comm x y
theorem M2 : (x * y) * z = x * (y * z) := Int.mul_assoc x y z
theorem M3 : 1 * x = x := Int.one_mul x
theorem M3r : x * 1 = x := Int.mul_one x
theorem D : x * (y + z) = x * y + x * z := Int.mul_add x y z
theorem Dr : (x + y) * z = x * z + y * z := Int.add_mul x y z
theorem O2 : (x < y) /\ (y < z) -> x < z := fun h => Int.lt_trans h.left h.right
theorem O3 : (x < y) -> (z + x < z + y) := fun h => Int.add_lt_add_left h _
```

Definition 3.3. Let n be an integer. n is *even* if there is an integer k such that $n = 2k$. n is *odd* if an integer k such that $n = 2k + 1$.

Definition of Even and Odd in Lean 4

```
def is_even (n : Int) := exists k, n = 2*k
def is_odd  (n : Int) := exists k, n = 2*k+1
```

thm:odd+even

Theorem 3.2. *Every integer is either odd or even but not both.*

eg:4.1

Example 3.3. Use the definitions of even and odd to justify your answers to the following questions.

1. Is 0 even?

Solution. Yes, $0 = 2 \times 0$.

□

2. Is -501 odd?

Solution. Yes, $-501 = 2(-251) + 1$.

□

3. If a and b are integers, is $8ab^3 + 6$ even?

Solution. Yes, $8ab^3 + 6 = 2(4ab^3 + 3)$, and since a and b are integers, so is $4ab^3$.

□

Definition 3.4. The *absolute value* of an integer n , $|n| := \begin{cases} n, & 0 \leq n, \\ -n, & n < 0. \end{cases}$

def:prime

Definition 3.5. A integer n is called a *prime (number)* if $n > 1$ and “ $\forall r \forall s ((r \in \mathbb{N} \wedge s \in \mathbb{N} \wedge n = r \cdot s) \rightarrow ((r = 1) \vee (s = 1)))$ ”. n is called *composite* if $n > 1$ and there are positive integers r and s such that $n = r \cdot s$ and $r \neq 1$ and $s \neq 1$, i.e. those numbers that have more than two factors. A negative integer n is *composite* if $|n|$ is composite.

Implementation of Even, Odd, Prime Checking Functions in Scheme

```
(define (isEven n) (zero? (modulo n 2)))
(define (isOdd n) (not (= 0 (modulo n 2))))

;
; 0 is neither prime nor composite. Since any number times zero equals zero,
; there are infinite factors for a product of zero.
; 1 has only 1 factor. For a number to be classified as a prime number,
; it should have exactly two factors.
;
(define (isPrime n)
  (and (> n 1)
    (or (= n 2)
      (let noDivisorsAbove ((m 2))
        (cond
          ((zero? (modulo n m)) #f)
          ((>= (* m m) n) #t)
          (else (noDivisorsAbove (+ m 1))))))))

(define (isComposite n)
  (and
    (> (abs n) 2)
    (let hasDivisorsAbove ((m 2))
      (cond
        ((zero? (modulo n m)) #t)
        ((>= (* m m) n) #f)
        (else (hasDivisorsAbove (+ m 1)))))))
```


Theorem 3.4. Every integer greater than one is either a prime or a composite but not both.

Example 3.5. 1. Write the first 6 prime numbers.

Solution. 2, 3, 5, 7, 11, 13. □

2. Write the first 6 non-negative composite numbers.

Solution. 4, 6, 8, 9, 10, 12 □

Example 3.6. Explain how the function `(isEven 3)` work.

Solution: $3 \bmod 2 = 1 \neq 0$ is false.

Example 3.7. Explain how the function `(isOdd 3)` work.

Solution. $3 \bmod 2 = 1 \neq 0$ is true. □

Example 3.8. Explain how the function `(isPrime 17)` work.

Solution:

`isPrime 17` = $17 > 1$ and `noDivisorAbove(2)`
`noDivisorAbove(2)` = $17 \bmod 2 \neq 0$ and $2*2 < 17$ calls `noDivisorAbove(3)`
`noDivisorAbove(3)` = $17 \bmod 3 \neq 0$ and $3*3 < 17$ calls `noDivisorAbove(4)`
`noDivisorAbove(4)` = $17 \bmod 4 \neq 0$ and $4*4 < 17$ calls `noDivisorAbove(5)`
`noDivisorAbove(5)` = $17 \bmod 5 \neq 0$ and $5*5 \geq 17$ returns true
 So `(isPrime 17)` returns (and #t #t) or true.

Example 3.9. Explain how the function `isComposite 16` work.

Solution. `isComposite 16` = $16 > 2$ and `hasDivisorAbove(2)`
`hasDivisorAbove(2)` = $16 \bmod 2 = 0$ is true.
 So `isComposite 16` = true and true = true □

Definition 3.6. A *counterexample* to the universal statement $\forall x P(x)$ is a value c for which $v(P(c)) = F$.

eg: 2.4a

Example 3.10. Determine the truth value of the following universal statements. If the universal statement is false, suggest a counterexample to the universal statement.

1. For every $x \in \{-2, 0, 4, 6, 8\}$, x^2 is even.

Solution. Let $E(x)$ be the predicate “ x is even”. Then the statement can be written as

$$\begin{aligned} \forall x(x \in \{-2, 0, 4, 6, 8\} \rightarrow E(x^2)) &\equiv \forall x((x = -2 \vee x = 0 \vee x = 4 \vee x = 6 \vee x = 8) \rightarrow E(x^2)) \\ &\equiv \forall x((x = -2 \rightarrow E(x^2)) \wedge (x = 0 \rightarrow E(x^2)) \\ &\quad \wedge (x = 4 \rightarrow E(x^2)) \wedge (x = 6 \rightarrow E(x^2)) \wedge (x = 8 \rightarrow E(x^2))) \\ &\equiv E((-2)^2) \wedge E(0^2) \wedge E(4^2) \wedge E(6^2) \wedge E(8^2) \\ &\equiv T \wedge T \wedge T \wedge T \wedge T \equiv T \end{aligned}$$

Hence, the statement is true. □

Note: The technique used to show the truth of the universal statement in this example is called the *method of exhaustion*. This technique is used when the domain has finite elements.

2. $\forall x(x \in \mathbb{Z} \rightarrow x + 1 < 4)$.

Solution. F. Counterexample: $x = 5, 5 + 1 = 6 > 4$. □

3. $\forall x(x \in \mathbb{Z}^+ \rightarrow (x - 1 \text{ is nonnegative}))$.

Solution. T. $x \geq 1 \rightarrow x - 1 \geq 0$ using O3. □

t5:q1

Example 3.11 (Tutorial 3, Q1). Assuming that m and n are particular integers, use the definitions of even, odd, prime and composite to answer the following questions.

1. If $m > n > 0$, is $m^2 - n^2$ composite?
2. Is $6m + 10n$ even?
3. Is $10mn + 13$ odd?
4. If $m > 0$ and $n > 0$, is $m^2 + 2mn + n^2$ composite?

Solution. 1. Not necessarily. E.g. When $m = 3, n = 2, m > n > 0$ but $m^2 - n^2 = 5$ is not a composite.

2. $6m + 10n = 2(3m + 5n)$ is even since $3m + 5n$ is an integer.

3. $10mn + 13 = 2(5mn + 6) + 1$ is odd since $5mn + 6$ is an integer.

4. $m^2 + 2mn + n^2 = (m + n)^2$ is a composite since $m + n > 1$. □

sec:real

3.1.3 Rational Numbers and Real Numbers

The set of rational numbers $\mathbb{Q}$ is the “smallest” Archimedean ordered field and the set of real numbers $\mathbb{R}$ is an Archimedean ordered field which is Cauchy complete. Hence, we will first define the notion of ordered field.

ordered_comm_field

Definition 3.7. A set F (the domain for the predicates) is called an **ordered field** if it is a commutative ordered ring with unit (Definition 3.1) and satisfies

def:ordered_comm_ring

M5. $\forall x(x \neq 0 \rightarrow (\exists y, y \cdot x = 1))$; multiplicative inverse

def:archimedean

Definition 3.8. We say an ordered field is **Archimedean** if for each $\epsilon \in F$ such that $\epsilon > 0$, there is positive integer N such that $1 < N\epsilon$.

ean-ordered-field

Definition 3.9. The rational number set $\mathbb{Q}$ is the smallest Archimedean ordered field. The member of $\mathbb{Q}$ is called **rational number**.

Theorem 3.12. A rational number $r \in \mathbb{Q}$ can be expressed as a quotient of two integers with a nonzero denominator, i.e. there are integers a and $b \neq 0$ such that $\frac{a}{b}$.

Theorem 3.13. Every integer is a rational number.

Proof. For any integer $n, n = n/1$. □

eg:4.3a

Example 3.14. Which of the following are rational numbers?

- $-(4/35)$

Solution. $-(4/35) = \frac{-4}{35}$, which is a quotient of the integers -4 and 35 . Hence it is rational. $\square$

- 0.211

Solution. Yes, $0.211 = 211/1000$ $\square$

- $0.\dot{1}\dot{2}$.

Solution. Yes. Let $x = 0.12121212 \dots$. Then $100x = 12.12121212 \dots$ and

$$100x - x = 12.12121212 \dots - 0.12121212 \dots \rightarrow 99x = 12 \rightarrow x = \frac{12}{99}$$

Therefore, $0.\dot{1}\dot{2} = \frac{12}{99}$ is a rational number. $\square$

eg:4.3b

Example 3.15. If m and n are integers and neither m nor n is zero, is $\frac{m+n}{mn}$ a rational number?

Solution. Yes, m and n are integers, so are $m+n$ and $mn \neq 0$. Note that $mn \neq 0$ because $m \neq 0$ and $n \neq 0$. $\square$

def:sup

Definition 3.10. Suppose F is an ordered field and $E \subset F$. If there exists a $\beta \in F$ such that $x \leq \beta, \forall x \in E$, we say that E is *bounded above* and call β an *upper bound* of E . If the upper bound β of E also satisfies the following property

- If $\gamma < \beta$ then γ is not an upper bound of E .

Then β is called the *least upper bound* of E or the *supremum* of E and we write $\beta = \sup E$.

om:completeness

Axiom 3.16 (Completeness Axiom). completeness

Let F be an ordered field and $E \subset F$. If $E \neq \emptyset$, and E is bounded above, then $\sup E$ exists in F .

def:real

Definition 3.11. The real number set $\mathbb{R}$ is an Archimedean ordered field that satisfies the Completeness Axiom. The members of $\mathbb{R}$ are called *real numbers*.

Theorem 3.17. A rational number is a real number.

def:4.3

Definition 3.12. An *irrational* number is a real number that is not rational.

By definition, every real number is either rational or irrational but not both.

eg:2.4b

Example 3.18. Determine the truth value of the following universal statements. If the universal statement is false, suggest a counterexample to the universal statement.

1. $\forall x((x \in \mathbb{R} \wedge 0 \leq x \wedge x \leq 1) \rightarrow x < 5x)$

Solution. F Counterexample: $x = 0$. $\square$

2. $\forall x(x \in \mathbb{R} \rightarrow x^2 + 1 = 0)$

Solution. F Counterexample: $x = 0$. $\square$

3. $\forall x(x \in \mathbb{R} \rightarrow x^2 - 1 = (x+1)(x-1))$

Solution. T Using distributive law. $\square$

t5:q9

Example 3.19 (Tutorial 3, Q2). (a) Assume that $a \neq 0$ and $b \neq 0$ are both integers. Is $(b - a)/(ab^2)$ a rational number?

Solution. Yes. Since $ab^2 \neq 0$ is an integer when $a, b \neq 0$ and $b - a$ is an integer. $(b - a)/(ab^2)$ is a well-defined ratio of two integers. Hence it is a rational number. $\square$

(b) Assume that a and $b > 0$ are both integers. Is $(5a + 12b)/4b$ a rational number?

Solution. Yes. Since $4b \neq 0$ is an integer when $b \neq 0$ and $5a + 12b$ is an integer. $(5a + 12b)/(4b)$ is a well-defined ratio of two integers. Hence it is a rational number. $\square$

t5:q11

Example 3.20 (Tutorial 3, Q3). Suppose a, b, c and d are integers and $a \neq c$. Suppose also x is a real number that satisfies the equation

$$\frac{ax + b}{cx + d} = 1. \quad (**)$$

Is x rational?

Solution. Yes. One can rewrite (**) as $x = \frac{d - b}{a - c}$, $a - c \neq 0$. Since x is written as the ratio of two integers $d - b$ and $a - c \neq 0$, it is a rational number. $\square$

t4:q3c

Example 3.21 (Tutorial 3, Q4). Is the following argument valid?

Any sum of two rational numbers is rational.
The sum $r + s$ is rational.

Therefore the numbers r and s are both rational.

Solution. No. Consider the model of real numbers and an assignment $r = \sqrt{2}$ and $s = -\sqrt{2}$, we have

Any sum of two rational numbers is rational $\equiv T$

The sum $r + s$ is rational $\equiv T$

Therefore the numbers r and s are both rational $\equiv F$

$\square$

3.2 Methods of Proof

sec:mproof

To **confirm** a mathematical statement is **true**, we need to **prove** it.

Let $A_1, \dots, A_n$ be the **axioms** of **mathematical objects** (e.g. $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}$).

We can derive theorems $T_1, \dots, T_m$ from the axioms $A_1, \dots, A_n$. For a new mathematical statement ψ , if we can prove it using the axioms and theorems, then it can be expressed as an argument below:

$$A_1, \dots, A_n, T_1, \dots, T_m \vdash \psi. \quad (3.1)$$

e:cond2

The techniques of *direct proof* and *proof by contradiction* are usually used to show the argument (3.1) syntactically.

For a new mathematical statement in conditional form, the argument can be expressed as

$$A_1, \dots, A_n, T_1, \dots, T_m \vdash (\phi \rightarrow \psi) \quad (3.2)$$

e:cond

where ϕ and ψ be two formulas. The techniques of *direct proof*, *contrapositive proof* and *proof by contradiction* are usually used to show the argument (3.1) syntactically. To know more about the different aspects of mathematical proofs, one may want to read Ording [2019].

3.2.1 Direct Proof

The *direct proof* of (3.2) ^{le:cond} starts from axioms, theorems and ϕ to reach ψ .

Example 3.22. 12 is an even number.

Proof: $12 = 2 \times 6$ and by definition, 12 is even.

Proving in Lean 4

```
example : is_even (12) :=
  have h : Int.ofNat 12 = 2*6 := rfl -- calculate & check equality
  Exists.intro 6 h -- k = 6
```

eg:4.4a

Example 3.23. Prove that “there is an even integer n that can be written in 2 ways as a sum of 2 prime numbers.”

Proof: Let $n = 10$. Then $10 = 5 + 5 = 3 + 7$ and 3, 5 and 7 are all prime numbers.

Let $P(n)$ be the predicate “ n is prime”. Then take $n = 10$, $p_1 = p_2 = 5$, $p_3 = 3$, $p_4 = 7$ and

$$\begin{array}{c} 10 = 2 \cdot 5 \wedge 5 \neq 3 \wedge 5 \neq 7 \wedge 5 \neq 7 \\ \hline \wedge P(5) \wedge P(5) \wedge P(3) \wedge P(7) \wedge 10 = 5 + 5 \wedge 10 = 3 + 7 \\ \hline \therefore \exists n, (\exists k(n = 2k) \wedge \exists p_1 \exists p_2 \exists p_3 \exists p_4 (p_1 \neq p_3 \wedge p_1 \neq p_4 \wedge p_2 \neq p_4 \\ \wedge P(p_1) \wedge P(p_2) \wedge P(p_3) \wedge P(p_4) \wedge n = p_1 + p_2 \wedge n = p_3 + p_4)) \end{array}$$

Due to the difficulty in manipulating formal proofs using appropriate tactics, we usually do not use formal proofs in deriving results in mathematics.

eg:4.6

Example 3.24. Use the **method of exhaustion** (useful for domain with finite values) to prove the following statements:

1. If n is even and $4 \leq n \leq 20$, then n can be written as a sum of 2 prime numbers.

Proof:

$$\begin{array}{cccccc} 4=2+2 & 8=3+5 & 12=5+7 & 16=5+11 & 20=7+13 \\ 6=3+3 & 10=5+5 & 14=11+3 & 18=7+11 & \end{array}$$

2. Every even positive integer n which are less than 26 can be written as a sum of less than or equal to 3 perfect squares.

$$\begin{array}{cccc} \text{Solution.} & 2 = 1^2 + 1^2 & 8 = 2^2 + 2^2 & 14 = 3^2 + 2^2 + 1^2 & 20 = 4^2 + 2^2 \\ & 4 = 2^2 & 10 = 3^2 + 1^2 & 16 = 4^2 & 22 = 3^2 + 3^2 + 2^2 \\ & 6 = 1^2 + 1^2 + 2^2 & 12 = 2^2 + 2^2 + 2^2 & 18 = 3^2 + 3^2 & 24 = 4^2 + 2^2 + 2^2 \end{array}$$

□

t5:q4

Example 3.25 (Tutorial 3, Q7). Prove the following universal statement by using method of exhaustion.

For each integer n with $1 \leq n \leq 10$, $n^2 - n + 11$ is a prime number.

$$\begin{array}{l} \text{Solution. } 1^2 - 1 + 11 = 11, 2^2 - 2 + 11 = 13, 3^2 - 3 + 11 = 17, \\ 4^2 - 4 + 11 = 23, 5^2 - 5 + 11 = 31, 6^2 - 6 + 11 = 41, \\ 7^2 - 7 + 11 = 53, 8^2 - 8 + 11 = 67, 9^2 - 9 + 11 = 83, \\ 10^2 - 10 + 11 = 101 \text{ are all prime numbers.} \end{array}$$

□

eg:4.4b

Example 3.26. Suppose r and s are integers. Prove that “there is an integer k such that $22r + 18s = 2k$ ”.

Proof: Let $k = 11r + 9s$. Then k is an integer because it is a sum of products of integers and $2k = 2(11r + 9s) = 22r + 18s$.

Example 3.27. Prove that the sum of any two even integers is even.

Remark: In this case you might imagine some pairs of even integers, say $2 + 4$, $6 + 10$, $12 + 12$, $28 + 54$, and mentally check that their sums are even. However, since you cannot possibly check all pairs of even numbers, you cannot know for sure that the statement is true in general by checking its truth in these particular instances. Many properties hold for a large number of examples and yet fail to be true in general.

To prove this statement in general, you need to show that no matter what even integers are given, their sum is even.

Proof: Suppose m and n are any even integers.

By definition of even, $m = 2r$ and $n = 2s$ for some integers r and s .

Then $m + n = 2r + 2s = 2(r + s)$.

Let $k = r + s$. k is an integer because it is a sum of integers r and s .

Hence $m + n = 2k$ where k is an integer. It follows by definition of even that $m + n$ is even.

Proving in Lean 4

```
theorem even_plus_even (h1 : is_even a) (h2 : is_even b) : is_even (a
+ b) :=
  Exists.elim h1 (fun w1 (hw1 : a = 2 * w1) =>
    Exists.elim h2 (fun w2 (hw2 : b = 2 * w2) =>
      Exists.intro (w1 + w2)
        (calc a + b
          _ = 2 * w1 + 2 * w2 := by rw [hw1, hw2]
          _ = 2 * (w1 + w2)   := by rw [Int.mul_add])))
```

eg:4.9b

Example 3.28. Prove that the product of any two even integers is even.

Proof: Suppose m and n are any even integers.

By definition of even, $m = 2r$ and $n = 2s$ for some integers r and s .

Then $mn = (2r) \cdot (2s) = 2(2rs)$.

Let $k = 2rs$. Note that k is an integer because it is a product of integers.

Hence $mn = 2k$ where k is an integer.

It follows by definition of even that mn is even.

Proving in Lean 4

```
theorem even_mul_even (h1 : is_even a) (h2 : is_even b) : is_even (a
* b) :=
  Exists.elim h1 (fun w1 (hw1 : a = 2 * w1) =>
    Exists.intro (w1*b)
      (calc a*b
        _ = (2 * w1) * b := by rw [hw1]
        _ = 2 * (w1 * b) := by rw [Int.mul_assoc]))
```

eg:4.9a

Example 3.29. Prove that the sum of any two odd integers is even.

Proof. Formally: $\forall m \in \mathbb{Z}, \forall n \in \mathbb{Z}, m \text{ and } n \text{ are odd} \rightarrow m + n \text{ is even.}$

Suppose m and n are any odd integers.

By definition of odd, $m = 2r + 1$ and $n = 2s + 1$ for some integers r and s .

Then $m + n = (2r + 1) + (2s + 1) = 2r + 2s + 2 = 2(r + s + 1)$.

Let $k = r + s + 1$. Note that k is an integer because it is a sum of integers.

Hence $m + n = 2k$ where k is an integer.

It follows by definition of even that $m + n$ is even. □

t5:q5

Example 3.30 (Tutorial 3, Q8). Prove the following universal statements:

1. For all integers n , if n is odd then n^2 is odd.
2. If n is any odd integer, then $(-1)^n = -1$.

Solution. 1. When n is odd, by definition, $n = 2k + 1$ for some integer k . Hence, $n^2 = (2k + 1)^2 = 2(2k^2 + 2k) + 1$ is odd where $2k^2 + 2k$ is integer.

2. When n is odd, by definition, $n = 2k + 1$ for some integer k . Hence, $(-1)^n = (-1)^{2k+1} = -(-1)^{2k} = -1^k = -1$. □

t5:q2

Example 3.31 (Tutorial 3, Q6). Prove the following existential statements:

1. There are distinct integers m and n such that $1/m + 1/n$ is an integer.
2. There are real numbers a and b such that $\sqrt{a+b} \neq \sqrt{a} + \sqrt{b}$.
3. There is an integer n such that $2n^2 - 5n + 2$ is prime.

Solution. 1. There are $m = 1$ and $n = -1$ such that $1/m + 1/n = 1 + (-1) = 0$.

2. There are $a = b = 2$ such that $\sqrt{2+2} = 2 \neq 2\sqrt{2}$ since $\sqrt{2} > 1$.

3. When $n = 3$, $2n^2 - 5n + 2 = 5$ is a prime number. □

Example 3.32 (Tutorial 3, Q5). Use the rules of inference and real number axioms to prove that $\forall x(3 < x \rightarrow 25 < x^2 + 5x + 2)$.

Solution.

1	x	$3 < x$	assumption, x is arbitrary real number
2		$3x < x^2$	1, $0 < x$, O4
3		$3 \times 3 < 3x$	1, $0 < 3$, O4
4		$9 < x^2$	2,3, O2
5		$5 \times 3 < 5x$	1, $0 < 5$, O4
6		$15 + 9 < 5x + 9 = 9 + 5x$	1, O3, A1
7		$9 + 5x < x^2 + 5x$	4, O3
8		$24 < x^2 + 5x$	6,7 O2
9		$24 + 2 < x^2 + 5x + 2$	8, O3
10		$25 < 26$	fact of $\mathbb{R}$
11		$25 < x^2 + 5x + 2$	9,10 O2
12		$\forall x(3 < x \rightarrow 25 < x^2 + 5x + 2)$	1–11, CP

□

eg:4.12

Theorem 3.33. Prove that the sum of any two rational numbers is rational.

Proof. Suppose r and s are rational numbers. By definition of rational number, $r = a/b$

and $s = c/d$ for some integers $a, b \neq 0, c$ and $d \neq 0$. Thus

$$r + s = \frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}$$

Since the products and sums of integers are integers, $ad + bc$ and bd are integers. Since $b \neq 0$ and $d \neq 0$, $bd \neq 0$.

Therefore, $r + s$ is a ratio of integers and it is rational by definition. $\square$

A **corollary** is a statement whose truth can be immediately deduced from a theorem that has already been proved.

Corollary 3.34. *The double of a rational number is rational.*

3.2.2 Proof by Contraposition/Contrapositive Proof

Proof by contraposition or **contrapositive proof** for a conditional statement $\phi \rightarrow \psi$ is based on the logical equivalence $\phi \rightarrow \psi \equiv \sim \psi \rightarrow \sim \phi$. Formally, contrapositive proof of (3.2) can be expressed as

$$A_1, \dots, A_n, T_1, \dots, T_m \vdash \sim \psi \rightarrow \sim \phi.$$

The outline of the proof is given below:

1. Rewrite the statement $\phi \rightarrow \psi$ to be proved in the contrapositive form $\sim \psi \rightarrow \sim \phi$
2. Prove the contrapositive form using direct proof.

eg:4.28

Example 3.35. Prove that for all integers n , if n^2 is even then n is even.

Proof: In contrapositive form: For all integers n , if n is odd then n^2 is odd.

Suppose n is any odd integer. Then $n = 2k + 1$ for some integer k . $n^2 = (2k + 1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1$ So n^2 is odd.

Remark 3.36. The proof is similar to the direct proof given in Example 3.30.

Example 3.37. Prove that for all integers m and n , if $m + n$ is odd, then m is odd or n is odd.

Proof: Suppose that m is even **and** n is even. By definition, $m = 2k_1$ for some k_1 and $n = 2k_2$ for some k_2 .

Therefore $m + n = 2k_1 + 2k_2 = 2(k_1 + k_2)$ where $k_1 + k_2$ is an integer. By definition, $m + n$ is even. $\square$

Example 3.38. Prove that for all integers n , if $3n + 2$ is odd then n is odd.

Proof. In contrapositive form: For all integers n , if n is even then $3n + 2$ is even.

Suppose n is any even integer. Then $n = 2k$ for some integer k . $3n + 2 = 6k + 2 = 2(3k + 1)$. Here $3k + 1$ is an integer. So $3n + 2$ is even. $\square$

Example 3.39. Prove that for any natural numbers n, a and b , if $n = ab$, then $a \leq \sqrt{n}$ or $b \leq \sqrt{n}$.

Proof. Suppose $a > \sqrt{n}$ and $b > \sqrt{n}$, then $a \cdot b > \sqrt{n} \cdot \sqrt{n} = n$ which means $a \cdot b \neq n$. $\square$

t6:q20

Example 3.40 (Tutorial 3, Q13). Prove the following statements by contraposition.

1. If a product of two positive real numbers is greater than 100, then at least one of the numbers is greater than 10.

Solution. Suppose $r \leq 10$ and $s \leq 10$ are positive real numbers. By arithmetic

$$0 \leq rs \leq 10 \times s \leq 10 \times 10 = 100.$$

□

2. If a sum of two real numbers is less than 50, then at least one of the numbers is less than 25.

Solution. Suppose a and b are two real numbers such that $a \geq 25$ and $b \geq 25$. Then $a + b \geq a + 25 \geq 25 + 25 = 50$. □

3.2.3 Proof by Contradiction

Proof by contradiction for (3.1) can be expressed as

$$A_1, \dots, A_n, T_1, \dots, T_m \vdash \sim \psi \rightarrow F. \quad (3.3)$$

e:contrad

The *proof by contradiction* for a conditional statement $\phi \rightarrow \psi$ can be expressed as

$$A_1, \dots, A_n, T_1, \dots, T_m \vdash (\phi \wedge \sim \psi) \rightarrow F. \quad (3.4)$$

e:contrad:cond

This technique is very similar to the tableaux method for checking the validity of an argument.

eg:4.10c

Example 3.41. For all integers m and n , if $mn = 1$ then $m = n = 1$ or $m = n = -1$.

This is of the form (3.4).

Domain: $\mathbb{Z}$. m and n are integers and axioms and theorems of integers must the obey.

Hypotheses ϕ : “ $mn = 1$ ”.

Conclusion ψ : “ $m = n = 1$ or $m = n = -1$ ”.

Proof by contradiction.

Suppose “ $mn = 1$ ” is true but “ $(m = 1 \wedge n = 1) \vee (m = -1 \wedge n = -1)$ ” is false.

So $(m, n \in \mathbb{Z}) \wedge ((m \neq 1) \vee (n \neq 1)) \wedge ((m \neq -1) \vee (n \neq -1))$. This is logically equivalent to

$$(m \neq 1 \wedge m \neq -1) \vee (n \neq 1 \wedge m \neq -1) \vee (m \neq 1 \wedge n \neq -1) \vee (n \neq 1 \wedge n \neq -1).$$

We can classify into the following cases:

1. $m \neq 1 \wedge m \neq -1$:

- (a) $m < -1$:

- i. $n = 0$: $mn = 0$. Contradicting with $mn = 1$.
- ii. $n \leq -1$: $mn > 1$. Contradicting with $mn = 1$.
- iii. $n \geq 1$: $mn < -1$. Contradicting with $mn = 1$.

- (b) $m = 0$: $mn = 0$. Contradicting with $mn = 1$.

- (c) $m > 1$:

- i. $n = 0$: $mn = 0$. Contradicting with $mn = 1$.
- ii. $n \leq -1$: $mn < -1$. Contradicting with $mn = 1$.
- iii. $n \geq 1$: $mn > 1$. Contradicting with $mn = 1$.

2. $n \neq 1 \wedge m \neq -1$:

- (a) $n > 1 \wedge m > -1$: $mn = 0 \vee mn > 1$. Contradicting with $mn = 1$.

- (b) $n > 1 \wedge m < -1$: $mn < -1$. Contradicting with $mn = 1$.

- (c) $n < 1 \wedge m > -1$: $mn = 0 \vee mn < 0$. Contradicting with $mn = 1$.

(d) $n < 1 \wedge m < -1$: $mn = 0 \vee mn > 1$. Contradicting with $mn = 1$.

3. $m \neq 1 \wedge n \neq -1$: This is similar to case 2., the only difference is m and n are exchanged.

4. $n \neq 1 \wedge n \neq -1$: This is similar to case 1., the only difference is m and n are exchanged.

All situation leads to contradiction. Hence, the conclusion cannot be false and hence the statement is proved.

eg:2.1b

Example 3.42. Show that the rational number $\frac{1}{4}$ is not an integer.

Proof: This is of the form (3.3). We will suppose $\sim \psi$ and then obtain a contradiction (with known results).

Suppose the rational number $\frac{1}{4}$ is an integer. Then

$$4 \times \frac{1}{4} = 1$$

and 1 can be factorised into two integers different from 1 and -1 . This is contradicting with Example 3.41.

eg:4.25

Theorem 3.43. *There is no greatest integer.*

Proof: Suppose there is a greatest integer N . Then $n \leq N$ for every integer n .

Let $M = N + 1$. Now M is an integer since it is a sum of integers. Also $N < N + 1 = M$ by O3 from Definition 3.1.

Thus M is an integer that is greater than N .

However, N is the greatest integer, so $M < N$. Hence $M < N \wedge N < M$ which is a contradiction violating O1 of Definition 3.1.

Thus the supposition is false and “there is no greatest integer” is true.

t6:q19c

Example 3.44 (Tutorial 3, Q14(a)). Use proof by contradiction to prove that there is no greatest even integer.

Solution. Suppose there is a greatest even integer. Let N be the greatest even integer. Then N is an even integer and $n \leq N$ for every even integer n .

Consider $M = N + 2$. M is even since both N and 2 are even. $N < M$ by axiom of integer.

However, from assumption $M < N < M$ which leads to $M < M$ which contradicts with the axiom of integer. $\square$

Theorem 3.45. *Using the method of proof by contradiction to prove Theorem 3.2, i.e. “there is no integer that is both even and odd.”*

Proof: Suppose there is an integer n that is both even and odd.

By definition of even, $n = 2a$ for some integer a . By definition of odd, $n = 2b + 1$ for some integer b .

By equality

$$2a = 2b + 1 \Rightarrow 2(a - b) = 1, \quad a - b \in \mathbb{Z}.$$

This is a contradiction since 1 cannot be factorised as proved in Example 3.41.

eg:4.29

Example 3.46. Use proof by contradiction to prove that for all integers n , if n^2 is even then n is even.

Proof. Suppose that n^2 is even but n is not even (therefore odd by Theorem [thm:odd+even](#) 3.2). Hence, $n = 2k + 1$ for some integer k and

$$n^2 = (2k + 1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1$$

So n^2 is both odd and even. A contradiction (to Theorem [thm:odd+even](#) 3.2). $\square$

t6:q19d

Example 3.47 (Tutorial 3, Q14(b)). Use proof by contradiction to prove that “for all real numbers x and y , if x is irrational and y is rational then $x - y$ is irrational.”

Solution. Suppose there are real numbers x and y such that x is irrational, y is rational and $x - y$ is rational. By definition, $y = a/b$ and $x - y = c/d$ for some integers $a, b \neq 0, c, d \neq 0$. Hence,

$$x - y = x - \frac{a}{b} = \frac{c}{d} \Rightarrow x = \frac{ad + bc}{bd}.$$

Since both $bc + ad$ and $bd \neq 0$ are integers, x is a rational number which contradicts with the assumption that x is irrational. $\square$

3.2.4 Classical Theorems: Irrationality of $\sqrt{2}$; Infinitude of Primes

There are two classical theorems which are derived using the proof by contradiction (using integer divisibility results from Section [sec:divisibility](#) 3.5).

thm:sqrt2

Theorem 3.48. $\sqrt{2}$ is irrational.

Proof: Suppose $\sqrt{2}$ is rational. Then there are integers m and n with no common factors and $n \neq 0$ such that

$$\sqrt{2} = \frac{m}{n}.$$

This implies

$$2 = \frac{m^2}{n^2} \Rightarrow 2n^2 = m^2$$

So m^2 is even and m must be even (otherwise it must be odd, but the square of odd number must be odd) and there is an integer k such that $m = 2k$. Hence

$$2n^2 = m^2 = 4k^2 \Rightarrow n^2 = 2k^2.$$

Now, n^2 is even and so n is even. This implies that both m and n have a common factor of 2, which contradicts the supposition that m and n have no common factors.

Using Theorem [thm:sqrt2](#) 3.48, we can show that something exists without finding the actual thing that matches the predicate. This is called the *non-constructive proof* and is demonstrated below.

Example 3.49. Prove that “There exist irrational numbers x and y s.t. x^y is rational”.

Proof. Consider the number $\sqrt{2}^{\sqrt{2}}$, it can either be (i) rational or (ii) irrational, but not both by definition.

Case (i)

Let $x = y = \sqrt{2}$, then x^y is rational.

Case (ii)

Let $x = \sqrt{2}^{\sqrt{2}}$ and $y = \sqrt{2}$, so both are irrational (the proof of $\sqrt{2}$ being irrational is given in Theorem [thm:sqrt2](#) 3.48). Then, $x^y = (\sqrt{2}^{\sqrt{2}})^{\sqrt{2}} = (\sqrt{2})^2 = 2$. Thus, x^y is rational. $\square$

thm:sqrt

Theorem 3.50 (Generalisation of Theorem 3.48). *For a positive integer k , if $\sqrt{k}$ is not integer, then $\sqrt{k}$ is irrational.*

Remark: In mathematics, a proof by infinite descent is a particular kind of proof by contradiction which relies on the facts that the natural numbers are well ordered and that there are only a finite number of them that are smaller than any given one. One typical application is to show that a given equation has no solutions.

Assuming an example with a particular property exists, one shows that another exists that is in some sense ‘smaller’ as measured by a natural number. Then by infinitely repeating the same step, one shows there are a yet smaller, then a yet even smaller, example, and hence there must be an infinitude of ever smaller examples. Since there are only a finite number of natural numbers smaller than the size of the initially postulated example, this is impossible — it is a contradiction, so no such initial example can exist.

Proof: Let $\sqrt{k}$ be a non-integer and rational. Then $E = \{b \in \mathbb{N}^* : \exists a(a \in \mathbb{N} \wedge \sqrt{k} = a/b)\} \neq \emptyset$ and by the Well-ordering principle 3.1, E contains the smallest value b_1 .

Let $\sqrt{k} = a_1/b_1$, $a_1 > 0$, q be the largest positive integer no greater than $\sqrt{k}$, i.e. $\sqrt{k} - 1 < q < \sqrt{k}$. Then

$$\sqrt{k} = \frac{a_1}{b_1} = \frac{a_1(\sqrt{k} - q)}{b_1(\sqrt{k} - q)} = \frac{a_1\sqrt{k} - a_1q}{b_1((\frac{a_1}{b_1}) - q)} = \frac{b_1\sqrt{k} \times \sqrt{k} - a_1q}{a_1 - b_1q} = \frac{b_1k - a_1q}{a_1 - b_1q}$$

Let $a_2 = b_1k - a_1q$ and $b_2 = a_1 - b_1q$. Since $q < \sqrt{k}$,

$$\begin{aligned} a_2 &= b_1k - a_1q = a_1\left(\frac{b_1}{a_1}k - q\right) = a_1\left(\frac{k}{\sqrt{k}} - q\right) = a_1(\sqrt{k} - q) > 0 \\ b_2 &= a_1 - b_1q = b_1\left(\frac{a_1}{b_1} - q\right) = b_1(\sqrt{k} - q) \end{aligned}$$

This implies $b_2 - b_1 = b_1(\sqrt{k} - q) - b_1 = b_1(\sqrt{k} - q - 1) < 0 \Rightarrow b_2 < b_1$.

The positive integer b_2 is smaller than b_1 . This contradicts with the definition of b_1 .

Remark: It is called “infinite descent” because, we can repeat the process above to get

$$\sqrt{k} = \frac{a_1}{b_1} = \frac{a_2}{b_2} = \frac{a_3}{b_3} = \dots$$

where $a_3 = b_2k - a_2q$ and $b_3 = a_2 - b_2q, \dots$. The contradiction happens because $b_1 > b_2 > b_3 > \dots$ but in $\mathbb{N}$, we cannot have a strictly decreasing sequence.

Example 3.51. Prove that $\log_2 3$ is rational.

Proof: There are integers m, n such that

$$\log_2 3 = \frac{m}{n}$$

with $n \neq 0$. That would imply that

$$2^{m/n} = 3 \Rightarrow 2^m = 3^n.$$

This implies that an even number equals an odd number. A contradiction.

In general, any $\log_a b$ where $a, b \in \mathbb{Z}$ and a, b are mutually prime is irrational, since $x^a = y^b$ and Unique Factorisation Theorem 3.90 leads to a contradiction.

Example 3.52. Prove that $\log_{10} 2$ is rational.

Proof. **Proof:** Suppose that it is rational, therefore there are integers $m, n \neq 0$ such that

$$\log_{10} 2 = \frac{m}{n} \Rightarrow 2 = 10^{m/n} \Rightarrow 2^n = 2^m \cdot 5^m \Rightarrow 2^{n-m} = 5^m.$$

Since the left hand side is even while the right hand side is odd, we have a contradiction. $\square$

eg: 4.27

Example 3.53. Use proof by contradiction to show that the sum of any rational number and any irrational number is irrational.

Proof. Suppose there is a rational number r and an irrational number s such that $r + s$ is rational.

By definition of rational, $r = a/b$ and $r + s = c/d$ for some integers a, b, c and d with $b \neq 0$ and $d \neq 0$. Then

$$r + s = \frac{a}{b} + s = \frac{c}{d} \Rightarrow s = \frac{c}{d} - \frac{a}{b} = \frac{bc - ad}{bd}$$

Both $bc - ad$ and bd are integers and $bd \neq 0$. Thus s is rational. This contradicts the supposition that s is irrational (by definition, irrational cannot be written as a ratio of integers).

Hence the sum of any rational number and any irrational number is irrational. $\square$

eg: 4.30

Example 3.54. Prove that $1 + 3\sqrt{2}$ is irrational.

Proof. Suppose $1 + 3\sqrt{2}$ is rational. Then

$$1 + 3\sqrt{2} = \frac{a}{b} \quad (*)$$

for some integers a and $b \neq 0$. Rearranging (*), we have

$$3\sqrt{2} = \frac{a}{b} - 1 = \frac{a - b}{b} \Rightarrow \sqrt{2} = \frac{a - b}{3b}$$

Both $a - b$ and $3b \neq 0$ are integers. Hence $\sqrt{2}$ is rational, this contradicts with the fact that $\sqrt{2}$ is irrational as demonstrated in Theorem 3.2. thm: odd+even

Hence $1 + 3\sqrt{2}$ is irrational. $\square$

Theorem 3.55 (The Infinitude of the Set of Prime Numbers). *The set of prime numbers is infinite.*

Proof: Suppose the set of prime numbers is finite. Then all the prime numbers can be listed, say, in ascending order:

$$p_1 = 2, p_2 = 3, p_3 = 5, p_4 = 7, p_5 = 11, \dots, p_n.$$

Consider the integer

$$N = p_1 p_2 p_3 \cdots p_n + 1 > 1$$

By Theorem of Divisibility by a Prime (Theorem 3.78), N is divisible by some prime number p . Since p is prime, p must equal one of the prime numbers $p_1, p_2, p_3, \dots, p_n$. Let $p = p_k$ thm: 4.4

for some $1 \leq k \leq n$ and

$$\begin{aligned} p &= p_k | (N = p_1 p_2 p_3 \cdots p_k \cdots p_n + 1) \\ \Rightarrow p_k m &= p_1 p_2 p_3 \cdots p_k \cdots p_n + 1 \\ \Rightarrow p_k (m - p_1 p_2 p_3 \cdots p_{k-1} p_{k+1} \cdots p_n) &= 1 \end{aligned}$$

for some integer m . This implies 1 can be factorised into two integers which are larger than 1, a contradiction.

3.3 Disproving Statements

Mathematical statements may not always be true. That is why we need to prove them! Sometimes, we may want to disprove them if we cannot prove them.

Disproving a universal statement may not be difficult if we are able to find a counterexample as illustrated in Section 3.3.1.

Disproving an existential statement would be tougher since we need to prove that its negation, which is a universal statement, is true, as demonstrated in Section 3.3.2.

3.3.1 Disproving Universal Statements with Counterexamples

Disproving a universal statement

$$\forall x(P(x) \rightarrow Q(x)) \quad (3.5)$$

is the same as assuming that

$$\forall x(P(x) \rightarrow Q(x)) \rightarrow F \Rightarrow \sim \forall x(P(x) \rightarrow Q(x)) \equiv \exists x(P(x) \wedge \sim Q(x)).$$

Hence, to disprove (3.5) becomes **finding** a value s , called the **counterexample**, such that $P(s) \wedge \sim Q(s)$ is true.

eg: 4.5a

Example 3.56. Disprove the statement “for real numbers n , if n is even, then $\frac{n+2}{2}$ is even” by finding a counterexample.

Solution: Formally, we have $\forall n(\text{even}(n) \rightarrow \text{even}(\frac{n+2}{2}))$.

To disprove this statement: Let $n = 4$. 4 is even but $\frac{4+2}{2} = 3$ is not even.

eg: 4.5b

Example 3.57. Disprove the following statement by finding a counterexample:

For all real numbers a and b , if $a^2 = b^2$ then $a = b$.

Solution: Formally, the statement can be expressed as

$$\forall a \forall b (a \in \mathbb{R} \wedge b \in \mathbb{R} \rightarrow (a^2 = b^2 \rightarrow a = b))$$

Though it is of different form from (3.5), to disprove it, we determine its negation

$$\begin{aligned} \sim \forall a \forall b (a \in \mathbb{R} \wedge b \in \mathbb{R} \rightarrow (a^2 = b^2 \rightarrow a = b)) &\equiv \exists a \exists b \sim (a \in \mathbb{R} \wedge b \in \mathbb{R} \rightarrow (a^2 = b^2 \rightarrow a = b)) \\ &\equiv \exists a \exists b (a \in \mathbb{R} \wedge b \in \mathbb{R} \wedge \sim (a^2 = b^2 \rightarrow a = b)) \\ &\equiv \exists a \exists b (a \in \mathbb{R} \wedge b \in \mathbb{R} \wedge a^2 = b^2 \wedge a \neq b) \end{aligned}$$

This statement is true, i.e. there are $a = 1$ and $b = -1$ such that $a^2 = 1^2 = 1$, $b^2 = (-1)^2 = 1$ and $a^2 = b^2$ but $a = 1 \neq b = -1$. The original statement is disproved with the counterexample $a = 1$ and $b = -1$.

eg: 4.5c

Example 3.58. Disprove the statement “For real numbers a and b , if $a > b$ then $a^2 > b^2$.”

Solution. Let $a = -1$, $b = -2$. Then $a > b$ but $a^2 = 1 < b^2 = 4$. $\square$

eg:4.5d

Example 3.59. Disprove the following statement

For integers m and n , if $2m + n$ is even, then m and n are both even.

Solution. Let $m = 1$ and $n = 2$. Then $2m + n = 4$ is even but $m = 1$ is not even. $\square$

t5:q13

Example 3.60 (Tutorial 3, Q18). Prove or disprove the following statements:

- (a) Every positive integer is the sum of the squares of three integers.

Solution. False. 7 cannot be written as $x^2 + y^2 + z^2$. [Try all possible values of $x, z, y = 0, 1, 2$.] $\square$

- (b) There are 100 consecutive positive integers that are not perfect squares (an integer which can be written as s^2 for some integer s).

Solution. True. 10001, $\dots$, 10100 are all not perfect squares since $100^2 = 10000$ and $101^2 = 10201$. $\square$

t5:q3

Example 3.61 (Tutorial 3, Q19). Disprove the following universal statements:

- (a) For all real numbers a and b , if $a < b$, then $a^2 < b^2$.

- (b) For all integers m and n , if $2m + n$ is odd, then m and n are both odd.

Solution. 1. When $a = -2$ and $b = -1$, $a < b$ but $a^2 = 4 > b^2 = 1$.
2. When $m = 2$, $n = 1$, $2m + n = 5$ is odd but m is even. $\square$

t5:q7

Example 3.62 (Tutorial 3, Q21). Determine whether the statement is true or false. Justify your answer with a proof or a counterexample, as appropriate.

- (a) The product of any two even integers is even.

Solution. True. Let m and n be any even integer. By definition, $m = 2k$ for some integer k . Since $mn = 2(kn)$ where kn is an integer. The product of two even integers m and n , mn is even. $\square$

- (b) For all integers m , if $m > 2$, then $m^2 - 4$ is composite.

Solution. False. When $m = 3$, $m^2 - 4 = 5$ is a prime number. $\square$

- (c) For all integers n and m , if $n - m$ is even, then $n^3 - m^3$ is even.

Solution. True. Since $n - m$ is even, $n - m = 2k$ for some integer k . From $n^3 - m^3 = (n - m)(n^2 + mn + m^2) = 2[k(n^2 + mn + m^2)]$, we know that $n^3 - m^3$ is even since $k(n^2 + mn + m^2)$ is an integer. $\square$

- (d) For all integers n , $n^2 - n + 11$ is a prime number.

Solution. False. When $n = 11$, $n^2 - n + 11 = 11^2$ is a composite number. $\square$

t5:q10b

- (e) The quotient of any two rational numbers is a rational number.

Solution. False. For a rational number 1 and 0, $1/0$ is undefined. $\square$

- (f) If r and s are any two rational numbers, then $\frac{r+s}{2}$ is rational.

Solution. True. Since $r = \frac{a}{b}$ and $s = \frac{c}{d}$ for some integers $a, c, b \neq 0$ and $d \neq 0$. Hence
$$\frac{r+s}{2} = \frac{\frac{a}{b} + \frac{c}{d}}{2} = \frac{\frac{ad+bc}{bd}}{2} = \frac{ad+bc}{2bd}.$$
Since $ad + bc$ is an integer and $2bd$ is a non-zero integer, $\frac{r+s}{2}$ is a rational number (the

ratio of integers).

□

3.3.2 Disproving Existential Statements

Disproving an existential statement

$$\exists x(P(x) \wedge Q(x)) \quad (3.6)$$

is the same as assuming that

$$\exists x(P(x) \wedge Q(x)) \rightarrow F \Rightarrow \sim \exists x(P(x) \wedge Q(x)) \equiv \forall x(P(x) \rightarrow \sim Q(x)).$$

Hence, to disprove (3.6) becomes **making sure that nothing can make $P(x) \wedge Q(x)$ true.**

Example 3.63. Disprove the following quantified statement over the real number domain:

$$\exists x((x \in \mathbb{R}) \wedge (x^2 + x + 1 = 0)).$$

Note: In Topic 1, I regard the $x^2 + x + 1 = 0$ as a predicate, not a statement. Once it is quantified (in this case), it becomes a statement.

Disproof: The negation is true: $\forall x((x \in \mathbb{R}) \rightarrow ((x + \frac{1}{2})^2 + \frac{3}{4}) > 0)$.

Example 3.64. Show that there is no positive integer n such that $n^2 + 3n + 2$ is prime.

Proof: Let $P(n)$ be the predicate denoting “ n is prime”. Then

$$\sim \exists n((n > 0) \wedge P(n^2 + 3n + 2)) \equiv \forall n((n > 0) \rightarrow \sim P(n^2 + 3n + 2))$$

is true since

$$n^2 + 3n + 2 = (n + 1)(n + 2), \quad n \geq 1$$

is a product of integers $n + 1 > 1$ and $n + 2 > 1$.

By Definition 3.5 (a prime number can only have two factors 1 and itself), $n^2 + 3n + 2$ has four factors 1, $n + 1$, $n + 2$, $n^2 + 3n + 2$ and is not prime.

Example 3.65 (Tutorial 3, Q17). Disprove the existential statement “There exists an integer n such that $6n^2 + 27$ is prime.”

Solution. For all integer n , $6n^2 + 27 = 3(n^2 + 9)$. Since 3 and $n^2 + 9$ are integers greater than 1, by definition, $6n^2 + 27$ is not a prime number. □

Example 3.66 (Tutorial 3, Q20). Consider the following existential statement:

There exists an integer x with $x \geq 4$ such that $2x^2 - 5x + 2$ is prime. (*)

1. Give a negation of the statement (*).

Solution. For all integers x with $x \geq 4$, $2x^2 - 5x + 2$ is not prime. □

2. Prove that the statement (*) is false by showing that its negation is true.

Solution. Suppose x is any integer with $x \geq 4$. Note that

$$2x^2 - 5x + 2 = (2x - 1)(x - 2) \quad (\dagger)$$

and $2x - 1 \geq 8 - 1 = 7$, $x - 2 \geq 2$ are integers. Since (†) is a product of two integers greater than 1, it is not a prime number. Hence the statement (*) is false. □

3.4 Mathematical Induction

sec:mind

Mathematical induction is a method of proof developed to check conjectures about the outcomes of processes that occur repeatedly and according to definite patterns based on the induction principle (page 67).

Let P be a predicate of integer k . Proving a statement

$$\forall n(n \geq a \rightarrow P(n))$$

by mathematical induction is a **two-step process**:

1. **Basis Step:** Show that the $P(a)$ is true for a particular integer a .
2. **Inductive Step:** Show that for all integers $k \geq a$, if $P(k)$ is true then $P(k + 1)$ is true.

To perform this step, assume that the property is true for $n = k$ for some integer $k \geq a$. This supposition is called the *inductive hypothesis*. Then show that the property is true for $n = k + 1$.

It is based on the following *principle of ordinary mathematical induction*.

Theorem 3.67 (Principle of Ordinary Mathematical Induction). *Let $P(n)$ be a predicate that is defined for integers n , and let a be a fixed integer. Suppose the following two statements are true:*

1. $P(a)$ is true.
2. For all integers $k \geq a$, if $P(k)$ is true then $P(k + 1)$ is true.

Then the statement “for all integers $n \geq a$, $P(n)$ ” is true.

eg:4.31

Example 3.68 (Induction involving equality). Use mathematical induction to prove that the sum of the first n odd positive integers is n^2 for $n \geq 1$.

Proof:

Let $P(n)$: “ $1 + 3 + \cdots + (2n - 1) = n^2$ ”.

Basis step: Show that $P(1)$ is true.

$$\text{LHS of } P(1) = 1, \quad \text{RHS of } P(1) = 1^2 = 1 \Rightarrow \text{LHS} = \text{RHS}$$

So $P(1)$ is true.

Inductive step: Suppose that $P(k)$ is true for a positive integer $k \geq 1$, that is, assume

$$1 + 3 + 5 + \cdots + (2k - 1) = k^2.$$

and we try to show

$$P(k + 1) : 1 + 3 + 5 + \cdots + (2k - 1) + (2(k + 1) - 1) = (k + 1)^2.$$

is true.

$$\begin{aligned} \text{LHS of } P(k + 1) &= \underbrace{1 + 3 + 5 + \cdots + (2k - 1)}_{\text{using assumption}} + (2k + 1) \\ &= k^2 + (2k + 1) = (k + 1)^2 = \text{RHS of } P(k + 1). \end{aligned}$$

Hence $P(k + 1)$ is true.

By mathematical induction, $P(n)$ is true for all $n \geq 1$.

eg:4.32

Example 3.69 (Induction involving equality). Use mathematical induction to prove that

$$\sum_{i=1}^n i = 1 + 2 + \cdots + n = \frac{n(n+1)}{2}.$$

Proof. Let $P(n)$: “ $1 + 2 + \cdots + n = \frac{n(n+1)}{2}$ ”.

Basis Step: Show that $P(1)$ is true.

LHS = 1, RHS = $\frac{1(1+1)}{2} = 1$.

So LHS of $P(1)$ = RHS of $P(1)$ and $P(1)$ is true.

Inductive Step: Suppose $P(k)$ is true for a positive integer $k \geq 1$ that is

$$\sum_{i=1}^k i = 1 + 2 + \cdots + k = \frac{k(k+1)}{2}.$$

Show that $P(k+1)$ is true:

$$\begin{aligned} \text{LHS of } P(k+1) &= \underbrace{1 + 2 + \cdots + k}_{\text{using assumption}} + k + 1 \\ &= \frac{k(k+1)}{2} + k + 1 = \frac{k(k+1) + 2(k+1)}{2} \\ &= \frac{(k+1)(k+1+1)}{2} = \text{RHS of } P(k+1). \end{aligned}$$

So $P(k+1)$ is true. By mathematical induction, $P(n)$ is true for all $n \geq 1$. □

eg:4.34

Example 3.70. Use mathematical induction to prove that the equality $\sum_{i=1}^n i^3 = \left[\frac{n(n+1)}{2} \right]^2$.

Proof. Let $P(n)$: “ $1^3 + 2^3 + \cdots + n^3 = \left[\frac{n(n+1)}{2} \right]^2$ ”.

Basis step: Show that $P(1)$ is true.

LHS = $1^3 = 1$ and RHS = $\left[\frac{1(1+1)}{2} \right]^2 = 1$. So LHS = RHS and $P(1)$ is true

Inductive step: Suppose $P(k)$ is true for a positive integer $k \geq 1$:

$$1^3 + 2^3 + \cdots + k^3 = \left[\frac{k(k+1)}{2} \right]^2$$

Show that $P(k+1)$ is true:

$$\begin{aligned} \text{LHS of } P(k+1) &= \underbrace{1^3 + 2^3 + \cdots + k^3}_{\text{using assumption}} + (k+1)^3 \\ &= \left[\frac{k(k+1)}{2} \right]^2 + (k+1)^3 = \frac{k^2(k+1)^2 + 4(k+1)^3}{4} = \frac{(k+1)^2[k^2 + 4(k+1)]}{2} \\ &= \frac{(k+1)^2(k+2)^2}{2} = \frac{(k+1)^2((k+1)+1)^2}{2} = \text{RHS of } P(k+1). \end{aligned}$$

Hence $P(k+1)$ is true.

By mathematical induction, $P(n)$ is true for all $n \geq 1$. □

We have learn the special cases of binomial theorem, i.e.

$$(x+y)^2 = x^2 + 2xy + y^2$$

$$(x+y)^3 = x^3 + 3x^2y + 3xy^2 + y^3$$

$$(x+y)^4 = x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4$$

where the coefficients above are related to **Pascal triangle**.

Theorem 3.71 (Binomial Theorem, must memorised). Let $n \geq 0$ [Epp, 2020, Chapter 9].

$$(x + y)^n = \sum_{i=0}^n \binom{n}{i} x^{n-i} y^i = x^n + \binom{n}{1} x^{n-1} y + \binom{n}{2} x^{n-2} y^2 + \cdots + \binom{n}{n-1} x y^{n-1} + y^n.$$

Proof:

Basis step

$$LHS = (x + y)^0 = 1; \quad RHS = \binom{0}{0} x^0 y^0 = 1$$

Inductive step

$$\text{Assume that } (x + y)^k = \sum_{i=0}^k \binom{k}{i} x^{k-i} y^i.$$

$$\begin{aligned} (x + y)^{k+1} &= (x + y)^k \cdot (x + y) = \left(\sum_{i=0}^k \binom{k}{i} x^{k-i} y^i \right) \cdot (x + y) = \sum_{i=0}^k \binom{k}{i} x^{k-i+1} y^i + \sum_{i=0}^k \binom{k}{i} x^{k-i} y^{i+1} \\ &= \binom{k}{0} x^{k+1} + \sum_{i=1}^k \binom{k}{i} x^{k-i+1} y^i + \sum_{i=0}^{k-1} \binom{k}{i} x^{k-i} y^{i+1} + \binom{k}{k} y^{k+1} \\ &= x^{k+1} + \sum_{i=0}^{k-1} \binom{k}{i+1} x^{k-(i+1)+1} y^{i+1} + \sum_{i=0}^{k-1} \binom{k}{i} x^{k-i} y^{i+1} + y^{k+1} \\ &= x^{k+1} + \sum_{i=0}^{k-1} \left[\binom{k}{i+1} + \binom{k}{i} \right] x^{k-i} y^{i+1} + y^{k+1} \\ &= \binom{k+1}{0} x^{k+1} + \sum_{i=0}^{k-1} \binom{k+1}{i+1} x^{k-i} y^{i+1} + \binom{k+1}{k+1} y^{k+1} \\ &= \binom{k+1}{0} x^{k+1} + \sum_{i'=1}^k \binom{k+1}{i'} x^{k-i'+1} y^{i'} + \binom{k+1}{k+1} y^{k+1} = \sum_{i'=0}^{k+1} \binom{k+1}{i'} x^{k+1-i'} y^{i'} \end{aligned}$$

$$\text{Note: } \binom{k}{0} = \binom{k}{k} = 1, \quad \binom{k+1}{0} = \binom{k+1}{k+1} = 1 \text{ and}$$

$$\begin{aligned} \binom{k}{i+1} + \binom{k}{i} &= \frac{k(k-1) \cdots (k-i)}{(i+1)i(i-1) \cdots 1} + \frac{k(k-1) \cdots (k-i+1)}{i(i-1) \cdots 1} = \frac{k(k-1) \cdots (k-i+1)}{i(i-1) \cdots 1} \left[\frac{k-i}{i+1} + 1 \right] \\ &= \frac{k(k-1) \cdots (k-i+1)}{i(i-1) \cdots 1} \left[\frac{k+1}{i+1} \right] = \binom{k+1}{i+1} \end{aligned}$$

eg: 4.35

Example 3.72. Use mathematical induction or direct proof to prove that for all integers $n \geq 1$, $2^{2n} - 1$ is divisible by 3.

Proof by mathematical induction:

Let $P(n)$: “ $3 \mid 2^{2n} - 1$ ”.

Basis Step: Show that $P(1)$ is true.

$2^{2(1)} - 1 = 3$ is divisible by 3. So $P(1)$ is true.

Inductive Step:

Suppose $P(k)$ is true for a positive integer $k \geq 1$, that is $3|2^{2k} - 1$. We must show that $2^{2(k+1)} - 1$ is divisible by 3.

$3|2^{2k} - 1 \Rightarrow 2^{2k} - 1 = 3a$ for some integer a .

$$\begin{aligned} 2^{2(k+1)} - 1 &= 2^{2k+2} - 1 = 4(2^{2k}) - 1 = 3(2^{2k}) + (2^{2k} - 1) \\ &= 3(2^{2k}) + 3a = 3(2^{2k} + a) \Rightarrow 3|2^{2(k+1)} - 1. \end{aligned}$$

Thus $P(k+1)$ is true.

Hence, by mathematical induction, $P(n)$ is true for all integers $n \geq 1$.

Proof: By using the Binomial Theorem, we have

$$\begin{aligned} 2^{2n} - 1 &= 4^n - 1 = (3+1)^n - 1 = 3^n + \binom{n}{1}3^{n-1} + \cdots + \binom{n}{n-1}3 + 1 - 1 \\ &= 3 \left[3^{n-1} + \binom{n}{1}3^{n-2} + \cdots + \binom{n}{n-1} \right] \end{aligned}$$

By definition $3|(2^{2n} - 1)$.

eg:4.36

Example 3.73 (Induction with inequality). Use mathematical induction to prove that

$$2n + 1 < 2^n, \quad \text{for all integers } n \geq 3.$$

Proof. Let $P(n)$: " $2n + 1 < 2^n$ ".

Basis Step: Show that $P(3)$ is true.

LHS of $P(3) = 2(3) + 1 = 7$, RHS of $P(3) = 2^3 = 8$.

Hence LHS of $P(3) < \text{RHS of } P(3)$ and so $P(3)$ is true.

Inductive Step: Suppose $P(k)$ is true for some integer $k \geq 3$:

$$2k + 1 < 2^k$$

Show that $P(k+1)$ is true:

$$\begin{aligned} \text{LHS of } P(k+1) &= 2(k+1) + 1 = \underbrace{(2k+1)}_{\text{using assumption}} + 2 \\ &< 2^k + 2 < 2^k + \underbrace{2^k}_{k \geq 3 \Rightarrow 2 < 2^k} = 2(2^k) = 2^{k+1} = \text{RHS of } P(k+1). \end{aligned}$$

So $P(k+1)$ is true.

Hence, by mathematical induction, $P(n)$ is true for all integers $n \geq 3$. □

Refer to Tutorial 3 for answers to June 2024 Final Exam Q3 solutions.

3.5 Divisibility

ec:divisibility

Divisibility is the basic property of integers. In this section, various concepts and proof techniques associated with divisibility used in Section [3.2](#) and Section [3.4](#) are discussed. sec:improof sec:mind

def:2.1

Definition 3.13. If n and $d \neq 0$ are integers, then n **is divisible by** d if there is an integer k such that $n = dk$. In this case, n is called the **multiple of** d . d is called the **factor** or **divisor** of n . We also say that d **divides** n and denote it by $d|n$. If d does not divide n , we denote it as $d \nmid n$.

Remark 3.74. Based on our definition, *divisors are assumed to be nonzero*. If d is a divisor of n , then n is also divisible by $-d$ (indeed, $n = dk$ implies that $n = (-d)(-k)$), so that the divisors of an integer always occur in pairs. To find all the divisors of a given integer, it is sufficient to obtain the positive divisors and then adjoin to them the corresponding negative integers. For this reason, we shall usually limit ourselves to a consideration of positive divisors.

eg:4.14

Example 3.75. 1. If a and b are integers, is $4a + 4b$ divisible by 2?

Solution. Yes. $4a + 4b = 2(2a + 2b)$, and $2a + 2b$ is an integer. $\square$

eg:4.13b

2. Does 4 divides 18?

Solution. No, since $18 = 4 \times 4 + 2 \neq 4k$ for any integer k . Hence 4 does not divide 18. $\square$

3. Is 32 a multiple of -16 ?

Solution. Yes, $32 = (-16)(-2)$. $\square$

4. Is -9 a factor of 54?

Solution. Yes, $54 = (-9)(-6)$. $\square$

5. Suppose a and b are positive integers and $a|b$. Is $a \leq b$?

Solution. Yes. When $a > 0$, $b > 0$ and $a|b$, $b = ka$ for some integer $k > 0$. Hence $b = a + (k - 1)a \geq a$, where $k \geq 1$. $\square$

3.5.1 Properties of Divisibility

It will be helpful to list some immediate consequences of Definition [3.13](#). ^{def:2.1}

thm:2.2

Theorem 3.76. For integers a, b, c , the following hold:

(a) $a|0$, $1|a$, $a|a$.

(b) $a|1$ iff $a = \pm 1$.

(c) If $a|b$ and $c|d$, then $ac|bd$.

(d) If $a|b$ and $b|c$, then $a|c$. (Transitivity of Divisibility)

(e) $a|b$ and $b|a$ iff $a = \pm b$.

(f) If $a|b$ and $b \neq 0$, then $|a| \leq |b|$.

(g) If $a|b$ and $a|c$, then $a|(bx + cy)$ for arbitrary integers x and y .

thm:div:trans

Proof:

(a) $0 = 0(a)$, $a = 1(a)$, $a = a(1)$.

(b) $a|1 \Rightarrow 1 = ak \Rightarrow a = \pm 1$.

$a = \pm 1 \Rightarrow 1 = a^2 \Rightarrow a|1$ by definition.

(c) Suppose $a|b$ and $c|d$, by definition, $b = ak_1$, $d = ck_2$. Hence, $bd = ac(k_1k_2)$. By definition, $ac|bd$.

(d) Suppose a, b and c are integers such that $a|b$ and $b|c$. Then there are integers r and s such that

$$b = ar, \quad c = bs$$

Hence, $c = (ar)s = a(rs) = ak$, where $k = rs$ is an integer since it is a product of integers r and s . Thus $a|c$.

(e) Suppose $a|b$, $b|a$, by definition, $b = ak_1$ and $a = bk_2$, so $b = bk_2k_1$. Hence $k_2k_1 = 1$ and $k_2 = k_1 = \pm 1$. Therefore, $a = bk_2 = \pm b$.

- (f) If $a|b$, then there exists an integer c such that $b = ac$; also, $b \neq 0$ implies that $c \neq 0$. Upon taking absolute values, we get $|b| = |ac| = |a||c|$. Because $c \neq 0$, it follows that $|c| \geq 1$, whence $|b| = |a||c| \geq |a|$.
- (g) The relations $a|b$ and $a|c$ ensure that $b = ar$ and $c = as$ for suitable integers r and s . But then whatever the choice of x and y ,

$$bx + cy = arx + asy = a(rx + sy).$$

Because $rx + sy$ is an integer, this says that $a|(bx + cy)$, as desired.

Remark 3.77. It is worth pointing out that property (g) of Theorem 3.76 extends by induction to sums of more than two terms. That is, if $a|b_k$ for $k = 1, 2, \dots, n$, then $a|(b_1x_1 + \dots + b_nx_n)$ for all integers $x_1, \dots, x_n$. thm:2.2

thm:4.4

Theorem 3.78 (Divisibility by a Prime). *Any integer $n > 1$ is divisible by a prime number.*

Proof: Suppose n is a integer that is greater than 1. If n is prime, then n is divisible by a prime number (namely itself), and we are done.

If n is not prime, then

$$n = r_0s_0$$

for some integers $1 < r_0 < n$ and $1 < s_0 < n$. By definition of divisibility, $r_0|n$.

If r_0 is prime, then r_0 is a prime number that divides n , and we are done.

If r_0 is not prime, then

$$r_0 = r_1s_1$$

for some integers $1 < r_1 < r_0$ and $1 < s_1 < r_0$. By definition of divisibility, $r_1|r_0$. But $r_0|n$. Consequently, by transitivity of divisibility (Theorem d), $r_1|n$. thm:div:trans

If r_1 is prime, then r_1 is a prime number that divides n , and we are done.

If r_1 is not prime, then

$$r_1 = r_2s_2$$

for some integers $1 < r_2 < r_1$ and $1 < s_2 < r_1$. By definition of divisibility, $r_2|r_1$. But $r_1|n$. Consequently, by transitivity of divisibility (Theorem d), $r_2|n$. thm:div:trans

If r_2 is prime, then r_2 is a prime number that divides n , and we are done.

If r_2 is not prime, then we may repeat the above process by factoring r_2 as r_3s_3 . Continuing in this way, factoring successive factors of n until we find a prime factor.

We must succeed in a finite number of steps because each new factor is both less than the previous one and greater than 1, and there are fewer than n integers strictly between 1 and n (justified by well-ordering principle). Thus we obtain a sequence $r_0, r_1, r_2, \dots, r_k$, where $k \geq 0$, $1 < r_k < r_{k-1} < \dots < r_2 < r_1 < r_0 < n$, and $r_i|n$ for $i = 0, 1, 2, \dots, k$. The condition for termination is that r_k should be prime. Hence r_k is a prime number that divided n .

sec:gcd

3.5.2 Greatest Common Divisor

An integer d is said to be a **common divisor/factor of two integers a and b** if $d|a$ and $d|b$. Because 1 is a divisor of every integer, 1 is a common divisor of a and b ; hence, the set of positive common divisors for a and b is nonempty.

The most important common divisor is the greatest common divisor/factor (abbreviated as $\gcd(a, b)$) which has applications in modular arithmetic (Section 3.6) and thus encryption algorithms such as RSA (Section 3.8.4). sec:modular sec:RSA

def:gcd

Definition 3.14. Let a and b be given integers, with at least one of them different from zero. The **greatest common divisor of a and b** , denoted $\gcd(a, b)$, is an integer d with the following properties:

1. d is a common divisor of both a and b , i.e. $d|a$ and $d|b$.
2. For all integers c , if $c|a$ and $c|b$, then $c \leq d$.

eg:2.2

Example 3.79. Determine the greatest common divisor of -12 and 30 .

Solution:

Positive divisors of $-12 = \{1, 2, 3, 4, 6, 12\}$

Positive divisors of $30 = \{1, 2, 3, 5, 6, 10, 15, 30\}$

Positive common divisors of -12 and 30 are $\{1, 2, 3, 6\}$.

Because 6 is the largest of these integers, $\gcd(-12, 30) = 6$ by definition.

Example 3.80. $\gcd(-5, 5) = 5$, $\gcd(8, 17) = 1$, $\gcd(-8, -36) = 4$.

The next theorem indicates that $\gcd(a, b)$ can be represented as a linear combination of a and b .

thm:2.3

Theorem 3.81. *Given integers a and b , not both of which are zero, there exist integers x and y such that*

$$\gcd(a, b) = ax + by.$$

Proof: Consider the set S of all positive linear combinations of a and b :

$$S = \{au + bv | au + bv > 0; u, v \text{ integers}\}$$

Notice first that $S \neq \emptyset$. For example, if $a \neq 0$, then the integer $|a| = au + b \cdot 0$ lies in S , where we choose $u = 1$ or $u = -1$ according as a is positive or negative.

By virtue of the Well-Ordering Principle, S must contain a smallest element d . Thus, from the very definition of S , there exist integers x and y for which $d = ax + by$.

We claim that $d = \gcd(a, b)$.

Taking stock of the Division Algorithm, we can obtain integers q and r such that $a = qd + r$, where $0 \leq r < d$. Then r can be written in the form

$$r = a - qd = a - q(ax + by) = a(1 - qx) + b(-qy).$$

If r were positive, then this representation would imply that r is a member of S , contradicting the fact that d is the least integer in S (recall that $r < d$). Therefore, $r = 0$, and so $a = qd$, or equivalently $d|a$.

By similar reasoning, $d|b$, the effect of which is to make d a common divisor of a and b .

Now if c is an arbitrary positive common divisor of the integers a and b , then part (g) of Theorem 3.76 allows us to conclude that $c|(ax + by)$; that is, $c|d$. By part (f) of the same theorem, $c = |c| \leq |d| = d$, so that d is greater than every positive common divisor of a and b . Piecing the bits of information together, we see that $d = \gcd(a, b)$ by Definition 3.14.

A perusal of the proof of Theorem 3.81 reveals that the greatest common divisor of a and b may be described as the smallest positive integer of the form $ax + by$.

Example 3.82. Consider the case in which $a = 6$ and $b = 15$. Here, the set S becomes

$$S = \{6u + 15v | u, v \in \mathbb{Z}\} = \{6(-2) + 15 \cdot 1, 6(-1) + 15 \cdot 1, 6 \cdot 1 + 15 \cdot 0, \dots\} = \{3, 9, 6, \dots\}.$$

We observe that 3 is the smallest integer in S , whence $3 = \gcd(6, 15)$.

The nature of the members of S appearing in this illustration suggests another result, which we give in the next corollary.

Corollary 3.83. *If a and b are given integers, not both zero, then the set $T = \{ax + by | x, y \in \mathbb{Z}\}$ is precisely the set of all multiples of $d = \gcd(a, b)$.*

Proof: $\Rightarrow$: Because $d|a$ and $d|b$, we know that $d|(ax + by)$ for all integers x, y . Thus, every member of T is a multiple of d .

$\Leftarrow$: Conversely, d may be written as $d = ax_0 + by_0$ for suitable integers x_0 and y_0 , so that any multiple nd of d is of the form

$$nd = n(ax_0 + by_0) = a(nx_0) + b(ny_0).$$

Hence, nd is a linear combination of a and b , and, by definition, lies in T .

It may happen that 1 and -1 are the only common divisors of a given pair of integers a and b , whence $\gcd(a, b) = 1$. E.g. $\gcd(2, 5) = \gcd(-9, 16) = \gcd(-27, -35) = 1$. This situation occurs often enough to prompt a definition.

def:4.13

Definition 3.15. Integers a and b , not both of which are zero, are called **relatively prime** or **mutually prime** or **coprime** if $\gcd(a, b) = 1$. Integers $a_1, a_2, a_3, \dots, a_n$ are **(pairwise) relatively prime** if $\gcd(a_i, a_j) = 1$ for all integers i and j with $1 \leq i, j \leq n$, and $i \neq j$.

The following theorem characterises relatively prime integers in terms of linear combinations.

thm:2.4

Theorem 3.84. *Let a and b be integers, not both zero. Then a and b are relatively prime iff there exist integers x and y such that $1 = ax + by$.*

Proof:

$\Rightarrow$: If a and b are relatively prime so that $\gcd(a, b) = 1$, then Theorem 3.81 guarantees the existence of integers x and y satisfying $1 = ax + by$.

$\Leftarrow$: As for the converse, suppose that $1 = ax + by$ for some choice of x and y , and that $d = \gcd(a, b)$. Because $d|a$ and $d|b$, Theorem 3.76 yields $d|(ax + by)$, or $d|1$. Inasmuch as d is a positive integer, this last divisibility condition forces d to equal 1 (part (b) of Theorem 3.76 plays a role here), and the desired conclusion follows.

This result leads to an observation that is useful in certain situations; namely,

Corollary 3.85. *If $\gcd(a, b) = d$, then $\gcd(a/d, b/d) = 1$.*

Proof: The fractions a/d and b/d are integers because d is a divisor both of a and of b .

Now, knowing that $\gcd(a, b) = d$, it is possible to find integers x and y such that $d = ax + by$. Upon dividing each side of this equation by d , we obtain the expression

$$1 = \left(\frac{a}{d}\right)x + \left(\frac{b}{d}\right)y.$$

Because a/d and b/d are integers, an appeal to the theorem is legitimate. The conclusion is that a/d and b/d are relatively prime.

Example 3.86. As an illustration of this corollary, observe that

$$\gcd(-12, 30) = 6 \Rightarrow \gcd(-12/6, 30/6) = \gcd(-2, 5) = 1$$

Corollary 3.87. *If $a|c$ and $b|c$, with $\gcd(a, b) = 1$, then $ab|c$.*

Proof: Inasmuch as $a|c$ and $b|c$, integers r and s can be found such that $c = ar = bs$. Now the relation $\gcd(a, b) = 1$ allows us to write $1 = ax + by$ for some choice of integers x and y . Multiplying the last equation by c , it appears that

$$c = c \cdot 1 = c(ax + by) = acx + bcy.$$

If the appropriate substitutions are now made on the right-hand side, then

$$c = a(bs)x + b(ar)y = ab(sx + ry) \text{ or, as a divisibility statement, } ab|c.$$

Our next result seems mild enough, but is of fundamental importance.

thm:euclid-lemma

Theorem 3.88 (Euclid's lemma). *If $a|bc$, with $\gcd(a, b) = 1$, then $a|c$.*

Proof: From Theorem [3.81](#), writing $1 = ax + by$, where x and y are integers. Multiplication of this equation by c produces

$$c = 1 \cdot c = (ax + by)c = acx + bcy.$$

Because $a|ac$ and $a|bc$, it follows that $a|(acx + bcy)$, which can be recast as $a|c$.

The subsequent theorem often serves as a definition of $\gcd(a, b)$. The advantage of using it as a definition is that order relationship is not involved. Thus, it may be used in algebraic systems having no order relation.

thm:2.6

Theorem 3.89. *Let a, b be integers, not both zero. For a positive integer d , $d = \gcd(a, b)$ iff*

(a) $d|a$ and $d|b$.

(b) Whenever $c|a$ and $c|b$, then $c|d$.

Proof:

$\Rightarrow$: Suppose that $d = \gcd(a, b)$. Certainly, $d|a$ and $d|b$, so that (a) holds. In light of Theorem [3.81](#), d is expressible as $d = ax + by$ for some integers x, y . Thus, if $c|a$ and $c|b$, then $c|(ax + by)$, or rather $c|d$. In short, condition (b) holds.

$\Leftarrow$: Let d be any positive integer satisfying the stated conditions. Given any common divisor c of a and b , we have $c|d$ from hypothesis (b). The implication is that $d \geq c$, and consequently d is the greatest common divisor of a and b .

3.5.3 Unique Factorisation Theorem & Standard Factor Form

Prime numbers are special factors because they are the only factors greater than 1. Hence, a natural question arises, how many prime numbers are there? Can integer greater than one be written as a product of prime(s)?

Proposition 3.1. *If p is a prime and $a, b \in \mathbb{Z}$ such that $p|ab$, then $p|a$ or $p|b$.*

Proof: Assume $p \nmid a$. Then $\gcd(a, p) = 1$. By Euclid's Lemma [3.88](#), $p|b$. □

Any integer > 1 is either prime or can be written as a product of primes. The following theorem characterises this property.

thm:uniq-fac

Theorem 3.90 (Unique Factorisation Theorem for the Integers). *Given any integer $n > 1$, there is a positive integer k , distinct prime numbers $p_1, p_2, \dots, p_k$, and there are positive integers $e_1, e_2, \dots, e_k$ such that*

$$n = p_1^{e_1} p_2^{e_2} p_3^{e_3} \cdots p_k^{e_k},$$

and any other expression of n as a product of prime numbers is identical to this except, perhaps, for the order in which the factors are written.

Definition 3.16. Given any integer $n > 1$, the **standard factored form** of n is an expression of the form

$$n = p_1^{e_1} p_2^{e_2} p_3^{e_3} \cdots p_k^{e_k}, \quad p_1 < p_2 < \cdots < p_k,$$

where k is a positive integer; $p_1, p_2, \dots, p_k$ are prime numbers; $e_1, e_2, \dots, e_k$ are positive integers.

Standard Factored Form in Scheme

```
;; https://wiki.haskell.org/99_questions/Solutions/35
(define (primeFactors n)
  (let primeFactorsAux ((n n) (f 2))
    (cond
      ((> (* f f) n) (list n)) ; n is a prime if it has no smaller factors
      ((= (modulo n f) 0) (cons f (primeFactorsAux (quotient n f) f)))
      (else (primeFactorsAux n (+ f 1)))))

(display (primeFactors 28234423783))
(newline)
```

eg:4.16

Example 3.91. Suppose m is an integer such that

$$8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot m = 17 \cdot 16 \cdot 15 \cdot 14 \cdot 13 \cdot 12 \cdot 11 \cdot 10.$$

Does $17|m$?

Solution. Since 17 is one of the prime factor of the right-hand side of the equation, it is also a prime factor of the left-hand side (by the unique factorisation theorem). But 17 does not equal any prime factor of 8, 7, 6, 5, 4, 3 or 2. Hence 17 must be a prime factor of m and so $17|m$. $\square$

Example 3.92. In Topic 1, Example 1.1.1, write 28234423783 in standard factored form.

Solution: Using the Standard ML program, we obtain a standard factored form

$$28234423783 = 229 * 641 * 192347$$

eg:4.17

Example 3.93. Write 3300 in standard factored form.

Solution. $3300 = 100 \cdot 33 = 4 \cdot 25 \cdot 3 \cdot 11 = 2 \cdot 2 \cdot 5 \cdot 5 \cdot 3 \cdot 11 = 2^2 \cdot 3^1 \cdot 5^2 \cdot 11^1$ $\square$

3.5.4 Quotient-Remainder Theorem, div and mod

sec:div-mod

thm:quo-rem

Theorem 3.94 (Quotient-Remainder Theorem). *Given any integer n and positive integer d , there exist unique integers q and r such that*

$$n = qd + r, \quad 0 \leq r < d.$$

Definition 3.17. With the notation of Theorem [3.94](#), r is called the **remainder** of the division of n by d . If $r = 0$, we say that n is a **multiple** of d , or that n is **divisible by** d , or d is a **divisor of** n , or that d **divides** n , or that d is a **factor of** n . The number q is called the **quotient** of n by d and is denoted $n \text{ div } d$.

We can also define the modulo operator:

$$n \bmod d = r.$$

eg:4.18

Example 3.95. For each values of n and d , find integers q and r such that $n = dq + r$ and $0 \leq r < d$.

	n	d	q	r	q, r with Standard ML
(i)	54	4			$54 \text{ div } 4, 54 \bmod 4$
(ii)	-54	4			$\sim 54 \text{ div } 4, \sim 54 \bmod 4$
(iii)	54	70			$54 \text{ div } 70, 54 \bmod 70$
(iv)	-54	70			$\sim 54 \text{ div } 70, \sim 54 \bmod 70$
(v)	32	9			$32 \text{ div } 9, 32 \bmod 9$

Solution. (i) $54 = 4 \cdot 13 + 2$; hence $q = 13$ and $r = 2$.

(ii) $-54 = 4 \cdot (-14) + 2$; hence $q = -14$ and $r = 2$.

(iii) $54 = 70 \cdot 0 + 54$; hence $q = 0$ and $r = 54$.

(iv) $-54 = 70 \cdot (-1) + 16$; hence $q = -1$ and $r = 16$.

(v) $32 = 9(3) + 5$, $q = 32 \text{ div } 9 = 3$ and $r = 32 \bmod 9 = 5$. □

Example 3.96. What day of the week will it be 1 year from today?

Solution. There are 365 days in a year that is not a leap year, and each week has 7 days. Now since $365 = 52(7) + 1$,

$$365 \text{ div } 7 = 52, \quad 365 \bmod 7 = 1$$

Thus 52 weeks, or 364 days, from today will be a today, and so 365 days from today will be 1 day later, namely tomorrow. □

3.5.5 Floor and Ceiling

It is often convenient to transform the problems of divisibility from the set of integers $\mathbb{Z}$ to the set of real numbers $\mathbb{R}$ and vice versa. The two operations, floor and ceiling, that relates $\mathbb{Z}$ and $\mathbb{R}$ are defined below.

Definition 3.18. Given any real number x , the *floor* of x , denoted $\lfloor x \rfloor$, is defined as follows:

$$\lfloor x \rfloor := \text{the unique integer } n \text{ such that } n \leq x < n + 1.$$

Symbolically,

$$\lfloor x \rfloor = n, \quad n \leq x < n + 1.$$

Definition 3.19. Given any real number x , the *ceiling* of x , denoted $\lceil x \rceil$, is defined as follows:

$$\lceil x \rceil := \text{the unique integer } n \text{ such that } n - 1 < x \leq n.$$

Symbolically,

$$\lceil x \rceil = n, \quad n - 1 < x \leq n.$$

eg:4.20

Example 3.97. Compute $\lfloor x \rfloor$ and $\lceil x \rceil$ for each of the following values of x .

	x	$\lfloor x \rfloor$	$\lceil x \rceil$	Standard ML
(i)	$25/4$			floor (25.0/4.0); ceil (25.0/4.0);
(ii)	0.999			floor (0.999); ceil (0.999);
(iii)	$0.999 \dots$			floor (1.0); ceil (1.0);
(iv)	-2.01			floor (~2.01); ceil (~2.01);
(v)	$\lfloor -\frac{1}{2} \rfloor + \lfloor \frac{2}{3} \rfloor$			already in integer form, not supported

Solution. (i) $25/4 = 6.25$ and $6 < 6.25 < 7$. Hence $\lfloor 25/4 \rfloor = 6$ and $\lceil 25/4 \rceil = 7$.
(ii) Since $0 < 0.999 < 1$, $\lfloor 0.999 \rfloor = 0$ and $\lceil 0.999 \rceil = 1$.
(iii) Since $0.999 \dots = 1$, $\lfloor 0.999 \dots \rfloor = 1$ and $\lceil 0.999 \dots \rceil = 1$.
(iv) $-3 < -2.01 < -2$. Hence $\lfloor -2.01 \rfloor = -3$ and $\lceil -2.01 \rceil = -2$.
(v) $\lfloor -\frac{1}{2} \rfloor + \lfloor \frac{2}{3} \rfloor = -1 + 0 = -1$. Hence, $\lfloor \lfloor -\frac{1}{2} \rfloor + \lfloor \frac{2}{3} \rfloor \rfloor = -1$ and $\lceil \lfloor -\frac{1}{2} \rfloor + \lfloor \frac{2}{3} \rfloor \rceil = -1$ $\square$

eg:4.21

Example 3.98. The 1,370 soldiers at a military base are given the opportunity to take buses into town for an evening out. Each bus holds a maximum of 40 passengers.

- For reasons of economy, the base commander will send only full buses. What is the maximum number of buses the base commander will send?

Solution. $\lfloor 1370/40 \rfloor = \lfloor 34.25 \rfloor = 34$ $\square$

- If the base commander is willing to send a partially filled bus, how many buses will the commander need to allow all the soldiers to take the trip?

Solution. $\lceil 1370/40 \rceil = \lceil 34.25 \rceil = 35$ $\square$

eg:4.22

Example 3.99. If k is an integer, simplify $\lfloor k \rfloor$ and $\lfloor k + \frac{1}{2} \rfloor$ as an expression of k .

Solution. Suppose k is an integer. Then $\lfloor k \rfloor = k$ because k is an integer and $k \leq k < k+1$, and $\lfloor k + \frac{1}{2} \rfloor = k$ because k is an integer and $k < k + \frac{1}{2} < k+1$. $\square$

eg:4.23

Example 3.100. Is the following statement true or false?

For all real numbers x and y , $\lfloor x + y \rfloor = \lfloor x \rfloor + \lfloor y \rfloor$.

Solution. The statement is false. Here's a counterexample: take $x = y = \frac{1}{2}$. Then

$$\lfloor x + y \rfloor = \lfloor 1 \rfloor = 1 \neq \lfloor x \rfloor + \lfloor y \rfloor = 0 + 0 = 0.$$

$\square$

thm:4.6

Theorem 3.101. $\forall x \in \mathbb{R}, \forall m \in \mathbb{Z}, \lfloor x + m \rfloor = \lfloor x \rfloor + m$.

Proof: For any given real number x and any integer m . Let $n = \lfloor x \rfloor$. By definition of floor, n is an integer and $n \leq x < n+1$. Add m to all sides gives

$$n + m \leq x + m < n + m + 1.$$

Now $n + m$ is an integer, and so, by definition of floor, the left-hand side of the equation to be shown is

$$\lfloor x + m \rfloor = n + m = \lfloor x \rfloor + m. \quad \square$$

Theorem 3.102. For any integer n ,

$$\left\lfloor \frac{n}{2} \right\rfloor = \begin{cases} \frac{n}{2}, & \text{if } n \text{ is even,} \\ \frac{n-1}{2}, & \text{if } n \text{ is odd.} \end{cases}$$

Proof: When n is odd, $n = 2k + 1$ for some integer k and

$$\left\lfloor \frac{n}{2} \right\rfloor = \left\lfloor \frac{2k+1}{2} \right\rfloor = \left\lfloor k + \frac{1}{2} \right\rfloor = k + \left\lfloor \frac{1}{2} \right\rfloor = k + 0 = k = \frac{n-1}{2}$$

using Theorem [thm:4.6](#) 3.101.

When n is even, $n = 2k$ for some integer k and

$$\left\lfloor \frac{n}{2} \right\rfloor = \left\lfloor \frac{2k}{2} \right\rfloor = \lfloor k \rfloor = k = \frac{n}{2}$$

□

[thm:4.8](#)

Theorem 3.103 (Raymond T. Boute). If n is an integer and d is a nonzero integer, and if $q = \lfloor \frac{n}{d} \rfloor$ and $r = n - d \lfloor \frac{n}{d} \rfloor$, then

$$n = dq + r, \quad 0 \leq r < d.$$

Proof: Suppose n is a nonnegative integer, d is a positive integer, $q = \lfloor \frac{n}{d} \rfloor$ and $r = n - d \lfloor \frac{n}{d} \rfloor$. Then

$$dq + r = d \left\lfloor \frac{n}{d} \right\rfloor + \left(n - d \left\lfloor \frac{n}{d} \right\rfloor \right) = n$$

and

$$q = \left\lfloor \frac{n}{d} \right\rfloor \Rightarrow q \leq \frac{n}{d} < q + 1 \Rightarrow dq \leq n < dq + d \Rightarrow 0 \leq n - dq < d.$$

But $n - dq = r$. Hence $0 \leq r < d$. □

Remark 3.104. From this we have $n \operatorname{div} d = \lfloor \frac{n}{d} \rfloor$ and $n \bmod d = n - d \lfloor \frac{n}{d} \rfloor$.

3.6 Modular Arithmetic

[sec:modular](#)

Using the modular arithmetic, we can define “congruence”, an equivalence relation over $\mathbb{Z}$ (Section [4.4](#)). Modular arithmetic also has applications in cryptography (Section [3.8](#)) and random number generation (Section [3.9](#)). [sec:equiv](#) [sec:rand-sim](#) [sec:crypt](#)

3.6.1 Properties of Congruence

[def:4.1](#)

Definition 3.20. Let $a, b \in \mathbb{Z}$ and $n \in \mathbb{Z}^+$. We say a and b are **congruent modulo n** provided that $n \mid (a - b)$. We write $a \equiv b \pmod{n}$ or $a \equiv_n b$ which means $a - b = kn$ for some integer k .

When $n \nmid (a - b)$, we say that a is **incongruent to b modulo n** , and in this case we write $a \not\equiv b \pmod{n}$.

Remark 3.105. Note that $\equiv$ for number theory is different from the logical equivalence found in mathematical logic which is used in Topics 1 and 2.

Example 3.106. For $n = 7$,

$$3 \equiv 24 \pmod{7}, \quad -31 \equiv 11 \pmod{7}, \quad -15 \equiv -64 \pmod{7}$$

because $3 - 24 = (-3)7$, $-31 - 11 = (-6)7$, and $-15 - (-64) = 7 \cdot 7$.

$25 \not\equiv 12 \pmod{7}$, because 7 fails to divide $25 - 12 = 13$.

Theorem 3.107. Let a , b , and $n > 1$ be any integers. The following statements are all equivalent:

1. $n \mid (a - b)$
2. $a \equiv b \pmod{n}$
3. $a = b + kn$ for some integer k
4. a and b have the same (nonnegative) remainder when divided by n
5. $a \bmod n = b \bmod n$

Remark 3.108. • Two integers are congruent modulo 2 when they are both even or both odd

- Any two integers are congruent modulo 1
- Congruence modulo 1 is not particularly interesting, the usual practice is to assume that $n > 1$.

Given an integer a , let q and r be its quotient and remainder upon division by n , so that

$$a = qn + r, \quad 0 \leq r < n.$$

Then, by definition of congruence, $a \equiv r \pmod{n}$. Because there are n choices for r , we see that every integer is congruent modulo n to exactly one of the values $0, 1, 2, \dots, n-1$; in particular, $a \equiv 0 \pmod{n}$ iff $n \mid a$. The set of n integers $0, 1, 2, \dots, n-1$ is called the set of **least nonnegative residues modulo n** .

thm:4.1

Theorem 3.109. $\forall a, b \in \mathbb{Z}$, $a \equiv b \pmod{n}$ iff a and b leave the same nonnegative remainder when divided by n .

Theorem 3.110 (Basic Arithmetic of Congruences). Let $n > 1$ be fixed and a, b, c, d be arbitrary integers. Then the following properties hold:

- (a) $a \equiv a \pmod{n}$.
- (b) If $a \equiv b \pmod{n}$, then $b \equiv a \pmod{n}$.
- (c) If $a \equiv b \pmod{n}$ and $b \equiv c \pmod{n}$, then $a \equiv c \pmod{n}$.
- (d) If $a \equiv b \pmod{n}$ and $c \equiv d \pmod{n}$, then $a + c \equiv b + d \pmod{n}$ and $ac \equiv bd \pmod{n}$.
- (e) If $a \equiv b \pmod{n}$, then $a + c \equiv b + c \pmod{n}$ and $ac \equiv bc \pmod{n}$.
- (f) If $a \equiv b \pmod{n}$, then $a^k \equiv b^k \pmod{n}$ for any positive integer k .

Proof:

- (a) For any integer a , we have $a - a = 0 \cdot n$, so that $a \equiv a \pmod{n}$.
- (b) If $a \equiv b \pmod{n}$, then $a - b = kn$ for some integer k . Hence, $b - a = -(kn) = (-k)n$ and because $-k$ is an integer, this yields property (b).
- (c) Suppose that $a \equiv b \pmod{n}$ and also $b \equiv c \pmod{n}$. Then there exist integers h and k satisfying $a - b = hn$ and $b - c = kn$. It follows that

$$a - c = (a - b) + (b - c) = hn + kn = (h + k)n$$

which is $a \equiv c \pmod{n}$ in congruence notation.

- (d) Similar to (c), if $a \equiv b \pmod{n}$ and $c \equiv d \pmod{n}$, then we are assured that $a - b = k_1n$ and $c - d = k_2n$ for some choice of k_1 and k_2 . Adding these equations, we obtain

$$(a + c) - (b + d) = (a - b) + (c - d) = k_1n + k_2n = (k_1 + k_2)n$$

or, as a congruence statement, $a + c \equiv b + d \pmod{n}$.

As regards the second assertion of property (d), note that

$$ac = (b + k_1n)(d + k_2n) = bd + (bk_2 + dk_1 + k_1k_2n)n$$

Because $bk_2 + dk_1 + k_1k_2n$ is an integer, this says that $ac - bd$ is divisible by n , whence $ac \equiv bd \pmod{n}$.

- (e) The proof of property (e) is covered by (d) and the fact that $c \equiv c \pmod{n}$.
- (f) The statement certainly holds for $k = 1$, and we will assume it is true for some fixed k . From (d), we know that $a \equiv b \pmod{n}$ and $a^k \equiv b^k \pmod{n}$ together imply that $aa^k \equiv bb^k \pmod{n}$, or equivalently $a^{k+1} \equiv b^{k+1} \pmod{n}$. Hence, the induction step is complete.

Corollary 3.111 (Congruence n is an Equivalence Relation). *The congruence relation $\equiv_n$ is an equivalence relation on $\mathbb{Z}$ and the map $\{0, 1, \dots, n-1\} \rightarrow \mathbb{Z}/\equiv_n, \quad r \mapsto \bar{r} = r + n\mathbb{Z}$ is a bijection.*

Proof: From the Basic Arithmetic Theorem 3.110 (a), (b) and (c), we know that $\equiv_n$ is reflexive, symmetric and transitive. Therefore $\equiv_n$ is an equivalence relation.

The map is well-defined since for every $r = 0, 1, \dots, n-1$, there is a set $r + n\mathbb{Z}$ corresponding to it.

To show that it is bijection, we show that it is an injection: assume $\bar{r} = \bar{s}$ with $0 \leq r, s < n$. Then, by definition, $r \equiv s \pmod{n}$, so $n|r - s$ and $|r - s| < n$, therefore $r = s$.

To show that the map is a surjection: Let $\bar{a} \in \mathbb{Z}/\equiv_n$, by definition, $\bar{a} = \{a + nk : k \in \mathbb{Z}\}$, by Quotient-Remainder Theorem, there is an $r \in \{0, 1, \dots, n-1\}$ such that $a = r + nm$, hence $r \equiv_n a$ and by definition, $\bar{r} = \bar{a}$.

In mathematics, $\mathbb{Z}/\equiv_n$ is denoted $\mathbb{Z}/n\mathbb{Z}$.

Modular arithmetic is important because all sorts of questions that are difficult to answer with respect to $\mathbb{Z}$ are effectively (though not necessarily efficiently, if n is large) decidable with respect to $\mathbb{Z}/n\mathbb{Z}$. If we can show in this way that something is impossible over $\mathbb{Z}/n\mathbb{Z}$, then this often implies a negative answer for $\mathbb{Z}$, too.

Some of the elementary properties of equality (reflexive, symmetry and transitivity) will make multiplication simple with congruence.

Corollary 3.112. *Let a, b , and n be integers with $n > 1$. Then*

$$ab \equiv [(a \pmod{n})(b \pmod{n})] \pmod{n}$$

or, equivalently,

$$ab \pmod{n} = [(a \pmod{n})(b \pmod{n})] \pmod{n}$$

In particular, if m is a positive integer, then

$$am \equiv [(a \pmod{n})m] \pmod{n}.$$

Note: Differentiate between the congruence relation $\pmod{n}$ and the operation “ $\pmod{n}$ ”.

eg: 4.2

Example 3.113. Show that 41 divides $2^{20} - 1$.

Proof: We begin by noting that $2^5 \equiv -9 \pmod{41}$, whence $(2^5)^4 \equiv (-9)^4 \pmod{41}$ by Theorem 3.110(f); in other words, $2^{20} \equiv 81 \cdot 81 \pmod{41}$. But $81 \equiv -1 \pmod{41}$, and so $81 \cdot 81 \equiv 1 \pmod{41}$. Using parts (b) and (e) of Theorem 3.110, we finally arrive at

$$2^{20} - 1 \equiv 81 \cdot 81 - 1 \equiv 1 - 1 \equiv 0 \pmod{41}$$

Thus, $41|2^{20} - 1$, as desired.

eg:4.38c

Example 3.114. Calculate $12^{43} \bmod 713$.

Solution: When the power is larger, we need to get the “binary” decomposition of the power:

$$43 = 32 + 8 + 2 + 1 = 2^5 + 2^3 + 2 + 1$$

so that we have

$$\begin{aligned} 12^{43} \bmod 713 &= [(12^{32})(12^8)(12^2)(12)] \bmod 713 \\ &= [(12^{32} \bmod 713)(12^8 \bmod 713)(12^2 \bmod 713)(12 \bmod 713)] \bmod 713 \\ &= (485)(629)(144)(12) \bmod 713 = 527152320 \bmod 713 = 48 \end{aligned}$$

where

$$\begin{aligned} 12^2 \bmod 713 &= 144 \\ 12^8 \bmod 713 &= 144^4 \bmod 713 = 629 \\ 12^{16} \bmod 713 &= (692)^2 \bmod 713 = 639 \\ 12^{32} \bmod 713 &= (639)^2 \bmod 713 = 485 \end{aligned}$$

The method mentioned above is not how real-world algorithm performs calculations.

An implementation of modular exponentiation

```
(define (powMod a n m)
  (define (auxPowMod x e acc)
    (cond
      ((zero? e) acc)
      (else (auxPowMod (modulo (* x x) m) (quotient e 2)
                        (if (zero? (modulo e 2)) acc (modulo (* x acc) m))))))
  (modulo (auxPowMod (modulo (abs a) m) n 1) m))

;;; Example 3.6.12
(display (powMod 12 43 713))
(newline)
```

Example 3.115. Work through Example 3.114 using powMod algorithm.

Solution: The initial values $x = 12$, $q = 43$, $m2=1$.

x^2	$q \text{ div } 2$	$q \bmod 2$	m2
$12^2 = 144 \equiv_{713} 144$	$43/2=21$	1	12
$144^2 = 20736 \equiv_{713} 59$	$21/2=10$	1	$12 \times 144 = 1728 \equiv_{713} 302$
$59^2 = 3481 \equiv_{713} 629$	$10/2=5$	0	302
$629^2 = 595641 \equiv_{713} 639$	$5/2=2$	1	$302 \times 629 = 189958 \equiv_{713} 300$
$639^2 = 595641 \equiv_{713} 485$	$2/2=1$	0	300
$485^2 = 595641 \equiv_{713} 648$	$1/2=0$	1	$485 \times 300 \equiv_{713} 48$

eg:4.38b

Example 3.116. Calculate $144^4 \bmod 713$

	x^2	$q \text{ div } 2$	$q \bmod 2$	m2
<i>Solution.</i>	$144^2 = 20736 \equiv_{713} 59$	$4/2 = 2$	0	1
	$59^2 = 3481 \equiv_{713} 629$	$2/2 = 1$	0	1
	$629^2 \equiv_{713} \dots$	$1/2 = 0$	1	$1 \times 629 = 629 \equiv_{713} 629$
$144^4 \bmod 713 = (144^2)^2 \bmod 713 = [144^2 \bmod 713]^2 \bmod 713$				
$= [20736 \bmod 713]^2 \bmod 713 = [59]^2 \bmod 713 = 3481 \bmod 713 = 629$				

□

Theorem 3.117 (Fermat's Little Theorem). *If p is any prime number and a is any integer, then $a^p \equiv a \pmod{p}$. If $p \nmid a$, then $a^{p-1} \equiv 1 \pmod{p}$.*

Example 3.118. Calculate $7^{11^{13}} \pmod{17}$.

Solution. Since 17 is a prime number, we know that $7^{16} \equiv 1 \pmod{17}$

What we need to calculate is

$$11^{13} = 11^{8+4+1} = 11^8 \cdot 11^4 \cdot 11 \equiv 9^4 \cdot 9^2 \cdot 11 = 11 \pmod{16}$$

where $11^2 \equiv 9 \pmod{16}$ and $9^2 \equiv 1$.

Therefore $11^{13} = 16k + 11$ for some integer k and

$$7^{11^{13}} = 7^{16k+11} = (7^{16})^k \cdot 7^{11} \equiv 1^k \cdot 7^{11} = 1977326743 \equiv 14 \pmod{17}.$$

□

3.6.2 Euclidean Algorithm

The **Euclidean algorithm** (also called *Euclid's algorithm*) is an efficient method for computing the greatest common divisor of two integers. It is named after the Greek mathematician Euclid, who described it in Books VII and X of his Elements (https://en.wikipedia.org/wiki/Euclid%27s_Elements).

The algorithm has many theoretical and practical applications. It is a key element of the RSA algorithm (Section 3.8.4). It is used to solve Diophantine equations, such as finding numbers that satisfy multiple congruences (Chinese remainder theorem) or multiplicative inverses of a finite field. It can also be used to construct continued fractions, in the Sturm chain method for finding real roots of a polynomial, and in several modern integer factorisation algorithms. Finally, it is a basic tool for proving theorems in modern number theory, such as Lagrange's four-square theorem and the fundamental theorem of arithmetic (unique factorisation).

The Euclidean algorithm is based on the following two properties: Let $a, b, q, r \in \mathbb{Z}$

- If $a > 0$ then $\gcd(a, 0) = a$.
- If $b \neq 0$, $q \geq 0$, $r \geq 0$, $a = bq + r$ then $\gcd(a, b) = \gcd(b, r)$

An implementation of Euclidean algorithm

```
;;; Euclidean Algorithm for GCD
;;; gcd is available in scheme, so we use mygcd
(define (mygcd a b)
  (if (zero? b) a (mygcd b (modulo a b))))
```

eg:4.39

Example 3.119. Use the Euclidean algorithm to find $\gcd(330, 156)$.

Solution:

$$\begin{aligned} \gcd(330, 156) &= \gcd(156, 18) & [330 = 156(2) + 18] \\ &= \gcd(18, 12) & [156 = 18(8) + 12] \\ &= \gcd(12, 6) & [18 = 12(1) + 6] \\ &= \gcd(6, 0) & [12 = 6(2) + 0] \\ &= 6 \end{aligned}$$

def:4.12

Definition 3.21. An integer d is said to be a *linear combination of integers a and b* if there exist integers s and t such that $as + bt = d$.

thm:4.16

Theorem 3.120. For all integers a and b , not both zero, if $d = \gcd(a, b)$, then there exist integers s and t such that $as + bt = d$.

eg:4.40

Example 3.121. Express $\gcd(330, 156)$ as a linear combination of 330 and 156.

Solution: From Example ^{eg:4.39}3.119, $6 = 18 - 12$

$$\begin{aligned} &= 18 - [156 - 8(18)] \\ &= 18 + (-1)(156) + 8(18) \\ &= 9(18) + (-1)(156) \\ &= 9[330 - 2(156)] + (-1)(156) \\ &= 9(330) + (-18)(156) + (-1)(156) \\ &= 9(330) + (-19)(156) \end{aligned}$$

Hence $\gcd(330, 156) = 9(330) + (-19)(156)$.

eg:4.41

Example 3.122. Show that 660 and 43 are relatively prime, and prove that the statement $\exists u \exists v (660u + 43v = 1)$ is true in the **integer** domain.

Solution.

$$\begin{aligned} \gcd(660, 43) &= \gcd(43, 15) & [660 = 43(15) + 15 \Rightarrow 15 = 660 - 43(15)] \\ &= \gcd(15, 13) & [43 = 15(2) + 13 \Rightarrow 13 = 43 - 15(2)] \\ &= \gcd(13, 2) & [15 = 13(1) + 2 \Rightarrow 2 = 15 - 13] \\ &= \gcd(2, 1) & [13 = 2(6) + 1 \Rightarrow 1 = 13 - 2(6)] \\ &= \gcd(1, 0) & [2 = 1(2) + 0] \\ &= 1 \end{aligned}$$

hence 660 and 43 are relatively prime and

$$\begin{aligned} 1 &= 13 - 2(6) \\ &= 13 - [15 - 13](6) \\ &= 7(13) - 6(15) \\ &= 7[43 - 15(2)] - 6(15) \\ &= 7(43) - 20(15) \\ &= 7(43) - 20[660 - 43(15)] \\ &= 307(43) + (-20)(660) \end{aligned}$$

□

thm:4.3

Theorem 3.123. If $ca \equiv cb \pmod{n}$, then $a \equiv b \pmod{n/d}$, where $d = \gcd(c, n)$.

Theorem ^{thm:4.3}3.123 gets its maximum force when the requirement that $\gcd(c, n) = 1$ is added, for then the cancellation may be accomplished without a change in modulus.

Corollary 3.124. If $ca \equiv cb \pmod{n}$ and $\gcd(c, n) = 1$, then $a \equiv b \pmod{n}$.

Corollary 3.125. If $ca \equiv cb \pmod{p}$ and $p \nmid c$, where p is prime, then $a \equiv b \pmod{p}$.

eg:4.4

Example 3.126. $33 \equiv 15 \pmod{9} \Rightarrow 3 \cdot 11 \equiv 3 \cdot 5 \pmod{9} \Rightarrow 11 \equiv 5 \pmod{3}$

cor:4.4

Corollary 3.127. For all integers a and n , if $\gcd(a, n) = 1$, then there exists an integer s such that $as \equiv 1 \pmod{n}$. The integer s is called the inverse of a modulo n .

eg:4.42

Example 3.128. 1. Find an inverse for 43 modulo 660. That is, find an integer s such that $43s \equiv 1 \pmod{660}$.

Solution. From Example [3.122](#), [eg:4.41](#),

$$307(43) + (-20)(660) = 1 \Rightarrow 307(43) = 1 + 20(660) \Rightarrow 307(43) \equiv 1 \pmod{660}$$

So 307 is an inverse for 43 modulo 660. $\square$

2. Find a positive inverse for 3 modulo 40. That is, find a positive integer s such that $3s \equiv 1 \pmod{40}$.

Solution.

$$\begin{aligned} 40 &= 3(13) + 1 \Rightarrow 1 = 40 - 3(13) \\ &\Rightarrow 3(-13) \equiv 1 \pmod{40} \end{aligned}$$

Thus -13 is an inverse for 3 modulo 40 and 27 is a positive integer that is an inverse for 3 modulo 40. $\square$

3.6.3 Linear Congruences

Solving a general equation in modular arithmetic is impossible. For a linear equation

$$ax \equiv b \pmod{n} \tag{3.7}$$

[e:lin-congr](#)

which is called a **linear congruence**, the theory of linear congruences allows us to find a solution $x_0 \in \mathbb{Z}$ for which $ax_0 \equiv b \pmod{n}$.

By definition, $ax_0 \equiv b \pmod{n}$ iff $n \mid ax_0 - b$ or $ax_0 - b = ny_0$ for some integer y_0 . Thus, the problem of finding all integers that will satisfy the linear congruence [\(3.7\)](#) is identical with that of obtaining all solutions of the **linear Diophantine equation** [e:lin-congr](#)

$$ax - ny = b. \tag{3.8}$$

[e:lin-diophant](#)

[thm:4.7](#)

Theorem 3.129. *The linear congruence [\(3.7\)](#) has a solution iff $d \mid b$, where $d = \gcd(a, n)$. If $d \mid b$, then it has d mutually incongruent solutions modulo n .* [e:lin-congr](#)

Proof: Note that the given congruence is equivalent to the linear Diophantine equation $ax - ny = b$. From Theorem 2.4.4 of Topic 2, it is known that the latter equation can be solved iff $d \mid b$; moreover, if it is solvable and x_0, y_0 is one specific solution, then any other solution has the form

$$x = x_0 + \frac{n}{d}t, \quad y = y_0 + \frac{a}{d}t$$

for some choice of t .

Among the various integers satisfying the first of these formulae, consider those that occur when t takes on the successive values $t = 0, 1, 2, \dots, d-1$:

$$x_0, x_0 + \frac{n}{d}, x_0 + \frac{2n}{d}, \dots, x_0 + \frac{(d-1)n}{d}.$$

We claim that these integers are incongruent modulo n , and all other such integers x are congruent to some one of them. If it happened that

$$x_0 + \frac{n}{d}t_1 \equiv x_0 + \frac{n}{d}t_2 \pmod{n}$$

where $0 \leq t_1 < t_2 \leq d-1$, then we would have

$$\frac{n}{d}t_1 \equiv \frac{n}{d}t_2 \pmod{n}$$

Now $\gcd(n/d, n) = n/d$, by Theorem [3.123](#) the factor n/d could be cancelled leading to [thm:4.3](#)

$$t_1 \equiv t_2 \pmod{d}$$

which is to say that $d|t_2 - t_1$. But this is impossible in view of the inequality $0 < t_2 - t_1 < d$.

It remains to argue that any other solution $x_0 + (n/d)t$ is congruent modulo n to one of the d integers listed above. The Division Algorithm permits us to write t as $t = qd + r$, where $0 \leq r \leq d - 1$. Hence

$$x_0 + \frac{n}{d}t = x_0 + \frac{n}{d}(qd + r) = x_0 + nq + \frac{n}{d}r \equiv x_0 + \frac{n}{d}r \pmod{n}$$

with $x_0 + (n/d)r$ being one of our d selected solutions.

The argument that we gave in Theorem 3.129 brings out a point worth stating explicitly: If x_0 is any solution of $ax \equiv b \pmod{n}$, then the $d = \gcd(a, n)$ incongruent solutions are given by

$$x_0, x_0 + \frac{n}{d}, x_0 + 2\frac{n}{d}, \dots, x_0 + (n-1)\frac{n}{d}.$$

Corollary 3.130. *If $\gcd(a, n) = 1$, then the linear congruence $ax \equiv b \pmod{n}$ has a unique solution modulo n .*

eg:4.7

Example 3.131. Solve the linear congruence $18x \equiv 30 \pmod{42}$.

Solution: Because $\gcd(18, 42) = 6$ and 6 surely divides 30, Theorem 3.129 guarantees the existence of exactly six solutions, which are incongruent modulo 42. By inspection, one solution is found to be $x = 4$.

The six solutions are as follows:

$$x \equiv 4 + (42/6)t = 4 + 7t \pmod{42}, \quad t = 0, 1, \dots, 5$$

or, plainly enumerated, $x \equiv 4, 11, 18, 25, 32, 39 \pmod{42}$.

Given relatively prime integers a and n , the congruence $ax \equiv 1 \pmod{n}$ has a unique solution. This solution is sometimes called the *(multiplicative) inverse of a modulo n* .

eg:4.8

Example 3.132. Solve the linear congruence $9x \equiv 21 \pmod{30}$.

Solution. Since $\gcd(9, 30) = 3$ and $3|21$, there must be three incongruent solutions.

One way to find these solutions is to divide the given congruence through by 3, thereby replacing it by the equivalent congruence $3x \equiv 7 \pmod{10}$. The relative primeness of 3 and 10 implies that the latter congruence admits a unique solution modulo 10. We could test the integers 0, 1, 2, ..., 9 in turn until the solution is obtained.

A better way is this: Multiply both sides of the congruence $3x \equiv 7 \pmod{10}$ by 7 to get

$$21x \equiv 49 \pmod{10}$$

which reduces to $x \equiv 9 \pmod{10}$. But the original congruence was given modulo 30, so that its incongruent solutions are sought among the integers 0, 1, 2, ..., 29.

Taking $t = 0, 1, 2$, in the formula $x = 9 + 10t$ we obtain 9, 19, 29, whence

$$x \equiv 9 \pmod{30}, \quad x \equiv 19 \pmod{30}, \quad x \equiv 29 \pmod{30}$$

are the required three solutions of $9x \equiv 21 \pmod{30}$.

A different approach to the problem is to use the method that is suggested in the proof of Theorem 3.129.

Because the congruence $9x \equiv 21 \pmod{30}$ is equivalent to the linear Diophantine equation $9x - 30y = 21$ we begin by expressing $3 = \gcd(9, 30)$ as a linear combination of 9 and 30. It is found, either by inspection or by using the Euclidean Algorithm, that $3 = 9(-3) + 30 \cdot 1$, so that

$$21 = 7 \cdot 3 = 9(-21) - 30(-7).$$

Thus, $x = -21$, $y = -7$ satisfy the Diophantine equation and, in consequence, all solutions of the congruence in question are to be found from the formula

$$x = -21 + (30/3)t = -21 + 10t.$$

The integers $x = -21 + 10t$, where $t = 0, 1, 2$, are incongruent modulo 30 (but all are congruent modulo 10); thus, we end up with the incongruent solutions

$$x \equiv -21 \pmod{30}, \quad x \equiv -11 \pmod{30}, \quad x \equiv -1 \pmod{30}$$

or, if one prefers positive numbers, $x = 9, 19, 29 \pmod{30}$. □

To solve a general linear congruence (3.7), we need the https://en.wikipedia.org/wiki/Chinese_remainder_theorem which is covered in “Number Theory” course.

3.7 Application: SMT For Number System

The Satisfiability Modulo Theories (SMT), which is mentioned in Topic 1, can combine the semantic logic from in Topics 1 and 2 and some basic arithmetic of the number systems in this topic to form linear integer arithmetic (LIA) theory and linear integer arithmetic (LRA) theory. In this section, explanations and examples from <https://jfmc.github.io/z3-play/> are adapted.

The basic building blocks of SMT formulas are constants and functions. Constants are just functions that take no arguments. So everything is really just a function.

An Integer Example in Z3 SMT-LIB

```
;;; https://jfmc.github.io/z3-play/
(declare-fun f (Int) Int)
(declare-fun a () Int) ; a is a constant
(declare-const b Int) ; syntax sugar for (declare-fun b () Int)
(declare-const u Int)
(declare-const v Int)
(assert (> a 20))
(assert (> b a)) ; Are there numbers (a,b) s.t. a>20 and b>a?
(assert (= (f 10) 1)) ; Is there a function f s.t. f(10) = 1?
(assert (and (not (= u 0)) (not (= v 0))))
(assert (= (+ (* 330 u) (* 156 v)) 6)) ; Example 3.6.20
(check-sat)
(get-model)
```

The SMT solver Z3 has builtin support for integer and real constants with support for some linear arithmetic (and extremely limited nonlinear arithmetic). These two types (sorts) represent the mathematical integers and reals rather than machine integers (32-bit or 64-bit) and floating point numbers.

After constants are declared, the user can assert formulas containing these constants. The formulas contain arithmetic operators such as $+$, $-$, $<$, and so on. The command `check-sat` will instruct Z3 to try to **find an interpretation/model** for the declared constants that makes all formulas true. The interpretation is basically assigning a number to each constant. If such interpretation exists, we say it is a model for the asserted formulas. The command `get-model` displays the model built by Z3.

Real constants should contain a decimal point. Unlike most programming languages, Z3 will not convert automatically integers into reals and vice-versa. The function `to-real` can be used to convert an integer expression into a real one.

Mixed Integer and Real Example in Z3 SMT-LIB

```
;;; https://jfmc.github.io/z3-play/
;;; https://smt-lib.org/examples.shtml
(declare-const a Int)
(declare-const b Int)
```

```

(declare-const c Int)
(declare-const d Real)
(declare-const e Real)
(assert (> a b)) ; a > b
(assert (> a (+ b 2))) ; a > b+2
(assert (= a (+ (* 2 c) 10))) ; a = 2c+10
(assert (<= (+ c b) 1000)) ; c+b <= 1000
(assert (>= d e)) ; d >= e
(assert (> e (+ (to_real (+ a b)) 2.0))) ; e > (a+b)+2
(assert (= d (+ (to_real c) 0.5))) ; d = c + 0.5
(check-sat)
(get-model)
(get-value (a b c d e))

```

Z3 also has support for division, integer division, modulo and remainder operators. Internally, they are all mapped to multiplication.

Mixed Integer and Real Example in Z3 SMT-LIB

```

;;; https://jfmc.github.io/z3-play/
(declare-const a Int)
(declare-const r1 Int)
(declare-const r2 Int)
(declare-const r3 Int)
(declare-const r4 Int)
(declare-const r5 Int)
(declare-const r6 Int)
(assert (= a 10))
(assert (= r1 (div a 4))) ; integer division
(assert (= r2 (mod a 4))) ; mod
(assert (= r3 (rem a 4))) ; remainder
(assert (= r4 (div a (- 4)))) ; integer division
(assert (= r5 (mod a (- 4)))) ; mod
(assert (= r6 (rem a (- 4)))) ; remainder
(declare-const b Real)
(declare-const c Real)
(assert (>= b (/ c 3.0)))
(assert (>= c 20.0))
(check-sat)
(get-model)

```

To illustrate the combination of semantic logic and model theory for number systems, we adopt the Job Shop Problem example from <https://www.cs.toronto.edu/~victorn/tutorials/z3/index.html>

Example 3.133 (Job Shop Problem). To schedule tasks for a problem (such as building a bike) on different work stations, with some constraints:

- Each task in a job must start only after the previous task has been completed.
- A task cannot be paused — the time it takes to complete cannot be divided.
- The work stations can only work on one task.

Suppose We have three jobs job0, job1, job2 containing task. We define the problem parameters in the following lists, where each element of the list is a pair (m, d) where m represents the machine where the task has to be executed and d is the duration of the task:

- job0 = [(0,3), (1,2), (2,2)];
- job1 = [(0,2), (2,1), (1,4)];
- job2 = [(1,4),(2,3)];

The jobs are a list of tasks, and each task is a pair where the first element represents the machine number that can execute the task and the second is the duration of the task.

```

;;; https://www.cs.toronto.edu/~victorn/tutorials/z3/SMT.html
;;; Variables representing the starting times of the task:
(declare-const t00 Int)
(declare-const t01 Int)
(declare-const t02 Int)
(declare-const t10 Int)
(declare-const t11 Int)
(declare-const t12 Int)
(declare-const t20 Int)
(declare-const t21 Int)
;;; A first constraint we need to add is that all starting times
;;; for the tasks are positive:
(assert (and (>= t00 0) (>= t01 0) (>= t02 0)
             (>= t10 0) (>= t11 0) (>= t12 0)
             (>= t20 0) (>= t21 0)))
;;; "All tasks in a job should be executed sequentially" constraints:
(assert (and (<= (+ t00 3) t01) (<= (+ t01 2) t02)))
(assert (and (<= (+ t10 2) t11) (<= (+ t11 1) t12)))
(assert (and (<= (+ t20 4) t21)))
;;; "A machine can execute only one task at a time" constraints:
;;; Machine 0- only 2 tasks, it's easy
(assert (or (and (<= (+ t00 3) t10))
            (and (<= (+ t10 2) t00))))
;;; Machine 1: three tasks, so 6 possible sequences!
(assert (or (and (<= (+ t01 2) t12) (<= (+ t12 4) t20))
            (and (<= (+ t01 2) t20) (<= (+ t20 4) t12))
            (and (<= (+ t12 4) t01) (<= (+ t01 2) t20))
            (and (<= (+ t12 4) t20) (<= (+ t20 4) t01))
            (and (<= (+ t20 4) t01) (<= (+ t01 2) t12))
            (and (<= (+ t20 4) t12) (<= (+ t12 4) t01))))
;;; Machine 2 : three tasks, so 6 possible sequences!
(assert (or (and (<= (+ t02 2) t11) (<= (+ t11 1) t21))
            (and (<= (+ t02 2) t21) (<= (+ t21 3) t11))
            (and (<= (+ t11 1) t02) (<= (+ t02 2) t21))
            (and (<= (+ t11 1) t21) (<= (+ t21 3) t02))
            (and (<= (+ t21 3) t02) (<= (+ t02 2) t11))
            (and (<= (+ t21 3) t11) (<= (+ t11 1) t02))))
(check-sat)
(get-model)    ;;; Not an optimal model
;;; define the max function (to have less idle time)
(define-fun max ((x Int) (y Int)) Int (ite (< x y) y x))
;;; optimisation problem
(minimize (max (+ t02 2) (max (+ t12 4) (+ t21 3))))
(check-sat)
(get-model)
(exit)

```

3.8 Application: Cryptography

sec:crypto

Cryptography is the study of methods for sending secret messages. It involves *encryption*, in which a message, called *plaintext*, is converted into a form, called *ciphertext*, that may be sent over channels possibly open to view by outside parties. The receiver of the ciphertext uses *decryption* to convert the ciphertext back into plaintext (Figure 3.1). ^{Fig:crypto}

Cryptography is important because it is one of the main pillar of computer security and it is crucial to business activities and computer communication. Modular arithmetic is frequently used to append an extra check digit to identification numbers, in order to recognise transmission errors or forgeries. Personal identification numbers of some kind appear on passports, credit cards, bank accounts, and a variety of other settings.

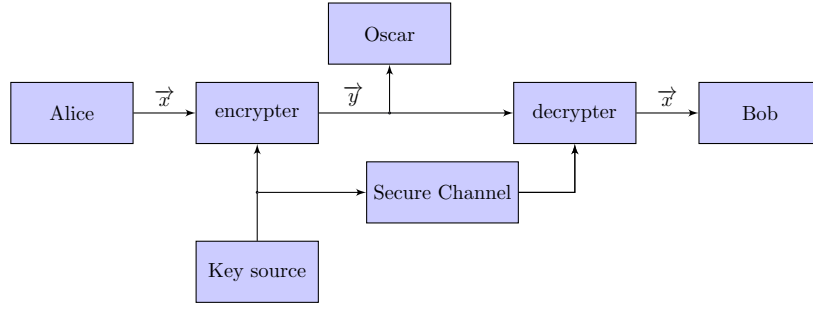


Figure 3.1: Cryptography process

fig:crypto

Definition 3.22. A *cryptosystem* is a tuple $(\mathcal{P}, \mathcal{C}, \mathcal{K}, \mathcal{E}, \mathcal{D})$ where

1. $\mathcal{P}$ is a finite set of possible plaintexts;
2. $\mathcal{C}$ is a finite set of possible ciphertexts;
3. $\mathcal{K}$ is a set of possible keys called the *keyspace*;
4. For each $k \in \mathcal{K}$, there is an encryption rule and a decryption rule respectively as follows:

$$e_K : \mathcal{P} \rightarrow \mathcal{C}, \quad d_K : \mathcal{C} \rightarrow \mathcal{P}$$

such that $d_K(e_K(x)) = x$ for every $x \in \mathcal{P}$. The set of e_K is denoted $\mathcal{E}$ and the set of d_K is denoted $\mathcal{D}$.

Remark 3.134. When $\mathcal{P} = \mathcal{C}$, then each encryption function is in fact a permutation.

In the following subsections, we will investigate three *classical* (or *private-key*, *symmetric-key*) *cryptosystem* and one public-key cryptosystem, the RSA cryptosystem. For simplicity, we limit ourselves to Latin characters A to Z ignoring the difference between capital and small letters. We assume each letter of the alphabet is coded by its position relative to the others as follows:

$$A=0, B=1, C=2, D=3, \dots, X=23, Y=24, Z=25.$$

3.8.1 Shift Cipher

Definition 3.23. Let $\mathcal{P} = \mathcal{C} = \mathcal{K} = \mathbb{Z}_{26}$. For $0 \leq k \leq 25$, define

$$e_k(x) = (x + k) \mod 26, \quad d_k(y) = (y - k) \mod 26$$

for every x, y in $\mathbb{Z}_{26}$. Then $(\mathcal{P}, \mathcal{C}, \mathcal{K}, \mathcal{E}, \mathcal{D})$ is called a *shift cipher*.

eg:Caesar

Example 3.135. An encryption system once used by Julius Caesar, and now called the Caesar cipher, encrypts messages by changing each letter of the alphabet to the one three places farther along, with X wrapping around to A, Y to B, and Z to C.

If the numerical version of the plaintext for a letter is denoted M and the numeric version of the ciphertext is denoted C , then

$$C = (M + 3) \mod 26$$

The receiver of such a message can easily decrypt it by using the formula

$$M = (C - 3) \mod 26$$

When a private key cryptosystem is used, a pair of people who wish to communicate in secret must have a separate key. Since anyone knowing this key can both encrypt and decrypt messages easily, these two people need to securely exchange the key.

eg:4.43b

Example 3.136. Use the Caesar cipher to encrypt the message HOW ARE YOU.

Solution: First translate the letters of HOW ARE YOU into their numeric equivalents:

7 14 22 0 17 4 24 14 20

Next encrypt the message by adding 3 to each number.

10 17 25 3 20 7 1 17 23

Finally, substitute the letters that correspond to these numbers. The encrypted messages becomes

KRZ DUH BRX

eg:4.43c

Example 3.137. Use the shift cipher with key 3 to decrypt the message L DP ILQH.

Solution. First translate the letters of L DP ILQH into their numeric equivalents:

11 3 15 8 11 16 7

Next decrypt the message by subtracting 3 from each number.

8 0 12 5 8 13 4

Then translate back into letters to obtain the original message: I AM FINE

□

Shift ciphers can be broken by what we call a *brute force attack*. It only takes 25 trials to guess the private key, hence, this is a useless cryptosystem in an information society.

Example 3.138. Consider Bob received a cipher text “haahjr ha khdu” encrypted using shift cipher. What is the original message?

Solution: Apply every possible decryption key from 1 through 25. Look at the results and see which one makes sense:

```
Shift: 0:  haahjr ha khdu
Shift: 1:  ibbiks ib liev
Shift: 2:  jccjlt jc mjfw
Shift: 3:  kddkmu kd nkgx
Shift: 4:  leelnv le olhy
Shift: 5:  mffmow mf pmiz
Shift: 6:  nggnpx ng qnja
Shift: 7:  ohhoqy oh rokb
Shift: 8:  piiprz pi splc
Shift: 9:  qjjqsa qj tqmd
Shift: 10: rkkrtb rk urne
Shift: 11: sllsuc sl vsof
Shift: 12: tmmtvd tm wtpg
Shift: 13: unnuwe un xugh
Shift: 14: voovxf vo yvri
Shift: 15: wppwyg wp zwsj
Shift: 16: xqqxzh xq atk
Shift: 17: yrryai yr byul
Shift: 18: zsszbj zs czvm
Shift: 19: attack at dawn
Shift: 20: buubdl bu ebxo
Shift: 21: cvvcem cv fcyp
Shift: 22: dwdfn dw gdzq
Shift: 23: exxego ex hear
```

Shift: 24: fyyfhp fy ifbs
Shift: 25: gzzgiq gz jgct

3.8.2 Affine Cipher

Definition 3.24. Let $\mathcal{P} = \mathcal{C} = \mathbb{Z}_{26}$ and $\mathcal{K} = \{(a, b) \in \mathbb{Z}_{26} \times \mathbb{Z}_{26} : \gcd(a, 26) = 1\}$. For each permutation $(a, b) \in \mathcal{K}$, define

$$e_{a,b}(x) = (ax + b) \mod 26, \quad d_{a,b}(y) = a^{-1}(y - b) \mod 26.$$

Since $\gcd(a, 26) = 1$, a can only take values from 1,3,5,7,9,11,15, 17,19,21,23,25.

Example 3.139. Encipher “ITS COOL” using an affine cipher with $a = 5$ and $b = 8$.

Solution: Using $e_{5,8}(x) = (5x + 8) \mod 26$, we fill in the following table

plaintext	I	T	S	C	O	O	L
x	8	19	18	2	14	14	11
$5x + 8$	48	103	98	18	78	78	63
$(5x + 8) \mod 26$	22	25	20	18	0	0	11
ciphertext	W	Z	U	S	A	A	L

Example 3.140. Decipher “HPCCXAQ” using an affine cipher with $a = 5$ and $b = 8$.

Solution. We begin by finding the decipherment function. Since $5x \equiv 1 \pmod{26}$ is solved with $x \equiv 21 \pmod{26}$, hence $5^{-1} \equiv 21 \pmod{26}$. Therefore,

$$d_{5,8}(y) = 21(y - 8) \mod 26$$

and so filling in our table gives

ciphertext	H	P	C	C	X	A	Q
y	7	15	2	2	23	0	16
$y - 8$	-1	7	-6	-6	15	-8	8
$21(y - 8)$	-21	147	-126	-126	315	-168	168
$21(y - 8) \mod 26$	5	17	4	4	3	14	12
plaintext	F	R	E	E	D	O	M

□

3.8.3 Substitution Cipher

A substitution cipher is one in which letters are represented by other letters; it can be deciphered by someone knowing the order of the cipher alphabet used. It is defined formally as follows.

Definition 3.25. Let $\mathcal{P} = \mathcal{C} = \mathbb{Z}_{26}$ and $\mathcal{K}$ be the set of all possible permutations of the 26 symbols in $\mathcal{P}$. For each substitution $\sigma \in \mathcal{K}$, define

$$e_{\sigma}(x) = \sigma(x), \quad d_{\sigma}(y) = \sigma^{-1}(y).$$

Remark 3.141. There are $26!$ permutations. Hence, finding the right private key may be difficult.

Example 3.142. Consider the following permutation for substitution cipher:

$$\begin{pmatrix} A & B & C & D & E & F & G & H & I & J & K & L & M & N & O & P & Q & R & S & T & U & V & W & X & Y & Z \\ R & Z & B & U & Q & K & F & C & P & Y & E & V & L & S & N & G & W & O & X & D & J & I & A & H & T & M \end{pmatrix}$$

Encode the word “HARDWORKING”.

Solution. The ciphertext is “CROUANOEPSF”.

□

Substitution ciphers are fairly easy to “crack” — the problem is that in English (or any language), certain letters are far more likely to appear. In English, for example, the letter “E” is far more likely to appear than the letter “Z”. In fact, we have the following English letter frequency table

A	8.2%	F	2.2%	K	0.8%	P	1.9%	U	2.8%	Z	0.1%
B	1.5%	G	2.0%	L	4.0%	Q	0.1%	V	1.0%		
C	2.8%	H	6.1%	M	2.4%	R	6.0%	W	2.3%		
D	4.3%	I	7.0%	N	6.7%	S	6.3%	X	0.1%		
E	12.7%	J	2.2%	O	7.5%	T	9.1%	Y	2.0%		

The approximate percentages for the first few letters in the list below are:

E: 12.7%, T: 9.1%, A: 8.2%, O: 7.5%

and the percentages for the last few are:

J: 0.2%, Q: 0.1%, Z: 0.1%.

3.8.4 RSA Cryptosystem

sec:RSA

RSA is a popular public-key encryption method used in electronic commerce. In what follows, we will learn how to encrypt and decrypt a message using RSA cryptography. First, we define RSA formally.

Definition 3.26. Let $n = pq$, where p and q different prime numbers. Let $\mathcal{P} = \mathcal{C} = \mathbb{Z}_n$, define

$$\mathcal{K} = \{(n, p, q, a, b) : ab \equiv 1 \pmod{\phi(n)}\}.$$

For every $k \in \mathcal{K}$, we define

$$e_k(x) = x^b \pmod{n}, \quad d_k(y) = y^a \pmod{n}$$

where $x, y \in \mathbb{Z}_n$ and ϕ is the **Euler phi function**, which is an arithmetic function that counts the number of positive integers less than or equal to n that are relatively prime to n . It is found mathematically to be

$$\phi(n) = n \prod_{\substack{p|n \\ p \text{ is prime}}} \left(1 - \frac{1}{p}\right).$$

The values n and b comprise the *public key* and the values p, q and a form the *private key*.

Example 3.143 (Getting familiar with Euler ϕ function). Find the number of integers relatively prime to 36.

Solution:

$$\phi(36) = \phi(2^2 3^2) = 36 \left(1 - \frac{1}{2}\right) \left(1 - \frac{1}{3}\right) = 36 \cdot \frac{1}{2} \cdot \frac{2}{3} = 12.$$

In words, this says that the distinct prime factors of 36 are 2 and 3; half of the thirty-six integers from 1 to 36 are divisible by 2, leaving eighteen; a third of those are divisible by 3, leaving twelve coprime to 36. And indeed there are twelve: 1, 5, 7, 11, 13, 17, 19, 23, 25, 29, 31, and 35.

Theorem 3.144. When p and q are prime numbers, $\phi(pq) = (p-1)(q-1)$.

To encrypt a message using the RSA cipher, a person needs to know the value of pq and of another number b , both of which are made publicly available. But only a person who knows the individual values of p , q and a can decrypt an encrypted message.

Example 3.145. Suppose Alice decides to set up an RSA cipher. She chooses two prime numbers, say $p = 5$ and $q = 11$, and computes $n = pq = 55$. She then chooses a positive integer b that is relatively prime to $(p-1)(q-1)$. In this case, $(p-1)(q-1) = 4(10) = 40$, so she may take $b = 3$ is relatively prime to 40. The two numbers $n = 55$ and $b = 3$ are the public key, which she may distribute widely. To decrypt the message, Alice needs to find the decryption key, a number a that is a positive inverse to b modulo $(p-1)(q-1)$. In this case, the key is

$$k = (55, 5, 11, a, 3).$$

eg: 4.44

1. Bobs wants to send Alice the message HA. Find the ciphertext for his message.

eg: 4.45

2. Find the value of a and decrypt the ciphertext 17.

Solution: Given $p = 5$, $q = 11$, $n = pq = 55$, $b = 3$, $e_k(x) = x^3 \pmod{n}$, $d_k(y) = y^a \pmod{n}$.

1. Bob will send his message in two blocks, one for the H and another for the A. The letters H and A are encoded as 7 and 0 respectively. The corresponding ciphertext is computed as follows:

$$\begin{aligned} e(7) &= 7^3 \pmod{55} = 343 \pmod{55} = 13, \\ e(0) &= 0^3 \pmod{55} = 0. \end{aligned}$$

Accordingly, Bob sends Alice the message: 13 0.

2. The integer a needs to satisfy

$$ab = 3a \equiv 1 \pmod{\phi(55)}$$

Here, $\phi(55) = (p-1)(q-1) = 40$. This problem is similar to Example 3.128(2), in which we found that

$$a \equiv 3^{-1} \equiv 27 \pmod{40}.$$

Then we compute

$$\begin{aligned} d(17) &\equiv 17^{27} \\ &\equiv 17^{16+8+2+1} \\ &\equiv [(17^{16} \pmod{55})(17^8 \pmod{55})(17^2 \pmod{55})(17 \pmod{55})] \pmod{55} \\ &\equiv (16 \cdot 26 \cdot 14 \cdot 17) \pmod{55} \\ &\equiv 99008 \equiv 8 \pmod{55} \end{aligned}$$

$$\text{where } \begin{cases} 17^2 \pmod{55} = 17^2 \pmod{55} = 14 \\ 17^4 \pmod{55} = (14)^2 \pmod{55} = 31 \\ 17^8 \pmod{55} = (31)^2 \pmod{55} = 26 \\ 17^{16} \pmod{55} = (26)^2 \pmod{55} = 16. \end{cases}$$

Thus the plaintext of Bob's message is 8. The letter corresponding to 8 is I.

In reality, RSA is used in setting up a secure communication channel as described in the following example. These days, a key length of at least 2048 bits is required (<https://stackoverflow.com/questions/589834/what-rsa-key-length-should-i-use-for-my-ssl-certificates>).

The Transport Layer Security, TLS for short (replaces the older Secure Socket Layer (SSL)), is a protocol by which many services that communicate over the Internet can do so in a secure fashion. Before we discuss how TLS works and what kinds of security it provides, let us first see what happens without TLS.

Life on the Internet without TLS

Let us compare communications between computers on the Internet and communications between people over the telephone. Without TLS, your computer-to-computer communications suffer from the same security problems from which your telephone communications suffer:

- **Who are you talking to?** In a phone conversation, how can you be sure that the person who picks up the phone at the other end is really the person you are trying to call (especially if you have never spoken to them before)? What if your phone call was intercepted or re-routed, or what if someone else is answering your call recipient's phone? There really is no way to be sure you have reached the right person, especially if they are actively trying to fool you (as what we seen in movies).
- **Eavesdropping?** As you are aware of from watching TV or reading, **it is very easy to tap phone lines**: the police and spies do this all the time to covertly gather information. It is not easy to detect if your lines are tapped. The same applies with communications over the Internet — how can you be sure that your communications are not being “tapped” and recorded? This is especially problematic in public wifi hotspots.

This results in two very real security issues for communications over the Internet: 1. knowing for sure that you are connecting to the right servers (i.e. those at your bank and not those at a hacker's or phisher's web site), and 2. knowing that your data is safe from prying eyes during transit to those computers. This is where TLS comes in.

Enter the TLS (https://en.wikipedia.org/wiki/Transport_Layer_Security)

To solve these problems to a large degree, most Internet services support use of TLS as a mechanism for securing communications. To illustrate how TLS works, let us use another analogy.

Client wants to communicate with a company to send important information back and forth. Client wants to be 100% sure that s/he is communicating with this particular company and that no one can eavesdrop on or intercept the communications. How can s/he do this?

- Client sends a courier to the company's address.
- The company has envelopes that, when closed, can only be opened by the company. The company and the courier go together to a trusted third party — a notary — which makes the company provide documentation to prove its identity. The notary certifies the company's secure envelopes and the courier takes these back to the client.
- The client gets the envelopes and, if it trusts the notary's reputation, can be sure that they are actually from the company indicated.
- The client also has secure envelopes that, once sealed, only the client can open. It puts some of these in one of the company's secure envelopes and sends them back to the company.
- The company gets the sealed secure envelope. It opens the envelope (as only it can). It now has the client's secure envelopes.
- The company has another kind of envelope that can be opened and sealed only by using a special combination. The company puts this special envelope with the combination lock, together with the combination, into one of the client's secure envelopes. The company seals the envelope.
- The company has another type of secure envelope that anyone can open, but which only the company can seal. If you open one of these sealed envelopes, you know for sure that it was sent by the company. The company puts the whole package inside this and sends it to the client.

- When the client gets the secure envelope, it opens it and thus knows that it came from the company. It then opens the next secure envelope inside that can only be opened by the client. Inside it gets out the combination-envelope and the combination itself.
- The client the puts his data in the combination envelope, seals it and sends it to the company.
- The company receives it, opens it, and puts the response in the same secure envelope and sends it back.
- The procedure is repeated as often as necessary for required communications.

TLS relies on **public key cryptography** (e.g. RSA) to accomplish these tasks. In normal encryption, the two parties communicating share a “password” and that password is used to both encrypt and decrypt messages. While this is fast and efficient, how do you communicate these passwords to people you have not yet met in a way that is itself secure?

In “public key cryptography”, each person has two keys — a **public** key and a **private** key. Anything encrypted with the user’s public key can only be decrypted with the private key and vice versa. Each person then tells the world what his public key is and keeps his private key safe and secure, and private.

If John sends Mary a message encrypted with Mary’s public key, then only Mary can open it, as only she has her private key. This is like an envelope that anyone can seal but which only Mary can open.

If John sends Mary a message encrypted with John’s private key, then anyone can open it, as everyone has access to John’s public key. However, successfully opening the message proves that it was sent by John and no one else, as only John has access to his private key. This is like an envelope that only John can seal, but which anyone can open and thus prove that John sealed it.

TLS in Action

So, let’s see how TLS actually works for securing your communications over the Internet. Before the communications occur, the following takes place:

- A company wishes to secure communications to their server `company.com`.
- They create a public and private key for `company.com` (this is also known as as **SSL Certificate**):

```
openssl req -out CSR.csr -new -newkey rsa:2048 -nodes -keyout privateKey.key
```

- They go to a trusted third party company such as Thawte or Verisign: Thawte makes the company prove its identity and right to use the `company.com` domain. This usually involves a lot of paperwork and paying a hefty fee.
- Once the verification is complete, Thawte gives the company a new public key that has some additional information in it. This information is the certification from Thawte that this public key is for the company and `company.com` and that this is verified by Thawte. This certification information is encrypted using Thawte’s private key.

Then, when Client wishes to communicate with the company at `company.com`,

- Client makes a connection to `company.com` with its computer. This connection is made to a special “port” (address) on `company.com` that is set up for TLS communications only.
- When Client connects to `company.com` on its TLS-secured port, the company sends back its public key (and some other information, like what Ciphers it supports).
- Client gets the public key and decides if it is OK
 - If the public key has expired, this could be a problem

- If the public key claims to be for some domain that is not `company.com` that could be a problem.
 - Client has the public key for Thawte (and many other third party companies) stored in its computer — because these come with the computer. Thus, client can decrypt the validation information, prove the validation is from Thawte and verify that the public key is certified by Thawte. If Client trusts Thawte, then Client can trust that he/she is really communicating with Company. If Client doesn't trust Thawte, or whatever Third Party company is actually being used, then the identity of who is running the computers to which Client is connecting is suspect.
- If the client doesn't trust the server, then the communication is terminated.
 - If the client has its own SSL certificate installed, it may send that to the server at this point to see if the server trusts the client. Client-side SSL certificates are not commonly used, but provide a good way for the client to authenticate itself with the server without using a username or password. In the case where this is used, the server would have to know about the client's certificate and verify it in a similar way to how the client verified the server. If this fails, the connection is terminated. If a client-side certificate is not needed, this step is skipped.
 - Once the client is happy with the server (and the server with the client, if needed), then the client choose an TLS Cipher to use from the list of encryption methods provided by the server, and generates a “symmetric key” (password) for use with that Cipher. The client encrypts this password using the server's public key and sends it back to the server. The server (and only the server) can decrypt this message and get this password, which is now shared by both the client and server.
 - The client will then start communicating with the company by encrypting all data using this password and the chosen Cipher. Normal “symmetric” (password-based) encryption takes place from this point forward because it is much faster than using the public and private keys for everything. These keys were needed to enable the company (and possibly the client) to prove its identity and right to `domain.com` and to enable the client and server to generate and securely communicate a common password.

3.9 Application: Random Number Generator and Simulation

sec:rand-sim

https://en.wikipedia.org/wiki/Random_number_generation is a process by which, often by means of a random number generator (RNG), a sequence of numbers is generated that cannot be reasonably predicted better than by random chance.

True random number generators can be **hardware random-number generators (HRNGs)**, wherein each generation is a function of the current value of a physical environment's attribute that is constantly changing in a manner that is practically impossible to model.

https://en.wikipedia.org/wiki/Pseudorandom_number_generator is a **computer algorithm for generating a sequence of numbers** whose properties **approximate** the properties of sequences of **random numbers**. It is important in practice for their speed in number generation and their reproducibility.

Random number generators have applications in statistical sampling, computer simulation, cryptography, completely randomised design (for agriculture/scientific experiments), and other areas where producing an unpredictable result is desirable. Generally, in applications having unpredictability as the paramount feature, such as in security applications, hardware generators are generally preferred over pseudorandom algorithms, where feasible.

http://en.wikipedia.org/wiki/Linear_congruential_generator (LCG) is the most popular random number generator, which is one of the oldest algorithm based on modular arithmetic

- `mersenne_twister_engine`: implements Mersenne twister algorithm (used in Python)
- `subtract_with_carry_engine`: implements a subtract-with-carry (lagged Fibonacci) algorithm
- `philox_engine` (C++26): a counter-based parallelizable generator

The following are predefined random number generators by C++:

- `default_random_engine` (C++11): It defaults to using `minstd_rand0` or CPU random generator (e.g. Intel <https://en.wikipedia.org/wiki/RDRAND> on GNU/Linux is shown as `/dev/random`).
- `minstd_rand0` (C++11): discovered in 1969 by Lewis, Goodman and Miller & adopted as “Minimal standard” in 1988 by Park and Miller `std::linear_congruential_engine<std::uint_fast32_t, 16807, 0, 2147483647>`
- `minstd_rand` (C++11): a newer “Minimum standard”, recommended by Park, Miller, and Stockmeyer in 1993 `std::linear_congruential_engine<std::uint_fast32_t, 48271, 0, 2147483647>`
- `mt19937` (C++11): a 32-bit Mersenne Twister by Matsumoto and Nishimura in 1998 `std::mersenne_twister_engine<std::uint_fast32_t, 32, 624, 397, 31, 0x9908b0df, 11, 0xffffffff, 7, 0x9d2c5680, 15, 0xefc60000, 18, 1812433253>`
- `mt19937_64` (C++11): a 64-bit Mersenne Twister by Matsumoto and Nishimura, 2000 `std::mersenne_twister_engine<std::uint_fast64_t, 64, 312, 156, 31, 0xb5026f5aa96619e9, 29, 0x5555555555555555, 17, 0x71d67fffed60000, 37, 0xffff7eee000000000, 43, 6364136223846793005>`
- `ranlux24_base` (C++11): `std::subtract_with_carry_engine<std::uint_fast32_t, 24, 10, 24>`
- `ranlux48_base` (C++11): `std::subtract_with_carry_engine<std::uint_fast64_t, 48, 5, 12>`
- `ranlux24` (C++11): a 24-bit RANLUX generator by Martin Lüscher and Fred James, 1994 `std::discard_block_engine<std::ranlux24_base, 223, 23>`
- `ranlux48` (C++11): a 48-bit RANLUX generator by Martin Lüscher and Fred James, 1994 `std::discard_block_engine<std::ranlux48_base, 389, 11>`
- `knuth_b` (C++11): `std::shuffle_order_engine<std::minstd_rand0, 256>`
- `philox4x32` (C++26): `std::philox_engine<std::uint_fast32_t, 32, 4, 10, 0xCD9E8D57, 0x9E3779B9, 0xD2511F53, 0xBB67AE85>`
- `philox4x64` (C++26): `std::philox_engine<std::uint_fast64_t, 64, 4, 10, 0xCA5A826395121157, 0x9E3779B97F4A7C15, 0xD2E7470EE14C6C93, 0xBB67AE8584CAA73B>`

In statistics, a *sample* is a subject chosen from a *population for investigation*; a *random sample* is one chosen by a method involving an unpredictable component. Random sampling can also refer to taking a number of independent observations from the same probability distribution, without involving any real population. The sample usually is not a representative of the population of people from which it was drawn — this random variation in the results is termed as *sampling error*. One method of producing random samples is by using LCG.

Example 3.148. Consider a small town of 25 people with the following salaries:

3000	2100	5100	3400	6000	2500	4800	3900	5100	3200
4900	5600	4300	5800	3400	3300	5300	5400	2100	2400
2300	4000	4100	3400	3600					

A company only has the budget to survey 5 person, use the LCG with $m = 25$, $c = 3$, $a = 6$ and the seed $X_0 = 2$ to find the five random values, the sample average and compare it to the population average. [Note: The first person is labelled with number 0, the second is labelled number 1, etc.]

Solution. Calculating the 5 random numbers:

$$X_0 = 2, X_1 = (6X_0 + 3) \bmod 25 = 15, X_2 = 18, X_3 = 11, X_4 = 19.$$

The sample average is

$$\frac{5100+3300+2100+5600+2400}{5} = 3700$$

compare to population average of

$$\frac{1}{25}(3000 + 2100 + 5100 + \cdots + 4100 + 3400 + 3600) = 3960$$

□

Example 3.149. A C++ program to generate 10000 normally distributed random numbers with a mean of 10 and a standard deviation of 3.

Solution

```
// https://en.cppreference.com/w/cpp/numeric/random.html
#include <random>
#include <map>          // for histogram
#include <iomanip>
#include <iostream>
#include <string>

int main()
{
    // std::seed_seq consumes a sequence of integer-valued data and
    // produces a requested number of 32-bit unsigned integer values,
    // based on the consumed data.
    std::seed_seq seedno{2025, 2035, 2015, 2075};
    std::mt19937 rand_engine(seedno);
    // Generate a normal distribution
    const double mean = 10, stdev = 3;
    std::normal_distribution<> rnorm(mean, stdev);

    std::map<int, int> hist;
    for (int n = 0; n < 10000; ++n)
        ++hist[std::round(rnorm(rand_engine))];

    std::cout << "Normal(" << mean << ", " << stdev << "):\n"
              << std::fixed << std::setprecision(1);
    for (auto [x, y] : hist)
        std::cout << std::setw(2) << x << ' ' << std::string(y / 100, '*') << '\n';
}
```

Output of the C++ program

Normal(10, 3):

```
-1
0
1
2
3
4 **
5 ***
6 *****
7 *****
8 *****
9 *****
10 *****
11 *****
12 *****
13 *****
14 *****
15 ***
16 *
17
18
19
20
```

Chapter 4

Sets and Binary Relations

ch:set-rel

4.1 Sets and Relations

Sets are used in mathematics to group elements (without repetition) of the “same property” (many of them can be characterised by predicates) and relations are used to relate multiple sets and one set. There are little characterisations for relations on multiple sets but equivalence relations and partial order relations are two important classes of relations on one set.

4.1.1 Definitions

sec:set-defn

Informally, a **set** is any well-defined list, collection, or class of objects. The objects comprising the set are called its **elements** or **members**. The statement “ p is an element of a set A ” is written as “ $p \in A$ ”.

The operations associated with a set A are defined below.

def:powerset

- Definition 4.1.**
1. B is a **subset** of A , denoted by $B \subset A$, if $x \in B$ then $x \in A$.
 2. The **complement (set)** of a subset B of A , denoted by B^c , is defined as $\{x \in A : x \notin B\}$.
 3. The **power set** of A , denoted by $\mathcal{P}(A)$, is the family of all the subsets of any set A .

Remark 4.1. Some books use $B \subseteq A$ to denote B is a subset of A . However, in this course, the notation $\subset$ will be used.

Example 4.2. Given $A = \{1, 2, 3\}$.

$\emptyset$ is a subset of A , i.e. $\emptyset \subset A$.

A is a subset of A , i.e. $A \subset A$.

$\{1\}$, $\{2\}$, $\{3\}$ are subsets of A .

The power set of A , $\mathcal{P}(A)$, is the set of all subsets of A , i.e.

$$\mathcal{P}(A) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

def:set-op2

Definition 4.2.

1. The **union** of A and B ,

$$A \cup B := \{x : x \in A \vee x \in B\}.$$

2. The **intersection** of A and B ,

$$A \cap B := \{x : x \in A \wedge x \in B\}$$

3. The **difference** of A and B ,

$$A \setminus B := \{x : x \in A \wedge x \notin B\}$$

4. The (*Cartesian*) **product** of A and B ,

$$A \times B := \{(a, b) : a \in A, b \in B\}$$

where (a, b) are ordered pairs of A and B .

5. The product can be generalised to the case of n sets $A_1, \dots, A_n$, i.e. the ***n*-ary Cartesian product over sets** $A_1, \dots, A_n$,

$$A_1 \times \dots \times A_n := (((A_1 \times A_2) \times A_3) \times \dots \times A_n) = \{(x_1, x_2), \dots, x_n) : x_i \in A_i\}.$$

Example 4.3. Given $A = \{1, 2, 3\}$, $B = \{1, a, b\}$.

$$A \cup B = \{1, 2, 3, a, b\}.$$

$$A \cap B = \{1\}.$$

$$A \setminus B = \{2, 3\}.$$

$$B \setminus A = \{a, b\}.$$

$$A \times B = \{(1, 1), (1, a), (1, b), (2, 1), (2, a), (2, b), (3, 1), (3, a), (3, b)\}.$$

Example 4.4. Let A and B be two sets. Show that $A \times B = \emptyset$ iff $A = \emptyset$ or $B = \emptyset$.

We have not shown the proof for “iff” in Topic 3, this is the first “iff” example.

Proof:

$\Rightarrow$: (Contrapositive Proof) If $A \neq \emptyset$ and $B \neq \emptyset$, $\exists x \in A$ and $\exists y \in B$ such that $(x, y) \in A \times B$. This implies $A \times B \neq \emptyset$.

$\Leftarrow$: (Contrapositive Proof) If $A \times B \neq \emptyset$, then $\exists z \in A \times B$, $z = (z_1, z_2)$, $z_1 \in A$ and $z_2 \in B$, hence $A \neq \emptyset$ and $B \neq \emptyset$. $\square$

Definition 4.3. A **bounded lattice** $((L, \vee, \wedge, 0, 1))$ is a lattice that has a greatest element (also called **maximum**, or **top element**, and denoted by 1 or by $\top$) and a least element (also called **minimum**, or **bottom**, denoted by 0 or by $\perp$), which satisfy

$$0 \preceq x \preceq 1, \quad x \vee 0 = x, \quad x \wedge 1 = x.$$

for every $x \in L$.

Two elements x and y of L are **complements** of each other if

$$x \vee y = 1 \quad \text{and} \quad x \wedge y = 0.$$

A bounded lattice for which every element has a complement is called a **complemented lattice** (https://en.wikipedia.org/wiki/Complemented_lattice).

A complemented lattice that is also distributive, i.e. for any $a, b, c \in L$

$$a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c); \quad a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c),$$

is a **Boolean algebra** $(B, \wedge, \vee, ', 0, 1)$ (https://en.wikipedia.org/wiki/Boolean_algebra_%28structure%29).

Theorem 4.5. For any set A , $(\mathcal{P}(A), \cap, \cup, ^c, \emptyset, A)$ is a Boolean algebra, i.e. for any $B, C, D \subset A$, we have

- double negative: $(B^c)^c = B$
- idempotent: $B \cap B = B$; $B \cup B = B$;
- universal bound: $B \cap \emptyset = \emptyset$; $B \cup A = A$;

- *identity*: $B \cap A = B$; $B \cup \emptyset = B$;
- *negation*: $B \cap B^c = \emptyset$; $B \cup B^c = A$;
- *commutative*: $B \cap C = C \cap B$; $B \cup C = C \cup B$;
- *absorption*: $B \cap (B \cup C) = B$; $B \cup (B \cap C) = B$;
- *associative*: $(B \cap C) \cap D = B \cap (C \cap D)$; $(B \cup C) \cup D = B \cup (C \cup D)$;
- *distributive*: $B \cap (C \cup D) = (B \cap C) \cup (B \cap D)$; $B \cup (C \cap D) = (B \cup C) \cap (B \cup D)$;
- *De Morgan*: $(B \cap C)^c = B^c \cup C^c$; $(B \cup C)^c = B^c \cap C^c$.

def:5.12

Definition 4.4. Let A and B be two nonempty sets. A **(binary) relation from A to B** , R , is a subset of $A \times B$, i.e. $R \subset A \times B$. In particular, when $B = A$, we have a **(binary) relation on A** , R , i.e. $R \subset A \times A$.

When $(a, b) \in R$, a **is related to b by R** , and can be written as aRb .

If a **is not related to b by R** , we denote it as $(a, b) \notin R$ or $a \not R b$.

The **domain of R** and **range of R** are defined respectively as

$$\begin{aligned}\text{Dom}(R) &= \{a \in A : \exists b \in B, aRb\}, \\ \text{Range}(R) &= \{b \in B : \exists a \in A, aRb\}.\end{aligned}$$

eg:5.12a

Example 4.6. The domain and range of the relation

$$R = \{(a, a), (a, b), (b, b), (b, c), (d, c), (d, e)\}$$

is

$$\text{Dom}(R) = \{a, b, d\}, \quad \text{Range}(R) = \{a, b, c, e\}.$$

eg:5.12b

Example 4.7. Given $A = \{-3, -2, -1, 0, 1, 2, 3, 4\}$, $B = \{0, 1, 4\}$ and define the relation R from A to B as $R = \{(x, y) : y = x^2\}$. Determine the domain and range of R .

Solution. By definition, the relation is

$$R = \{(-2, 4), (-1, 1), (0, 0), (1, 1), (2, 4)\}$$

So $\text{Dom}(R) = \{-2, -1, 0, 1, 2\}$, $\text{Range}(R) = \{0, 1, 4\}$. □

eg:5.9

Example 4.8. Let $A = \{3, 4, 5, 6, 7\}$ and let R be the relation $<$ (less than) on A . List the elements of R .

Solution. Since $aRb = a < b$,

$$R = \{(3, 4), (3, 5), (3, 6), (3, 7), (4, 5), (4, 6), (4, 7), (5, 6), (5, 7), (6, 7)\}.$$

□

eg:5.10

Example 4.9. Let R be a relation on $A = \{1, 2, 3, 4\}$ defined by: $xRy = x$ divides y and $x < y$. List all elements in R .

Solution. $R = \{(1, 2), (1, 3), (1, 4), (2, 4)\}$ □

Definition 4.5 (https://en.wikipedia.org/wiki/Finitary_relation). R is an **n -ary relation** or a **finitary relation** over a sequence of sets $A_1, \dots, A_n$ if it is a subset of the Cartesian product $A_1 \times \dots \times A_n$, i.e. $R \subset A_1 \times \dots \times A_n$.

def:function

Definition 4.6. A **function from A to B** , f , is a relation from A to B such that for every $x \in \text{Dom}(f)$, there is a **unique** value y such that xfy or what we usually write $y = f(x)$. It is normally written as

$$f : A \rightarrow B, \quad x \mapsto y.$$

A function f is called a **bijection** if $\text{Dom}(f) = A$ and for every $y \in B$, there is a unique $x \in A$ such that $f(x) = y$.

A function f is called an **injection** if for every x_1 and x_2 in A , if $f(x_1) = f(x_2)$ then $x_1 = x_2$.

A function f is called an **surjection** if for every y in B , there is an x in A such that $y = f(x)$. In other words, $\text{Range}(f) = B$.

:bourbaki-set-p81

Definition 4.7. Let $X : I \rightarrow S$ be a function from the **index set** I to the set of sets S . The function X is called a **family of sets**, which is normally denoted as $\{X_i := X(i) | i \in I\}$ or $(X_i)_{i \in I}$.

Using logic and the notion of function/family defined in Definition 4.7, one can generalise the set operations in Definition 4.2 as follows.

def:bourbaki-set-p81

def:set-op2

:bourbaki-set-p90

Definition 4.8. Let $(X_i)_{i \in I}$ be a family of subsets of a set A . Then its

1. **union**, $\bigcup_{i \in I} X_i := \{x \in A : \exists i(i \in I \wedge x \in X_i)\}$.
2. **intersection**, $\bigcap_{i \in I} X_i := \{x \in A : \forall i(i \in I \rightarrow x \in X_i)\}$.
3. **product**, $\prod_{i \in I} X_i := \{f : I \rightarrow \bigcup_{i \in I} X_i : f(i) \in X_i\}$.

Remark 4.10. Note that when the index set I is empty, we have the following:

$$\bigcup_{i \in I} X_i = \emptyset, \quad \bigcap_{i \in I} X_i = A, \quad \prod_{i \in I} X_i = \{\emptyset\}.$$

Theorem 4.11 (De Morgan). Let $(X_i)_{i \in I}$ be a family of subsets of a set A . Then

$$\left(\bigcap_{i \in I} X_i\right)^c = \bigcup_{i \in I} X_i^c, \quad \left(\bigcup_{i \in I} X_i\right)^c = \bigcap_{i \in I} X_i^c.$$

4.1.2 Properties of Relations

:relation:prop

There is very little mathematical characterisation about n -ary relation. In computer science, it is common to refer to a Boolean valued function as an **n -ary predicate**. In formal logic, the relation R constitutes a “model” or a “relational structure” or an “interpretation” for some n -ary predicate symbol.

The only relation that have good mathematical characterisation is the **binary relation on a set**. This is because we can generalise the equality relation and the ordering relation.

def:5.21

Definition 4.9. A relation R on a set A is called **reflexive** if $(a, a) \in R$ for every $a \in A$. A relation R on a set A is called **irreflexive** if $(a, a) \notin R$ for every $a \in A$.

eg:5.25

Example 4.12. Let $A = \{1, 2, 3, 4, 5, 6\}$. Determine whether the following relations are reflexive, irreflexive or neither.

1. $R_1 = \{(1, 1), (1, 2), (2, 2), (3, 3), (4, 4), (5, 5), (5, 6), (6, 6)\}$

Solution: R_1 is reflexive since $(a, a) \in R_1$ for all $a \in A$. But it is not irreflexive.

2. $R_2 = \{(x, y) \in A \times A | x > y\}$

Solution. R_2 is irreflexive since $(a, a) \notin R_2$ for every $a \in A$. $\square$

eg:5.24

3. $R_3 = \{(1, 1), (1, 3), (2, 2)\}$

Solution. R_3 is not reflexive since $(3, 3) \notin R_3$. R_3 is not irreflexive since $(1, 1) \in R_3$. $\square$

def:5.22

Definition 4.10. Given a relation R on a set A .

1. R is called **symmetric** if when $(a, b) \in R$, $(b, a) \in R$.
2. R is called **asymmetric** if when $(a, b) \in R$, $(b, a) \notin R$.
3. R is called **antisymmetric** if when $(a, b) \in R$ and $(b, a) \in R$, $a = b$.

Remark 4.13. Beware that not symmetric does not mean asymmetric or antisymmetric. However, asymmetric implies antisymmetric but not vice versa.

eg:5.27

Example 4.14. Let $A = \{1, 2, 3, 4, 5, 6\}$. Determine whether the following relations are symmetric, asymmetric and anti-symmetric.

• $R_0 = \emptyset$

Solution: By definition, since R_0 does not have any element, $(a, b) \in R_0$ is always false, it satisfies all rules. It is symmetric, asymmetric and anti-symmetric.

• $R_1 = \{(1, 2), (2, 1), (2, 3), (3, 2), (3, 4), (5, 5)\}$.

Solution: R_1 is *not symmetric* since $(3, 4) \in R_1$ but $(4, 3) \notin R_1$.

R_1 is *not asymmetric* since $(1, 2) \in R_1$ and $(2, 1) \in R_1$ but $1 \neq 2$.

R_1 is *not antisymmetric* since $1 \neq 2$ but $(1, 2) \in R_1$ and $(2, 1) \in R_1$.

• $R_2 = \{(1, 2), (2, 1), (2, 4), (3, 4), (4, 2), (4, 3), (5, 6), (6, 5), (6, 6)\}$.

Solution. R_2 is *symmetric* since when $(x, y) \in R_2$, $(y, x) \in R_2$.

R_2 is *not asymmetric* since $(1, 2) \in R_2$ and $(2, 1) \in R_2$.

R_2 is *not antisymmetric* since $1 \neq 2$ but $(1, 2) \in R_2$ and $(2, 1) \in R_2$. $\square$

• $R_3 = \{(a, b) \in A \times A : a > b\}$.

Solution. R_3 is *not symmetric* since $(2, 1) \in R_3$ but $(1, 2) \notin R_3$.

R_3 is *asymmetric* since $(x, y) \in R$ but $(y, x) \notin R$ for all $(x, y) \in R$.

R_3 is *antisymmetric* since for all $x \neq y$, $(x, y) \notin R$ or $(y, x) \notin R$. $\square$

def:5.24

Definition 4.11. A relation on R on a set A is **transitive** if whenever $(a, b) \in R$ and $(b, c) \in R$, then $(a, c) \in R$.

A relation R on a set A is **not transitive** if there exist $a, b, c \in A$ such that $(a, b) \in R$ and $(b, c) \in R$ but $(a, c) \notin R$.

Example 4.15. $<$, $\leq$, $>$, $\geq$ are transitive relations on $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$ because we have

- if $a < b$, $b < c$, then $a < c$
- if $a \leq b$, $b \leq c$, then $a \leq c$
- if $a > b$, $b > c$, then $a > c$
- if $a \geq b$, $b \geq c$, then $a \geq c$

4.1.3 Application of Sets and Relations

The following are some areas where sets and relations are applied:

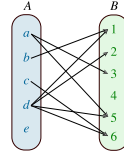
- The relation R , a subset of $A_1 \times \cdots \times A_n$ (Definition [4.2](#)), is a mathematical representation of predicate $P(x_1, \cdots, x_n)$ as in Topic 1's Definition 1.8.1 (terms and formulas).
- Some specification languages such as Z notation (uses a mixed of sets and types?), TLA+ language (for concurrent and distributed programs) use sets in **specification** and **modelling**.
- Denotational semantics, also known as *mathematical semantics* or *Scott-Strachey semantics* is an approach for formalising the “meanings” of programming languages by constructing mathematical objects (called denotations, e.g. domains, a special kinds of partial ordered set (Section [4.5](#))) that describe the meanings of the expressions from programming languages.
- Relational programming with <https://minikanren.org>: Set relations allows programming language to deal with predicates rather than just functions.

4.2 Representing Relations: Boolean Matrix and Digraph

To make binary relations easier to understand, various forms of representations are proposed:

- **Listing Tuples (Roster Notation):** E.g. $R = \{(x_1, y_1), \dots, (x_n, y_n)\}$.

This is useful when the domain and range are finite.



- **Arrow Diagram Representation:** E.g.

This is useful when the domain and range are finite.

- **(Boolean/Logical) Matrix Representation:** Section [4.2.1](#)

This is useful when the domain and range are the same and finite.

- **Directed Graph (Digraph) Representation:** Section [4.2.2](#)

This is useful when the domain and range are the same and finite.

- **Set Builder Notation:** E.g. $\{(x, y) : y = x^2\}$

This is useful when the mathematical formulation is clear.

- **Cartesian Plane Graph:**

This is useful for visualisation for domain and range which are numbers.

4.2.1 (Boolean/Logical) Matrix

A relation from a finite set A to a finite set B has a “Boolean” (or binary or logical) matrix representation.

def:5.15

Definition 4.12. Let $A = \{a_1, a_2, \cdots, a_m\}$ and $B = \{b_1, b_2, \cdots, b_n\}$ and let R be a relation from A to B . Then R can be represented by an $m \times n$ **Boolean matrix** or **logical matrix** or **0-1 integral matrix**

$$M_R = [m_{ij}]$$

called the *matrix representation of R* , where

$$m_{ij} = \begin{cases} 1 \text{ or T,} & \text{if } (a_i, b_j) \in R, \\ 0 \text{ or F,} & \text{if } (a_i, b_j) \notin R. \end{cases}$$

We can regard the Boolean matrix as usual matrix and apply the usual matrix multiplication or define matrix operations specific for logical operations generalising from the logical connectives $\wedge$ and $\vee$.

Definition 4.13. Let $M_R = [r_{ij}]$ and $M_S = [s_{ij}]$ be two $m \times n$ Boolean matrices. Let $M_T = [t_{jk}]$ be an $n \times k$ Boolean matrix.

- The **meet** (Conjunction or AND) of M_R and M_S , $M_R \wedge M_S = [r_{ij} \wedge s_{ij}]$;
- The **join** (Disjunction or OR) of M_R and M_S , $M_R \vee M_S = [r_{ij} \vee s_{ij}]$;
- The **product** of M_R and M_T , $M_R \odot M_T = [\bigvee_j (r_{ij} \wedge t_{jk})]$

Note that $0 \wedge 0 = 0$, $0 \wedge 1 = 0$, $1 \wedge 0 = 0$, $1 \wedge 1 = 1$; $0 \vee 0 = 0$, $0 \vee 1 = 1$, $1 \vee 0 = 1$, $1 \vee 1 = 1$. We can regard $0 = F$, $1 = T$.

Theorem 4.16. Let A and B be defined in Definition [4.12](#) and $C = \{c_1, c_2, \dots, c_k\}$. Let $R_1, R_2 \subset A \times B$, $R_3 \subset B \times C$ and the matrix representations of R_i be M_{R_i} , $i = 1, 2, 3$.

- The matrix representation of $R_1 \cap R_2$, $M_{R_1 \cap R_2} = M_{R_1} \wedge M_{R_2}$;
- The matrix representation of $R_1 \cup R_2$, $M_{R_1 \cup R_2} = M_{R_1} \vee M_{R_2}$;
- The matrix representation of $R = \{(a, c) \mid \exists b (aR_1b \wedge bR_3c)\} \subset A \times C$, $M_R = M_{R_1} \odot M_{R_3}$.

eg:5.14a

Example 4.17. Let $A = \{a, b, c, d, e, f\}$, $B = \{1, 2, 4, 5, 6\}$ and let $R = \{(a, 2), (a, 5), (b, 5), (b, 6), (c, 1), (c, 4), (e, 5), (f, 6)\}$ be the relation from A to B . Find the matrix representation of R .

Solution: $M_R =$

		1	2	4	5	6
a	$\left(\begin{array}{cccccc}$	0	1	0	1	0
b		0	0	0	1	1
c		1	0	1	0	0
d		0	0	0	0	0
e		0	0	0	1	0
f		0	0	0	0	1

eg:5.14b

Example 4.18. Let $M_R = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}$ be the matrix representation of a relation R from $A = \{a, b, c\}$ to $B = \{1, 2, 3, 4\}$. Write down the set R .

Solution.

$$R = \{(a, 1), (a, 2), (a, 4), (b, 2), (b, 3), (b, 4), (c, 1), (c, 4)\}.$$

□

In the following we will limit ourselves to the discussion of (binary) relations on a set. For a relation on a set, it can be represented by a square matrix.

Example 4.19. Let $A = \{1, 2, 3, 4, 5\}$ and $R = \{(1, 1), (1, 2), (2, 1), (3, 4), (3, 5), (4, 5)\}$. Write down the matrix representation M_R of R .

Solution: The matrix representation $M_R = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

Square Matrix in Scheme

```
(define (display-M M m n)
  (for-each
    (lambda (i)
      (for-each
        (lambda (j)
          (display (if (vector-ref M (+ (* i n) j)) 1 0))
          (display " ")
          (set! j (+ j 1)))
        (iota n))
      (newline))
    (iota m)))

(define (warshall M N)
  (for-each
    (lambda (k)
      (for-each
        (lambda (i)
          (if (vector-ref M (+ (* i N) k))
              (for-each
                (lambda (j)
                  (let ((pi (+ (* i N) j))
                        (pk (+ (* k N) j)))
                    (vector-set! M pi (or (vector-ref M pi)
                                           (vector-ref M pk))))
                  (iota N))))
              (iota N)))
        (iota N))
      (display (make-string 60 #\.)
        (newline)
        (display-M M N N))
      (iota N)))

;;; Discrete Maths (UECM1304) Test 2 Q2 on Warshall Algorithm for Transitive Closure
(define N 6)
(define M (make-vector (* N N) #f))

(for-each
  (lambda (i j) (vector-set! M (+ (* i N) j) #t))
  (list 0 0 1 1 2 3 4 5)
  (list 2 5 2 4 1 2 3 0))

(display-M M N N)
(display (make-string 60 #\=))
(newline)

(warshall M N)
```

rem:trans

Remark 4.20. Given a relation R on a finite set A .

- R is reflexive iff *the diagonal element of matrix representation M_R are all 1.*
- R is irreflexive iff *the diagonal element of M_R are all 0.*
- R is neither reflexive nor irreflexive iff *0's and 1's co-exist on the main diagonal of M_R .*

- R is **symmetric** iff M_R *is a symmetric matrix*, i.e. $M_R = M_R^T$.
- R is **asymmetric** iff M_R *has no symmetric pair* and its *diagonal contains all 0's*.
- R is **antisymmetric** iff M_R *has no symmetric pair*.
- R is **transitive** iff M_R *has the property if $m_{ij} = 1$ and $m_{jk} = 1$ then $m_{ik} = 1$* , i.e. if every non-zero element in the matrix M_R^2 corresponds to a non-zero element in M_R , then R is transitive.

eg: 5.29

Example 4.21. Let $A = \{1, 2, 3, 4, 5\}$ and R_i , $i = 1, 2, 3, 4$ be relations on A . Determine whether the following relations are reflexive, irreflexive, symmetric, anti-symmetric and transitive.

- $R_1 = \{(1, 2), (3, 4)\}$

Solution: R_1 is not reflexive because $(1, 1) \notin R_1$.

R_1 is irreflexive because the diagonal elements are all zeros.

R_1 is not symmetric because $(1, 2) \in R_1$ but $(2, 1) \notin R_1$.

R_1 is antisymmetric because there is no symmetric pair.

R_1 is transitive since there are no elements a, b and c in A such that $(a, b) \in R_1$ and $(b, c) \in R_1$, but $(a, c) \notin R_1$. We can also show this by comparing $M_{R_1}^2$ to M_{R_1} :

$$M_{R_1} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad M_{R_1}^2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Since there is no non-zero element in $M_{R_1}^2$, the rule in Remark 4.20 is satisfied, R_1 is transitive.

- $R_2 = \{(1, 1), (1, 2), (2, 1), (3, 4), (3, 5), (4, 5)\}$

Solution:

R_2 is not reflexive because $(5, 5) \notin R_2$.

R_2 is not irreflexive because $(1, 1) \in R_2$.

R_2 is not symmetric because $(3, 4) \in R_2$ but $(4, 3) \notin R_2$.

R_2 is not antisymmetric because we have a symmetric pair $(1, 2) \in R_2$ and $(2, 1) \in R_2$.

R_2 is not transitive since $(2, 1) \in R_2$ and $(1, 2) \in R_2$ but $(2, 2) \notin R_2$.

- $R_3 = \{(1, 2), (1, 3), (1, 4), (4, 4)\}$

Solution. R_3 is not reflexive because $(1, 1) \notin R_3$.

R_3 is not irreflexive because $(4, 4) \in R_3$.

R_3 is not symmetric because $(1, 2) \in R_3$ but $(2, 1) \notin R_3$.

R_3 is antisymmetric because there is no symmetric pair.

R_3 is transitive since we have $(1, 4) \in R_3$ and $(4, 4) \in R_3$ implying $(1, 4) \in R_3$. We can also show this by comparing $M_{R_3}^2$ to M_{R_3} :

$$M_{R_3} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad M_{R_3}^2 = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Since there are two non-zero element in $M_{R_3}^2$ corresponding to the two non-zero in M_{R_3} , the rule in Remark 4.20 is satisfied, R_3 is transitive. $\square$

- $R_4 = \{(1, 2), (2, 1), (2, 2), (2, 3), (3, 2)\}$

Solution. R_4 is not reflexive because $(1, 1) \notin R_4$.

R_4 is not irreflexive because $(2, 2) \in R_4$.

R_4 is symmetric because $M_{R_4} = M_{R_4}^T$.

R_4 is not antisymmetric because there is a symmetric pair $(1, 2)$ and $(2, 1)$.

R_4 is not transitive since we have $(1, 2), (2, 1) \in R_4$ but $(1, 1) \notin R_4$. We can also show this by comparing $M_{R_4}^2$ to M_{R_4} :

$$M_{R_4} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad M_{R_4}^2 = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 3 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Since there is a non-zero element (in fact, 4 non-zero elements) in $M_{R_4}^2$ which does not correspond to a non-zero element in M_{R_4} , according to the rule in Remark 4.20, R_4 is not transitive. $\square$

4.2.2 Digraphs

The digraph representation for the binary relation from A to B is the arrow diagram representation (or bipartite graph (https://en.wikipedia.org/wiki/Bipartite_graph) representation) mentioned earlier.

When $A = B = S$, a relation $R \subset S \times S$ becomes a binary relation on the set S . We can use a **directed graph (or digraph) representation** for R which is a pair (S, R) where S is the set of **nodes** and R is the set of **edges**.

Digraph is an abstraction of oriented connections of objects. The object is marked as a node while the connection is marked as an edge in a digraph. Some applications of digraph are shown in the table below:

digraph	node	edge
financial	stock, currency	transaction
transportation	street intersection, airport	highway, airway route
scheduling	task	precedence constraint
WordNet	synset	hypernym
Web	web page	hyperlink
game	board position	legal move
telephone	person	placed call
food web	species	predator-prey relation
infectious disease	person	infection
citation	journal article	citation
object graph	object	pointer
inheritance hierarchy	class	inherits from
control flow	code block	jump

Definition 4.14. A *directed graph* or *digraph* is a pair $G = (V, E)$ such that

- V is a set whose elements are called **nodes** or **vertices**;
- E is a set of ordered pairs of nodes, called (*directed*) **edges**, **arcs** or **arrows**.

Definition 4.15. The **adjacency matrix** of a digraph $G = (V, E)$ (with loops and multiple arcs) is the integer-valued matrix with rows and columns corresponding to the digraph nodes, where a non-diagonal entry a_{ij} is the number of arcs from node i to node j , and the diagonal entry a_{ii} is the number of loops at node i . The adjacency matrix for a digraph is unique up to the permutations of rows and columns. It is denoted $Adj(G)$.

The digraph representation of a relation R on a set A is defined below.

def:5.16

Definition 4.16. If R is a relation on a finite set A , we can construct a **digraph representation of R** , denoted by G_R , as follows:

1. Draw a small circle for each element of A and label the circle with the corresponding element of A . Such a circle is the node or vertex of the graph.
2. Draw a line with an arrow from a node v_i to a node v_j iff $v_i R v_j$. This line (v_i, v_j) is the edge when $i \neq j$ and it is a loop when $i = j$.

Remark 4.22. The matrix representation M_R of a relation R on a set A is equivalent to the adjacency matrix of the digraph representation G_R .

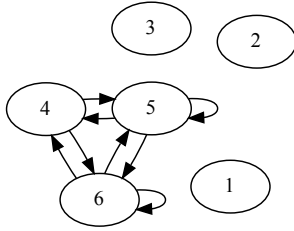
eg:5.15

Example 4.23. Let $A = \{1, 2, 4, 5, 6\}$ and a relation R on A be defined by $xRy = x + y > 8$. Draw the digraph representation for R and find the matrix representation of R .

Solution:

$$R = \{(4, 5), (4, 6), (5, 4), (5, 5), (5, 6), (6, 4), (6, 5), (6, 6)\}.$$

The digraph representation of R is



The matrix representation/adjacency matrix is

$$M_R = \begin{matrix} & \begin{matrix} 1 & 2 & 4 & 5 & 6 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 4 \\ 5 \\ 6 \end{matrix} & \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix} \end{matrix}$$

def:5.17

Definition 4.17. Given a digraph $G = (V, E)$. For a node $v \in V$, the **in-degree of v** , denoted as $\deg^-(v)$, is the number of head endpoints adjacent to the node v and the **out-degree of v** , denoted as $\deg^+(v)$, is the number of tail endpoints to v .

A node with $\deg^-(v) = 0$ is called a **source** whereas a node with $\deg^+(v) = 0$ is called a **sink**.

Definition 4.18. A **weighted digraph** or **network** or **labelled digraph** is a digraph in which the edges are attached with a **weight** or **label**. **Weights** are usually real numbers while labels can be real numbers or strings. The **(weighted) adjacency matrix** for a weighted digraph is a matrix with the weights of edges.

eg:5.16

Example 4.24. Given $A = \{a, b, c, d, e\}$ and a matrix representation of R below:

$$M_R = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{pmatrix}.$$

Draw the digraph representation for R and find the in-degree and out-degree of each node of the digraph representation.

The digraph representation is

Solution:

Node	a	b	c	d	e
In-degree	3	2	2	1	2
Out-degree	3	1	3	1	2

eg:5.17

Example 4.25. Consider a digraph $G = (V, E)$ where $V = \{1, 2, 3, 4, 5\}$ and

$$E = \{(1, 2), (1, 4), (2, 1), (2, 3), (2, 4), (2, 5), (3, 5), (4, 5), (5, 1), (5, 5)\}.$$

Find $Adj(G)$ and the in-degree and out-degree of each node of G .

Solution.

$$Adj(G) = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The in-degrees and out-degrees are listed in the table below:

Node	1	2	3	4	5
In-degree	2	1	1	2	4
Out-degree	2	4	1	1	2

□

An important concept in the study of digraph is the path, which is defined as below. The concept of path is used to characterise the concept of “reachability”, which is important in many areas of computer science. For example, a program using pointer-based data structures is free of memory leaks if every allocated segment of memory can be reached from a program variable. Finding solutions to solitaire puzzles, or more generally, goal search and motion planning can be expressed in terms of reachability.

def:5.19

Definition 4.19. For a digraph $G = (V, E)$, a **path from a node a to a node b** is a sequence of nodes

$$a, v_1, v_2, \dots, v_{n-1}, b$$

such that $(a, v_1), (v_1, v_2), (v_2, v_3), \dots, (v_{n-1}, b) \in E$. The **length of a path** is the number of edges in that path. A **cycle** is a path begins and ends at the same node, i.e. $a = b$. A **non-null path** is a path in which the length is positive. For any two nodes $s, t \in V$, t is called **reachable** from s if there is a path from s to t .

Given a digraph G with n nodes, how many paths of length k are there from one node to another? $[Adj(G)^2]_{ij}$ is the number of length 2 paths from node i to node j ! Continuing in this fashion, the answer we desire can be found by computing $Adj(G)^k$!

eg:digraph2

Example 4.26. Consider a set $A = \{1, 2, 3, 4\}$ and its relation $R = \{(1, 2), (2, 1), (2, 3), (3, 4)\}$. Write down all the paths of length 1 to length 4.

Solution: The digraph representation $G = (A, R)$ of R is



The adjacency matrix $Adj(G)$ of the digraph representation of R ,

$$Adj(G) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \text{ gives all length 1 paths: } (1, 2), (2, 1), (2, 3), (3, 4)$$

$$Adj(G)^2 = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \text{ gives all length 2 paths: } (1, 2, 1), (1, 2, 3), (2, 1, 2), (2, 3, 4)$$

$$Adj(G)^3 = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \text{ gives all length 3 paths: } (1, 2, 1, 2), (1, 2, 3, 4), (2, 1, 2, 1),$$

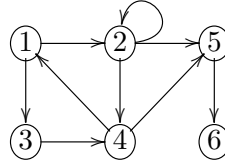
(2, 1, 2, 3)

$$Adj(G)^4 = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \text{ gives all length 4 paths:}$$

(1, 2, 1, 2, 1), (1, 2, 1, 2, 3), (2, 1, 2, 1, 2), (2, 1, 2, 3, 4)

eg:5.20

Example 4.27. Consider the following digraph



1. Find path(s) from the node u to the node v with given length.

(a) $u = 1, v = 4$, length=2.	1, 2, 4 and 1, 3, 4
(b) $u = 2, v = 6$, length=3.	2, 2, 5, 6 and 2, 4, 5, 6
(c) $u = 5, v = 3$, length=2.	No path from 5 to 3

2. Find all cycles with length 3.

Solution. C1: 1, 2, 4, 1; C2: 1, 3, 4, 1; C3: 2, 4, 1, 2;
 C4: 2, 2, 2, 2; C5: 3, 4, 1, 3; C6: 4, 1, 2, 4;
 C7: 4, 1, 3, 4.

□

Digraph theory and algorithms are used to solve practical problems related to connected nodes/states:

- *Transitive closure*: Is there a directed path from v to w ?
- *Strong connectivity*: Are all nodes mutually reachable?
- *Topological sort*: Can another digraph be drawn so that all edges point from left to right?
- *PERT/CPM*¹: Given a set of tasks with precedence constraints, how we can we best complete them all?
- *Shortest path*: Find best route from s to t in a weighted digraph.
- *PageRank*: What is the importance of a web page?

¹Project Evaluation Review Technique/Critical Path Method

4.3 Closures of (Binary) Relations

For a set $A = \{1, 2, 3\}$, if we are considering the addition relation on A , we find that $1 + 3$ gives 4, which is not inside A .

This kind of relation (the addition on A) is said to be **not closed**.

By adding enough elements into A , we can form $\bar{A} = \{1, 2, 3, 4, \dots\}$ which is closed for addition relation. Any two numbers adding together is still in A .

With the same rationale, we wish to turn some properties (e.g. reflexive, symmetry, etc.) of a binary relation to become “closed”, i.e. all necessary elements to meet the property to be in R .

In the theory of rewriting systems (<https://en.wikipedia.org/wiki/Rewriting>), one often uses more wordy notions such as the reflexive transitive closure R^* — the smallest preorder containing R , or the reflexive transitive symmetric closure $R^\equiv$ — the smallest equivalence relation containing R , and therefore also known as the **equivalence closure**.

When considering a particular term algebra (https://en.wikipedia.org/wiki/Term_algebra), an equivalence relation that is compatible with all operations of the algebra is called a *congruence relation*. The **congruence closure** of R is defined as the smallest congruence relation containing R .

For arbitrary property P and relation R , the **P closure of R need not exist**.

The properties reflexivity, transitivity and symmetry are **closed under arbitrary intersections**. In such cases, the P closure can be directly defined as the intersection of all sets with property P containing the relation R .

Some important particular closures can be constructively obtained as follows:

- $cl_{ref}(R) = R \cup \{(x, x) : x \in S\}$ is the **reflexive closure** of R ,
- $cl_{sym}(R) = R \cup \{(y, x) : (x, y) \in R\}$ is its **symmetric closure**,
- $cl_{trn}(R) = R \cup \{(x_1, x_n) : n > 1 \wedge (x_1, x_2), \dots, (x_{n-1}, x_n) \in R\} = \bigcup_{i=1}^n R^i$ is its **transitive closure**,
- $cl_{emb, \Sigma}(R) = R \cup \{ \langle f(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n), f(x_1, \dots, x_{i-1}, y, x_{i+1}, \dots, x_n) \rangle : \langle x_i, y \rangle \in R \wedge f \in \Sigma \text{ } n\text{-ary} \wedge 1 \leq i \leq n \wedge x_1, \dots, x_n \in S \}$ is its **embedding closure** with respect to a given set Σ of operations on S , each with a fixed arity.

The relation R is said to **have closure under some** cl_{xxx} , if $R = cl_{xxx}(R)$; for example R is called symmetric if $R = cl_{sym}(R)$.

In the latter case, the nesting order does matter; e.g. if S is the set of terms over $\Sigma = \{a, b, c, f\}$ and $R = \{\langle a, b \rangle, \langle f(b), c \rangle\}$, then the pair $\langle f(a), c \rangle$ is contained in the congruence closure of R , i.e. $cl_{trn}(cl_{emb, \Sigma}(cl_{sym}(cl_{ref}(R))))$, but not in the relation $cl_{emb, \Sigma}(cl_{trn}(cl_{sym}(cl_{ref}(R))))$.

Any of the four closures preserves symmetry, i.e., if R is symmetric, so is any $cl_{xxx}(R)$. Similarly, all four preserve reflexivity. Moreover, cl_{trn} preserves closure under $cl_{emb, \Sigma}$ for arbitrary Σ . As a consequence, the equivalence closure of an arbitrary binary relation R can be obtained as

$$cl_{trn}(cl_{sym}(cl_{ref}(R))),$$

and the congruence closure with respect to some Σ can be obtained as

$$cl_{trn}(cl_{emb, \Sigma}(cl_{sym}(cl_{ref}(R)))).$$

4.3.1 Reflexive Closure

The reflexive closure of a relation R on a set A is the smallest set $cl_{ref}(R)$ that contains R and is reflexive.

Example 4.28. Consider a set $A = \{1, 2, 3, 4\}$ and its relation $R = \{(1, 2), (2, 1), (2, 3), (3, 4)\}$. Find the reflexive closure $cl_{ref}(R)$ of R .

$$\text{Solution: } M_R = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{matrix} \Rightarrow M_{cl_{ref}(R)} = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}$$

We have $cl_{ref}(R) = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (1, 2), (2, 1), (2, 3), (3, 4)\}$

4.3.2 Symmetric Closure

A binary relation R may not be symmetric, but R is always contained in some symmetric relation $cl_{sym}(R)$. This is because we can add symmetric elements into R until we get the **smallest** relation R^{sym} is a symmetric closure.

Example 4.29. Consider a set $A = \{1, 2, 3, 4\}$ and its relation $R = \{(1, 2), (2, 1), (2, 3), (3, 4)\}$. Find the symmetric closure $cl_{sym}(R)$ of R .

$$\text{Solution: } M_R = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{matrix} \Rightarrow M_{cl_{sym}(R)} = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{matrix}$$

We have $cl_{sym}(R) = \{(1, 2), (2, 1), (2, 3), (3, 4), (3, 2), (4, 3)\}$

4.3.3 Transitive Closure and Warshall's Algorithm

A binary relation R may not be transitive but it is always contained in some transitive relation $cl_{trn}(R)$, which is the smallest called the **transitive closure**.

The **intersection** of two transitive relations is transitive. However, the **union** of two transitive relations need not be transitive. To preserve transitivity, one must take the transitive closure. This occurs, for example, when taking the union of two equivalence relations or two preorders. To obtain a new equivalence relation (Definition 4.20) or a preorder relation (Definition 4.22) one must take the transitive closure (reflexivity and symmetry — in the case of equivalence relations — are automatic).

In logic, the transitive closure of a binary relation cannot solely be expressed in first-order logic (FO). This means that one cannot write a formula using predicate symbols R and T that will be satisfied in any model iff T is the transitive closure of R . In finite model theory, first-order logic (FO) extended with a transitive closure operator is usually called transitive closure logic, and abbreviated FO(TC) or just TC. TC is a sub-type of fixpoint logic. The fact that FO(TC) is strictly more expressive than FO was discovered by Ronald Fagin in 1974; the result was then rediscovered by Alfred Aho and Jeffrey Ullman in 1979, who proposed to use fixpoint logic as a database query language.

Finding the transitive closure of a digraph is an practical problem in many computational tasks. It is used, for instance, in the reachability analysis of transition networks representing distributed and parallel systems and in the construction of parsing automata in compiler construction. Recently, efficient transitive closure computation has been recognised as a significant sub-problem in evaluating recursive database queries, since almost all practical recursive queries are transitive.

Let M be the matrix representation of the binary relation R of a finite set A of n elements.
Naïve Algorithm for getting transitive closure:

The following matrix

$$M^+ = M \vee M^2 \vee \dots \vee M^n.$$

gives the matrix representation of the transitive closure of R .

The naïve algorithm is very computationally inefficient! requiring n matrix multiplications!

(Floyd-)Warshall's Algorithm for getting transitive closure:

```

1 C := copy(M)
2 for k = 1, ..., n
3   for i = 1, ..., n
4     if i ≠ k and Cik ≠ 0:
5       for j = 1, ..., n
6         Cij := Cij ∨ Ckj

```

eg: 5.42

Example 4.30. Consider a set $A = \{1, 2, 3, 4\}$ and its relation $R = \{(1, 2), (2, 1), (2, 3), (3, 4)\}$ from Example 4.26. Find the transitive closure $cl_{trn}(R)$ of R .

Solution: The digraph representation of R is



Since $(1, 2) \in R$ and $(2, 1) \in R$ but $(1, 1) \notin R$, so R is not transitive.

The matrix representation of R is

$$M_R = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{matrix}$$

Step $k = 1$: look for 1 in first column.

Step $k = 2$: look for 1 in second column.

$$M_R^{(1)} = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & \boxed{1} & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{matrix}$$

$$M_R^{(2)} = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} \boxed{1} & 1 & \boxed{1} & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{matrix}$$

Step $k = 3$: look for 1 in third column.

Step $k = 4$: look for 1 in forth column.

$$M_R^{(3)} = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} 1 & 1 & 1 & \boxed{1} \\ 1 & 1 & 1 & \boxed{1} \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{matrix}$$

$$M_R^{(4)} = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{matrix}$$

The transitive closure of R is $cl_{trn}(R) = \{(1, 1), (1, 2), (1, 3), (1, 4), (2, 1), (2, 2), (2, 3), (2, 4), (3, 4)\}$.

Warshall's Algorithm in Scheme

```

(define (display-M M m n)
  (for-each
    (lambda (i)
      (for-each
        (lambda (j)
          (display (if (vector-ref M (+ (* i n) j)) 1 0))
          (display " ")
          (set! j (+ j 1)))
        (iota n))
      (newline))

```

```

        (iota m)))

(define (warshall M N)
  (for-each
    (lambda (k)
      (for-each
        (lambda (i)
          (if (vector-ref M (+ (* i N) k))
              (for-each
                (lambda (j)
                  (let ((pi (+ (* i N) j))
                      (pk (+ (* k N) j)))
                    (vector-set! M pi (or (vector-ref M pi)
                                           (vector-ref M pk))))))
                (iota N))))
          (iota N)))
    (iota N))
  (display (make-string 60 #\.)
    (newline)
    (display-M M N N))
  (iota N))

;;; Create 6 x 6 matrix (stored in 1-D array)
(define N 6)
(define M (make-vector (* N N) #f))
;;; Update the 6 x 6 matrix with the edges
(for-each
  (lambda (i j) (vector-set! M (+ (* i N) j) #t))
  (list 0 0 1 1 2 3 4 5)
  (list 2 5 2 4 1 2 3 0))

(display-M M N N)
(display (make-string 60 #\=))
(newline)

(warshall M N)

```

eg:5.44

Example 4.31. Consider a relation $R = \{(1, 1), (1, 2), (2, 1), (2, 2), (3, 3), (3, 4), (4, 3), (4, 4), (4, 5), (5, 4), (5, 5)\}$ on $A = \{1, 2, 3, 4, 5\}$. Find the transitive closure of R .

$$\begin{aligned}
 \text{Solution. } M_R &= \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \\
 \text{Step 2: } M_R^{(2)} &= \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \\
 \text{Step 4: } M_R^{(4)} &= \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix} \\
 \text{Step 1: } M_R^{(1)} &= \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \\
 \text{Step 3: } M_R^{(3)} &= \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \\
 \text{Step 5: } M_R^{(5)} &= \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix} = cl_{trn}(R)
 \end{aligned}$$

□

4.4 Equivalence Relation

sec:equiv

Equivalence relation is the **generalisation** of the equality of numbers “=” to sets with the following properties:

- $x = x$ for all x
- if $x = y$, then $y = x$
- if $x = y$ and $y = z$, then $x = z$.

def:equivalence

Definition 4.20. A binary relation R on a set A is called

- **reflexive**: xRx for all $x \in A$ (Definition [4.9](#));
- **symmetric**: if xRy , then yRx for all x, y (Definition [4.10](#));
- **transitive**: whenever $(x, y) \in R$ and $(y, z) \in R$, then $(x, z) \in R$ (Definition [4.11](#));
- a **compatibility relation** (<https://math.stackexchange.com/questions/508167/how-to-define-a-compatible-relation-using-a-cover>) or a **tolerance relation** (https://en.wikipedia.org/wiki/Tolerance_relation) if it is reflexive and symmetric;
- a **equivalence relation** if it is a compatibility relation which is transitive.

eg:5.32b

Example 4.32. Is the inequality relation “ $<$ ” an equivalence relation on $\mathbb{Z}$?

Solution: No, since the relation $<$ is not reflexive (e.g. $1 \not< 1$).

eg:5.33

Example 4.33. Let $A = \mathbb{Z}$ and R be the relation on A defined by $aRb = a + b$ is even. Show that R is an equivalence relation.

Proof:

- $a + a = 2a$ is even, so $(a, a) \in R$. R is reflexive.
- If $a + b$ is even then $b + a$ is also even. Thus $(a, b) \in R \Rightarrow (b, a) \in R$. So R is symmetric.
- Let $a, b, c \in A$. Let $(a, b) \in R$ and $(b, c) \in R$, then $a + b = 2k_1$ and $b + c = 2k_2$ for some integer k_1 and k_2 . Then $a + c = 2k_1 + 2k_2 - 2b = 2(k_1 + k_2 - b)$ is even. So R is transitive.

Hence R is an equivalence relation.

def:5.27

Definition 4.21. Let R be an equivalence relation on a set A and let $a \in A$. The set of all elements that related to a , is called the **equivalence class of a** and is denoted as

$$[a] \text{ or } \bar{a} = \{x \in A : (x, a) \in R\}.$$

The set of all the distinct equivalence classes of R form a set A/R , called the **quotient set of A** .

Example 4.34 (Finite Set). Consider the set

$$A = \{a, b, c, d, e, f, g\}$$

and the equivalence relation

$$R = \{(a, a), (a, b), (b, a), (b, b), (c, c), (c, d), (d, c), (d, d), (e, e), (e, f), (e, g), (f, e), (f, f), (f, g), (g, e), (g, f), (g, g)\}.$$

Find the equivalence class of each element in A and the quotient set A/R .

Solution: The equivalence class of

- a, b is $\{a, b\}$
- c, d is $\{c, d\}$
- e, f, g is $\{e, f, g\}$

The quotient set $A/R = \{\{a, b\}, \{c, d\}, \{e, f, g\}\}$

eg:5.36

Example 4.35. Let $A = \{1, 2, 3, 4\}$ and $R = \{(1, 1), (1, 2), (2, 1), (2, 2), (3, 3), (4, 4)\}$ be an equivalence relation on A . Find the equivalence class of each element in A and the quotient set A/R .

Solution. From $[1] = \{1, 2\} = [2]$, $[3] = \{3\}$, $[4] = \{4\}$,
 $A/R = \{[1], [3], [4]\}$. □

Example 4.36. Let $A = \mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$ and R be defined by $aRb = a + b$ is even. Find the equivalence class of each element in A and the quotient set A/R .

Solution: We put in some values to learn R concretely:

$$0R0, 0R2, 0R4, \dots, 0R(-2), 0R(-4), \dots$$

$$1R1, 1R3, 1R5, \dots, 1R(-1), 1R(-3), 1R(-5), \dots$$

Unfolding: It just means

$0+0$ is even, $0+2$ is even, $0+4$ is even, ... etc.

There are two equivalence classes:

- $[0] = \{0, \pm 2, \pm 4, \dots\}$
- $[1] = \{\pm 1, \pm 3, \pm 5, \dots\}$

The quotient is

$$A/R = \{[0], [1]\} = \{\{0, \pm 2, \pm 4, \dots\}, \{\pm 1, \pm 3, \pm 5, \dots\}\}.$$

eg:equiv3

Example 4.37. Let $A = \mathbb{Z}^+$ and R be the relation on A defined by

$$R = \{(x, y) \in A \times A : x - y \mod 3 = 0\}.$$

Find the equivalence class (called the “modulus” or “modular numbers”) of each element in A and the quotient set of A .

Solution: Since

- for all $x, y \in \{1, 4, 7, 10, \dots\}$, $x - y \mod 3 = 0$;
- for all $x, y \in \{2, 5, 8, 11, \dots\}$, $x - y \mod 3 = 0$;
- for all $x, y \in \{3, 6, 9, 12, \dots\}$, $x - y \mod 3 = 0$;

we have

$$[1] = \{1, 4, 7, 10, \dots\} = [4] = [7] = \dots = [3k - 2]$$

$$[2] = \{2, 5, 8, 11, \dots\} = [5] = [8] = \dots = [3k - 1]$$

$$[3] = \{3, 6, 9, 12, \dots\} = [6] = [9] = \dots = [3k]$$

where $k = 1, 2, \dots$.

So there are only three distinct equivalent classes of R , and hence $A/R = \{[1], [2], [3]\}$.

Remark 4.38. Example [4.37](#) is just a special case of the fact that the congruence modulo n is an equivalence relation.

- reflexive: For any $x \in \mathbb{Z}$, $x \equiv x \pmod{n}$;

- symmetric: For any $x, y \in \mathbb{Z}$, if $x \equiv y \pmod{n}$, then $y \equiv x \pmod{n}$;
- transitive: For any $x, y, z \in \mathbb{Z}$, if $x \equiv y \pmod{n}$ and $y \equiv z \pmod{n}$, then $x \equiv z \pmod{n}$.

The quotient $\mathbb{Z}/n\mathbb{Z} = \mathbb{Z}/\equiv_n$ is $\{[0], [1], \dots, [n-1]\}$

where

$$[0] = \{0, \pm n, \pm 2n, \dots\} = n\mathbb{Z}$$

$$[1] = \{1, \pm n + 1, \pm 2n + 1, \dots\} = n\mathbb{Z} + 1$$

...

$$[n-1] = \{n-1, \pm n + (n-1), \pm 2n + (n-1), \dots\} = n\mathbb{Z} + (n-1)$$

Mathematicians construct new mathematical objects from known mathematical objects using **(set) product**, **(set) subset**, **(set) union** and **(set) quotient** (based on some kind of equivalence relations).

Construction of the Set of Integers $\mathbb{Z}$

Assume we have a natural number set $\mathbb{N}$. The integer set $\mathbb{Z}$ is defined as the quotient set $(\mathbb{N} \times \mathbb{N})/\sim$ with

$$(m_1, n_1) \sim (m_2, n_2) \equiv m_1 + n_2 = m_2 + n_1$$

where $m_1, m_2, n_1, n_2 \in \mathbb{N}$. So $0 = \{(i, i) : i \in \mathbb{N}\}$, $1 = \{(i+1, i) : i \in \mathbb{N}\}$, $-1 = \{(i, i+1) : i \in \mathbb{N}\}$, etc.

Construction of the Set of Rational Numbers $\mathbb{Q}$

$\mathbb{Q} = \mathbb{Z} \times \mathbb{Z}/\sim$ where the equivalence relation $\sim$ is given by

$$(p_1, q_1) \sim (p_2, q_2) \equiv p_1 q_2 = p_2 q_1.$$

So we have $0 = \{(0, i) : i \in \mathbb{Z}\}$, $\frac{1}{2} = \{(i, 2i) : i \in \mathbb{Z}\}$, etc.

Construction of the Set of Real Numbers $\mathbb{R}$

There are many methods to construct $\mathbb{R}$ from a lot of $\mathbb{Q}$ which are introduced in https://en.wikipedia.org/wiki/Construction_of_the_real_numbers

1. Equivalent Cauchy sequence
2. Dedekind cuts
3. ... less popular.

Construction of the Euclidean Space $\mathbb{R}^n$ (Linear Algebra)

By using set product, we can obtain the Euclidean Space

$$\mathbb{R}^n = \mathbb{R} \times \mathbb{R} \times \dots \times \mathbb{R}.$$

However, we need to define the necessary operations such as equality, addition, subtraction, scalar multiplication, dot product, etc.

Construction of the Geometrical Objects

We construct geometrical objects (e.g. a circle) using subset in a product space like the Euclidean space or using equivalence relations on various sets to **glue** them together like a sculptor (these mathematicians are called topologists).

Example 4.39. $S^1 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 - 1 = 0\} \subset \mathbb{R}^2$

Example 4.40. $S^n = \{(x_1, \dots, x_n) \in \mathbb{R}^n : x_1^2 + \dots + x_n^2 - 1 = 0, j = 1, \dots, n\}.$

If X is a topological space (Definition [1.1](#)), gluing the points x and y in X means considering the quotient space obtained from the equivalence relation

$$a \sim b \text{ iff } a = b \text{ or } a = x, b = y.$$

For example, consider $[0, 1]$ and the equivalence relation $\sim$ such that $0 \sim 1$ while $x \sim x$ for $0 < x < 1$, then $[0, 1]/\sim$ is homeomorphic to the circle S^1 .

Consider the unit square $I^2 = [0, 1] \times [0, 1]$ and the equivalence relation $\sim$ generated by the requirement that all boundary points be equivalent, thus identifying all boundary points to a single equivalence class. Then $I^2/\sim$ is homeomorphic to the sphere S^2 .

More generally, suppose X is a space and A is a subspace of X . One can identify all points in A to a single equivalence class and leave points outside of A equivalent only to themselves. The resulting quotient space is denoted X/A (as an abbreviation of X/R).

Consider the set $\mathbb{R}$ of real numbers with the ordinary topology, and write $x \sim y$ iff $x - y \in \mathbb{Z}$. Then the quotient space $\mathbb{R}/\sim$ is homeomorphic to the unit circle S^1 via the homeomorphism which sends the equivalence class of x to $\exp(2\pi i x)$.

A generalisation of the previous example is the following: Suppose a topological group G acts continuously on a space X . One can form an equivalence relation on X by saying points are equivalent iff they lie in the same orbit. The quotient space under this relation is called the **orbit space**, denoted X/G .

There are YouTube videos on the application of equivalence relation to construct interesting geometrical objects:

- <https://www.youtube.com/watch?v=lmcT2mP2bfE>

Most of them are boring mathematics because it is difficult to create pretty video explaining how to use product, quotient, etc. to create new mathematical objects called manifolds.

Many equivalence relations which are just “impossible” to visualise.

Example 4.41 (https://en.wikipedia.org/wiki/Projective_space). A **projective space** $\mathbb{P}(\mathbb{R}^n)$ of dimension n is the quotient set of $\mathbb{R}^{n+1} \setminus \{0\}$ by the equivalence relation “being on the same vector line”, i.e. $x \sim y$ if there is a nonzero element $\lambda \in \mathbb{R}$ such that $x = \lambda y$.

A generalisation of projective space is the Grassmannian $Gr(k, \mathbb{R}^n)$ which is a set that parameterises all k -dimensional linear subspaces of the n -dimensional vector space $\mathbb{R}^n$. When $k = 1$, $Gr(1, \mathbb{R}^n)$ is the projective space $\mathbb{P}(\mathbb{R}^{n-1})$.

There are many equivalence relations where the quotient is **impossible** to describe such as the logical equivalence relation: Let ϕ, ψ, ξ be any propositions.

- $\phi \leftrightarrow \phi$, therefore $\phi \equiv \phi$ and $\equiv$ is reflexive;
- If $\phi \equiv \psi$, then $\phi \leftrightarrow \psi$ is a tautology, then $\psi \leftrightarrow \phi$ is a tautology, therefore $\psi \equiv \phi$ and $\equiv$ is symmetric.
- If $\phi \equiv \psi$ and $\psi \equiv \xi$, then $\phi \rightarrow \psi$, $\phi \leftarrow \psi$, $\psi \rightarrow \xi$, $\psi \leftarrow \xi$ are tautologies, then $\phi \rightarrow \xi$ and $\phi \leftarrow \xi$ by modus ponens and $\phi \leftrightarrow \xi$ is a tautology, therefore $\phi \equiv \xi$ and $\equiv$ is transitive.

However, the quotient of all propositions under logical equivalence is **too complex** to describe.

4.5 Partial Order Relation

sec:poset

Partial order relation is the **generalisation** of the ordering “ $\leq$ ” to sets with the following properties:

- $x \leq x$
- if $x \leq y$ and $y \leq x$, then $x = y$
- if $x \leq y$ and $y \leq z$, then $x \leq z$.

for all x, y, z . The investigation of partial order relations has lead to the development of lattice theory in mathematical science but computer scientists have extended the theory to domain theory, which can be applied to denotational semantics of (functional) programming languages.

def:poset

Definition 4.22. A binary relation R on a set A is called

- **reflexive:** xRx for all $x \in A$ (Definition [4.9](#));
- **antisymmetric:** if xRy and yRx , then $x = y$ for all $x, y \in A$ (Definition [4.10](#));
- **transitive:** whenever $(x, y) \in R$ and $(y, z) \in R$, then $(x, z) \in R$ (Definition [4.11](#));
- **preorder** if it is reflexive and transitive (<https://en.wikipedia.org/wiki/Preorder>);
- **(partial) order** or **approximation ordering** if it is reflexive, antisymmetric and transitive (https://en.wikipedia.org/wiki/Partially_ordered_set).

The set A together with the partial order relation $\preceq$, $(A, \preceq)$, is called a **partially ordered set**, or **poset**. For $x, y \in A$, y **approximates** x if $x \preceq y$.

Remark 4.42. Computer scientists use $\sqsubseteq$ rather than $\preceq$.

eg:poset1

Example 4.43. Typical examples of partial order relation and posets:

- $\subset$ and $(\mathcal{P}(S), \subset)$;
- $\leq$ and $(\{1, 2, 3\}, \leq)$;
- $\geq$ and $(\{1, 2, 3\}, \geq)$;
- $\leq$ and $(\mathbb{Z}, \leq)$;
- $\geq$ and $(\mathbb{Z}, \geq)$;
- $\leq$ and $(\mathbb{R}, \leq)$;
- $\geq$ and $(\mathbb{R}, \geq)$;
- (A, R) where $A = \text{English dictionary}$, $xRy = x$ comes before y in dictionary order;
- (A, R) where $A = \text{set of project tasks}$, xRy means that x must be completed before y begins where x and y are tasks in a project A .

eg:5.45b

Example 4.44. $(\mathbb{Z}, <)$ and $(\mathbb{Z}, >)$ are **not** posets since the relations “ $<$ ” and “ $>$ ” are not reflexive: $1 \not< 1$ and $1 \not> 1$.

Example 4.45. Show that $(\mathcal{P}(\{1, 2, 3\}), \subset)$, a special case of Example [4.43](#) is a poset.

Proof:

$$\mathcal{P}(\{1, 2, 3\}) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$$

The matrix representation

$$M_{\subset} = \begin{matrix} & \emptyset & \{1\} & \{2\} & \{3\} & \{1, 2\} & \{1, 3\} & \{2, 3\} & \{1, 2, 3\} \\ \begin{matrix} \emptyset \\ \{1\} \\ \{2\} \\ \{3\} \\ \{1, 2\} \\ \{1, 3\} \\ \{2, 3\} \\ \{1, 2, 3\} \end{matrix} & \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}$$

The $\subset$ is reflexive because the diagonal elements of $M_{\subset}$ are all 1.

The $\subset$ is antisymmetric because the lower triangular part of $M_{\subset}$ are all 0 so not symmetric pair can be formed.

The $\subset$ is transitive since if $A \subset B$ and $B \subset C$, then $A \subset C$.

Attempting to use M_R^2 to verify transitivity would be too painful:

$$M_{\subset}^2 = \begin{matrix} & \emptyset & \{1\} & \{2\} & \{3\} & \{1,2\} & \{1,3\} & \{2,3\} & \{1,2,3\} \\ \emptyset & 1 & 2 & 2 & 2 & 4 & 4 & 4 & 8 \\ \{1\} & 0 & 1 & 0 & 0 & 2 & 2 & 0 & 4 \\ \{2\} & 0 & 0 & 1 & 0 & 2 & 0 & 2 & 4 \\ \{3\} & 0 & 0 & 0 & 1 & 0 & 2 & 2 & 4 \\ \{1,2\} & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 2 \\ \{1,3\} & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 2 \\ \{2,3\} & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 \\ \{1,2,3\} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{matrix}$$

Since every non-zero element in the matrix $M_{\subset}^2$ corresponds to a non-zero element in $M_{\subset}$, $\subset$ is transitive.

The “divisibility” in Topic 3 are reflexive, antisymmetric and transitive and is therefore a partial order relation.

eg:5.45c

Example 4.46. Show that $(\mathbb{Z}^+, |)$ a poset where $|$ is the ‘integral divisibility’.

Proof:

- The relation $|$ is reflexive because any positive integer divides itself.
- Note that if $a|b$ and $b|a$ then $a = \pm b$. Since $a, b \in \mathbb{Z}^+$, the negative sign will not happen and $a = b$. So the relation $|$ is antisymmetric.
- The relation $|$ is transitive because $a|b \Rightarrow b = k_1a$, $b|c \Rightarrow c = k_2b$, therefore $c = k_2k_1a \Rightarrow a|c$.

Hence the relation of divisibility $|$ is a partial order and $(\mathbb{Z}^+, |)$ is a poset.

In the poset $(\mathbb{Z}^+, |)$

- $2|4$, therefore $2 \preceq 4$
- $2|6$, therefore $2 \preceq 6$
- $4 \nmid 6$, therefore $4 \not\preceq 6$
- $6 \nmid 4$, therefore $6 \not\preceq 4$

Note that **neither** $4 \preceq 6$ **nor** $6 \preceq 4$, we say that 4 and 6 are **incomparable** in $(\mathbb{Z}^+, \preceq)$.

Definition 4.23. In a poset $(A, \preceq)$, for any $a, b \in A$, if neither $a \preceq b$ nor $b \preceq a$, then a and b are said to be **incomparable**.

Definition 4.24. We say $a \prec b$ if $a \preceq b$ and $a \neq b$. Similarly, $a \succ b$ if $a \succeq b$ and $a \neq b$.

Definition 4.25. Given a poset $(A, \preceq)$. For $a \in A$,

- If there is no $b \in A$ such that $b \prec a$, we say that a is a **minimal element (of A)** or **least element (of A)**.
- If there is some $b \in A$ so that $b \prec a$ and there is no $c \in A$ so that $b \prec c \prec a$, we say that b is an **immediate predecessor of a** .
- If there is no $b \in A$ such that $a \prec b$, we say that a is a **maximal element (of A)**.
- If there is some $b \in A$ so that $a \prec b$ and there is no $c \in A$ so that $a \prec c \prec b$, we say that b is an **immediate successor of a** .

Example 4.47. In $(\mathbb{Z}^+, |)$:

- 1 is the minimal element of $\mathbb{Z}^+$ because $1|a$, $a \in \mathbb{Z}^+$.
- 1 is the immediate predecessor of any prime number
- 2 is the immediate predecessor of 4, 6 but not 8 because $2 \prec 4 \prec 8$; 4 is the immediate successor of 2, 6 is the immediate successor of 2 and 3, etc.
- There is no maximal element in $\mathbb{Z}^+$

Example 4.48. $(\mathbb{Q}, \leq)$ has no minimal element, maximal element and any $a \in \mathbb{Q}$ does not have immediate predecessor or successor.

A special case of poset is chain.

wiki:total_order

Definition 4.26. A **total order** or **linear order** is a partial order relation on some set A which is **total**, i.e. for all $x, y \in A$:

- $x \preceq y$ or $y \preceq x$ (totality).

A set paired with a total order (A, R) is called a **totally ordered set**, a **linearly ordered set**, or a **chain**.

Example 4.49. $\leq$ and $\geq$ are total order on $\mathbb{Z}$. $(\mathbb{Z}, \leq)$ and $(\mathbb{Z}, \geq)$ are chains.

A particular case of total order is the well-order.

Definition 4.27. A **well-order** is a total order such that

- every non-empty subset has a least element.

A set A with its well-order $\preceq$ forms a **well-ordered set** $(A, \preceq)$.

Example 4.50 (<https://en.wikipedia.org/wiki/Well-order>). • $\leq$ on $\mathbb{R}$, $\mathbb{Z}$ is not well-order because the non-empty subset $\{-2, -4, -6, \dots\}$ does not have a least element.

- $\leq$ on $\mathbb{N}$ is well order.
- The order $\preceq$ on $\mathbb{N}$ such that

$$0 \preceq 2 \preceq 4 \preceq 6 \preceq 8 \preceq \dots \preceq 1 \preceq 3 \preceq 5 \preceq 7 \preceq \dots$$

is a well order.

- The order $\preceq$ on $\mathbb{Z}$ such that

$$0 \preceq 1 \preceq 2 \preceq 3 \preceq 4 \preceq \dots \preceq -1 \preceq -2 \preceq -3 \preceq -4 \preceq \dots$$

is a well order.

- The order $\preceq$ on $\mathbb{Z}$ such that

$$0 \preceq -1 \preceq 1 \preceq -2 \preceq 2 \preceq -3 \preceq 3 \preceq -4 \preceq 4 \preceq \dots$$

is a well order.

Theorem 4.51 (Well-Ordering Principle). *For any non-empty set A , there is a well-order $\prec$.*

Partial orders form a natural setting for monotonic functions.

def:monotonic

Definition 4.28. Given two posets $(A, \preceq_A)$ and $(B, \preceq_B)$, a function, $f : A \rightarrow B$, is

- **monotonic, increasing or order-preserving** if for all $a, b \in A$, $a \preceq_A b \Rightarrow f(a) \preceq_B f(b)$.

When $\preceq$ is changed to $\prec$, we call f **strictly increasing**.

- **decreasing** if for all $a, b \in A$, $a \preceq_A b \Rightarrow f(b) \preceq_B f(a)$.

When $\preceq$ is changed to $\prec$, we call f **strictly decreasing**.

The following is a generalisation of mathematical induction.

Theorem 4.52 (Principle of Transfinite Induction). *Let $(A, \preceq)$ be a well-ordered set with a minimal element x_0 .*

1. *Let $B \subset A$ and the initial segment of the set A with respect to the element z be $A_z := \{x \in A : x \prec z\}$. For any $z \in A$, when $A_z \subset B$, $z \in B$. Then $B = A$.*

2. *Suppose $P(x)$ is a predicate with $x \in A$ such that*

(a) *$P(x_0)$ is true;*

(b) *For any $x \in A$, if $\forall y(y \prec x \rightarrow P(y))$, then $P(x)$ is true.*

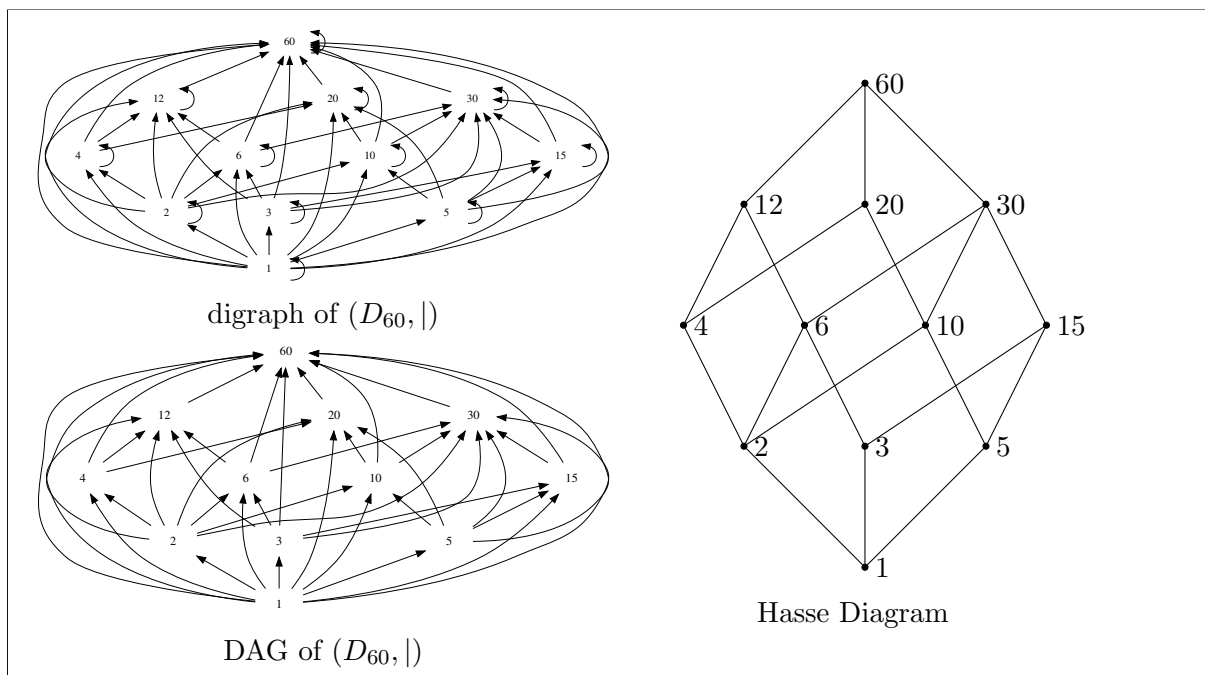
Then $\forall x \in A$, $P(x)$.

Definition 4.29. A digraph (Definition 4.14) that contains no sequence of edges leading back to its starting point is called a **directed acyclic graph** or **DAG**.

Definition 4.30. A **Hasse diagram** is a “reduced” DAG to represent the **transitive reduction**. For a finite poset $(S, \preceq)$, one represents each element of S as a vertex in the plane and draws a line segment or curve that goes upward from x to y whenever $x \preceq y$ and there is no z such that $x \preceq z \preceq y$. These curves may cross each other but must not touch any vertices other than their endpoints. Such a diagram, with labelled vertices, uniquely determines its partial order.

Let D_n denote the **set of all divisors of the integer n** .

Example 4.53. Sketch the digraph representation, DAG and Hasse diagram of the poset $(D_{60}, |)$.

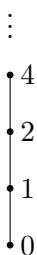


g:hasse-diagram

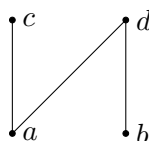
Example 4.54.

eg:5.45a

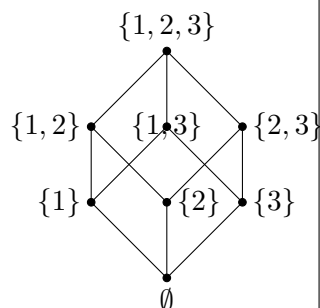
$(\mathbb{Z}^{\geq 0}, \leq)$ and $(\mathbb{Z}^{\geq 0}, \geq)$ are posets since the relations “ $\leq$ ” and “ $\geq$ ” are reflexive, antisymmetric and transitive on $\mathbb{Z}$.



$(\{a, b, c, d\}, \preceq)$ where $a \preceq c$, $a \preceq d$, $b \preceq d$ and $e \preceq e$ for every $e = a, b, c, d$.



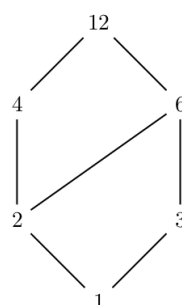
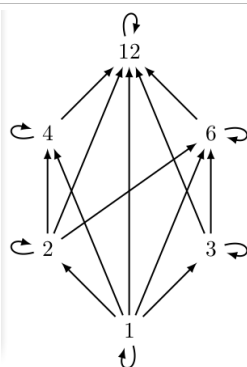
$(\mathcal{P}(\{1, 2, 3\}), \subset)$ is a poset.



Example 4.55. $(D_{12} = \{1, 2, 3, 4, 6, 12\}, |)$ is a poset. Let the elements of $|$ and sketch the digraph representation and the Hasse diagram.

Solution.

$| = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 6), (1, 12), (2, 2), (2, 4), (2, 6), (2, 12), (3, 3), (3, 6), (3, 12), (4, 4), (4, 12), (6, 6), (6, 12), (12, 12)\}$



□

It would be interesting to generalise the factors of an integer to “factors of polynomials”.

The first difficulty we would face would be factors for a polynomial are not unique. For example,

$$x^2 + 5x + 6 = (x + 2)(x + 3) = 6(1 + \frac{x}{2})(1 + \frac{x}{3}).$$

To make the factorisation unique, we have to limit ourselves to polynomials with the coefficient of the higher degree being 1. In such a case, the factorisation will be the middle term and is unique.

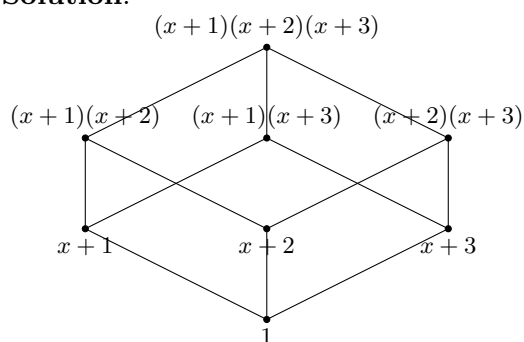
Definition 4.31. A **monic polynomial** is a nonzero univariate polynomial in with the leading coefficient is equal to 1, i.e. it has the form

$$x^n + c_{n-1}x^{n-1} + \cdots + c_2x^2 + c_1x + c_0$$

for some $n \geq 0$.

Example 4.56 (Monic Factors of a Monic Polynomial). Let A be the set all the monic polynomial factors of the monic polynomial $(x+1)(x+2)(x+3)$ and the partial order relation $p \preceq q$ be p divides q for any monic polynomials p, q . Sketch the Hasse diagram of the poset $(A, \preceq)$.

Solution:



It seems that we can obtain a Hasse diagram representing a distributive lattice isomorphic to the lattice given by the Boolean algebra generated by $(\mathcal{P}(\{1, 2, 3\}), \subset)$.

Remark 4.57. The lack of uniqueness in factorisation of algebraic objects such as the set of polynomials $\mathbb{R}[x]$ forces mathematicians to develop “abstract modern algebra” to characterise the components of algebraic objects. This is also why the notion of lattice in [Section ??](#) is not popular in mathematics.

Example 4.58 (Abstraction from English dictionary). Let Σ be some **alphabet** and consider the set

$$\Sigma^* = \Sigma^0 \cup \Sigma^1 \cup \Sigma^2 \cup \cdots$$

of all **finite words** drawn from Σ .

Given two words x and y , we define the partial order

$$x \preceq y := x \text{ is a prefix of } y,$$

i.e. if there is some word z such that $xz = y$.

$(\Sigma^*, \preceq)$ can be shown to be a poset.

Methods for constructing new poset from two posets $(A, \preceq_A)$ and $(B, \preceq_B)$ are given below.

Method 1

Define $\preceq$ on $A \times B$ to be

$$(a, b) \preceq (a', b') := a \preceq_A a' \text{ and } b \preceq_B b'.$$

It can be proved that $\preceq$ is a partial order. The poset $(A \times B, \preceq)$ defined in this way is called the **product poset** of A and B .

Method 2

Define $\preceq$ on $A \times B$ to be

$$(a, b) \preceq (a', b') := (a \preceq_A a') \vee (a = a' \wedge b \preceq_B b').$$

It is a partial order on $A \times B$ called the **lexicographic order**.

The useful property of lexicographic order (lex order for short) is that if the original partial orders are total, so is the lex order: this is why dictionary-makers use it. This also gives a source of very difficult-to-visualise total orders, like lex order on $\mathbb{R} \times \mathbb{R}$, which looks like the classic real number line where every point is replaced by an entire copy of the reals.

zorn-lemma

Axiom 4.59 (Zorn’s Lemma). Every non-empty poset in which each chain has an upper bound has a maximal element.

4.5.1 Topological Sorting

We can study topological sorting on a poset or a directed acyclic graph. The former provides us a mathematical way to transform *partial order structure* into a *linear order structure*. The latter is more appealing from the visualisation aspect, it transforms an *acyclic digraph* into a *linear structure* and this section study topological sorting from this point of view.

Definition 4.32. A *directed acyclic graph* (DAG) or *acyclic digraph* is a directed graph with no directed cycles. Special cases of directed acyclic graphs include the *multi-trees* (graphs in which no two directed paths from a single starting node meet back at the same ending node), *oriented trees* or *poly-trees* (the digraphs formed by orienting the edges of undirected acyclic graphs), and the *rooted trees* (oriented trees in which all edges of the underlying undirected tree are directed away from the root).

To construct a topological sorting for a DAG, first, find a list of “start nodes” which have no incoming edges and insert them into a set S . For acyclic digraph, at least one such node must exist.

Find the Topological Sort of a DAG $G = (V, E)$

$S \leftarrow$ Set of all nodes with no outgoing edges

$L \leftarrow []$

$\triangleright$ An Empty List

procedure visit(Node n)

if n has not been visited yet **then**

 mark n as visited

for each node m such that $(m, n) \in E$ **do**

 visit(m)

 add n to L

for each node n in S **do**

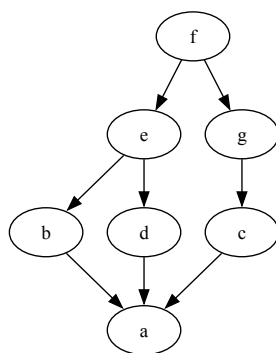
 visit(n)

return L

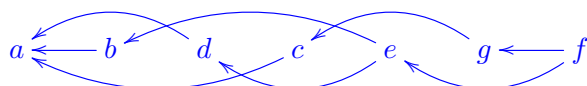
$\triangleright$ A list that contains the sorted elements

eg:5.52

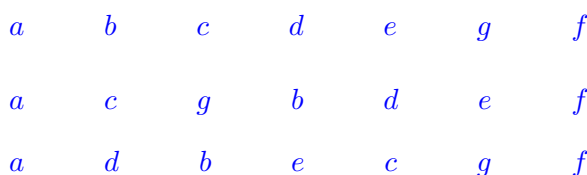
Example 4.60. Give a topological sorting for the DAG below.



Solution.



Other possible answers are



□

Remark 4.61. There are many ways of topologically sorting of a given DAG (and poset).

Topological sorting has an application to the scheduling of projects.

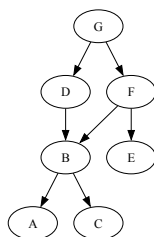
eg: 5.54

Example 4.62. A development project at a computer company requires the completion of seven tasks. Some of these tasks can be started only after other tasks are finished. A partial ordering on tasks is set up by considering task $X \prec$ task Y if task Y cannot be started until task X has been completed. The dependency for the seven tasks, is shown below:

- G depends on D and F
- D depends on B
- F depends on B and E
- B depends on A and C

Find a linear order in which these tasks can be carried out to complete the project.

Solution. We would first construct the digraph of the dependency.



The result of this sort is

$A \prec C \prec B \prec E \prec F \prec D \prec G.$

□

Part II

Real and Complex Analysis

Chapter 5

Real Number System

The set of real numbers $\mathbb{R}$ (Section [3.1.3](#) together with the arithmetic and relational operations (Definition [3.11](#)) make it a complete (Definition [3.16](#)) Archimedean ordered field (Definition [3.9](#)).

I have not taught real analysis before, so I need to work on the notes when I have the time.

I will follow the following references:

BEST!!! Edwin Hewitt, Karl Stromberg, Real And Abstract Analysis, GTM25, 1975

Tom Apostol-Calculus-v1,2

Bridger-Real Analysis A Constructive Approach-2006

Schechter E-Handbook Of Analysis And Its Foundations-AP1997

Peng Yee Lee-Rudolf Vyborny-The Integral An Easy Approach After Kurzweil And Henstock-Australian Mathematical Society Series on Mathematics and its Applications
Cambridge University Press 2000

Russell A Gordon-The Integrals Of Lebesgue Denjoy Perron And Henstock-Graduate Studies In Mathematics 4-AMS 1994

Frank E Burk-Garden Of Integral-MAA 2007

5.1 Topological Properties of Real Line

- Well-ordering property of natural numbers: Refer to Section [3.1.1](#)
 - Supremum (Definition [3.10](#)), infimum (need to state definition)
 - Archimedean property (Definition [3.8](#))
 - Open sets, closed sets
 - Interior, closure, boundary
 - Accumulation points
 - Compactness
 - Heine-Borel Theorem

5.2 Real-Valued Functions

Polynomial functions and n -root function?

Exponential and Logarithmic functions

Trigonometric functions

Inverse trigonometric functions

Definition 5.1.

- (a) $\sin^{-1} x$ is a number in $[-\frac{\pi}{2}, \frac{\pi}{2}]$, whose sine value is x .
- (b) $\cos^{-1} x$ is a number in $[0, \pi]$, whose cosine value is x .
- (c) $\tan^{-1} x$ is a number in $(-\frac{\pi}{2}, \frac{\pi}{2})$, whose tangent value is x .

- (d) $\cot^{-1} x$ is a number in $(0, 2\pi)$, whose cotangent value is x .
- (e) $\sec^{-1} x$ is a number in $[0, \frac{\pi}{2}) \cup (\frac{\pi}{2}, \pi]$, whose secant value is x .
- (f) $\csc^{-1} x$ is a number in $[-\frac{\pi}{2}, 0) \cup (0, \frac{\pi}{2}]$, whose cosecant value is x .

If $f(x) = \sin^{-1} x$, then the domain of f is $[-1, 1]$ and its range is $[-\frac{\pi}{2}, \frac{\pi}{2}]$. To obtain the derivative of $f(x) = \sin^{-1} x$, we let $y = \sin^{-1} x$. By definition, we have $\sin y = x$.

Hyperbolic functions are obtained through various algebraic combinations of e^x and e^{-x} . They have identities with ‘forms’ similar to their trigonometric counterparts.

Definition 5.2. Hyperbolic sine, cosine, tangent, cotangent, secant and cosecant functions are stated below.

- (a) $\sinh : \mathbb{R} \rightarrow \mathbb{R}, \quad x \mapsto \sinh x = \frac{1}{2}(e^x - e^{-x})$.
- (b) $\cosh : \mathbb{R} \rightarrow [1, \infty), \quad x \mapsto \cosh x = \frac{1}{2}(e^x + e^{-x})$.
- (c) $\tanh : \mathbb{R} \rightarrow (-1, 1), \quad x \mapsto \tanh x = \frac{\sinh x}{\cosh x}$.
- (d) $\coth : \mathbb{R}^* \rightarrow (-\infty, -1) \cup (1, \infty), \quad x \mapsto \coth x = \frac{\cosh x}{\sinh x}$.
- (e) $\operatorname{sech} : \mathbb{R} \rightarrow (0, 1], \quad x \mapsto \operatorname{sech} x = \frac{1}{\cosh x}$.
- (f) $\operatorname{csch} : \mathbb{R}^* \rightarrow \mathbb{R} \setminus \{0\} \quad x \mapsto \operatorname{csch} x = \frac{1}{\sinh x}$.

We include here a list of identities involving the hyperbolic functions for reference.

$$\begin{array}{ll}
 \sinh(-x) = -\sinh x & \cosh(-x) = \cosh x \\
 \sinh(x+y) = \sinh x \cosh y + \cosh x \sinh y & \cosh(x+y) = \cosh x \cosh y + \sinh x \sinh y \\
 \cosh^2 x - \sinh^2 x = 1 & \tanh^2 x + \operatorname{sech}^2 x = 1 \\
 \coth^2 x - \operatorname{csch}^2 x = 1 &
 \end{array}$$

Similar to the inverse trigonometric functions, we may also define inverse hyperbolic functions.

Definition 5.3.

- (a) $\sinh^{-1} : \mathbb{R} \rightarrow \mathbb{R}, \quad x \mapsto \sinh^{-1} x = \ln(x + \sqrt{x^2 + 1})$.
- (b) $\cosh^{-1} : [1, \infty) \rightarrow [0, \infty), \quad x \mapsto \cosh^{-1} x = \ln(x + \sqrt{x^2 - 1})$.
- (c) $\tanh^{-1} : (-1, 1) \rightarrow \mathbb{R}, \quad x \mapsto \tanh^{-1} x = \frac{1}{2} \ln \left(\frac{1+x}{1-x} \right)$.
- (d) $\coth^{-1} : (-\infty, -1) \cup (1, \infty) \rightarrow \mathbb{R}^*, \quad x \mapsto \coth^{-1} x = \frac{1}{2} \ln \left(\frac{x+1}{x-1} \right)$.
- (e) $\operatorname{sech}^{-1} : (0, 1] \rightarrow (0, \infty), \quad x \mapsto \operatorname{sech}^{-1} x = \ln \left(\frac{1+\sqrt{1-x^2}}{x} \right)$.
- (f) $\operatorname{csch}^{-1} : \mathbb{R}^* \rightarrow \mathbb{R}^*, \quad x \mapsto \operatorname{csch}^{-1} x = \ln \left(\frac{1}{x} + \frac{1+\sqrt{1+x^2}}{|x|} \right)$.

5.3 Sequences and Series

def:sequence

Definition 5.4. A **sequence** is a function with a subset of natural numbers $I \subset \mathbb{N}$ as its domain. It is usually denoted as $\{x_i\}$ or $\{x_i\}_{i=0}^{\infty}$ (when $I = \mathbb{N}$) instead of $x : I \subset \mathbb{N} \rightarrow S$ where S is a set. When $S = \mathbb{R}$, we have a real sequence; when $S = \{f : \mathbb{R} \rightarrow \mathbb{R}\}$, we have a function sequence.

Definition 5.5. · subsequences, Cauchy sequences

- Limit of sequences
 - Limit theorems
 - Monotone Convergence Theorem
 - Bolzano-Weierstrass Theorem

def:series

Definition 5.6. A series of a sequence $\{x_i\}$ is defined to be the sum of the elements in the order where the sequence is defined, i.e.

$$\sum_{i=0}^{\infty} = x_0 + x_1 + x_2 + \cdots .$$

- Numerical series: convergence tests for positive series, absolute and conditional convergence, alternating series, alternating series test
- Rearrangement of series
- Review of convergence tests on series of numbers

Chapter 6

Limits and Continuity

6.1 Limit of Functions

- Non-existence criterion for limits
- Limit theorems

6.1.1 The limit of a function

Definition 6.1. A *function* is a rule that assigns each element x in a set X to exactly one element, called $f(x)$, in a set Y .

We usually consider functions for which the sets X and Y are sets of real numbers. The set X is called the domain of the function f . When x varies throughout the domain of f , the range of f is the set of all possible values of $f(x)$.

The most common method for visualizing a function is its graph. The graph of f consists of all points (x, y) in the coordinate plane such that $y = f(x)$ and x is in the domain of f .

Definition 6.2. We write

$$\lim_{x \rightarrow a} f(x) = L$$

and say “the limit of $f(x)$, as x approaches a , equals to L ”

if the values of $f(x)$ are arbitrarily close to L (as close to L as we like) by taking x to be sufficiently close to a (on both sides of a) but not equal to a .

In other words, the values of $f(x)$ become closer to the number L as x approaches the number a (from either side of a) but not equal to a .

Notice the phrase “but $x \neq a$ ” in the definition of limit. In finding the limit of $f(x)$ as x approaches a , we never consider $x = a$. Furthermore, $f(x)$ need not even be defined when $x = a$ and in fact the main thing is how f defined near a .

Figure below shows the graphs of three functions. Note that in part (c), $f(a)$ is not defined and in part (b), $f(a) \neq L$. But in each case, regardless of what happens at a , $\lim_{x \rightarrow a} f(x) = L$.

Definition 6.3. We write

$$\lim_{x \rightarrow a^-} f(x) = L$$

and say the left-hand limit of $f(x)$ as x approaches a [or the limit of $f(x)$ as x approaches a from the left] is equal to L if we can make the values of $f(x)$ as close to L as we like by taking x to be sufficiently close to a and x less than a .

Notice that Definition 2 differs from Definition 1 only in that we require x to be less than a . Similarly, if we require that x be greater than a , we get “the right-handed limit of $f(x)$ as x approaches a is equal to L ” and we write

$$\lim_{x \rightarrow a^+} f(x) = L$$

Thus, the symbol “ $x \rightarrow a^+$ ” means that we consider only $x > a$. These definitions are illustrated in the figure below.

By comparing the Definition 1 with the definitions of one-sided limits, we see that the following is true.

$$\lim_{x \rightarrow a} f(x) = L \quad \text{if and only if} \quad \lim_{x \rightarrow a^-} f(x) = L \quad \text{and} \quad \lim_{x \rightarrow a^+} f(x) = L$$

Example 6.1. The graph of a function f is shown in the following figure.

Use it to state the values (if they exist) of the following:

(a) $\lim_{x \rightarrow 1^-} f(x)$

(b) $\lim_{x \rightarrow 1^+} f(x)$

(c) $\lim_{x \rightarrow 1} f(x)$

(d) $\lim_{x \rightarrow 6^-} f(x)$

(e) $\lim_{x \rightarrow 6^+} f(x)$

(f) $\lim_{x \rightarrow 6} f(x)$

Solution. (a) $\lim_{x \rightarrow 1^-} f(x) = 4$ and (b) $\lim_{x \rightarrow 1^+} f(x) = 3$
 (c) $\lim_{x \rightarrow 1} f(x)$ does not exist since the left and right limits are different.
 (d) $\lim_{x \rightarrow 6^-} f(x) = 5$
 (e) $\lim_{x \rightarrow 6^+} f(x) = 5$
 (f) Since the left and right limits are the same, we have $\lim_{x \rightarrow 6} f(x) = 5$. □

Example 6.2. $\lim_{x \rightarrow 3} (4 - x^2) = -5$.

As x approaches 3, x^2 approaches 9, and hence $4 - x^2$ approaches -5 .

Example 6.3. Find $\lim_{x \rightarrow 1} \frac{x^2 - x - 2}{x - 1}$ if it exists.

Solution. Let $f(x) = \frac{x^2 - x - 2}{x - 1}$. $f(2)$ is not defined since both numerator and denominator are 0. For all $x \neq 1$,

$$f(x) = \frac{x^2 - x - 2}{x - 1} = x + 2.$$

Hence, $\lim_{x \rightarrow 1} \frac{x^2 - x - 2}{x - 1} = \lim_{x \rightarrow 1} (x + 2) = 3$. □

Example 6.4. Find $\lim_{x \rightarrow -1} f(x)$ if $f(x) = \begin{cases} 10, & x = -1; \\ 2x + 3, & x \neq -1. \end{cases}$

Solution. When x approaches -1 , $f(x) = 2x + 3$ and hence

$$\lim_{x \rightarrow -1} f(x) = \lim_{x \rightarrow -1} (2x + 3) = 1. \quad \square$$

x2 **Example 6.5.** Find $\lim_{x \rightarrow 0} \frac{1}{x^2}$ if it exists.

Solution. As x approaches 0, x^2 also approaches to 0, and $\frac{1}{x^2}$ becomes very large. From the graph of the function $f(x) = \frac{1}{x^2}$, we observe that the values of $f(x)$ can be made arbitrarily large by taking x close enough to 0. Therefore the values of $f(x)$ do not approach a number, so $\lim_{x \rightarrow 0} \frac{1}{x^2}$ does not exist. We can use the notation

$$\lim_{x \rightarrow 0} \frac{1}{x^2} = \infty. \quad \square$$

It doesn't mean that we are regarding ∞ as a number nor the limit exists. It simply expresses the particular way in which the limit does not exist:

$\frac{1}{x^2}$ can be made as large as we like by taking x close enough to 0.

Definition 6.4. Let f be a function defined on both sides of a , except possibly at a itself. Then

$$\lim_{x \rightarrow a} f(x) = \infty$$

means that the values of $f(x)$ can be made arbitrarily large (as large as we please) by taking x sufficiently close to a but not equal to a .

Definition 6.5. Let f be a function defined on both sides of a , except possibly at a itself. Then

$$\lim_{x \rightarrow a} f(x) = -\infty$$

means that the values of $f(x)$ can be made arbitrarily large negative by taking x sufficiently close to a , but not equal to a .

Similar definitions can be given for the one-sided infinite limits

$$\begin{array}{ll} \lim_{x \rightarrow a^-} f(x) = \infty & \lim_{x \rightarrow a^+} f(x) = \infty \\ \lim_{x \rightarrow a^-} f(x) = -\infty & \lim_{x \rightarrow a^+} f(x) = -\infty \end{array}$$

Example 6.6. Find $\lim_{x \rightarrow 2^-} \frac{x}{2-x}$ and $\lim_{x \rightarrow 2^+} \frac{x}{2-x}$.

Solution. If x is close to 2 but less than 2, then the denominator $2-x$ is a small positive number. So the quotient $\frac{x}{2-x}$ is a large positive number. Hence

$$\lim_{x \rightarrow 2^-} \frac{x}{2-x} = \infty$$

and similarly we have

$$\lim_{x \rightarrow 2^+} \frac{x}{2-x} = -\infty \quad \square$$

$\square$

6.1.2 Some limit laws

In this subsection we use the following properties of limits, called the limit laws, to calculate limits.

Theorem 6.7. (Limit Laws) Suppose that c is constant and the limits

$$\lim_{x \rightarrow a} f(x) \quad \text{and} \quad \lim_{x \rightarrow a} g(x)$$

exist. Then

$$(i) \text{ (Sum law) } \lim_{x \rightarrow a} [f(x) + g(x)] = \lim_{x \rightarrow a} f(x) + \lim_{x \rightarrow a} g(x)$$

$$(ii) \text{ (Difference law) } \lim_{x \rightarrow a} [f(x) - g(x)] = \lim_{x \rightarrow a} f(x) - \lim_{x \rightarrow a} g(x)$$

$$(iii) \text{ (Constant multiple law) } \lim_{x \rightarrow a} [cf(x)] = c \lim_{x \rightarrow a} f(x)$$

$$(iv) \text{ (Product law) } \lim_{x \rightarrow a} [f(x)g(x)] = \lim_{x \rightarrow a} f(x) \cdot \lim_{x \rightarrow a} g(x)$$

$$(v) \text{ (Quotient law) } \lim_{x \rightarrow a} \frac{f(x)}{g(x)} = \frac{\lim_{x \rightarrow a} f(x)}{\lim_{x \rightarrow a} g(x)} \text{ if } \lim_{x \rightarrow a} g(x) \neq 0$$

Theorem 6.8. Suppose that c is constant and the limits $\lim_{x \rightarrow a} f(x)$ exist. Then

$$(vi) \text{ (Power law) } \lim_{x \rightarrow a} [f(x)]^n = [\lim_{x \rightarrow a} f(x)]^n$$

$$(vii) \lim_{x \rightarrow a} c = c$$

$$(viii) \lim_{x \rightarrow a} x = a$$

$$(ix) \lim_{x \rightarrow a} x^n = \lim_{x \rightarrow a} a^n$$

$$(x) \lim_{x \rightarrow a} \sqrt[n]{x} = \sqrt[n]{a} \text{ where } n \text{ is a positive integer. (If } n \text{ is even, we assume that } a > 0.)$$

$$(xi) \text{ (Root law) } \lim_{x \rightarrow a} \sqrt[n]{f(x)} = \sqrt[n]{\lim_{x \rightarrow a} f(x)} \text{ where } n \text{ is a positive integer. (If } n \text{ is even, we assume that } \lim_{x \rightarrow a} f(x) > 0.)$$

direct1

Example 6.9. Evaluate the following limits:

$$(a) \lim_{x \rightarrow 1} (5 - 3x - x^2);$$

$$(b) \lim_{x \rightarrow 2} \sqrt{x^2 + 2x + 8}.$$

Solution. (a)

$$\begin{aligned} \lim_{x \rightarrow 1} (5 - 3x - x^2) &= \lim_{x \rightarrow 1} 5 - \lim_{x \rightarrow 1} (3x) - \lim_{x \rightarrow 1} x^2 \\ &= \lim_{x \rightarrow 1} 5 - 3 \lim_{x \rightarrow 1} x - \lim_{x \rightarrow 1} x^2 \\ &= 5 - 3(1) - 1 = 1 \end{aligned}$$

(b)

$$\begin{aligned} \lim_{x \rightarrow 2} \sqrt{x^2 + 2x + 8} &= \sqrt{\lim_{x \rightarrow 2} (x^2 + 2x + 8)} \\ &= \sqrt{\lim_{x \rightarrow 2} x^2 + \lim_{x \rightarrow 2} (2x) + \lim_{x \rightarrow 2} 8} \\ &= \sqrt{2^2 + 2(2) + 8} = 4 \end{aligned}$$

□

Theorem 6.10. If $\lim_{x \rightarrow a} f(x) = l$ with $l \neq 0$ and $\lim_{x \rightarrow a} g(x) = 0$, then $\lim_{x \rightarrow a} \frac{f(x)}{g(x)}$ does not exist.

direct2

Example 6.11. Evaluate the following limits:

(a) $\lim_{x \rightarrow 3} \frac{x^2 + 1}{x - 3};$

(b) $\lim_{x \rightarrow 1} \frac{x - 1}{x^3 - 3x + 2}.$

Solution. (a) $\lim_{x \rightarrow 3} \frac{x^2 + 1}{x - 3}$ does not exist since $\lim_{x \rightarrow 3} (x^2 + 1) = 10$ but $\lim_{x \rightarrow 3} (x - 3) = 0$.

(b) For all $x \neq 1$,

$$\frac{x - 1}{x^3 - 3x + 2} = \frac{x - 1}{(x - 1)^2(x + 2)} = \frac{1}{(x - 1)(x + 2)}.$$

Thus, $\lim_{x \rightarrow 1} \frac{x - 1}{x^3 - 3x + 2} = \lim_{x \rightarrow 1} \frac{1}{(x - 1)(x + 2)}$ does not exist since $\lim_{x \rightarrow 1} (x - 1)(x + 2) = 0$. □

Theorem 6.12. *If the limits of f and g both exist as x approaches a , and $f(x) \leq g(x)$ when x approaches a (except possibly at a), then*

$$\lim_{x \rightarrow a} f(x) \leq \lim_{x \rightarrow a} g(x).$$

Theorem 6.13 (The Squeeze Theorem). *If $f(x) \leq g(x) \leq h(x)$ when x approaches a (except possibly at a) and*

$$\lim_{x \rightarrow a} f(x) = \lim_{x \rightarrow a} h(x) = L$$

then

$$\lim_{x \rightarrow a} g(x) = L.$$

The Squeeze Theorem also called the Sandwich Theorem or the Pinching Theorem. It says that if $g(x)$ is squeezed between $f(x)$ and $h(x)$ as x approaches a , and if f and h have the same limit L at a , then g is forced to have the same limit L at a .

Example 6.14. Show that $\lim_{x \rightarrow 0} x^2 \cos \frac{1}{x^2} = 0$.

Solution. Since

$$-1 \leq \cos \frac{1}{x^2} \leq 1$$

we have

$$-x^2 \leq x^2 \cos \frac{1}{x^2} \leq x^2$$

We observe that

$$\lim_{x \rightarrow 0} x^2 = 0 \quad \text{and} \quad \lim_{x \rightarrow 0} -x^2 = 0$$

Let $f(x) = -x^2$, $g(x) = x^2 \cos \frac{1}{x^2}$, and $h(x) = x^2$ in the Squeeze Theorem, we have

$$\lim_{x \rightarrow 0} x^2 \cos \frac{1}{x^2} = 0. \quad \square$$

6.2 Continuity

- Continuous functions
 - Discontinuity criterion
 - Preservation of compactness

- Intermediate Value Theorem
- Uniform continuity

cont

Definition 6.6. A function f is said to be continuous at a number a if

$$\lim_{x \rightarrow a} f(x) = f(a).$$

If f is not continuous at a , then f is discontinuous at a , or f has discontinuity at a . Note that Definition 6.6 requires three things if f is continuous at a :

1. $f(a)$ is defined;
2. $\lim_{x \rightarrow a} f(x)$ exists;
3. $\lim_{x \rightarrow a} f(x) = f(a)$.

Geometrically, a function is said to be continuous at every number in an interval if there is has no break in its graph or the graph can be drawn without removing our pen from the paper.

Example 6.15. Determine the discontinuous of the following functions:

(a) $f(x) = \frac{x^2 - 1}{x + 1}$

(b) $f(x) = \begin{cases} \frac{3}{\cos x}, & \text{if } x \neq 0; \\ 2, & \text{if } x = 0. \end{cases}$

(c) $f(x) = \begin{cases} \frac{x^2}{x^2 + 2}, & \text{if } x \leq 2; \\ \sqrt{x + 2}, & \text{if } x > 2. \end{cases}$

Solution. (a) Note that $f(-1)$ is not defined, so f is discontinuous at $x = -1$.

(b) $f(0) = 2$ is defined but

$$\lim_{x \rightarrow 0} f(x) = \lim_{x \rightarrow 0} \frac{3}{\cos x} = 3.$$

So f is discontinuous at -1 .

(c) Here $f(2) = \frac{2^2}{2^2 + 2} = \frac{2}{3}$ is defined and

$$\lim_{x \rightarrow 2^-} f(x) = \lim_{x \rightarrow 2^-} \frac{x^2}{x^2 + 2} = \frac{2}{3}$$

and

$$\lim_{x \rightarrow 2^+} f(x) = \lim_{x \rightarrow 2^+} \sqrt{x + 2} = 2$$

But

$$\lim_{x \rightarrow 2^-} f(x) \neq \lim_{x \rightarrow 2^+} f(x)$$

so f is not continuous at 2. □

Definition 6.7. A function f is called continuous from the right at a number a if

$$\lim_{x \rightarrow a^+} f(x) = f(a)$$

and f is called continuous from the left at a if

$$\lim_{x \rightarrow a^-} f(x) = f(a).$$

=====

Example 6.16. At each integer n , the function $f(x) = [x]$ is continuous from the right but discontinuous from the left because

$$\lim_{x \rightarrow n^+} f(x) = \lim_{x \rightarrow n^+} [x] = n = f(n)$$

but

$$\lim_{x \rightarrow n^-} f(x) = \lim_{x \rightarrow n^-} [x] = n - 1 \neq f(n)$$

Definition 6.8. A function f is continuous on an interval if it is continuous at every number in the interval. (If f is defined only on one side of an endpoint of the interval, we understand continuous at the endpoint to mean continuous from the right or continuous from the left.)

Example 6.17. Show that the function $f(x) = 1 - \sqrt{1 - x^2}$ is continuous on the interval $[-1, 1]$.

Solution. If $-1 < a < 1$, then using the Limit Laws, we have

$$\begin{aligned} \lim_{x \rightarrow a} f(x) &= \lim_{x \rightarrow a} (1 - \sqrt{1 - x^2}) \\ &= 1 - \lim_{x \rightarrow a} \sqrt{1 - x^2} \\ &= 1 - \sqrt{\lim_{x \rightarrow a} (1 - x^2)} \\ &= 1 - \sqrt{1 - a^2} \\ &= f(a) \end{aligned}$$

Thus, by Definition [6.6](#), f is continuous at a if $-1 < a < 1$. Similar calculations show that

$$\lim_{x \rightarrow -1^+} f(x) = 1 = f(-1) \quad \text{and} \quad \lim_{x \rightarrow 1^-} f(x) = 1 = f(1)$$

so f is continuous from the right at -1 and continuous from the left at 1 . Therefore, f is continuous on $[-1, 1]$. □

Theorem 6.18. If f and g are continuous at a and c is a constant, then the following functions are also continuous at a :

1. $f + g$
2. $f - g$
3. cf
4. fg
5. $\frac{f}{g}$ if $g(a) \neq 0$

Each of the five parts of this theorem follows from the corresponding Limit Law in Section 1.2. For instance, we give the proof of part 1. Since f and g are continuous at a , we have

$$\lim_{x \rightarrow a} f(x) = f(a) \quad \text{and} \quad \lim_{x \rightarrow a} g(x) = g(a)$$

Therefore

$$\begin{aligned} \lim_{x \rightarrow a} (f + g)(x) &= \lim_{x \rightarrow a} [f(x) + g(x)] \\ &= \lim_{x \rightarrow a} f(x) + \lim_{x \rightarrow a} g(x) \\ &= f(a) + g(a) \\ &= (f + g)(a) \end{aligned}$$

This shows that $f + g$ is continuous at a . $\square$

Theorem 6.19. (a) Any polynomial is continuous everywhere; that is, it is continuous on $\mathbb{R}$.

(b) Any rational function is continuous wherever it is defined; that is, it is continuous on its domain.

Proof. (a) A polynomial is a function of the form

$$P(x) = c_n x^n + c_{n-1} x^{n-1} + \cdots + c_1 x + c_0$$

where $c_0, c_1, \dots, c_n$ are constants. We know that

$$\lim_{x \rightarrow a} c_0 = c_0$$

and

$$\lim_{x \rightarrow a} x^m = a^m \quad m = 1, 2, \dots, n$$

This equation is precisely the statement that the function $f(x) = x^m$ is a continuous function. Thus, the function $g(x) = cx^m$ is continuous. Since P is a sum of functions of this form and a constant function, it follows that P is continuous.

(b) A rational function is a function of the form

$$f(x) = \frac{P(x)}{Q(x)}$$

where P and Q are polynomials. The domain of f is $D = \{x \in \mathbb{R} | Q(x) \neq 0\}$. We know from part (a) that P and Q are continuous everywhere. Thus, f is continuous at every number in D . $\square$

Example 6.20. Find $\lim_{x \rightarrow -2} \frac{x^3 + 2x^2 - 1}{5 - 3x}$.

Solution. The function

$$f(x) = \frac{x^3 + 2x^2 - 1}{5 - 3x}$$

is rational, so it is continuous on its domain, which is $\{x | x \neq \frac{5}{3}\}$. Therefore

$$\begin{aligned} \lim_{x \rightarrow -2} \frac{x^3 + 2x^2 - 1}{5 - 3x} &= \lim_{x \rightarrow -2} f(x) = f(-2) \\ &= \frac{(-2)^3 + 2(-2)^2 - 1}{5 - 3(-2)} \\ &= -\frac{1}{11} \quad \square \end{aligned}$$

$\square$

Theorem 6.21. Polynomial functions, sine, cosine, exponentiation, hyperbolic sine, hyperbolic cosine are continuous on $\mathbb{R}$.

Rational functions, root functions, tangent, cotangent, secant, cosecant, inverse trigonometric functions, logarithmic functions are continuous on their real domain.

Example 6.22. Where is the function $f(x) = \frac{\ln x + \tan^{-1} x}{x^2 - 1}$ continuous?

Solution. We know from theorem that the function $y = \ln x$ is continuous for $x > 0$ and $y = \tan^{-1} x$ is continuous on $\mathbb{R}$. Thus, $y = \ln x + \tan^{-1} x$ is continuous on $(0, \infty)$. The denominator, $y = x^2 - 1$, is a polynomial, so it is continuous everywhere. Therefore, $\frac{1}{x^2 - 1}$ is continuous at all positive number x except where $x^2 - 1 = 0$. So f is continuous on the intervals $(0, 1)$ and $(1, \infty)$. □

Example 6.23. Evaluate $\lim_{x \rightarrow \pi} \frac{\sin x}{2 + \cos x}$.

Solution. The function $y = \sin x$ is continuous. The denominator, $y = 2 + \cos x$, is the sum of two continuous functions and is therefore continuous. Notice that this function is never 0 because $\cos x \geq -1$ for all x and so $2 + \cos x > 0$ everywhere. Thus, the ratio

$$f(x) = \frac{\sin x}{2 + \cos x}$$

is continuous everywhere. Hence, by definition of a continuous function,

$$\lim_{x \rightarrow \pi} \frac{\sin x}{2 + \cos x} = \lim_{x \rightarrow \pi} f(x) = f(\pi) = \frac{\sin \pi}{2 + \cos \pi} = \frac{0}{2 - 1} = 0 \quad \square$$

Another way of combining continuous functions f and g to get a new continuous function is to form the composite function $f \circ g$. This fact is a consequence of the following theorem.

Theorem 6.24. If f is continuous at b and $\lim_{x \rightarrow a} g(x) = b$, then $\lim_{x \rightarrow a} f(g(x)) = f(b)$. In other words,

$$\lim_{x \rightarrow a} f(g(x)) = f(\lim_{x \rightarrow a} g(x))$$

Chapter 7

Sequences and Series of Functions

- Pointwise convergence of sequences of functions
 - Cauchy condition for uniform convergence
 - WeierstrassM-Test for uniform convergence of series of functions
 - Term-by-term differentiation and integration
- Power series
 - Power series : Radius and interval of convergence
 - Algebraic operations on power series
 - Differentiation and integration of power series
 - Binomial Theorem and Binomial Series

Chapter 8

Measures and Integration

Lebesgue Integrals on $\mathbb{R}$ · Integrable functions · Properties of integrals · Numerical computation of integrals

UECM2713 Fundamentals of Financial Mathematics

1. Shreve, S. 2004. Stochastic calculus for finance I. New York: Springer. 2. Steven, S. 2004. Stochastic calculus for finance II. New York: Springer. 'Additional References: 3. Rosenthal, J. S. 2006. A first look at rigorous probability theory. 2nd ed. Singapore: World Scientific. 4. Chung, K. and AitSahlia, F. 2006. Elementary probability theory: With stochastic processes and an Introduction to Mathematical Finance. 4th ed. New York: Springer.

https://www.youtube.com/watch?v=_yiW6XC6rN4: A Nonintegrable Derivative

Very interesting used of integral to estimate limit!!! <https://www.youtube.com/watch?v=Eu40jV036yE>

Euler's limit:

$$\lim_{n \rightarrow \infty} \left(\frac{n!}{n^n} \right)^{1/n} = \frac{1}{e}$$

<https://mathoverflow.net/questions/6711/integrability-of-derivatives> <https://math.stackexchange.com/que>
for-what-makes-the-volterra-function-non-riemann-integrable https://en.wikipedia.org/wiki/Volterra%27s_function

math1a2024-lecture33.pdf

Integration is the process of “measuring” area of region bounded by two curves. There are two types of integral, indefinite integral (or antiderivative) and definite integral. These two seemingly unrelated notions are linked by the Fundamental Theorem of Calculus. Therefore, you can see that many rules in differentiation have similar counterparts in integration.

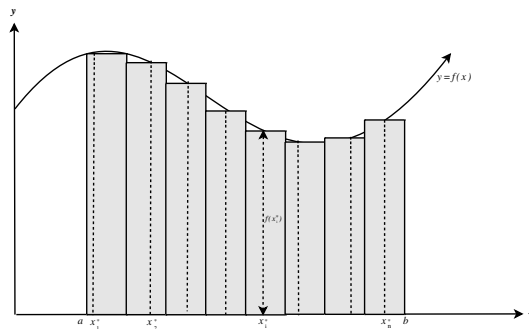
In this chapter, the rules and basic formulae in integration will first be introduced, followed by the study of various techniques of integration, e.g. integration by partial fraction which handles rational functions that arise naturally from the analysis of linear systems in engineering. Finally, definite integral and its applications in the representation of physical quantities such as volume, energy, work, etc. are discussed. This chapter is confined to the discussion of integrals that can be expressed in closed form, the numerical evaluation of integral will be discussed in Chapter ??.

8.1 Riemann Sum and Definite Integral

In the following diagram, suppose we want to find the area of the region under the curve $y = f(x)$ on the interval $[a, b]$. We may first partition the interval into n subintervals of lengths equal to $\Delta x = \frac{b-a}{n}$.¹ The area of the region under the curve can be approximated by the sum of areas of the rectangles. Let x_i^* be any value in the interval $x_{i-1} \leq x_i^* \leq x_i$, the sum of areas of the rectangles may be expressed as a *Riemann sum*

$$\sum_{i=1}^n f(x_i^*) \Delta x.$$

¹In general theory, the partition need not be equal.



If the partition of the interval is taken to be infinitesimal small, then n approaches ∞ . The area of the region under the curve $y = f(x)$ on the interval $[a, b]$ will be denoted by a *definite integral*

$$\int_a^b f(x)dx = \lim_{n \rightarrow \infty} \sum_{i=1}^n f(x_i^*)\Delta x.$$

Example 8.1. What is represented by

$$\int_{-2}^2 \sqrt{4 - x^2} dx?$$

Solution. The definite integral represent the area of the region under the semicircle $y = \sqrt{4 - x^2}$ and above the x -axis. Therefore

$$\int_{-2}^2 \sqrt{4 - x^2} dx = 2\pi$$

□

Interpreting the definite integrals as the limit of Riemann sum or area, we can see the following properties easily:

- (a) $\int_a^b f(x)dx = -\int_b^a f(x)dx$;
- (b) $\int_a^a f(x)dx = 0$;
- (c) $\int_a^b c dx = c(b - a)$;
- (d) $\int_a^b [f(x) + g(x)]dx = \int_a^b f(x)dx + \int_a^b g(x)dx$;
- (e) $\int_a^b cf(x)dx = c \int_a^b f(x)dx$.
- (f) If $f(x) \geq g(x)$ on the interval $[a, b]$ then $\int_a^b f(x)dx \geq \int_a^b g(x)dx$;
- (g) $\int_a^c f(x)dx = \int_a^b f(x)dx + \int_b^c f(x)dx$.

The rules above provide us the way to evaluate only the simplest type of definite integral. As you may expect, evaluating definite integral using Riemann sum though fundamental, is not practical. We must find a way to evaluate the definite integral. Fortunately, the following theorem allow us to evaluate it avoiding the Riemann sum and limit.

Chapter 9

Differentiation

Differentiation provides an approach to obtaining certain integrals in closed forms, i.e. the integrals can be expressed as a finite combination (with addition, multiplication, roots, etc.) of elementary or special functions or numbers associated with them.

Volterra's function V is differentiable everywhere and its derivative V' is bounded everywhere but its derivative is not Riemann-integrable!!! $\Rightarrow$ Show the significance of Fundamental Theorem of Calculus

<https://web.archive.org/web/20201123131325/http://www.macalester.edu/~bressoud/talks/AlleghenyCollege/Wrestling.pdf>

<https://en.namu.wiki/w/%EB%B3%BC%ED%85%8C%EB%9D%BC%20%ED%95%A8%EC%88%98>

More references to read:

JohnDePree-CharlesSwartz-IntroductionToRealAnalysis-Wiley1988

NicolasBourbaki-FunctionsOfARealVariable-2003

LangS-RealAnalysis-AW1983

Rudin, W., 1976. Principles of mathematical analysis. 3rd ed. New York: McGraw-Hill.

9.1 Derivative and Differentiation Rules

· Chain rule

9.2 Derivative of Some Analytic Functions

Using the derivative formulae of e^x and e^{-x} , it is easy to get the following:

$$\frac{d}{dx} \sinh x = \cosh x,$$

$$\frac{d}{dx} \tanh x = \operatorname{sech}^2 x,$$

$$\frac{d}{dx} \operatorname{sech} x = -\operatorname{sech} x \tanh x,$$

$$\frac{d}{dx} \cosh x = \sinh x$$

$$\frac{d}{dx} \coth x = -\operatorname{csch}^2 x$$

$$\frac{d}{dx} \operatorname{csch} x = -\operatorname{csch} x \coth x$$

9.3 Mean Value Theorem?

· L'Hospital's rule

· Taylor's theorem

Prove that

$$L_n(x) = e^x \frac{d^n}{dx^n} (x^n e^{-x})$$

is a polynomial and all its roots are positive.

Use mathematical induction ???

Calculate $\arctan^{(n)}(0)$ and $\arcsin^{(n)}(0)$

Use their Taylor series at 0???

Example 9.1. V. I. Arnold, Hyugens and Barrow, Newton & Hooke, Birkhauser 1990 page 28:
Find

$$\lim_{x \rightarrow 0} \frac{\sin(\tan(x)) - \tan(\sin(x))}{\arcsin(\arctan(x)) - \arctan(\arcsin(x))}$$

Note: L'hospital rule will work terribly on this example.

Hint: Calculate the mean value of $1/r$ over the sphere $(x-a)^2 + (y-b)^2 + (z-c)^2 = R^2$ and the mean value of $\log(1/r)$ over the circle $(x-a)^2 + (y-b)^2 = R^2$.

Prove that e and π are irrational!!!

<https://pawel-czyz.github.io/posts/irrational-pi-and-e.html>

Show that e and π are transcendental ???

https://en.wikipedia.org/wiki/Lindemann%E2%80%93Weierstrass_theorem

https://en.wikipedia.org/wiki/Gelfond%E2%80%93Schneider_theorem

9.4 Antiderivative and Methods of Integration

Not every antiderivative can be obtained by using the integral table directly. There are some integrals that need special techniques to evaluate, there are also integrals that cannot be expressed using elementary functions. We are going to introduce antiderivative and three methods for obtaining the antiderivative form for the integration of “almost” analytic functions in this section.

Definition 9.1. Let $f : A \rightarrow \mathbb{R}$ be a function, if there exists a function g such that every $x \in A$ satisfies $g'(x) = f(x)$, then g is an *antiderivative* of f .

Example 9.2.

$$\frac{d}{dx} \sin x = \cos x \text{ is true for all } x \in \mathbb{R}.$$

Therefore $\sin x$ is an antiderivative of $\cos x$. However, you may notice that for any constant c , $\frac{d}{dx}(\sin x + c)$ is also equal to $\cos x$, we may say that $\sin x + c$ is an antiderivative of $\cos x$ for arbitrary constant c .

From the example above, we observe that a function can have infinitely many antiderivatives.

Definition 9.2. If the function $f : A \rightarrow \mathbb{R}$ has antiderivatives, we call the set of all its antiderivatives the *indefinite integral* of f , denoted by

$$\int f(x) dx.$$

$f(x)$ is called the *integrand*.

If $g : A \rightarrow \mathbb{R}$ is an antiderivative of f , then we usually write the indefinite integral of f as

$$\int f(x) dx = g(x) + C$$

where C is called the constant of integration.

Example 9.3.

$$\frac{d}{dx} \sec x = \sec x \tan x \Leftrightarrow \int \sec x \tan x dx = \sec x + C$$

As a consequence of the linearity of differentiation rules, we have the following results:

(a)

$$\int [f_1(x) + f_2(x)]dx = \int f_1(x)dx + \int f_2(x)dx.$$

(b) For any constant $c \neq 0$,

$$\int cf(x)dx = c \int f(x)dx.$$

We also have the following formulas which are corresponding to the formulas of derivatives:

$\frac{d}{dx}c = 0,$	$\int 0dx = c;$
$\frac{d}{dx}x^n = nx^{n-1},$	$\int x^n dx = \frac{x^{n+1}}{n+1} + C, n \neq -1;$
$\frac{d}{dx}\sin ax = a \cos ax,$	$\int \cos ax dx = \frac{1}{a} \sin ax + C;$
$\frac{d}{dx}\cos ax = -a \sin ax,$	$\int \sin ax dx = -\frac{1}{a} \cos ax + C;$
$\frac{d}{dx}\tan ax = a \sec^2 ax,$	$\int \sec^2 ax dx = \frac{1}{a} \tan ax + C;$
$\frac{d}{dx}\cot ax = -a \csc^2 ax,$	$\int \csc^2 ax dx = -\frac{1}{a} \cot ax + C;$
$\frac{d}{dx}\sec ax = a \sec ax \tan ax,$	$\int \sec ax \tan ax dx = \frac{1}{a} \sec ax + C;$
$\frac{d}{dx}\csc ax = -a \csc ax \cot ax,$	$\int \csc ax \cot ax dx = -\frac{1}{a} \csc ax + C;$
$\frac{d}{dx}\ln x = \frac{1}{x},$	$\int \frac{1}{x} dx = \ln x + C;$
$\frac{d}{dx}a^x = a^x \ln a,$	$\int a^x dx = \frac{1}{\ln a} a^x + C, a > 0, a \neq 1;$
$\frac{d}{dx}e^{ax} = ae^{ax},$	$\int e^{ax} dx = \frac{1}{a} e^{ax} + C.$

In the process of integration, we may need to apply some algebraic or trigonometric transformations to simplify the integrand.

Example 9.4. Evaluate $\int 5x^2 - x^{-1} + 2dx$.

Solution.

$$\begin{aligned}\int 5x^2 - x^{-1} + 2dx &= 5 \int x^2 dx - \int x^{-1} dx + 2 \int dx \\ &= \frac{5}{3}x^3 - \ln |x| + 2x + C\end{aligned}$$

□

Example 9.5. Evaluate $\int \sin^2 x dx$.

Solution. Since $\sin^2 x = \frac{1}{2}(1 - \cos 2x)$, we have

$$\begin{aligned}\int \sin^2 x dx &= \int \frac{1}{2}(1 - \cos 2x) dx \\ &= \frac{x}{2} - \frac{1}{4} \sin 2x + C\end{aligned}$$

□

Example 9.6. Evaluate $\int \sin 2x \cos 3x dx$.

Solution. Using the product formula, we have $\sin 2x \cos 3x = \frac{1}{2}(\sin 5x - \sin x)$.

$$\begin{aligned} \int \sin 2x \cos 3x dx &= \int \frac{1}{2}(\sin 5x - \sin x) dx \\ &= \frac{1}{2} \int \sin 5x dx - \frac{1}{2} \int \sin x dx \\ &= -\frac{1}{10} \cos 5x + \frac{1}{2} \cos x + C \end{aligned}$$

□

9.4.1 Integration by Parts

Theorem 9.7. Let $f, g : A \rightarrow \mathbb{R}$ be differentiable functions. If $f'g$ has antiderivative, then fg' also has antiderivative, and their relationship is given by

$$\int f(x)g'(x)dx = f(x)g(x) - \int f'(x)g(x)dx.$$

Example 9.8. Evaluate $\int xe^x dx$.

Solution. Take $f(x) = x, g'(x) = e^x$, we have $f'(x) = 1, g(x) = e^x$. Using integration by parts formula we have

$$\begin{aligned} \int xe^x dx &= xe^x - \int e^x dx \\ &= xe^x - e^x + C \end{aligned}$$

□

Example 9.9. Evaluate $\int x \ln x dx$.

Solution. Take $f(x) = \ln x, g'(x) = x$, we have $f'(x) = \frac{1}{x}, g(x) = \frac{1}{2}x^2$. Using integration by parts formula we have

$$\begin{aligned} \int x \ln x dx &= \frac{1}{2}x^2 \ln x - \frac{1}{2} \int x dx \\ &= \frac{1}{2}x^2 \ln x - \frac{1}{4}x^2 + C \end{aligned}$$

□

Sometimes special techniques are needed in finding the integrals as shown in the following examples.

Example 9.10. Evaluate $\int e^x \sin 2x dx$.

Solution. Take $f(x) = e^x, g'(x) = \sin 2x$, we have $f'(x) = e^x, g(x) = -\frac{1}{2} \cos 2x$. Using integration by parts formula we have

$$\int e^x \sin 2x dx = -\frac{1}{2}e^x \cos 2x + \frac{1}{2} \int e^x \cos 2x dx.$$

Using integration by parts formula again to the integral on the right hand side. Taking

$f_1(x) = e^x, g_1'(x) = \cos 2x$, we have $f_1'(x) = e^x, g_1(x) = \frac{1}{2} \sin 2x$.

$$\begin{aligned} \int e^x \sin 2x dx &= -\frac{1}{2}e^x \cos 2x + \frac{1}{2}\left(\frac{1}{2}e^x \sin 2x - \frac{1}{2} \int e^x \sin 2x dx\right) \\ &= -\frac{1}{2}e^x \cos 2x + \frac{1}{4}e^x \sin 2x - \frac{1}{4} \int e^x \sin 2x dx \end{aligned}$$

We may now treat the result as an equation in terms of $\int e^x \sin 2x dx$ and solve it to get

$$\int e^x \sin 2x dx = \frac{1}{5}e^x(\sin 2x - 2 \cos 2x) + C.$$

□

Example 9.11. Evaluate $\int \tan^4 x dx$.

Solution. Making use of the trigonometric identity $\tan^2 x = \sec^2 x - 1$, we have

$$\begin{aligned} \int \tan^4 x dx &= \int \tan^2 x \sec^2 x dx - \int \tan^2 x dx \\ &= \int \tan^2 x \sec^2 x dx - (\tan x - x) \end{aligned}$$

Using integration by parts we have

$$\int \tan^2 x \sec^2 x dx = \tan^3 x - 2 \int \tan^2 x \sec^2 x dx.$$

Hence

$$\int \tan^4 x dx = \frac{1}{3} \tan^3 x - \tan x + x + C.$$

Using similar method as in the previous example, ones can prove that

$$\int \tan^n x dx = \frac{1}{n-1} \tan^{n-1} x - \int \tan^{n-2} x dx.$$

Equations in this form are called *reduction formula*.

□

9.4.2 Integration by Substitution

Theorem 9.12. Let $f : A \rightarrow B$ and $g : B \rightarrow \mathbb{R}$ be two functions. If g has antiderivative h , i.e.

$$\int g(u) du = h(u) + C_1$$

and f is a differentiable function, then the composite function $g(f(x))f'(x)$ also has antiderivative and is given by

$$\int g(f(x))f'(x) dx = h(f(x)) + C.$$

Suppose you want to evaluate $\int k(x)dx$. If we can find two functions f and g such that $k(x) = g(f(x))f'(x)$, where f is a differentiable function, then

$$\begin{aligned}\int k(x)dx &= \int g(f(x))f'(x)dx \\ &= \int g(u)du \\ &= h(u) + C \\ &= h(f(x)) + C\end{aligned}$$

Example 9.13. Evaluate $\int 2x(1+x^2)^3dx$.

Solution. Observe that if we take $u = 1 + x^2$, then $du = 2xdx$. Hence

$$\begin{aligned}\int 2x(1+x^2)^3dx &= \int u^3du \\ &= \frac{1}{4}u^4 + C \\ &= \frac{1}{4}(1+x^2)^4 + C\end{aligned}$$

□

Example 9.14. Evaluate $\int \sec x dx$.

Solution.

$$\begin{aligned}\int \sec x dx &= \int \frac{\sec x(\sec x + \tan x)}{\sec x + \tan x} dx \\ &= \int \frac{(\sec x + \tan x)'}{\sec x + \tan x} dx \\ &= \ln |\sec x + \tan x| + C\end{aligned}$$

□

Example 9.15. Evaluate $\int \sin^2 x \cos^4 x dx$.

Solution. One of the ways to solve the problem is to make use of the trigonometric identities $\sin^2 x \cos^2 x = \frac{1}{4} \sin^2 2x$ and $\cos^2 x = \frac{1}{2}(1 + \cos 2x)$.

$$\begin{aligned}\int \sin^2 x \cos^4 x dx &= \frac{1}{8} \int \sin^2 2x (1 + \cos 2x) dx \\ &= \int \sin^2 2x dx + \frac{1}{8} \int \sin^2 2x \cos 2x dx \\ &= \frac{1}{16} \int (1 - \cos 4x) dx + \frac{1}{16} \int \sin^2 2x d(\sin 2x) \\ &= \frac{1}{16} x - \frac{1}{64} \sin 4x + \frac{1}{48} \sin^3 2x + C\end{aligned}$$

□

Use the chain rule to differentiate both sides of $\sin y = x$ with respect to x , we have

$$\begin{aligned}\frac{d}{dx} \sin y &= \frac{d}{dx} x \\ \cos y \frac{dy}{dx} &= 1 \\ \frac{dy}{dx} &= \frac{1}{\cos y} \\ &= \frac{1}{\sqrt{1 - \sin^2 y}} \\ &= \frac{1}{\sqrt{1 - x^2}}\end{aligned}$$

Similarly, we can get

$$\begin{aligned}\frac{d}{dx} \sin^{-1} x &= \frac{1}{\sqrt{1 - x^2}}, & \frac{d}{dx} \cos^{-1} x &= \frac{-1}{\sqrt{1 - x^2}} \\ \frac{d}{dx} \tan^{-1} x &= \frac{1}{1 + x^2}, & \frac{d}{dx} \cot^{-1} x &= \frac{-1}{1 + x^2} \\ \frac{d}{dx} \sec^{-1} x &= \frac{1}{|x|\sqrt{x^2 - 1}}, & \frac{d}{dx} \csc^{-1} x &= \frac{-1}{|x|\sqrt{x^2 - 1}}\end{aligned}$$

Example 9.16. Evaluate $\int \frac{1}{9 + x^2} dx$.

Solution.

$$\begin{aligned}\frac{1}{9 + x^2} &= \frac{1}{9(1 + \frac{x^2}{9})} \\ \int \frac{1}{9 + x^2} &= \frac{1}{9} \int \frac{1}{1 + \frac{x^2}{9}} dx \\ &= \frac{1}{9} \int \frac{3}{1 + u^2} du \\ &= \frac{1}{3} \tan^{-1} \frac{x}{3} + C\end{aligned}$$

□

Example 9.17. Evaluate $\int \tan^{-1} x dx$.

Solution. Using integration by parts formula,

$$\begin{aligned}\int \tan^{-1} x dx &= x \tan^{-1} x - \int \frac{x}{1 + x^2} dx \\ &= x \tan^{-1} x - \frac{1}{2} \int \frac{1}{u} du \\ &= x \tan^{-1} x - \frac{1}{2} \ln(1 + x^2) + C\end{aligned}$$

□

Example 9.18. Evaluate $\int \sqrt{2x - x^2} dx$.

Solution. By completing square, $2x - x^2 = 1 - (x - 1)^2$ and let $\sin u = x - 1$, $\cos u du = dx$.

In fact, from $\sin u = x - 1$, we also have $\cos u = \sqrt{2x - x^2}$

$$\begin{aligned}
 \int \sqrt{2x - x^2} dx &= \int \sqrt{1 - (x - 1)^2} dx \\
 &= \int \cos^2 u du \\
 &= \frac{1}{2} \int \cos 2u + 1 du \\
 &= \frac{1}{2} \left(\frac{\sin 2u}{2} + u \right) + C \\
 &= \frac{1}{2} \left[(x - 1) \sqrt{2x - x^2} + \sin^{-1}(x - 1) \right] + C
 \end{aligned}$$

□

It is important to note that, the substitution that we use in the last example is not at all arbitrary. From the original integrand, we must have $-1 \leq x - 1 \leq 1$, and it is therefore make sense to let $\sin u = x - 1$.

9.4.3 Hyperbolic Functions

Using the definitions, one can easily get the derivatives of the inverse hyperbolic functions.

$$\begin{aligned}
 \frac{d}{dx} \sinh^{-1} x &= \frac{1}{\sqrt{x^2 + 1}} & \frac{d}{dx} \cosh^{-1} x &= \frac{1}{\sqrt{x^2 - 1}}, x > 1 \\
 \frac{d}{dx} \tanh^{-1} x &= \frac{1}{1 - x^2}, |x| < 1 & \frac{d}{dx} \coth^{-1} x &= \frac{1}{1 - x^2}, |x| > 1 \\
 \frac{d}{dx} \operatorname{sech}^{-1} x &= \frac{-1}{x\sqrt{x^2 + 1}}, 0 < x < 1 & \frac{d}{dx} \operatorname{csch}^{-1} x &= \frac{-1}{|x|\sqrt{x^2 + 1}}, x \neq 0
 \end{aligned}$$

The derivatives above provide us another family of integration formulas.

Example 9.19. Evaluate $\int \sqrt{2x - x^2} dx$.

Solution. By completing square, $2x - x^2 = 1 - (x - 1)^2$ and let $\tanh u = x - 1$, $\operatorname{sech}^2 u du = dx$

we have

$$\begin{aligned}
\int \sqrt{2x - x^2} dx &= \int \sqrt{1 - (x - 1)^2} dx \\
&= \int \operatorname{sech}^3 u du \\
&= \operatorname{sech} u \tanh u + \int \tanh^2 \operatorname{sech} u du \\
&= \operatorname{sech} u \tanh u + \int \operatorname{sech} u du - \int \operatorname{sech}^3 u du \\
&= \frac{1}{2} \left(\operatorname{sech} u \tanh u + \int \operatorname{sech} u du \right) \\
&= \frac{1}{2} \left(\operatorname{sech} u \tanh u + \int \frac{\operatorname{sech}^2 u}{\sqrt{1 - \tanh^2 u}} du \right) \\
&= \frac{1}{2} \left(\operatorname{sech} u \tanh u + \int \frac{d(\tanh u)}{\sqrt{1 - \tanh^2 u}} \right) \\
&= \frac{1}{2} (\operatorname{sech} u \tanh u + \sin^{-1}(\tanh u)) + C \\
&= \frac{1}{2} [(x - 1)\sqrt{2x - x^2} + \sin^{-1}(x - 1)] + C
\end{aligned}$$

□

9.4.4 Integration using Partial Fractions

Example 9.20. Evaluate $\int \frac{1}{x^2 - 1} dx$.

Solution.

$$\begin{aligned}
\frac{1}{x^2 - 1} &= \frac{1}{2} \left(\frac{1}{x + 1} - \frac{1}{x - 1} \right) \\
\int \frac{1}{x^2 - 1} dx &= \frac{1}{2} \int \left(\frac{1}{x + 1} - \frac{1}{x - 1} \right) dx \\
&= \frac{1}{2} \ln \left| \frac{x + 1}{x - 1} \right| + C
\end{aligned}$$

□

The integral above is evaluated by first expressing $\frac{1}{x^2 - 1}$ in partial fractions. This is a useful technique to deal with integral of rational functions. We know that all rational function can be expressed as the sum of a polynomial and a proper fraction using long division. It can be shown that a proper fractions can always be written as the sum of the fractions of the forms

$$\frac{c}{(x - a)^n}; n \in \mathbb{N}; a, c \in \mathbb{R}$$

and

$$\frac{dx + e}{(ax^2 + bx + c)^n}; n \in \mathbb{N}; a \neq 0, b, c, d, e \in \mathbb{R}.$$

We may summarise the rules as follows:

1. When the denominator has the form $(ax + b)^n \rightarrow \frac{A_1}{ax + b} + \frac{A_2}{(ax + b)^2} + \dots + \frac{A_n}{(ax + b)^n}$
2. When the denominator has the form $(ax^2 + bx + c)^n \rightarrow \frac{A_1 x + B_1}{ax^2 + bx + c} + \frac{A_2 x + B_2}{(ax^2 + bx + c)^2} + \dots + \frac{A_n x + B_n}{(ax^2 + bx + c)^n}$

We illustrate this method by a few examples.

Example 9.21. Evaluate $\int \frac{dx}{(x-1)(x-2)(x-3)}$.

Solution. The denominator can be factorised into product of linear factor $(x-1)$, $(x-2)$ and $(x-3)$. We may express $\frac{1}{(x-1)(x-2)(x-3)}$ as

$$\begin{aligned} \frac{1}{(x-1)(x-2)(x-3)} &= \frac{A}{x-1} + \frac{B}{x-2} + \frac{C}{x-3} \\ &= \frac{A(x-2)(x-3) + B(x-1)(x-3) + C(x-1)(x-2)}{(x-1)(x-2)(x-3)} \end{aligned}$$

We have

$$1 \equiv A(x-2)(x-3) + B(x-1)(x-3) + C(x-1)(x-2).$$

By putting $x = 1$, we have $A = \frac{1}{2}$ By putting $x = 2$, we have $B = -1$ By putting $x = 3$, we have $C = \frac{1}{2}$ Therefore,

$$\begin{aligned} \int \frac{1}{(x-1)(x-2)(x-3)} dx &= \int \frac{1}{2(x-1)} - \frac{1}{x-2} + \frac{1}{2(x-3)} dx \\ &= \frac{1}{2} [\ln|x-1| + \ln|x-3|] - \ln|x-2| + C \end{aligned}$$

□

Can the solution above be written as $\ln \left| \frac{\sqrt{(x-1)(x-3)}}{x-2} \right| + C$?

Example 9.22. Evaluate $\int \frac{dx}{x^3-1}$.

Solution. The denominator can be factorised into product of $x-1$ and x^2+x+1 . We may express $\frac{1}{x^3-1}$ as

$$\frac{1}{x^3-1} = \frac{A}{x-1} + \frac{Bx+C}{x^2+x+1}.$$

Comparing coefficients, we get $A = \frac{1}{3}$, $B = \frac{-1}{3}$, $C = \frac{-2}{3}$ Therefore,

$$\begin{aligned} \int \frac{1}{x^3-1} dx &= \int \frac{1}{3(x-1)} + \frac{-x-2}{3(x^2+x+1)} dx \\ &= \int \frac{1}{3(x-1)} - \frac{x+1}{3(x^2+x+1)} - \frac{1}{3(x^2+x+1)} dx \\ &= \int \frac{1}{3(x-1)} - \frac{2x+1}{6(x^2+x+1)} - \frac{1}{2(x^2+x+1)} dx \\ &= \frac{1}{3} \ln|x-1| - \frac{1}{6} \ln|x^2+x+1| - \int \frac{1}{2(x^2+x+1)} dx \\ &= \frac{1}{3} \ln|x-1| - \frac{1}{6} \ln|x^2+x+1| - \int \frac{2}{3((x+\frac{1}{2})^2 + \frac{3}{4})} dx \\ &= \frac{1}{3} \ln|x-1| - \frac{1}{6} \ln|x^2+x+1| - \frac{\sqrt{3}}{3} \tan^{-1} \left(\frac{(2x+1)\sqrt{3}}{3} \right) \end{aligned}$$

□

Example 9.23. Evaluate $\int \frac{dx}{x^2(x-1)}$

Solution. The denominator can be factorised into product of x, x^2 and $(x - 1)$. We may express $\frac{1}{x^2(x-1)}$ as

$$\begin{aligned}\frac{1}{x^2(x-1)} &= \frac{A}{x} + \frac{B}{x^2} + \frac{C}{x-1} \\ &= \frac{Ax(x-1) + B(x-1) + Cx^2}{(x-1)(x-2)(x-3)}\end{aligned}$$

We have

$$1 \equiv (A + C)x^2 - (A - B)x - B$$

By comparing coefficient we have $B = -1$ $A - B = 0 \Rightarrow A = -1$ $A + C = 0 \Rightarrow C = 1$ Therefore,

$$\begin{aligned}\int \frac{1}{x^2(x-1)} dx &= \int \frac{-1}{x} + \frac{-1}{x^2} + \frac{1}{x-1} dx \\ &= \ln \left| \frac{x-1}{x} \right| + \frac{1}{x} + C\end{aligned}$$

□

9.5 Definite Integral and Applications

Theorem 9.24 (Fundamental Theorem of Calculus). (a) Suppose f is a continuous function on $[a, b]$ and F is any antiderivative of f on $[a, b]$, then

$$\int_a^b f(x) dx = F(x)|_a^b = F(b) - F(a).$$

(b) Let f be a continuous function on an interval I and $a \in I$. Define the function F by

$$F(x) = \int_a^x f(t) dt$$

then $F'(x) = f(x)$ at each point in the interval I .

The theorem above points out that the problem of evaluating a definite integral is essentially the problem of finding antiderivative.

Example 9.25. Evaluate $\int_0^3 3x^2 + \sin x - e^{2x} dx$.

Solution. The function $f(x) = 3x^2 + \sin x - e^{2x}$ has antiderivative $x^3 - \cos x - \frac{1}{2}e^{2x}$ on the interval. Therefore,

$$\int_0^3 3x^2 + \sin x - e^{2x} dx = (x^3 - \cos x - \frac{1}{2}e^{2x})|_0^3 = 27\frac{1}{2} - \frac{e^6}{2} - \cos 3.$$

□

Example 9.26. The velocity of a particle which is moving in a straight line is given by $v(t) = 3t^2 - 5t$. Find the displacement of the particle during the first 2 seconds.

Solution. We want to find $S(2) - S(0)$. Since $S'(t) = v(t)$,

$$\begin{aligned} S(2) - S(0) &= \int_0^2 v(t) dt \\ &= \int_0^2 (3t^2 - 5t) dt \\ &= \left(t^3 - \frac{5}{2}t^2 \right) \Big|_0^2 \\ &= -2. \end{aligned}$$

□

The upper and lower limits in the definite integral might change when performing integration by substitution as shown in the following theorem.

Theorem 9.27. If $f : [c, d] \rightarrow \mathbb{R}$ is continuous on $[c, d]$ and $g : [a, b] \rightarrow [c, d]$ has continuous derivative on $[a, b]$. Let $u = g(x)$, then

$$\int_a^b f(g(x))g'(x)dx = \int_{g(a)}^{g(b)} f(u)du.$$

Example 9.28. Evaluate $\int_0^2 x^2 \sqrt{1+x^3} dx$.

Solution. Using the substitution $u = 1 + x^3$, we have $du = 3x^2 dx$ or $\frac{1}{3}du = x^2 dx$. Observe that when $x = 0$, $u = 1$ and $u = 9$ when $x = 2$, we have

$$\begin{aligned} \int_0^2 x^2 \sqrt{1+x^3} dx &= \frac{1}{3} \int_1^9 \sqrt{u} du \\ &= \frac{1}{3} \times \frac{2}{3} u^{\frac{3}{2}} \Big|_1^9 \\ &= \frac{52}{9} \end{aligned}$$

□

9.5.1 Areas, Volumes, Arc Lengths and Area of Surface of Revolution

Definite integrals solved the problem of finding area under a curve which in turn solve various problems in finding physical quantities which can be interpreted as areas such as work done, volume, revenue, total product etc. We will see several applications of definite integrals in this section.

Example 9.29. Find the area between $y = \sin x$ and $y = \cos x$ on the interval $[0, \pi]$

Solution. If we look at the graphs of sine and cosine functions, we see $\sin x \leq \cos x$ on the

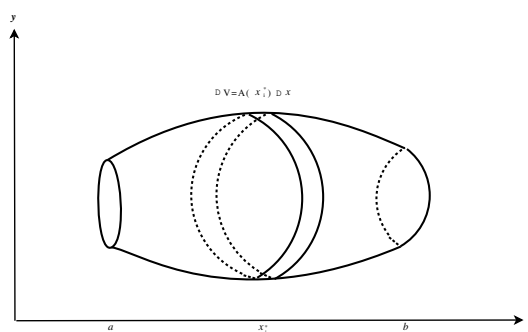
interval $[0, \frac{\pi}{4}]$ and $\sin x \geq \cos x$ on $[\frac{\pi}{4}, \pi]$.

$$\begin{aligned}\text{Area} &= \int_0^{\frac{\pi}{4}} \cos x - \sin x dx + \int_{\frac{\pi}{4}}^{\pi} \sin x - \cos x dx \\ &= (\sin x - (-\cos x))\Big|_0^{\frac{\pi}{4}} + (-\cos x - \sin x)\Big|_{\frac{\pi}{4}}^{\pi} \\ &= 2\sqrt{2} \text{ unit}^2\end{aligned}$$

□

From the diagram below, we see that to find the volume of a solid, we may use the area of the cross section $A(x)$ as our integrand and integrate over the specific range.

$$\text{Volume} = \lim_{n \rightarrow \infty} \sum_{i=0}^n A(x_i) \Delta x = \int A(x) dx$$



Example 9.30. Find the volume of the solid generated when the part of the curve $y = e^x$ on $[0, 1]$ is revolved through 360° about the line $y = -1$.

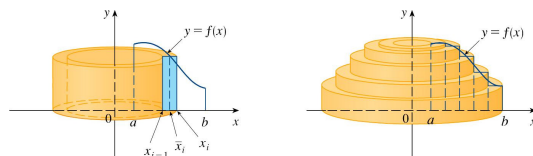
Solution. If we cut the solid with plane perpendicular to x -axis. The area of the circular cross section is a function of x given by $A(x) = \pi(e^x + 1)^2$. Therefore

$$\begin{aligned}\text{Volume} &= \int_0^1 \pi(e^x + 1)^2 dx \\ &= \pi\left(\frac{1}{2}e^{2x} + 2e^x + x\right)\Big|_0^1 \\ &\doteq 23.97 \text{ unit}^3\end{aligned}$$

□

Usually we find the cross section by cutting the solid with a plane perpendicular to the axis of revolution. However, the radius of the circular cross section may not be easily obtained as shown in the following example. In this case we may approximate the volume by expressing it as the sum of many thin cylindrical shell, similar to a roll of 'toilet paper'. As you can see from the following diagram, the volume of every shell may be written as $2\pi x f(x) \Delta x$. This diagram needs to be redrawn..... Therefore, the volume of the solid may be expressed as

$$\int_a^b 2\pi x f(x) dx.$$



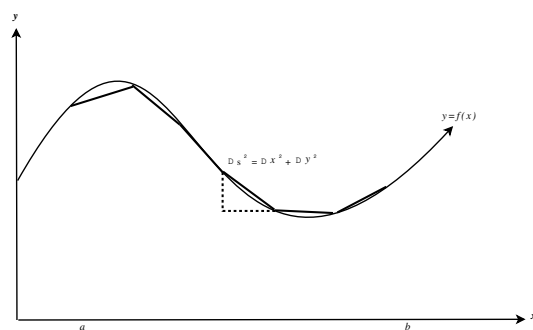
Example 9.31. Find the volume of the solid generated when the part of the curve $y = e^x$ on $[0, 1]$ is revolved through 360° about the line $x = 1$.

Solution. If we cut the solid with plane perpendicular to y -axis, we need to find the inverse function in order to obtain the radius. We may use the shell method as described above.

$$\begin{aligned} \text{Volume} &= \int_0^1 2\pi(1-x)e^x dx \\ &= 2\pi[-e^x(-2+x)]_0^1 \\ &= 2\pi(e-2) \text{ unit}^3 \end{aligned}$$

□

Suppose we want to find the length of a curve $y = f(x)$ on the interval $[a, b]$. We may treat the length of the curve as the limit of the sum of short line segment as shown in the following diagram.



$$\Delta s_i = \sqrt{\Delta x^2 + \Delta y_i^2}.$$

If the function f is continuous on $[a, b]$ and has continuous derivative on (a, b) , by the mean value theorem, there exists an x_i^* in (x_{i-1}, x_i) such that $\Delta y_i = f'(x_i^*)\Delta x$. Therefore

$$\begin{aligned} \text{Arc length} &= \lim_{\Delta x \rightarrow 0} \sum \Delta s_i \\ &= \lim_{\Delta x \rightarrow 0} \sum \sqrt{\Delta x^2 + \Delta y_i^2} \\ &= \lim_{\Delta x \rightarrow 0} \sum \sqrt{1 + [f'(x_i^*)]^2} \Delta x \\ &= \int_a^b \sqrt{1 + f'(x)^2} dx \end{aligned}$$

Example 9.32. Show that the length of the circumference of a circle with radius r is $2\pi r$.

Solution. Consider the quarter of a circle of radius r with its centre at the origin on the interval $[0, r]$ which is given by the equation $y = \sqrt{r^2 - x^2}$. We have

$$\frac{dy}{dx} = -\frac{x}{\sqrt{r^2 - x^2}}.$$

$$\begin{aligned} \text{Length of arc} &= \int_0^r \sqrt{1 + \frac{x^2}{r^2 - x^2}} dx \\ &= \int_0^r \frac{r}{\sqrt{r^2 - x^2}} dx \\ &= r \int_0^{\frac{\pi}{2}} d\theta \quad (\text{Let } x = r \sin \theta, \, dx = r \cos \theta d\theta) \\ &= \frac{\pi r}{2} \end{aligned}$$

Therefore, the length of the circumference of a circle with radius r is $2\pi r$. □

If the equation of the arc above is given in a parametric form $(f(t), g(t))$ such that $a = f(t_1)$ and $b = f(t_2)$, then

$$\text{Length of arc} = \int_{t_1}^{t_2} \sqrt{[f'(t)]^2 + [g'(t)]^2} dt.$$

When an arc is revolved about an axis, the locus of the arc forms a surface of revolution. To calculate the area of the surface of revolution, we may partition use the following theorem.

Theorem 9.33. *If the function $f(x)$ is nonnegative and has continuous derivative $f'(x)$ on the interval $[a, b]$. The area of the surface of revolution generated when the arc $y = f(x)$ on $[a, b]$ is rotated 360° about the x -axis is given by*

$$\int_a^b 2\pi f(x) \sqrt{1 + [f'(x)]^2} dx.$$

Example 9.34. Show that the surface area of a sphere is $4\pi r^2$.

Solution. Consider the hemisphere generated when the quarter of the circle $y = \sqrt{r^2 - x^2}$ on the interval $[0, r]$ is revolved through 360° about the x -axis. The area of the surface area generated by the arc is given by

$$\begin{aligned} \text{Area of surface of revolution} &= \int_0^r 2\pi \sqrt{r^2 - x^2} \sqrt{1 + \frac{x^2}{r^2 - x^2}} dx \\ &= 2\pi \int_0^r r dx \\ &= 2\pi r^2 \end{aligned}$$

Therefore, the surface area of a sphere is $4\pi r^2$. □

9.5.2 Work Done

The work done W of a constant force F to an object is defined by

$$W = Fd$$

where d is the displacement of the object under the force. When the force is changing when a particle is moving from $x = a$ to $x = b$, we may partition the displacement interval into small intervals. If F is continuous, on each of this small interval the work done is given by $F(x_i^*)\Delta x$. Therefore the total work done is

$$W = \lim_{\Delta x \rightarrow 0} \sum_{i=1}^n F(x_i^*)\Delta x = \int_a^b F(x)dx.$$

Example 9.35. Consider a spring with natural length of 10 cm, a force of 1 N is needed to maintain the spring at 15 cm. Find the work done needed to press the spring such that the length reduces to 8 cm.

Solution. By the Hooke's law,

$$\begin{aligned} F &= kx \\ 1 &= 0.05k \\ k &= 20Nm^{-1} \end{aligned}$$

$$\begin{aligned} \text{Work done} &= \int_0^{0.02} F(x)dx \\ &= \int_0^{0.02} 20x dx \\ &= 10x^2 \Big|_0^{0.02} \\ &= 0.004J \end{aligned}$$

□

Example 9.36. Consider a cylindrical tank with base radius equals to 1 m and height 1 m that is half filled with water. How much work is needed to pump all the water to the top and drain out. (Let the density of the water $\rho = 1000kg/m^3$)

Solution. Let x be the distance from the top to the water surface. If we slice the water into thin layer with thickness Δx . The work done needed to pump that surface layer to the top is given by $\pi \times 1^2 \Delta x \times 1000 \times 9.8x$. Therefore the total work done needed is

$$\begin{aligned} \lim_{\Delta x \rightarrow 0} \sum_{i=1}^n \pi \times 1^2 \Delta x \times 1000 \times 9.8x &= 9800\pi \int_{0.5}^1 x dx \\ &= 3675\pi J \end{aligned}$$

□

9.5.3 Centre of Mass

In order to introduce the centre of mass or centroid, we need to know the concept of moment. Consider n particles on the a number line position at $x_1, x_2, \dots, x_n$ with masses $m_1, m_2, \dots, m_n$ respectively. We call m_1x_1 the moment of the particle with respect to the origin, in similar way we may say that $\sum m_i x_i$ is the moment of the system with respect to the origin. If the points are on the cartesian plane, we may call $\sum m_i x_i$ the moment of the system about the y -axis. Similarly we may define the moment of the system about the x -axis, i.e. $\sum m_i y_i$, where $y_1, y_2, \dots, y_n$ are the ordinates of the points.

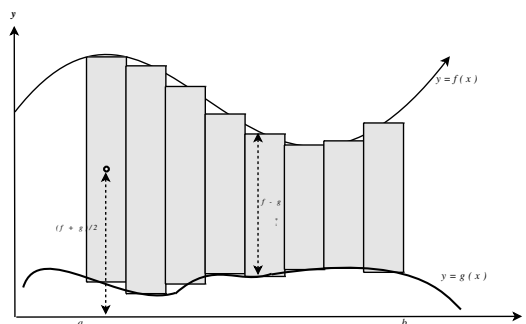
For a system of n particles, the centroid of the system $(\bar{x}, \bar{y})$ is given by

$$(\bar{x}, \bar{y}) = \left(\frac{\sum m_i y_i}{\sum m_i}, \frac{\sum m_i x_i}{\sum m_i} \right)$$

For a lamina of uniform density which is bounded by two curves $y = f(x)$ and $y = g(x)$ as shown in the following figure, using areas as analogue of masses in the discrete case; the centroid of the

lamina is

$$(\bar{x}, \bar{y}) = \left(\frac{\int_a^b x(f - g)dx}{\int_a^b (f - g)dx}, \frac{\frac{1}{2} \int_a^b (f^2 - g^2)dx}{\int_a^b (f - g)dx} \right)$$



The centroid can be thought as a point such that if a fulcrum is put at the centroid, the lamina will remain balance on the fulcrum.

Example 9.37. Find the centroid of the region bounded by the semicircle $y = \sqrt{1 - x^2}$ and the x -axis.

Solution. We may take $f(x) = \sqrt{1 - x^2}$ and $g(x) = 0$. By symmetry, we know that $\bar{x} = 0$.

$$\begin{aligned} \bar{y} &= \frac{\frac{1}{2} \int_{-1}^1 f^2 dx}{\int_{-1}^1 f dx} \\ &= \frac{\frac{1}{2} \int_{-1}^1 (1 - x^2) dx}{\frac{1}{2} \pi} \\ &= \frac{4}{3\pi} \end{aligned}$$

□

9.6 Improper Integrals????

9.7 Calculus with Maxima

batch(), append(), concat(), cons(), kill(variable), makelist(), map(), lhs(), rhs()

, numer can convert algebraic expressions to floating point values

bfloat(expr) and fpprec : 32 allows user to change the precision to higher prec.

expr, x=5/z performs substitution of $x = 5/z$. Or subst()

expand, ratsimp, factor, allroots, denom, ... for polynomial manipulations

trigexpand, trigreduce, trigsimp for simplifying some integral???

diff, integrate, taylor(f, x, x0, n), limit(f, x, x0)

coeff

realpart, imagpart, demoivre, exponentialize

Let $R(x, y)$ be a rational function of x and y

Technique for solving

$$\int R(x, \sqrt{ax^2 + bx + c}) dx \quad (9.1)$$

e:int:quadrati

Technique for solving

$$\int R(\cos(x), \sin(x)) dx$$

is to introduce $t = \tan(x/2)$ to reduce to it to (9.1). e:int:quadratic-rat

Calculate

$$\int \frac{dx}{x(x^n + a)} \quad (n \geq 1)$$

Calculate

$$\int \frac{dx}{1 + \cos(x)}$$

$$\int \frac{\ln(\tan(x))}{\cos(x) \sin(x)} dx$$

Q: Use $t = \tan(x/2)$, $\cos(x) = \frac{1-t^2}{1+t^2}$, $\sin(x) = \frac{2t}{1+t^2}$???

9.8 Summary of Results

classical Legendre polynomials $P_n(x)$:

$$1, x, \frac{3x^2 - 1}{2}, \frac{5x^3 - 3x}{2}, \dots$$

are orthogonal on $[-1, 1]$ with respect to the Lebesgue measure:

$$\int_{-1}^1 P_m(x) P_n(x) dx = 0, \text{ for } m \neq n.$$

$$\int (ax + b)^n dx = \frac{1}{(n+1)a} (ax + b)^{n+1}, \quad a \neq 0, n \neq -1$$

$$\int \frac{1}{ax + b} dx = \frac{1}{a} \ln |ax + b|, \quad a \neq 0$$

$$\int \frac{1}{x^2 + a^2} dx = \frac{1}{a} \tan^{-1} \frac{x}{a}$$

$$\int \frac{1}{\sqrt{x^2 + A}} dx = \ln |x + \sqrt{x^2 + A}|$$

$$\int \sqrt{x^2 + A} dx = \frac{x}{2} \sqrt{x^2 + A} + \frac{1}{2} A \ln |x + \sqrt{x^2 + A}|$$

$$\int \frac{1}{\sqrt{a^2 - x^2}} dx = \sin^{-1} \frac{x}{a}.$$

$$\int \sqrt{a^2 - x^2} dx = \frac{a^2}{2} \sin^{-1} \frac{x}{a} + \frac{x}{2} \sqrt{a^2 - x^2}.$$

$$\int e^{bx} dx = \frac{1}{b} e^{bx}, \quad \int a^{bx} dx = \frac{1}{b \ln a} e^{bx}$$

$$\int \ln x dx = x \ln x - x$$

$$\int \sin ax dx = -\frac{\cos ax}{a}$$

$$\int \cos ax dx = \frac{\sin ax}{a}$$

$$\int \sin^m x \cos^n x dx = \dots \quad m + n \neq 0$$

$$\int \tan ax dx = -\frac{1}{a} \ln |\cos ax|$$

$$I_n = \int \tan^n x dx = \frac{\tan^{n-1} x}{n-1} - I_{n-2}, \quad n \neq 1$$

$$\int \cot ax dx = \frac{1}{a} \ln |\sin ax|$$

$$\int \sec^2 ax dx = \frac{1}{a} \tan ax$$

$$\int \csc^2 ax dx = \frac{1}{a} \cot ax$$

$$\int \frac{1}{\sin ax} dx = \frac{1}{a} \ln \left| \tan \frac{ax}{2} \right|$$

$$\int \frac{1}{\cos ax} dx = \frac{1}{a} \ln \left| \tan \frac{ax}{2} + \frac{\pi}{4} \right|$$

$$\int f(x, \sqrt[n]{ax+b}) dx$$

$$\int f(x, \sqrt{ax^2+bx+c}) dx$$

$$\int x^m (ax^n + b)^{p/q} dx, \quad m, n, p, q \in \mathbb{Z}, \quad a \neq 0, b \neq 0, q > 0, p/q \notin \mathbb{Z}$$

$$\int p(x)T(x)dx = p(x) \int T(x)dx - \int p'(x) \int T(u)du dx$$

$$\int f(\sin x, \cos x) dx = \int f\left(\frac{2t}{1+t^2}, \frac{1-t^2}{1+t^2}\right) \frac{2}{1+t^2} dt, \quad t = \tan \frac{x}{2}$$

Chapter 10

Theory of Interest

Theory of interest can be regarded as the application of series (Definition ^{def:series}5.6) and calculus to calculate the present value and future value for a simple financial asset which is a finite sequence $\{(t_i, v_i)\}_{i=0}^n$. This chapter is based on the main reference book “The Theory of Interest (Third Edition)” by Kellison, S. G. (2009).

10.1 Measurement of Interest

10.1.1 Accumulation and Amount Functions

Definition 10.1. A common financial transaction is the investment of an amount of money at interest. E.g. we invest in a savings account at bank. The initial amount of money, *capital*, invested is called the *principal* and the total amount received after a period of time is called the *accumulated value*. The difference between the accumulated value and the principal is the *amount of interest*, or just, *interest*.

Let t (units can be days, months, decades, etc.) measure time from the date of investment. The unit in which time is measured is called the (*measurement*) *period*. The most common measurement period is one year (unless otherwise stated).

Definition 10.2. The *accumulation function* $a(t)$ is the accumulated value at time $t \geq 0$ of an original investment of 1.

Properties of $a(t)$:

1. $a(0) = 1$.
2. $a(t)$ is increasing w.r.t t in general (decreasing happens when losing money).
3. If interest rate is discontinuous, $a(t)$ is discontinuous.
4. The *amount function* $A(t) = k \times a(t)$.

Remark 10.1. The definition of $a(t)$ and $A(t)$ assumes that the fund is only growing through interest, i.e. no principal is added or withdrawn!

Definition 10.3. The amount of interest earned during the n th period from the date of investment, $I_n = A(n) - A(n - 1)$, $n = 1, 2, 3, \dots$.

10.1.2 Effective Rate of Interest

Definition 10.4. The *effective rate of interest*, i , is the amount of money that one unit invested at the beginning of a period will earn during the period, where interest is paid at the end of the period.

Remark 10.2. By definition, $i = a(1) - a(0)$, i.e. $a(1) = 1 + i$.

Effective rates of interest can be calculated over any measurement period. Let i_n be the *effective rate of interest during the n th period from the date of investment*. Then

$$i_n = \frac{a(n) - a(n-1)}{a(n-1)} = \frac{A(n) - A(n-1)}{A(n-1)} = \frac{I_n}{A(n-1)}, \quad n = 1, 2, \dots$$

10.1.3 Simple Interest

Definition 10.5. When the accumulation function, $a(t) = 1 + it$, $t = 1, 2, \dots$, the accruing of interest according to this pattern is called *simple interest*.

In this case, the effective rate of interest,

$$i_n = \frac{(1 + in) - (1 + i(n-1))}{1 + i(n-1)} = \frac{i}{1 + i(n-1)}.$$

This is a decreasing function of n .

Remark 10.3. Its application can be found in:

1. Some bank charges simple interest for *short term* loan.
2. Car mortgages.
3. US Treasury bills are quoted by simple interest.

eg:1.3

Example 10.4. Find the accumulated value of RM2000 invested for 4 years, if the rate of simple interest is 8% per annum.

Solution. $A(4) = \text{RM}2000 \times (1 + 4 \times 0.08) = \text{RM}2640$

Note that $i_4 = \frac{0.08}{1+0.08(4-1)} = 0.0645 = 6.45\% < 8\%$

□

10.1.4 Compound Interest

Definition 10.6. When the accumulation function, $a(t) = (1+i)^t$, $t = 1, 2, \dots$, the accruing of interest according to this pattern is called *compound interest*.

In this case, the effective rate of interest is a constant,

$$i_n = \frac{(1+i)^n - (1+i)^{n-1}}{(1+i)^{n-1}} = (1+i) - 1 = i.$$

The use of compound interest for completed periods and simple interest for a final fractional period can be analysed as follows: Consider the investment of 1 for $n + k$ periods, where $n \in \mathbb{Z}$ and $0 < k < 1$. Then the linear interpolation gives

$$(1+i)^{n+k} \approx (1-k)(1+i)^n + k(1+i)^{n+1} = (1+i)^n(1-k+k(1+i)) = \underline{(1+i)^n(1+ki)}$$

This is useful when you just have a normal (non-scientific) calculator.

eg:1.4

Example 10.5. Rework Example 10.4 using compound interest.

Solution. $A(4) = \text{RM}2000 \times (1 + 0.08)^4 = \text{RM}2720.98 > \text{RM}2640$ □

10.1.5 Present Value

Definition 10.7. An investment of 1 will accumulate to $1 + i$ at the end of one period. The term $1 + i$ is called an *accumulation factor*. We define the term

$$v = \frac{1}{1 + i}$$

as a *discount factor*. This is because it “discounts” the value of an investment at the end of a period to its value at the beginning of the period.

We called $a^{-1}(t) = \frac{1}{a(t)}$ (this is confusing mathematically but is used in “theory of interest”) the *discount function*.

For simple interest, $a^{-1}(t) = \frac{1}{1+it}$.

For compound interest, $a^{-1}(t) = \frac{1}{(1+i)^t} = v^t$.

eg:1.7

Example 10.6. Find the amount which must be invested at 9% per annum in order to accumulate RM1000 at the end of 3 years: (a) assuming simple interest. (b) assuming compound interest.

Solution.

(a) $\frac{\text{RM}1000}{1 + (0.09)(3)} = \text{RM}787.40$

(b) $\frac{\text{RM}1000}{(1 + 0.09)^3} = \text{RM}772.18$ □

10.1.6 Effective Rate of Discount

In real life, when you borrow RM1, there are two ways to deal with the interest:

- Interest Model: Borrow RM1 at $t = 0$, return $\text{RM}(1 + i)$ at $t = 1$.
- Discount Model: Borrow $\text{RM}(1 - d)$ at $t = 0$, return RM1 at $t = 1$. (Money is deducted from the beginning.)

Let us introduce a concept opposite to the effective rate of interest:

Definition 10.8. The *effective rate of discount*, d , is the ratio of the amount of interest earned during the period to the amount invested at the end of the period, i.e.

$$d_n = \frac{A(n) - A(n-1)}{A(n)} = \frac{I_n}{A(n)}, \quad 1, 2, \dots$$

Assume that a person borrows 1 at an effective rate of discount d , then, in effect, the original principal is $1 - d$ and the amount of interest (discount) is d . From the definition of i , we have

$$i = \frac{\text{amount of interest (discount)}}{\text{principal}} = \frac{d}{1 - d}. \quad (10.1)$$

eg:i+d

From here, we have $d = \frac{i}{1+i} = iv = 1 - v$.

eg:1.8

Example 10.7. For Example 10.6, we calculate (a) assuming simple discount instead of simple interest. (b) Assuming compound discount.

Solution.

(a) $PV = RM1000(1 - (0.09)(3)) = RM730.$

(b) $PV = RM1000(1 - 0.09)^3 = RM753.57.$

□

10.1.7 Nominal Rates of Interest and Discount

Definition 10.9. The frequency with which interest is paid and reinvested to earn additional interest is called *the interest conversion period*.

Remark 10.8. Vague terminologies occur in many finance related subjects, so beware!

“Compounded”: the interest being earned is reinvested to earn additional interest.

“Payable”: the interest is paid out in instalments as it is earned.

Definition 10.10. The symbol for a *nominal rate of interest* payable m times per period is $i^{(m)}$, it's related to amount or accumulation function as follows:

$$\frac{i^{(m)}}{m} = \frac{A(t + \frac{1}{m}) - A(t)}{A(t)}$$

where m is a positive integer > 1 . By a nominal rate of interest $i^{(m)}$, we mean a rate payable m thly, i.e. the rate of interest is $i^{(m)}/m$ for each m th of a period and not $i^{(m)}$.

Remark 10.9. For example, a nominal rate of 8% convertible quarterly does not mean an interest rate of 8% per quarter but rather an interest rate of 2% per quarter.

From the definition, we have

$$1 + i = \left(1 + \frac{i^{(m)}}{m}\right)^m.$$

Definition 10.11. The symbol for a *nominal rate of discount* payable m times per period is $d^{(m)}$. By a nominal rate of discount $d^{(m)}$, we mean a rate payable m thly, i.e. the effective rate of discount is $d^{(m)}/m$ for each m th of a period.

From the definition, we have

$$1 - d = \left(1 - \frac{d^{(m)}}{m}\right)^m.$$

and $d^{(m)}$ is related to v by

$$d^{(m)} = m(1 - v^{1/m}).$$

In pictorial form,

Time:	0	$\frac{1}{m}$	$\dots$	$\frac{m-2}{m}$	$\frac{m-1}{m}$	1
Balance:	$1 - d = (1 - \frac{d^{(m)}}{m})^m$	$(1 - \frac{d^{(m)}}{m})^{m-1}$	$\dots$	$(1 - \frac{d^{(m)}}{m})^2$	$1 - \frac{d^{(m)}}{m}$	1

There is a close relationship between nominal rates of interest and nominal rates of discount, given by

$$1 + i = \left(1 + \frac{i^{(m)}}{m}\right)^m = 1 + \frac{d}{1 - d} = \frac{1}{1 - d} = \left(1 - \frac{d^{(p)}}{p}\right)^{-p}$$

using Equation (10.1).

eg:1.11

Example 10.10. Find the accumulated value of RM500 invested for 5 years at 8% per annum convertible quarterly.

Solution. $A(5) = 500(1 + \frac{0.08}{4})^{4 \times 5} = 500 \times 1.02^{20} = \text{RM}742.97.$ $\square$

Remark 10.11. *Payable in advance* means our money is discounted in the beginning.

eg:1.12

Example 10.12. Find the present value of RM1000 to be paid at the end of 6 years at 6% per annum payable in advance and convertible semiannually.

Solution. $PV = 1000(1 - \frac{0.06}{2})^{2 \times 6} = 1000 \times 0.97^{12} = \text{RM}693.84.$ $\square$

eg:1.13

Example 10.13. Find the nominal rate of interest convertible quarterly which is equivalent to a nominal rate of discount of 6% per annum convertible monthly.

Solution. $(1 + \frac{i^{(4)}}{4})^4 = (1 - \frac{0.06}{12})^{-12}$
 $\Rightarrow 1 + \frac{i^{(4)}}{4} = (0.995)^{-3}$
 $\Rightarrow i^{(4)} = 4((0.995)^{-3} - 1) = 6.06\%$ $\square$

10.1.8 Force of Interest and Discount

The measures of interest defined in previous sections are useful for measuring interest over specified intervals of time. To measure the intensity of interest at each “moment of time” (over infinitesimally small intervals of time), the *force of interest* (as a limit of the nominal rate of interest) is introduced.

Definition 10.12. The *force of interest* at time t , denoted by δ_t is defined as

$$\delta_t = \lim_{h \rightarrow 0} \frac{A(t+h) - A(t)}{hA(t)} = \frac{A'(t)}{A(t)} = \frac{a'(t)}{a(t)} = \frac{d}{dt} \ln a(t) = \lim_{m \rightarrow \infty} i^{(m)}$$

where $h = \frac{1}{m}$, $i^{(m)} = \frac{A(t + \frac{1}{m}) - A(t)}{\frac{1}{m}} \times \frac{1}{A(t)}.$

From the definition and using calculus, we obtain

$$\int_0^t \delta_r dr = [\ln a(r)]_0^t = \ln a(t) \Rightarrow a(t) = \exp \int_0^t \delta_r dr.$$

and

$$\int_0^n A(t) \delta_t dt = [A(t)]_0^n = A(n) - A(0).$$

Definition 10.13. Analogously, we can define the *force of discount* as

$$\delta'_t = -\frac{\frac{d}{dt} a^{-1}(t)}{a^{-1}(t)}.$$

Using calculus, we can show that “force of discount = force of interest”:

$$\delta'_t = \frac{a^{-2}(t) \frac{d}{dt} a(t)}{a^{-1}(t)} = a^{-1}(t) \times a(t) \delta_t = \delta_t.$$

In theory, the force of interest may vary instantaneously. However, in practice it is often constant, i.e. $\delta_t = \delta$ (recall the definition of compound interest). In this case,

$$a(n) = (1 + i)^n = \exp(\int_0^n \delta dt) = \exp(\delta n) \Rightarrow \delta = \ln(1 + i).$$

eg:1.14

Example 10.14. Find the accumulated value of RM1000 invested for 10 years if the force of interest is 5%.

Solution. From $1 + i = e^\delta = e^{0.05}$, we have

$$A(10) = \text{RM}1000(1 + i)^{10} = \text{RM}1000e^{0.05 \times 10} = \text{RM}1648.72.$$

□

Example 10.15. Find the accumulated value of 1 at the end of n years if $\delta_t = \frac{1}{1+t}$.

Solution. $a(t) = \exp \int_0^n \frac{1}{1+t} dt = \exp \ln(1+n) = 1+n$

□

Remark 10.16. For simple interest, $a(t) = 1 + it$ and $\delta_t = \frac{i}{1+it}$.

Example 10.17. Money accumulates at a varying force of interest $\delta_t = 0.05 + 0.01t$ for $0 \leq t \leq 4$. Find the present value at time $t = 0$ of two payments of RM100 each to be paid at times $t = 2$ and $t = 4$.

Solution. We use the fact that “force of discount = force of interest” to do the calculation.

(a) $PV = 100 \exp\left(-\int_0^2 (0.05 + 0.01t) dt\right) = 100 \exp(-(0.05 \times 2 + 0.005 \times 4))$
 $= 100e^{-0.12} = \text{RM}88.69.$

(b) $PV = 100 \exp\left(-\int_0^4 (0.05 + 0.01t) dt\right) = 100 \exp(-(0.05 \times 4 + 0.005 \times 16))$
 $= 100e^{-0.28} = \text{RM}75.58.$

□

10.1.9 Varying Interest

Two types of varying interest are commonly encountered in practice.

The **first type of variation** involves changes in the effective rate of interest over a period of time. Let i_k denote the effective rate of interest during the k th period from the date of investment. Then

$$a(t) = (1 + i_1)(1 + i_2) \cdots (1 + i_t) = \prod_{k=1}^t (1 + i_k).$$

and

$$a^{-1}(t) = (1 + i_1)^{-1}(1 + i_2)^{-1} \cdots (1 + i_t)^{-1} = \prod_{k=1}^t (1 + i_k)^{-1} = \prod_{k=1}^t v_k.$$

Example 10.18. Find the accumulated value of RM1000 at the end of 15 years if the effective rate of interest is 4% for the first 5 years, 4.5% for the second 5 years and 5% for the third 5 years.

Solution. $A(20) = 1000(1.04)^5(1.045)^5(1.05)^5 = \text{RM}1935.05.$

□

Example 10.19. An investor in common stock measures investment returns annually using an effective rate of interest. The investor earns 15% during the first year, -5% during the second year, and 8% during the third year. Determine the equivalent level annual effective rate of interest over the 3-year period, i .

Solution. $a(3) = (1 + i)^3 = (1 + 0.15)(1 + (-0.05))(1 + 0.08)$
 $\Rightarrow i = (1.05 \times 0.95 \times 1.08)^{1/3} - 1 = 5.67\%.$

□

The **second type of variation** involves a force of interest which varies from period to period, but is constant during each period. Let i_k denote the effective rate of interest during the k th period from the date of investment where k is a positive integer. We define $\delta_{[k]}$ as the

(level) force of interest equivalent to i_k during the k th period. From $\delta_{[k]} = \ln(1 + i_k)$, we have the accumulated values and present values as

$$a(t) = \exp\left(\int_0^t \delta_r dr\right) = \exp\left(\sum_{k=1}^t \delta_{[k]}\right)$$

and

$$a^{-1}(t) = \exp\left(-\sum_{k=1}^t \delta_{[k]}\right)$$

respectively for $t \geq 1$.

Example 10.20. An investor in common stock measures investment returns annually using an effective rate of interest. The force of interest during the 1st, 2nd and 3rd year is $\delta_{[1]} = 0.15$, $\delta_{[2]} = -0.05$ and $\delta_{[3]} = 0.08$ respectively. Find the equivalent level force of interest over the 3-year period, δ .

Solution. $a(3) = (e^\delta)^3 = \exp(0.15 + (-0.05) + 0.08) = \exp(0.18)$
 $\Rightarrow \delta = 0.18/3 = 0.06.$

□

10.2 Solutions to Problems in Interest

An interest problem involves 4 basic quantities:

1. The principal originally invested.
2. The length of the investment period.
3. The rate (or force) of interest (or discount).
4. The accumulated value of the principal at the end of the investment period.

In practical applications involving interest the terminology can become confusing. Many terms have ambiguous meaning. Furthermore, some terms used do not convey an intuitive description of the transactions involved (e.g. “annuity-immediate” and “annuity-due” in Topic 3). Many parties involved in financial transactions simply do not use terms with textbook precision.

10.2.1 Equations of Value

Principle of Time Value of Money. The value of an amount of money at any given point in time depends upon the time elapsed since the money was paid in the past or upon the time which will elapse in the future before it is paid.

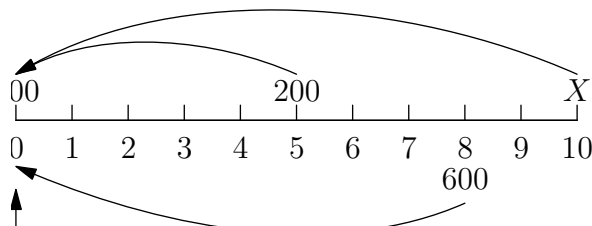
Caution: The “time value of money” reflects the effect of interest but not the effect of inflation which reduces the purchasing power of money over time.

Definition 10.14. The Principle of Time Value of Money implies that two or more amounts of money payable at different points in time cannot be compared until all the amounts are accumulated or discounted to a common date. This common date is called the *comparison date*, and the equation which accumulates or discounts each payment to the comparison date is called the *equation of value*.

One useful technique in the solutions of equations of value is the *time diagram*, a 1-D diagram in which units of time are measured along the 1-D and payments are placed on the diagram at the appropriate points. Note that payments in one direction are placed on the top of the diagram and payments in the other direction are placed on the bottom of the diagram.

Example 10.21. In return for a promise to receive RM600 at the end of 8 years, a person agrees to pay RM100 at once, RM200 at the end of 5 years, and to make a further payment at the end of 10 years. Find the payment at the end of 10 years if the normal rate of interest is 8% convertible semiannually.

Solution. First, sketch the time diagram:



Since the interest is convertible semiannually, we will count time periods in half-years. The equation of value is

$$100 + 200v^{2 \times 5} + Xv^{2 \times 10} = 600v^{2 \times 8}$$

where

$$v = \frac{1}{1+i} = \frac{1}{1+0.08/2} = 0.961538461538$$

Hence,

$$X = \frac{600v^{16} - 100 - 200v^{10}}{v^{20}} = \text{RM186.75}$$

One can take another approach by calculating using interest instead of discounting:

$$100(1.04)^{20} + 200(1.04)^{10} + X = 600(1.04)^4 \Rightarrow X = \text{RM186.75}$$

□

10.2.2 Unknown Time

Case 1

Consider a situation in which several payments made at various points in time are to be replaced by one payment numerically equal to the sum of the other payments. The problem is to

find the point in time at which the one payment should be made such that it is equivalent in value to the payments made separately.

The fundamental equation of value is

$$(s_1 + s_2 + \cdots + s_n)v^t = s_1v^{t_1} + \cdots + s_nv^{t_n} \quad (10.2) \quad \text{e:2.1}$$

which is one equation in one unknown t . This equation can readily be solved using logarithms.

As a first approximation, we can estimate t by

$$\bar{t} = \frac{s_1t_1 + \cdots + s_nt_n}{s_1 + \cdots + s_n}. \quad (10.3) \quad \text{e:2.2}$$

This estimation is called the *method of equated time*. It can be shown that $\bar{t} > t$.

eg:2.3

Example 10.22. Payments of RM100, RM200, RM500 are due at the ends of years 2, 3, and 8, respectively. Assuming an effective rate of interest of 5% per annum, find the point in time at which a payment of RM800 would be equivalent: (1) By the method of equated time. (2) By an exact method.

Solution. (1) $\bar{t} = \frac{100 \times 2 + 200 \times 3 + 500 \times 8}{100 + 200 + 500} = 6$ years.

$$(2) v = \frac{1}{1.05} = 0.95238, v^t = \frac{100v^2 + 200v^3 + 500v^8}{100 + 200 + 500} = \frac{601.89}{800} = 0.75236$$

This implies $t = \frac{\ln 0.75236}{\ln 0.95238} = 5.83185$. □

Case 2

Consider how long it takes money to double at a given rate of interest.

$$(1+i)^n = 2 \Rightarrow n = \frac{\ln 2}{\ln(1+i)} \approx \frac{0.06831}{i} \frac{i}{\ln(1+i)} = \frac{0.06831}{i} (1.0395) = \frac{0.72}{i}$$

where $i = 0.08 \Rightarrow \frac{i}{\ln(1+i)} = 1.0395$. This is called the *rule of 72*.

eg:2.2

Example 10.23. Find the length of time necessary for RM1000 to accumulate to RM1500 if invested at 6% per annum compounded semiannually.

Solution. From the equation of value $1000(1 + \frac{0.06}{2})^{2n} = 1500$, $n = \frac{\ln 1.5}{2 \ln(1.03)} = 6.85861871$ (years). □

10.2.3 Unknown Rate of Interest

.....Some Problems requires Financial Calculator

The unknown rate of interest is usually described by an equation $f(i) = 0$. To solve this, we need to use *numerical successive iteration* such as Newton's iteration:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

eg:2.4

Example 10.24. At what interest rate *convertible quarterly* would RM1000 accumulate to RM1600 in 6 years?

Solution. The equation of value is

$$1000(1 + \frac{i^{(4)}}{4})^{4 \times 6} = 1600 \Rightarrow i^{(4)} = 4(\exp(\frac{\ln 1.6}{24}) - 1) = 0.0791$$

□

eg:2.5

Example 10.25. At what effective rate of interest will the present value of RM2000 at the end of two years and RM3000 at the end of four years be equal to RM4000?

Solution. $4000 = 2000v^2 + 3000v^4 \Rightarrow 3000v^4 + 2000v^2 - 4000 = 0$

This is a quadratic equation with the variable v^2 , hence,

$$v^2 = \frac{-2 + \sqrt{2^2 - 4 \times 3 \times (-4)}}{2 \times 3} = 0.868517091822 \Rightarrow v = 0.931942644062 = \frac{1}{1+i}$$

This implies the effective rate of interest is 0.07302740825. □

10.2.4 Determining Time Periods

In practical problems involving interest it is necessary to determine the exact time period of an investment. Three methods of counting the days in a *period of investment* have arisen in practice:

1. "actual/actual": 365 days/year — In counting days between two dates, the last but not the first, date is included.
2. "30/360": Assumes each calendar month has 30 days and entire calendar year has 360 days. — Public Security Association has a guideline for this.
3. "actual/360": Exact number of days for the period of investment and 360 days in a year.

Example 10.26. Find the amount of interest that RM2000 deposited on 17 June will earn if the money is withdrawn on 10 September in the same year (not a leap year) and if the rate of simple interest is 8%, on the following bases:

- (1) Exact simple interest (actual/actual).
- (2) Ordinary simple interest (30/360).
- (3) The Banker's Rule (actual/360).

Solution.

- (1) 10 September is day 253, June 17 is day 168. Actual number of days = $253 - 168 = 85$.

Hence, interest earn = $2000 \times 0.08 \times \frac{85}{365} = \text{RM}37.26$.

- (2) $360(0) + 30(9 - 6) + (10 - 17) = 83$.

Hence, interest earn = $2000 \times 0.08 \times \frac{83}{360} = \text{RM}36.89$.

- (3) Interest earn = $2000 \times 0.08 \times \frac{85}{360} = \text{RM}37.78$.

□

Example 10.27. Treasury Bill with face value is a promise by the government to pay on the maturity date. A treasury bill with face value \$1000 is sold on 19 August 2004 and matures 16 September 2004 (28 days term). The bill is sold for \$998.94. What is the effective annual rate of the investment, assuming simple interest? Round to 4 digits. Assume that there are 366 days (2004 is a leap year) in the year.

Solution.

$$1000 = 998.94(1 + \frac{28}{366}i) \Rightarrow i = 0.0138704169276$$

□

10.3 Basic Annuities

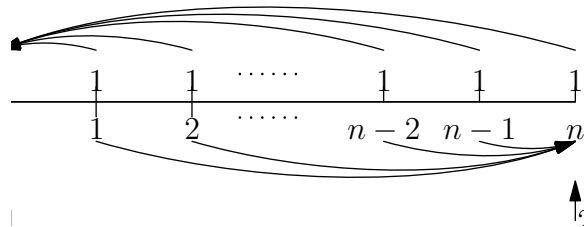
10.3.1 Annuity-immediate, Annuity-due

Definition 10.15. An *annuity* is a series of payments made at equal intervals of time. An annuity s.t. payments are certain to be made for a fixed period of time is called an *annuity-certain*. The fixed period of time for which payments are made is called the *term* of the annuity. An annuity under which the payments are not certain to be made is called a *contingent annuity*.

Remark 10.28. We will just concerntrate on *annuity-certain*, abbreviated *annuity*. The interval between annuity payments is called the *payment period*.

Example 10.29. House rents, mortgage payments (annuity-certain), monthly retirement benefits from a pension plan (contingent annuity), instalment payments on automobiles, and interest payments on money invested.

Consider an annuity under which payments of 1 are made at the end of each period for n periods, where n is a positive integer. Such an annuity is called an *annuity-immediate* (or *ordinary annuity*). Assuming that the rate of interest is i per period, where $i > 0$. The present value of the annuity at the beginning is denoted by $a_{\overline{n}|}$. The accumulated value of the annuity at time n is denoted by $s_{\overline{n}|}$.



Time diagram for annuity-immediate

We can derive the formula for $a_{\overline{n}|}$ and $s_{\overline{n}|}$ w.r.t. i and v :

$$a_{\overline{n}|} = v + v^2 + \cdots + v^n = v \frac{1 - v^n}{1 - v} = \frac{1 - v^n}{i} \quad (10.4) \quad \text{e:3.2}$$

and

$$s_{\overline{n}|} = 1 + (1 + i) + \cdots + (1 + i)^{n-2} + (1 + i)^{n-1} = \frac{(1 + i)^n - 1}{(1 + i) - 1} = \frac{(1 + i)^n - 1}{i} \quad (10.5) \quad \text{e:3.4}$$

The following relations link $a_{\overline{n}|}$ and $s_{\overline{n}|}$:

$$s_{\overline{n}|} = a_{\overline{n}|}(1 + i)^n$$

and

$$\frac{1}{s_{\overline{n}|}} + i = \frac{1}{a_{\overline{n}|}}.$$

Numerical values for annuities can be calculated in two ways:

1. Using a pocket calculator with exponential and logarithmic functions and formulae (10.4) and (10.5); or
2. Using a financial calculator (e.g. BA II Plus) with five keys:

N	– the number of payments, i.e. n
I/Y	– the interest rate as a percentage, i.e. $100i$
PV	– present value, i.e. $a_{\overline{n} }$
PMT	– annuity payment
FV	– future value, i.e. $s_{\overline{n} }$

eg:3.1

Example 10.30. Find the present value of an annuity which pays RM500 at the end of each half-year for 20 years if the rate of interest is 9% convertible semiannually (nominal by default!).

Solution. Present Value = $500a_{\overline{40}|0.045} = 500 \frac{1 - \frac{1}{1.045^{40}}}{0.045} = 500 \times 18.40 = \text{RM}9200.79$ □

Solution. Using financial calculator, we set

- 2ND & CLR
- N = 40
- I/Y = 4.5 (percent)
- PMT = -500
- CPT & PV $\Rightarrow$ 9200.79

□

eg:3.2

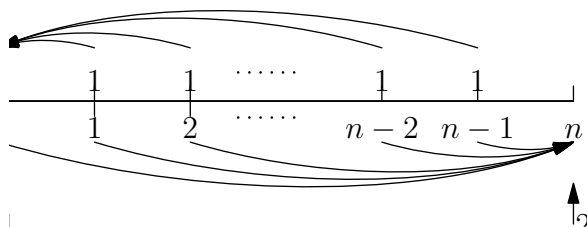
Example 10.31. If a student invests RM10,000 at 8% per annum convertible quarterly, how much can be withdrawn at the end of every quarter to use up the money exactly at the end of 4 years of university study?

Solution. Let R be the amount of each withdrawal. The equation of value is

$$Ra_{\overline{40}|0.02} = 10000 \Rightarrow R = \frac{10000}{13.57771} = \text{RM}736.50$$

□

Definition 10.16. Consider an annuity under which payments of 1 are made at the *beginning* of each period for n periods, where n is a positive integer. Such an annuity is called an *annuity-due*.



Time diagram for annuity-due

The present value of annuity-due (at arrow 1) is denoted by $\ddot{a}_{\overline{n}|}$. The accumulated value of the annuity-due (at arrow 2) is denoted by $\ddot{s}_{\overline{n}|}$. They can be calculated as follows (recall that $d = iv = 1 - v$):

$$\ddot{a}_{\overline{n}|} = 1 + v + \cdots + v^{n-1} = \frac{1 - v^n}{d} \quad (10.6) \quad \text{e:3.8}$$

and

$$\ddot{s}_{\overline{n}|} = (1 + i) + (1 + i)^2 + \cdots + (1 + i)^n = \frac{(1 + i)^n - 1}{d}. \quad (10.7) \quad \text{e:3.10}$$

The present value and accumulated value are linked by the following relations:

$$\ddot{s}_{\overline{n}|} = \ddot{a}_{\overline{n}|}(1 + i)^n \quad (10.8) \quad \text{e:3.11}$$

and

$$\frac{1}{\ddot{a}_{\overline{n}|}} = \frac{1}{\ddot{s}_{\overline{n}|}} + d. \quad (10.9) \quad \text{e:3.12}$$

The following are the relations between annuity-immediate and annuity-due:

$$\ddot{a}_{\overline{n}|} = a_{\overline{n}|}(1 + i), \quad \ddot{s}_{\overline{n}|} = s_{\overline{n}|}(1 + i), \quad \ddot{a}_{\overline{n}|} = a_{\overline{n-1}|} + 1, \quad \ddot{s}_{\overline{n}|} = s_{\overline{n+1}|} - 1. \quad (10.10) \quad \text{e:3.15}$$

eg:3.4

Example 10.32. An investor wishes to accumulate RM100,000 in a fund for retirement at the end of 12 years. To accomplish this the investor plans to make deposits at the end of each year, the final payment to be made one year prior to the end of the investment period. How large should each deposit be if the fund earns 7% effective.

Solution. There will only be 11 payments since we are interested in the accumulated value one year after the last payment, the equation of value is

$$R\ddot{s}_{\overline{11}|} = 100,000$$

where R is the annual deposit. Solving for R gives

$$R = \frac{100,000(1 - \frac{1}{1.07})}{1.07^{11} - 1} = 100,000 \times 0.05921206 = \text{RM}5921.21$$

□

Solution. Using financial calculator

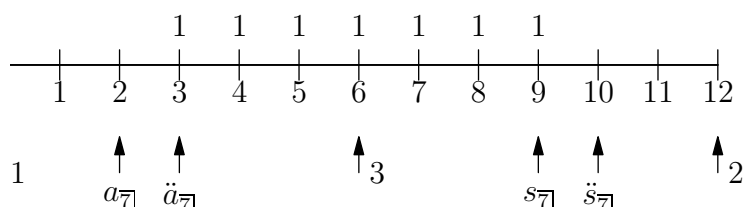
- 2ND & BGN annuity-due
- 2ND & SET
- N = 11
- I/Y = 7 (%)
- FV = -100000
- CPT & PMT $\Rightarrow$ 5921.21

□

10.3.2 Annuity Values on Any Date

Previously, we have considered evaluating annuities only at the beginning of the term or at the end of the term. However, it is often necessary to evaluate annuities on other dates.

Consider an annuity under which seven payments of 1 are made at the end of the 3rd through the 9th periods inclusive:



1. Present values more than one period before the first payment due date: In this case, the present value at the end of the 2nd period discounted for two periods,

$$v^2 a_{\overline{7}|} = a_{\overline{9}|} - a_{\overline{2}|}.$$

This type of annuity is often called a *deferred annuity*, since payments commence only after a deferred period. A symbol given to an annuity-immediate deferred for m periods with a term of n periods after the deferral period is ${}_m|a_{\overline{n}|}$. Thus, the annuity being illustrated above could be labelled ${}_2|a_{\overline{7}|}$.

2. Accumulated values more than one period after the last payment date: In this case, the accumulated value of the annuity at the end of 12th period is seen to be the accumulated value at the end of the 9th period, accumulated for three periods, i.e.

$$s_{\overline{7}|}(1+i)^3 = s_{\overline{10}|} - s_{\overline{3}|}.$$

3. Current values between the first and last payment dates: In this case, the current value of the annuity at the end of the 6th period is seen to be the present value at the end of the 2nd period accumulated for four periods or the accumulated value at the end of the 9th period discounted for three periods, i.e.

$$a_{\overline{7}|}(1+i)^4 = v^3 s_{\overline{7}|} = s_{\overline{4}|} + a_{\overline{3}|}.$$

10.3.3 Perpetuities

Definition 10.17. A *perpetuity* is an annuity whose payments continue forever, i.e. the term of the annuity is not finite.

Example 10.33. Though it seems unrealistic to have an annuity with payments continuing forever. The *dividends on the preferred stock with no redemption provision* and the *British consols*, which are nonredeemable obligations of the British government, are examples of perpetuities.

The present value of a *perpetuity-immediate* (assuming $i > 0$ so that $v < 1$) is

$$a_{\infty|} = v + v^2 + \cdots = \frac{v}{1-v} = \frac{1}{i}. \quad (10.11) \quad \text{e:3.18}$$

Similarly, the present value of a *perpetuity-due* (assuming $i > 0$ so that $v < 1$) is

$$\ddot{a}_{\infty|} = 1 + v + \cdots = \frac{1}{1-v} = \frac{1}{d}. \quad (10.12) \quad \text{e:3.19}$$

eg:3.5

Example 10.34. A leaves an estate of RM100,000. Interest on the estate is paid to beneficiary B for the first 10 years, to beneficiary C for the second 10 years, and to charity D thereafter. Find the relative shares of B, C, and D in the estate, if it is assumed the estate will earn a 7% annual effective rate of interest.

Solution. The value of B's share ($i = 0.07$, $v = 0.93457944$) is

$$100,000 \times 7\% \times a_{\overline{10}|} = 7000 \times \frac{1-v^{10}}{i} = \text{RM}49165.07.$$

The value of C's share is

$$7000 \times v^{10} a_{\overline{10}|} = \text{RM}24993.03$$

The value of D's share is

$$7000 \times v^{20} a_{\infty|} = 7000 \times \frac{(1.07)^{-20}}{0.07} = \text{RM}25841.90$$

□

Note that the sum of the shares of B, C and D is equal to RM100,000 as expected. Also note that the present value of the estate at the end of 20 years is $100,000(1.07)^{-20} = 25,841.90$, which is equal to D's share. This confirms the fact that charity D continuing to receive the interest into perpetuity or receiving the estate value in a lump-sum at the end of 20 years are equivalent in value.

10.3.4 Unknown Time

So far, we assume that n and i are known. This section consider the case when n is unknown. However, the n obtained from the equation of value may not be exact integer, so we need to consider what $a_{\overline{n+k}|}$ means, where n is a positive integer and $0 < k < 1$:

$$a_{\overline{n+k}|} = \frac{1-v^{n+k}}{i} = \frac{1-v^n + v^n - v^{n+k}}{i} = a_{\overline{n}|} + v^{n+k} \frac{(1+i)^k - 1}{i} \quad (10.13) \quad \text{e:3.20}$$

This means that

$a_{\overline{n+k}|}$ = the present value of an n -period annuity-immediate of 1 per period
+ a final payment at time $n+k$ of $\frac{(1+i)^k - 1}{i}$.

eg:3.6

Example 10.35. An investment of RM1000 is to be used to make payments of RM100 at the end of every year for as long as possible. If the fund earns an annual effective rate of interest of 5%, find how many regular payments can be made and find the amount of the smaller payment:

- To be paid on the date of the last regular payment.
- To be paid one year after the last regular payment.

- (c) To be paid during the year following the last regular payment, consistent with formula (10.13). eg: 3.20

Solution. Using financial calculator to determine n :

- $I/Y = 5$ (%)
- $PV = 1000$
- $PMT = -100$
- $CPT \& N \Rightarrow$ A value between 14 & 15. Different financial calculator gives a slightly different result.

□

Solution. The equation of value is

$$100a_{\overline{n}|} = 1000 \Rightarrow a_{\overline{n}|} = \frac{1 - v^n}{i} = 10 \Rightarrow 1 - 1.05^{-n} = 0.5 \Rightarrow 1.05^n = \frac{1}{0.5} = 2$$

Hence,

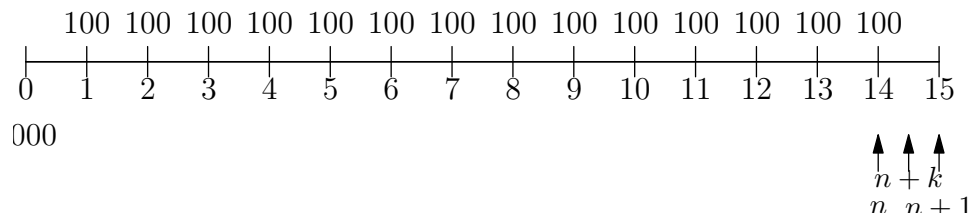
$$n = \frac{\ln 2}{\ln 1.05} = 14.2067.$$

- (a) The equation of value at the end of the 14th year (the date of the last regular payment) is

$$100s_{\overline{14}|} + X_1 = 1000(1.05)^{14} \Rightarrow X_1 = \text{RM}20.07$$

- (b) The equation of value at the end of the 15th year is

$$100\ddot{s}_{\overline{14}|} + X_2 = 1000(1.05)^{15} \Rightarrow X_2 = \text{RM}21.07$$



It should be noted that $X_1(1+i) = X_2$.

- (c) In this case equation of value gives $k = 0.2067$ and

$$X_3 = 100 \frac{1.05^{0.2067} - 1}{0.05} = \text{RM}20.27.$$

□

eg: 3.7

Example 10.36. A fund of RM25000 is to be accumulated by means of deposits of RM1000 made at the end of every year as long as necessary. If the fund earns an effective rate of interest of 8%, find (a) how many regular deposits will be necessary and (b) the size of a final deposit to be made one year after the last regular deposit.

Solution. The equation of value is

$$1000s_{\overline{n}|} = 25000 \Rightarrow \frac{1.08^n - 1}{0.08} = 25 \Rightarrow n = 14.2749$$

Thus, (a) it takes 14 regular deposit plus a smaller final deposit, X .

(b) The equation of value at the end of the 15th year is

$$1000\ddot{s}_{\overline{14}|} + X = 25000 \Rightarrow X = 25000 - 26152 = -\text{RM}1152.$$

□

10.3.5 Unknown Rate of Interest

This section consider the case when the rate of interest, i , is unknown.

By far, the best way to determine an unknown rate of interest for a basic annuity is to use a financial calculator. The calculator internally uses an algorithm for iteration. However, there is a simple approximation formula given $a_{\overline{n}|i} = g$:

$$i \approx \frac{2(n-g)}{g(n+1)}. \quad (10.14) \quad \text{e:3.21}$$

and given $s_{\overline{n}|i} = g$:

$$i \approx \frac{2(g-n)}{g(n-1)}. \quad (10.15) \quad \text{e:3.22}$$

Proof. From $a_{\overline{n}|i} = g$, assuming $i \ll 1$, we have

$$\begin{aligned} \frac{1}{g} &= \frac{1}{a_{\overline{n}|i}} = \frac{i}{1 - (1+i)^{-n}} = \frac{i}{1 - [1 + (-n)i + \frac{(-n)(-n-1)}{2}i^2 + o(i^2)]} \\ &\approx \frac{1}{n} \times \frac{1}{1 - \frac{(n+1)i}{2}} \approx \frac{1}{n} \left[1 + \frac{(n+1)i}{2} + o(i) \right]. \end{aligned}$$

This implies

$$\frac{n-g}{ng} = \frac{1}{g} - \frac{1}{n} \approx \frac{(n+1)i}{2n} \Rightarrow i \approx \frac{2(n-g)}{g(n+1)}.$$

From $s_{\overline{n}|i} = g$,

$$\frac{1}{g} = \frac{1}{s_{\overline{n}|i}} = \frac{1}{a_{\overline{n}|}} - i \approx \frac{1}{n} \left[1 + \frac{(n+1)i}{2} \right] - i$$

This implies

$$\frac{n-g}{ng} \approx \frac{(n+1)i - 2ni}{2n} = \frac{(1-n)i}{2n} \Rightarrow i \approx \frac{2(n-g)}{(1-n)g}.$$

□

eg:3.8

Example 10.37. At what rate of interest, convertible quarterly, is RM16,000 the present value of RM1000 paid at the end of every quarter for 5 years?

Solution. Let $j = \frac{i^{(4)}}{4}$. The equation of value is

$$1000a_{\overline{20}|j} = 1000 \frac{1 - (\frac{1}{1+j})^{20}}{j} = 16,000 \Rightarrow 1 - 16j = \frac{1}{(1+j)^{20}} \Rightarrow f(j) = 1 - 16j - (1+j)^{-20} = 0$$

Using Newton iteration on this, we have

$$j_{n+1} = j_n - \frac{1 - 16j_n - (1+j_n)^{-20}}{-16 + 20(1+j_n)^{-21}}, \quad n = 0, 1, 2, \dots$$

Let $j_0 = 1$ and we can obtain the following numerical iteration to an accuracy of 10^{-6} :

- $j_1 = 0.0624993816014$
- $j_2 = 0.0339007037116$
- $j_3 = 0.0247115477377$

- $j_4 = 0.0224486432071$
- $j_5 = 0.0222636088821$
- $j_6 = 0.0222623113723$
- $j_7 = 0.0222623113084$

Hence, $i^{(4)} = 4(0.0222623113) = 0.089049$ □

Solution. Using a financial calculator, we just need to set

- $N = 20$
- $PV = 16$
- $PMT = -1$
- $CPT \ \& \ I/Y \Rightarrow 2.2262$ (in percentage)

Hence, $i^{(4)} = 4 \times 2.2262/100 = 0.08905$. □

Solution. Using (e:3.21) (10.14), we have $j \approx \frac{2(20-16)}{16 \times 21} = 0.238$. □

In very rare cases, we can solve using algebraic approach.

eg:3.10

Example 10.38. Find the rate of interest at which $\ddot{s}_{\overline{2}|} = 2.5$.

Solution. $\ddot{s}_{\overline{2}|} = (1+i)^2 + (1+i) = 2.5 \Rightarrow i^2 + 3i - 0.5 = 0$. This is a quadratic equation which has the following solution

$$i = \frac{-3 \pm \sqrt{11}}{2} = 0.1583 \quad (\text{only positive result is reasonable})$$

□

10.3.6 Varying Interest

Thus far we have assumed a level rate of interest throughout the term of the annuity. This section consider the situation in which the rate of interest can vary each period, but compound interest is still in effect.

Let i_k denote the rate of interest applicable for period k , i.e. for the interval from time $k-1$ to time k .

There are two patterns of variation for the present value of an n -period annuity-immediate:

- *Portfolio rate method:* i_k is the applicable rate for period k regardless of when the payment is made. In this case, the present value

$$a_{\overline{n}|} = (1+i_1)^{-1} + (1+i_1)^{-1}(1+i_2)^{-1} + \cdots + (1+i_1)^{-1}(1+i_2)^{-1} \cdots (1+i_n)^{-1} = \sum_{t=1}^n \prod_{s=1}^t (1+i_s)^{-1}. \quad (10.16) \quad \text{e:3.23}$$

- *Yield curve method:* i_k is applicable rate for payment made at time k over all k periods. In this case, the present value

$$a_{\overline{n}|} = (1+i_1)^{-1} + (1+i_2)^{-2} + \cdots + (1+i_n)^{-n} = \sum_{t=1}^n (1+i_t)^{-t}. \quad (10.17) \quad \text{e:3.24}$$

Present values of the annuity-due can be obtained from present values of the annuity-immediate by using (e:3.15) (10.10), i.e. $\ddot{a}_{\overline{n}|} = 1 + a_{\overline{n-1}|}$.

Similarly, there are two patterns of variation for the accumulated value of an n -period annuity-due:

- *Portfolio rate method*: The accumulated value,

$$\ddot{s}_{\overline{n}|} = (1 + i_n) + (1 + i_n)(1 + i_{n-1}) + \cdots + (1 + i_n)(1 + i_{n-1}) \cdots (1 + i_1) = \sum_{t=1}^n \prod_{s=1}^t (1 + i_{n-s+1}). \quad (10.18) \quad \text{e:3.25}$$

- *Yield curve method*: The accumulated value,

$$\ddot{s}_{\overline{n}|} = (1 + i_n) + (1 + i_{n-1})^2 + \cdots + (1 + i_1)^n = \sum_{t=1}^n (1 + i_{n-t+1})^t. \quad (10.19) \quad \text{e:3.26}$$

Accumulated values of the annuity-immediate can be obtained from accumulated values of the annuity-due by using (10.10), i.e. $s_{\overline{n+1}|} = 1 + \ddot{s}_{\overline{n}|}$.

eg:3.11

Example 10.39. Find the accumulated value of a 10-year annuity-immediate of RM100 per year if the effective rate of interest is 5% for the first 6 years and 4% for the last 4 years.

Solution. We could apply formula (10.18) but the following approach is simpler:

- Accumulated value for first 6 payments after 6 years = $100s_{\overline{6}|0.05}$
- It is accumulated to the end of the 10 years at 4%, giving

$$100s_{\overline{6}|0.05}(1.04)^4$$

- The accumulated value of the last 4 payments = $100s_{\overline{4}|0.04}$

The accumulated value = $100[s_{\overline{6}|0.05}(1.04)^4 + s_{\overline{4}|0.04}] = 100[(6.8019)(1.16986) + 4.2465] = \text{RM1220.38}$ □

10.3.7 Extra: Annuities Not Involving Compound Interest

Formula (10.4) can be generalised to two different non-equivalent formula

$$a_{\overline{n}|} = \sum_{t=1}^n \frac{1}{a(t)} \quad (10.20) \quad \text{e:3.27}$$

or

$$a_{\overline{n}|} = \sum_{t=0}^{n-1} \frac{a(t)}{a(n)} \quad (10.21) \quad \text{e:3.30}$$

while the accumulated value generalises to two different non-equivalent formula

$$s_{\overline{n}|} = \sum_{t=1}^n \frac{a(n)}{a(t)} \quad (10.22) \quad \text{e:3.28}$$

or

$$s_{\overline{n}|} = \sum_{t=0}^{n-1} a(t). \quad (10.23) \quad \text{e:3.29}$$

10.4 More General Annuities

In the previous topic, the payment period and the interest conversion period of annuities are equal and coincide. This topic introduces the annuities for which payments are made more or less frequently than interest is convertible and annuities with varying payments.

10.4.1 Differing Payment and Interest Conversion Period

There are two approaches to handling annuities for which the payment period and the interest conversion period differ and for which the payments are level.

1. To compute the numerical value of an annuity:
 - (a) Find the rate of interest, convertible at the same frequency as payments are made, that is equivalent to the given rate of interest.
 - (b) Using this new rate of interest, find the value of the annuity using the techniques discussed in the Basic Annuity.
2. Algebraic analysis of an annuity: Develop algebraic expressions for such annuities in terms of annuity symbols, optionally with adjustment factors.

eg:4.1

Example 10.40. Find the accumulated value at the end of 4 years of an investment fund in which RM100 is deposited at the beginning of each quarter for the first two years and RM200 is deposited at the beginning of each quarter for the second two years, if the fund earns 12% convertible monthly.

Solution. We are given an interest rate of 1% per month. Let j be the equivalent rate of interest per quarter, which is the payment period. We have $j = (1.01)^3 - 1 = 0.030301$. The value of the annuity in symbols is

$$100(\ddot{s}_{\overline{16}|j} + \ddot{s}_{\overline{8}|j})$$

which evaluates to $100(20.81704 + 9.17157) = \text{RM}2998.86$ □

eg:4.2

Example 10.41. A loan of RM3000 is to be repaid with quarterly instalments at the end of each quarter for 5 years. If the rate of interest charged on the loan is 10% convertible semiannually, find the amount of each quarterly payment.

Solution. We are given an interest rate of 5% per half-year. Let j be the equivalent rate of interest per quarter, which is the payment period. We have

$$j = (1.05)^{1/2} - 1 = 0.024695.$$

Let the quarter payment be denoted by R . Then the equation of value is

$$Ra_{\overline{5 \times 4}|j} = 3000 \Rightarrow R = \frac{3000}{15.63417} = \text{RM}191.89. \quad \square$$

eg:4.3

Example 10.42. At what annual effective rate of interest will payments of RM100 at the end of every quarter accumulate to RM2500 at the end of 5 years?

Solution. Let $j = i^{(4)}/4$ be the interest rate per quarter which accomplishes the above. Then the equation of value at the end of 5 years is

$$100s_{\overline{20}|j} = 2500 \Rightarrow s_{\overline{20}|j} = 25.$$

Using a financial calculator, we set:

- $N = 20$
- $FV = 25$
- $PMT = -1$
- $\text{CPT } \& \text{ I/Y} \Rightarrow 2.2854.$

Thus $j = 0.022854$. The annual effective rate of interest i is given by $i = (1.022854)^4 - 1 = 0.0946$ or 9.46%. □

10.4.2 Annuities Payable More Frequently Than Interest is Convertible

In practice annuities payable more frequently than interest is convertible are more common than annuities payable less frequently than interest is convertible.

Let m be the number of payment periods in one interest conversion period, let n be the term of the annuity measured in interest conversion periods, and let i be the interest rate per interest conversion period. Assume that each interest conversion period contains an integral number of payment periods, i.e. m and n are both positive integers. The number of annuity payments made is mn .

Example 10.43. • Real estate mortgage is a loan where the collateral is real estate property.

- Knowledge of mortgage terminology is required neither for SOA, nor UECM1403. However, it is a very important example.
- The amount of mortgage is equal to the present value of a series of equal repayments.
- Mortgage can be repaid by monthly, biweekly, weekly instalments.
- In the U.S., the quoted mortgage rate is $i^{(12)}$, the nominal annual rate of interest convertible monthly.
- In Canada, the quoted rate is $i^{(2)}$, the nominal annual rate of interest convertible semiannually.

Annuity-immediate

The present value of an annuity which pays $1/m$ at the end of each m th of an interest conversion period for a total of n interest conversion periods is denoted by $a_{\overline{n}|}^{(m)}$:

$$a_{\overline{n}|}^{(m)} = \frac{1}{m} [v^{1/m} + v^{2/m} + \cdots + v^n] = \frac{1}{m} \frac{v^{1/m} - v^{n+1/m}}{1 - v^{1/m}} = \frac{1 - v^n}{m[(1+i)^{1/m} - 1]} = \frac{1 - v^n}{i^{(m)}} \quad (10.24) \quad \text{e:4.7}$$

The accumulated value of this annuity immediately after the last payment is made is denoted by $s_{\overline{n}|}^{(m)}$:

$$s_{\overline{n}|}^{(m)} = a_{\overline{n}|}^{(m)} (1+i)^n = \frac{(1+i)^n - 1}{i^{(m)}}. \quad (10.25) \quad \text{e:4.8}$$

Formulae (10.24) and (10.25) should be compared to formulae (3.1) and (3.2) in Topic 3, respectively. They are identical except the denominators are $i^{(m)}$ instead of i . Since $i^{(m)}$ is a measure of interest paid at the end of m ths of an interest conversion periods, the points at which interest is paid under this measure are consistent with the points at which payments are made.

Annuity-due

The present value of annuity which pays $1/m$ at the beginning of each m th of an interest conversion period for a total of n interest conversion periods is denoted by $\ddot{a}_{\overline{n}|}^{(m)}$:

$$\ddot{a}_{\overline{n}|}^{(m)} = \frac{1 - v^n}{d^{(m)}} \quad (10.26) \quad \text{e:4.9}$$

The accumulated value of this annuity one m th of an interest conversion periods after the last payment is made is denoted by $\ddot{s}_{\overline{n}|}^{(m)}$:

$$\ddot{s}_{\overline{n}|}^{(m)} = \ddot{a}_{\overline{n}|}^{(m)} (1+i)^n = \frac{(1+i)^n - 1}{d^{(m)}}. \quad (10.27) \quad \text{e:4.10}$$

Other considerations

On occasion a perpetuity payable more frequently than interest is convertible is encountered. The following formulae analogous formulae of (3.8) and (3.9) in Topic 3:

$$a_{\infty|}^{(m)} = \frac{1}{i^{(m)}} \quad (10.28) \quad \text{e:4.11}$$

and

$$\ddot{a}_{\infty|}^{(m)} = \frac{1}{d^{(m)}}. \quad (10.29) \quad \text{e:4.12}$$

A second special case, rarely encountered in practice, is a situation in which each interest conversion period does not contain an integral number of payment periods, i.e. $m > 1$ but $m \notin \mathbb{Z}$. In this case, the best approach would be to resort to basic principles, i.e. to write an expression as the sum of present values or accumulated values of each separate payment, and then to sum this expression as a geometric progression.

One important consideration in practical applications involves the proper coefficients for annuities payable m thly. Each payment made is of amount $1/m$, while the coefficient of the symbol is 1. In general, the proper coefficient is the amount paid during one interest conversion period, and not the amount of each actual payment. The amount paid during one interest conversion period is often called the *annual rent* of the annuity. This is an appropriate term if the interest conversion period is one year, as is often the case; but it is confusing if the interest conversion period is other than one year. Sometimes term *periodic rent* is used instead of “annual rent” to avoid this confusion.

eg:4.6

Example 10.44. A loan of RM5000 is to be repaid with quarterly instalments at the end of each quarter for 5 years. If the rate of interest charged on the loan is 10% convertible semiannually, find the amount of each quarterly payment.

Solution. The rate of interest is 5% per half-year, the term of the loan is ten interest conversion periods, and each interest conversion period contains two payment periods. Thus the equation of value is

$$2Ra_{10|0.05}^{(2)} = 5000 \Rightarrow R \frac{1 - (1 + 0.05)^{-10}}{i^{(2)}} = 2500$$

Since $(1 + \frac{i^{(2)}}{2})^2 = 1.05$, $i^{(2)} = 0.0493901532$,

$$R = \text{RM}319.81$$

□

eg:Ordine

Example 10.45. Suppose that a mortgage of \$100,000 is repaid monthly over 30 years, at the quoted rate of 10%. Calculate the monthly payment for Canada and for the U.S.

Solution. In both cases, the monthly payment is

$$\frac{100,000}{12a_{30|}^{(12)}} = \frac{100,000i^{(12)}}{12(1 - (1 + i)^{-30})}$$

Use the following conversion formulae:

- US: $1 + i = (1 + \frac{i^{(12)}}{12})^{12} = 1.1047130674$
Hence, payment = \$877.571570141
- Canada: $1 + i = (1 + \frac{i^{(2)}}{2})^2 = 1.1025 = (1 + \frac{i^{(12)}}{12})^{12} \Rightarrow i^{(12)} = 0.0979781526$
Hence, payment = \$ $\frac{816.484605}{0.946464476254} = 862.667987531$

□

eg:4.7

Example 10.46. At what annual effective rate of interest is the present value of a series of payments of RM1 every six months forever, with the first payments made immediately, equal to RM10?

Solution. The equation of value is

$$10 = 1 + v^{0.5} + v^1 + v^{1.5} + \dots = \frac{1}{1 - v^{0.5}} \Rightarrow (1 + i)^{-0.5} = 0.9$$

Solving it gives $i = 0.23456790123$, i.e. 23.46%. □

10.4.3 Optional: Annuities Payable Less Frequently Than Interest is Convertible

Annuity-Immediate

Let k be the number of interest conversion periods in 1 payment period, n be term of the annuity measure in interest conversion periods and i be the rate of interest per interest conversion period.

Assume that each payment period contains an integral number of interest conversion periods, i.e. $k, n \in \mathbb{Z}^{*+}$ and the number of annuity payments made is $n/k \in \mathbb{Z}^{*+}$.

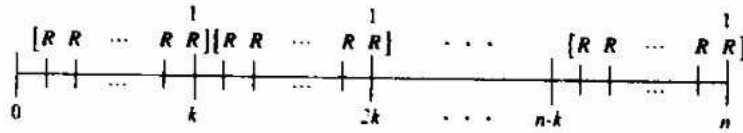
The present value of an annuity which pays 1 at the end of each k interest conversion periods for a total of n interest conversion periods is

$$v^k + v^{2k} + \dots + v^{n/k \cdot k} = \frac{v^k - v^{n+1}}{1 - v^k} = \frac{1 - v^n}{(1 + i)^k - 1} = \frac{a_{\overline{n}|}}{s_{\overline{k}|}}.$$

The accumulated value of this annuity immediately after the last payment is

$$\frac{a_{\overline{n}|}}{s_{\overline{k}|}}(1 + i)^n = \frac{s_{\overline{n}|}}{s_{\overline{k}|}}$$

Note that $\frac{1}{s_{\overline{k}|}}$ can be interpreted as a payment R s.t. the series of payments of 1 at the end of each k interest conversion periods for n interest conversion periods can be replaced by a series of payments of R at the end of each interest conversion period so that the present value are equal.



Annuity-Due

The present value of an annuity that pays at the beginning of each k interest conversion periods for a total of n interest conversion periods is

$$1 + v^k + v^{2k} + \dots + v^{n-k} = \frac{1 - v^n}{1 - v^k} = \frac{a_{\overline{n}|}}{a_{\overline{k}|}}$$

Accumulated value of this annuity k interest conversion periods after the last pays is

$$\frac{a_{\overline{n}|}}{a_{\overline{k}|}}(1 + i)^n = \frac{s_{\overline{n}|}}{s_{\overline{k}|}}.$$

Example 10.47. An investment of RM1000 is used to make payments of RM100 at the end of each year for as long as possible with a smaller final payment to be made at the time of the last regular payment. If interest is 7% convertible semiannually, find the number of payments and the amount of the total final payment.

Solution. This is an annuity-immediate with

$$100 \frac{a_{\overline{n}|}}{s_{\overline{2}|}} = 1000.$$

Since $i = 0.035$, we have $a_{\overline{n}|} = 10s_{\overline{2}|}20.35$. Solving it we have

$$\frac{1 - 1.035^{-n}}{0.035} = 20.35 \Rightarrow n = 36.21.$$

Thus, we have 18 regular payments and a smaller final payment can be made. Let the smaller additional payment at the time of the final regular payment be R . Then, an equation of value at the end of 18 years is

$$R + 100 \frac{s_{\overline{36}|}0.035}{s_{\overline{2}|}0.035} = 1000(1.035)^{36} \Rightarrow R = 10.09.$$

Thus the total final payment would be RM110.09. □

10.4.4 Continuous Annuities

A special case of annuities payable more frequently than interest is convertible is one in which the frequency of payment becomes infinite, i.e. payments are made continuously. This is useful as an approximation to annuities payable with great frequency, such as daily.

We will denote the present value of an annual payable continuously for n interest conversion periods, such that total amount paid during each interest conversion period is 1, by the symbol

$$\bar{a}_{\overline{n}|} = \int_0^n v^t dt = \left[\frac{v^t}{\ln v} \right]_0^n = \frac{1 - v^n}{\delta}, \quad (10.30) \quad \text{e:4.14}$$

where the differential $v^t dt$ is the present value of the payment dt made at exact moment t .

Note that (10.30) can have been obtained as follows:

$$\bar{a}_{\overline{n}|} = \lim_{m \rightarrow \infty} a_{\overline{n}|}^{(m)} = \frac{1 - v^n}{\lim_{m \rightarrow \infty} i^{(m)}} = \frac{1 - v^n}{\delta}$$

or

$$\bar{a}_{\overline{n}|} = \lim_{m \rightarrow \infty} \ddot{a}_{\overline{n}|}^{(m)} = \frac{1 - v^n}{\lim_{m \rightarrow \infty} d^{(m)}} = \frac{1 - v^n}{\delta}.$$

Thus, the continuous annuity is seen to be the limiting case of annuities payable m thly as $m \rightarrow \infty$.

The accumulated value of a continuous annuity at the end of the term of the annuity is denoted by

$$\bar{s}_{\overline{n}|} = \int_0^n (1 + i)^t dt = \frac{(1 + i)^n - 1}{\delta} = \lim_{m \rightarrow \infty} s_{\overline{n}|}^{(m)} = \lim_{m \rightarrow \infty} \ddot{s}_{\overline{n}|}^{(m)} \quad (10.31) \quad \text{e:4.16}$$

Both (10.30) and (10.31) can be expressed in terms of force of interest as follows:

$$\bar{a}_{\overline{n}|} = \frac{1 - e^{-n\delta}}{\delta} \quad (10.32) \quad \text{e:4.19}$$

$$\bar{s}_{\overline{n}|} = \frac{e^{n\delta} - 1}{\delta} \quad (10.33) \quad \text{e:4.20}$$

eg:4.8

Example 10.48. Find the force of interest at which $\bar{s}_{\overline{20}|} = 3\bar{s}_{\overline{10}|}$.

Solution. From (10.33),

$$\frac{e^{20\delta} - 1}{\delta} = 3 \frac{e^{10\delta} - 1}{\delta} \Rightarrow e^{20\delta} - 3e^{10\delta} + 2 = 0 \Rightarrow e^{10\delta} = \frac{3 \pm \sqrt{9 - 4(2)}}{2} = 2, 1$$

A reasonable solution is $\delta = 0.069314718056$. □

10.4.5 Increasing and Decreasing Annuities

Thus far all the annuities considered have had level payments. This section assumed that the payment period and interest conversion period are equal and coincide but payments vary.

Any type of *varying annuity* can be evaluated by taking the present value or the accumulated value of each payment separately and summing the results. On occasion, this may be the only feasible approach. However, there are two types of commonly encountered varying annuities for which relatively simple expressions can be developed:

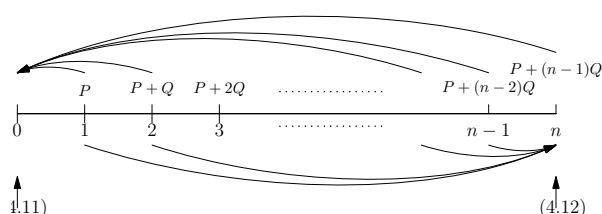
- annuities with payments varying in arithmetic progression
- annuities with payments varying in geometric progression.

In practical applications the value of varying annuities in which other payment patterns exist are generally best handled from basic principles. The annuity value can be computed by finding the present value or accumulated value of each payment separately and summing the results.

Remark 10.49. Beware the distinction between “varying annuity” and “variable annuity”. A *variable annuity* is a type of life annuity in which payments vary according to the investment experience of an underlying investment account, usually one invested in common stocks. The varying annuities considered here have variations which are specified in advance and *do not depend on some external event*.

..... Arithmetic Progression

Consider first a general annuity-immediate with a term of n periods in which payments begin at P and increase by Q per period thereafter. The interest rate is i per period.



Note that P must be positive, Q can be either positive or negative as long as $P + (n - 1)Q > 0$ to avoid negative payments.

The present value of the annuity, PV , is

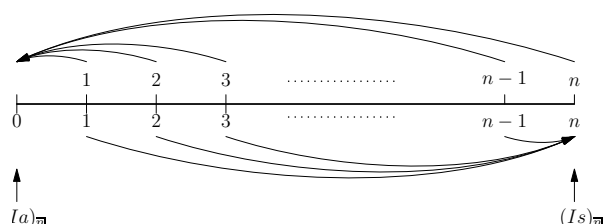
$$\begin{aligned} PV &= Pv + (P + Q)v^2 + \cdots + [P + (n - 1)Q]v^n \\ &= P \frac{1 - v^n}{i} + Q \frac{a_{\overline{n}|} - nv^n}{i} = Pa_{\overline{n}|} + Q \frac{a_{\overline{n}|} - nv^n}{i} \end{aligned} \quad (10.34) \quad \text{e:4.21}$$

The accumulated value is given by

$$AV = Ps_{\overline{n}|} + Q \frac{s_{\overline{n}|} - n}{i}. \quad (10.35) \quad \text{e:4.22}$$

Formulae (10.34) and (10.35) can be used in solving any problem in which payments vary in arithmetic progression. However, there are two special cases which often appear and have special notation.

- When $P = Q = 1$, we have an *increasing annuity*.



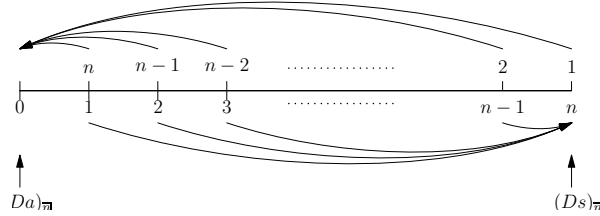
The present value is

$$(Ia)_{\overline{n}|} = a_{\overline{n}|} + \frac{a_{\overline{n}|} - nv^n}{i} = \frac{\ddot{a}_{\overline{n}|} - nv^n}{i} \quad (10.36) \quad \text{e:4.23}$$

The accumulated value is

$$(Is)_{\overline{n}|} = (Ia)_{\overline{n}|}(1+i)^n = \frac{\ddot{s}_{\overline{n}|} - n}{i}. \quad (10.37) \quad \text{e:4.24}$$

- When $P = n$, $Q = -1$, we have a *decreasing annuity*.



The present value is

$$(Da)_{\overline{n}|} = na_{\overline{n}|} - \frac{a_{\overline{n}|} - nv^n}{i} = \frac{n - a_{\overline{n}|}}{i} \quad (10.38) \quad \text{e:4.25}$$

The accumulated value is

$$(Ds)_{\overline{n}|} = (Da)_{\overline{n}|}(1+i)^n = \frac{n(1+i)^n - s_{\overline{n}|}}{i}. \quad (10.39) \quad \text{e:4.26}$$

Remark 10.50. All the above formulae are for annuities-immediate. Changing i in the denominator of any of the above formulae to d will produce values for annuities-due.

To calculate varying perpetuities, we take the limit of the perpetuity-immediate (e:4.21):

$$\lim_{n \rightarrow \infty} PV = P \lim_{n \rightarrow \infty} a_{\overline{n}|} + Q \lim_{n \rightarrow \infty} \frac{a_{\overline{n}|} - nv^n}{i} = \frac{P}{i} + \frac{Q}{i^2} \quad (10.40) \quad \text{e:4.27}$$

since $\lim_{n \rightarrow \infty} a_{\overline{n}|} = \frac{1}{i}$ and $\lim_{n \rightarrow \infty} nv^n = 0$.

eg:4.9

Example 10.51. Find the present value of a perpetuity-immediate whose successive payments are 1, 2, 3, 4, ..., at an effective rate of interest of 5%.

Solution. From (e:4.27), $PV = \frac{1}{i} + \frac{1}{i^2} = \frac{1}{0.05} + \frac{1}{0.05^2} = 420$. □

eg:4.10

Example 10.52. Find the present value of an annuity-immediate such that payments start at 1, increase by annual amounts of 1 to a payment of n , and then decrease by annual amounts of 1 to a final payment of 1.

Solution. The present value is

$$\begin{aligned} (Ia)_{\overline{n}|} + v^n (Da)_{\overline{n-1}|} &= \frac{\ddot{a}_{\overline{n}|} - nv^n}{i} + v^n \frac{(n-1) - a_{\overline{n-1}|}}{i} \\ &= \frac{1}{i} [a_{\overline{n-1}|} + 1 - nv^n + nv^n - v^n - v^n a_{\overline{n-1}|}] \\ &= \frac{1}{i} [a_{\overline{n-1}|}(1 - v^n) + (1 - v^n)] = \frac{1}{i} (1 - v^n)(a_{\overline{n-1}|} + 1) \\ &= a_{\overline{n}|} \times \ddot{a}_{\overline{n}|}. \end{aligned}$$

□

eg:4.11

Example 10.53. Find the present value of an annuity-immediate such that payments start at 1, each payment thereafter increases by 1 until reaching 10, and then remain at that level until 25 payments in total are made.

Solution. The present value is

$$(Ia)_{\overline{10}|} + 10v^{10}a_{\overline{15}|}.$$

Other valid expressions are $(Ia)_{\overline{25}|} - v^{10}(Ia)_{\overline{15}|}$, $10a_{\overline{25}|} - (Da)_{\overline{9}|}$ and $\sum_{t=0}^9 v^t a_{25-t}$. $\square$

..... Geometric Progression

Consider an annuity-immediate with a term of n periods in which the first payments is 1 and successive payments increase in geometric progression with common ratio $1 + k$.

The present value is

$$PV = v + v^2(1 + k) + \cdots + v^n(1 + k)^{n-1} = v \frac{1 - (\frac{1+k}{1+i})^n}{1 - (\frac{1+k}{1+i})} = \frac{1 - (\frac{1+k}{1+i})^n}{i - k} \quad (10.41) \quad \text{e:4.29}$$

and the accumulated value is

$$AV = (1 + i)^{n-1} + (1 + i)^{n-2}(1 + k) + \cdots + (1 + i)(1 + k)^{n-2} + (1 + k)^{n-1} = \frac{(1 + i)^n - (1 + k)^n}{i - k}. \quad (10.42) \quad \text{e:4.33}$$

In the case of annuity-due, the present value is

$$\begin{aligned} PV &= 1 + v(1 + k) + \cdots + v^{n-1}(1 + k)^{n-1} = 1 + \frac{1 + k}{1 + i} + \cdots + \left(\frac{1 + k}{1 + i}\right)^{n-1} \\ &= 1 + \frac{1}{1 + i'} + \cdots + \frac{1}{(1 + i')^{n-1}} = \ddot{a}_{\overline{n}|i'} \end{aligned} \quad (10.43) \quad \text{e:4.30}$$

where $i' = \frac{i-k}{1+k}$; the accumulated value is

$$AV = \ddot{a}_{\overline{n}|i'}(1 + i)^n. \quad (10.44) \quad \text{e:4.35}$$

When $k < i$, the present value of perpetuity-immediate is

$$PV = \frac{1}{i - k}, \quad (10.45) \quad \text{e:4.36}$$

the present value of perpetuity-due is

$$PV = \frac{1 + i}{i - k}. \quad (10.46) \quad \text{e:4.37}$$

When $k > i$, the infinite geometric progression diverges and the present value of the perpetuity does not exist.

eg:4.12

Example 10.54. An annuity provides for 20 annual payments, the first payment a year hence being RM1000. The payments increase in such a way that each payment is 4% greater than the preceding payment. Find the present value of this annuity at an annual effective rate of interest of 7%.

Solution. The present value of this annuity is

$$1000 \left[\frac{1}{1.07} + \frac{1.04}{1.07^2} + \cdots + \frac{1.04^{19}}{1.07^{20}} \right] = \frac{1000}{1.07} \times \frac{1 - (\frac{1.04}{1.07})^{20}}{1 - \frac{1.04}{1.07}} = \text{RM}14459.07$$

$\square$

eg:4.13

Example 10.55. A perpetuity-due makes annual payments which begin at RM100 for the first year, then increase at 6% per year through the 10th year, and then remain level thereafter. Calculate the present value of this perpetuity, if the annual effective rate of interest is equal to 8%.

Solution. The present value of the first 10 payments is

$$100 \left[1 + \frac{1.06}{1.08} + \cdots + \left(\frac{1.06}{1.08}\right)^9 \right] = 100 \frac{1 - (\frac{1.06}{1.08})^{10}}{1 - \frac{1.06}{1.08}} = 920.65$$

The present value of the rest of the payments is

$$100(1.06)^9 \left[\frac{1}{1.08^{10}} + \frac{1}{1.08^{11}} + \frac{1}{1.08^{12}} + \cdots \right] = 1056.45$$

The present value of the perpetuity is $920.65 + 1056.45 = \text{RM}1977.10$. □

eg:4.14

Example 10.56. A participant in a pension plan aged 45 currently earns RM50,000 per year. This participant plans to contribute 5% of salary at the end of each year into the pension fund for 20 years. Find the accumulated value at age 65 of these contributions, if the participant receives annual 3% salary increases until retirement and if the pension fund earns 7% effective per year.

Solution. The accumulated value of the contribution made by this participant at the end of 20 years, i.e. at the age of 65, is

$$\begin{aligned} & 50000 \times 0.05 \times [1.07^{19} + (1.03)(1.07)^{18} + \cdots + (1.03)^{18}(1.07) + (1.03)^{19}] \\ &= 2500 \frac{1.07^{20} - 1.03^{20}}{1.07 - 1.03} = 128973.33. \end{aligned}$$

□

10.4.6 More General Varying Annuities

The varying annuities considered in the previous section assumed that the payment period and the interest conversion period are equal and coincide. In practice, varying annuities with payments made more or less frequently than interest is convertible occur infrequently.

Consider the case in which payments are made less frequently than interest is convertible. Let k be the number of interest conversion periods in one payment period, let n be the term of the annuity measured in interest conversion periods, and let i be the rate of interest per interest conversion period. The number of payments is an integer $\frac{n}{k}$.

The present value of the generalised increasing annuity is

$$PV = v^k + 2v^{2k} + \cdots + \left(\frac{n}{k} - 1\right)v^{n-k} + \frac{n}{k}v^n$$

Multiply PV by $(1+i)^k$ and subtracting by the above equation results in

$$PV[(1+i)^k - 1] = 1 + v^k + v^{2k} + \cdots + v^{n-k} - \frac{n}{k}v.$$

This implies

$$PV = \frac{\frac{a_{\overline{n}|}}{a_{\overline{k}|}} - \frac{n}{k}v^n}{is_{\overline{k}|}}. \quad (10.47) \quad \text{e:4.38}$$

Consider the case in which payments are made more frequently than interest is convertible. If the rate of payment is constant during each interest conversion period with increases occurring only once per interest conversion period, then

$$(Ia)_{\overline{n}|}^{(m)} = \frac{\ddot{a}_{\overline{n}|} - nv^n}{i^{(m)}}. \quad (10.48) \quad \text{e:4.39}$$

If the rate of payment changes with each payment period, i.e. an increases annuity is payable at the rate of $1/m$ per interest conversion period at the end of the first m th of an interest conversion period, $2/m$ period interest conversion period at the end of the second m th of an interest conversion period, etc. Then the first payment will be $1/m^2$, the second will be $2/m^2$, etc. The present value, $(I^{(m)}a)_{\overline{n}|}^{(m)}$, is

$$(I^{(m)}a)_{\overline{n}|}^{(m)} = \frac{1}{m^2} [v^{1/m} + 2v^{2/m} + \cdots + nmv^{nm/m}] = \frac{\ddot{a}_{\overline{n}|}^{(m)} - nv^n}{i^{(m)}}. \quad (10.49) \quad \text{e:4.40}$$

Annuities in which the payments vary in geometric progression and in which the payment period and the interest conversion period differ are occasionally encountered. They can be handled by expressing the annuity value as a summation of the present value or accumulated value of each payment. This technique will be illustrated below.

eg:4.15

Example 10.57. Find the present value of a perpetuity which pays 1 at the end of the third year, 2 at the end of the sixth year, 3 at the end of the ninth year, etc.

Solution. $PV = v^3 + 2v^6 + 3v^9 + \dots$
 $v^3 PV = v^6 + 2v^9 + 3v^{12} + \dots$
 $(1 - v^3)PV = v^3 + v^6 + v^9 + \dots = \frac{v^3}{1 - v^3}$
Hence, $PV = \frac{v^3}{(1 - v^3)^2}$.

□

eg:4.16

Example 10.58. Find the accumulated value at the end of 10 years of an annuity in which payments are made at the beginning of each half-year for 5 years. The first payment is RM2000, and each payment is 98% of the prior payment. Interest is credited at 10% convertible quarterly.

Solution. The accumulated value is

$$\begin{aligned} & 2000[(1.025)^{40} + (0.98)(1.025)^{38} + (0.98)^2(1.025)^{36} + \dots + (0.98)^9(1.025)^{22}] \\ &= 2000(1.025)^{22}[(1.025)^{18} + (0.98)(1.025)^{16} + (0.98)^2(1.025)^{14} + \dots + (0.98)^9] \\ &= 2000(1.025)^{22} \frac{1.025^{20} - 0.98^{10}}{1.025^2 - 0.98} = \text{RM}40052.28 \end{aligned}$$

□

10.4.7 Continuous Varying Annuities

The last type of varying annuity is one in which payments are being made continuous at a varying rate. Such annuities are primarily of theoretical interest.

Consider an increasing annuity for n interest conversion periods in which payments are being made continuously at the rate of t per period at exact moment t . The present value of this annuity, $(\bar{I}a)_{\overline{n}|}$, is

$$(\bar{I}a)_{\overline{n}|} = \int_0^n tv^t dt \quad (10.50) \quad \text{e:4.41}$$

It can be simplified to by using integration by part

$$(\bar{I}a)_{\overline{n}|} = \left[\frac{tv^t}{\ln v} \right]_0^n - \left[\frac{v^t}{(\ln v)^2} \right]_0^n = \frac{\bar{a}_{\overline{n}|} - nv^n}{\delta} \quad (10.51) \quad \text{e:4.42}$$

or by taking limit of (10.49) ^{e:4.40}

$$(\bar{I}a)_{\overline{n}|} = \lim_{m \rightarrow \infty} (I^{(m)}a)_{\overline{n}|}^{(m)} = \frac{\lim_{m \rightarrow \infty} \ddot{a}_{\overline{n}|}^{(m)} - \lim_{m \rightarrow \infty} nv^n}{\lim_{m \rightarrow \infty} i^{(m)}} = \frac{\bar{a}_{\overline{n}|} - nv^n}{\delta}.$$

In general, if the amount of payment being made at exact moment t is $f(t)dt$ and the interest varies continuously by a force of interest δ_r , then an expression for the present value of an n -period continuous varying annuity would be

$$\int_0^n f(t)e^{-\int_0^t \delta_r dr} dt. \quad (10.52) \quad \text{e:4.44}$$

eg:4.18

Example 10.59. Find an expression for the present value of a continuously increasing annuity with a term of n years if the force of interest is δ and if the rate of payment at time t is t^2 per annum.

Solution. According to (10.52) and applying integration by part, the present value is

$$\int_0^n t^2 e^{-\delta t} dt = \left[\frac{t^2 e^{-\delta t}}{-\delta} - \frac{2te^{-\delta t}}{\delta^2} + \frac{2e^{-\delta t}}{-\delta^3} \right]_0^n = \frac{2}{\delta^3} - e^{-\delta n} \left(\frac{n^2}{\delta} + \frac{2n}{\delta^2} + \frac{2}{\delta^3} \right).$$

□

10.4.8 Summary

The following table summarises the expressions for present values of level annuities for various payment periods and interest conversion periods. The annuity illustrated pays 60 per annum for 10 years at 12% per annum.

Interest conversion period	Payment period			
	Annual	Quarterly	Monthly	Continuous
Annual	$60a_{\overline{10} 0.12}$	$60a_{\overline{10} 0.12}^{(4)}$	$60a_{\overline{10} 0.12}^{(12)}$	$60\bar{a}_{\overline{10} 0.12}$
Quarterly	$60\frac{a_{\overline{40} 0.03}}{s_{\overline{4} 0.03}}$	$15a_{\overline{40} 0.03}$	$15a_{\overline{40} 0.03}^{(3)}$	$15\bar{a}_{\overline{40} 0.03}$
Monthly	$60\frac{a_{\overline{120} 0.01}}{s_{\overline{12} 0.01}}$	$15\frac{a_{\overline{120} 0.01}}{s_{\overline{3} 0.01}}$	$5a_{\overline{120} 0.01}$	$5\bar{a}_{\overline{120} 0.01}$
Continuous	$60\frac{1-e^{-1.2}}{e^{0.12}-1}$	$15\frac{1-e^{-1.2}}{e^{0.03}-1}$	$5\frac{1-e^{-1.2}}{e^{0.01}-1}$	$60\frac{1-e^{-1.2}}{0.12}$

Refer to “The Theory of Interest (Third Edition)” by Kellison, S. G. (2009) for the topics on “Amortisation Schedules and Sinking Funds”, “Bonds”, “Yield Rates” and “Practical Applications”.

10.5 Computations with Calculators and Excel

10.5.1 BA II Plus

Calculator Formats

You can change these calculator formats:

To Select	Press	Display	Default
Number of decimal places	2ND [FORMAT]	DEC 09 (Press 9 for floating-decimal)	2
Angle units	↓	DEG (degrees) RAD (radians)	DEG
Dates	↓	US (mm-dd-yyyy) Eur (dd-mm-yyyy)	US
Number separators	↓	US (1,000.00) Eur (1.000,00)	US
Calculation method	↓	Chn (chain) AOS (algebraic operating system)	Chn

In the chain (Chn) calculation mode, the calculator solves problems in the order you enter them:

$$3 + 2 * 4 = (3 + 2) * 4 = 20.$$

Note that Scientific Calculators use AOS, i.e.

$$3 + 2 * 4 = 3 + (2 * 4) = 11.$$

To reset default values for all formats, press **2ND** [CLR WORK]

Using Worksheets

The calculator contains worksheets with embedded formulae to solve specific problems. You apply settings or assign known values to worksheet variables and then compute the unknown value. Changing the values lets you ask what if questions and compare results.

Except for TVM variables, accessed in the standard-calculator mode, all variables are *prompted*. For example, to assign values to amortization variables, you must first press **2ND** **[AMORT]** to access the Amortization worksheet.

Each worksheet is independent of the others: operations in a worksheet do not affect variables in other worksheets. When you exit a worksheet or turn off the calculator, the calculator retains all worksheet data.

To select	Function	Press
TVM worksheet	Analyses equal cash flows, for example, annuities, loans, mortgages, leases, and savings	N , I/Y , PV , PMT , FV , or 2ND [P/Y]
Amortization worksheet	Performs amortization calculations and generates an amortization schedule	2ND [AMORT]
Cash Flow worksheet	Analyses unequal cash flows by calculating net present value and internal rate of return	2ND [CF]
Bond worksheet	Computes bond price and yield to maturity or call	2ND [BOND]
Depreciation worksheet	Generates a depreciation schedule using one of six depreciation methods	2ND [DEPR]
Statistics worksheet	Analyses statistics on one- or two-variable data using four regression analysis options	2ND [STAT]
Percent Change/Compound Interest worksheet	Computes percent change, compound interest, and cost-sell markup	2ND [Δ%]
Interest Conversion worksheet	Converts interest rates between nominal rate (or annual percentage rate) and annual effective rate	2ND [ICONV]
Date worksheet	Computes number of days between two dates, or date/day of the week a specified number of days is from a given date	2ND [DATE]
Profit Margin worksheet	Computes cost, selling price, and profit margin	2ND [PROFIT]
Breakeven worksheet	Analyses relationship between fixed cost, variable cost, price, profit, and quantity	2ND [BRKEVN]
Memory worksheet	Accesses storage area for up to 10 values	2ND [MEM]

To return to STANDARD MODE, press **2ND** **[QUIT]**.

Time-Value-of-Money and Amortization Worksheets

Variable	Key	Display	Type of Variable
Number of periods	N	N	Enter-or-compute
Interest rate per year	I/Y	I/Y	Enter-or-compute
Present value	PV	PV	Enter-or-compute
Payment	PMT	PMT	Enter-or-compute
Future value	FV	FV	Enter-or-compute

Number of payments per year	2ND [P/Y]	P/Y	Enter-only
Number of compounding periods per year	↓	C/Y	Enter-only
End-of-period payments	2ND [BGN]	END	Setting
Beginning-of-period payments	2ND [SET]	BGN	Setting
Starting payment	2ND [AMORT]	P1	Enter-only
Ending payment	↓	P2	Enter-only
Balance	↓	BAL	Auto-compute
Principal paid	↓	PRN	Auto-compute
Interest paid	↓	INT	Auto-compute

To *reset*, press **2ND** **[RESET]** **ENTER**.

Time Value of Money formula:

$$\begin{aligned} \text{PMT} = 0: & \quad i = \sqrt[N]{FV/PV} - 1 \\ \text{PMT} \neq 0: & \quad i = \exp[y \times \ln(1 + x)] - 1 \end{aligned}$$

where

- $y = C/Y \div P/Y$
- $x = (0.01 \times I/Y) \div C/Y$
- C/Y = compounding periods per year
- P/Y = payment periods per year
- I/Y = interest rate per year.

The iteration used to compute i :

$$0 = PV + PMT \times G_i \left[\frac{1 - (1 + i)^{-N}}{i} \right] + FV \times (1 + i)^{-N}$$

where $G_i = 1 + i \times k$, $k = 0$ for end-of-period payments, $k = 1$ for beginning-of-period payments.

eg:p30

Example 10.60. If you make a monthly payment of RM425.84 on a 30-year mortgage for RM75,000, what is the interest rate on your mortgage?

Solution. (A) Equation of value

$$425.84a_{\overline{30 \times 12}|j} = 75,000 \Rightarrow j =$$

(B) Using calculator,

- 12 payment per year: **2ND** **[P/Y]** 12 **ENTER**
- standard mode: **2ND** **[QUIT]**
- number of years: 30 **2ND** **[xP/Y]** **N**
- loan amount: 75000 **PV**
- payment amount: 425.84 **+/-** **PMT**

- compute interest rate: $\boxed{\text{CPT}} \boxed{\text{I/Y}} \Rightarrow 5.50(\% \text{ per year})$

□

Amortization Algorithm:

INPUT: PMT, PV, I/Y, PMT_1 and PMT_2

If computing BAL, $PMT_2 = NPMT$

Let $BAL_0 = RND(PV)$

From $m=1$ to PMT_2 :

$$I_m = RND(RND12(-i \times BAL_{m-1}))$$

$$BAL_m = BAL_{m-1} - I_m + RND(PMT)$$

$$BAL = BAL_{PMT_2}$$

$$\text{Sum of Prn} = BAL_{PMT_2} - BAL_{PMT_1}$$

$$\text{Sum of Int} = (PMT_2 - PMT_1 + 1) \times RND(PMT) - \text{Sum of Prn}$$

where

- RND = round the display to the number of decimal places selected
- $RND12$ = round to 12 decimal places

eg:p31

Example 10.61. Compute basic loan payments on a \$75,000 mortgage at 5.5% for 30 years.

- Computing Monthly Payments

- $\boxed{2ND} \boxed{P/Y} \boxed{12} \boxed{ENTER} \quad \boxed{2ND} \boxed{QUIT}$
- $30 \boxed{2ND} \boxed{X P/Y} \boxed{N}$
- $5.5 \boxed{I/Y}$
- $75000 \boxed{PV}$
- $\boxed{CPT} \boxed{PMT} \Rightarrow -425.84.$

- Computing Quarterly Payments

- $\boxed{2ND} \boxed{P/Y} \boxed{12} \boxed{ENTER} \quad \boxed{2ND} \boxed{QUIT} \dots$

eg:p35

Example 10.62. The ABC Company purchased a machine that will save these end-of-year amounts:

Year	1	2	3	4
Amount	\$5000	\$7000	\$8000	\$10000

Given a 10% discount rate, does the present value of the cash flows exceed the original cost of \$23,000?

- $\boxed{2ND} \boxed{RESET} \boxed{ENTER}$
- $10 \boxed{I/Y} \quad -5000 \boxed{FV} \quad 1$

eg:p40

Example 10.63. You consider buying a car for \$15,100. The finance company charges 7.5% APR compounded monthly on a 48-month loan. If you can afford a monthly payment of \$325, how much can you borrow? How much do you need for a down payment?

To Press Display	
Set all variables to defaults.	RST 0.00
Set payments per year to 12.	P/Y= 12.00
Return to standard-calculator mode	0.00
Enter number of payments using payment multiplier	4 N= 48.00
Enter interest rate.	7.5 - I/Y= 7.50
Enter payment.	325 S / PMT= -325.00
Compute loan amount.	PV= 13,441.47
Compute down payment	H 15,100 S N -1,658.53

Answer: You can borrow \$13,441.47 with a down payment of \$1,658.53.

eg:p42

Example 10.64. Calculate the monthly payments on a 30-year loan and generate an amortization schedule for the first three years of the loan.

1. Calculate the monthly payment with a loan amount of \$120,000 and 6.125% APR.

- Set all variables to defaults: RST 0.00
- Set payments per year to 12: P/Y= 12.00
- Return to standard-calculator mode.
- Enter number of payments using payment multiplier.
- Enter interest rate
- Enter loan amount
- Compute payment

Answer: The computed monthly payment, or outflow, is \$729.13.

2. Generate an amortization schedule for the first three years of the loan. If the first payment is in April, the first year has nine payment periods. (Following years have 12 payment periods each.)

- Select the Amortization worksheet: **2ND**[AMORT]
- Set beginning period to 1: 1**ENTER** P1=1.00
- Set ending period to 9: **↓** 9**ENTER** P2=9.00
- Display 1st year amortization data: BAL= 118,928.63* **↓** PRN= -1071.37* **↓** INT= -5,490.80*
- Change beginning period to 10
- Change ending period to 21
- Display 2nd year amortization data BAL= 117,421.60* PRN= -1,507.03* INT= -7,242.53*
- Move to P1 and press **CPT** to enter next range of payments:
- Display P2 P2= 33.00
- Display 3rd year amortization data. BAL= 115,819.62* PRN= -1601.98 INT= -7,147.58*

eg:p43

Example 10.65. A group of sellers considers financing the sale price of a property for \$82,000 at 7% annual interest, amortized over a 30-year term with a balloon payment due after five years. They want to know:

- Amount of the monthly payment
- Amount of interest they will receive
- Remaining balance at the end of the term (balloon payment)

1. Computing the Monthly Payment

- Set all variables to defaults
- Set payments per year to 12
- Return to standard-calculator mode.
- Enter number of payments using payment multiplier.
- Enter interest rate
- Enter loan amount
- Compute payment

2. Generating an Amortization Schedule for Interest and Balloon Payment

- Select Amortization worksheet.
- Enter end period (five years)
- View balance due after five years (balloon payment).
- View interest paid after five years:

If the sellers financed the sale, they would receive:

- Monthly payment: \$545.55 for five years
- Interest: \$27,790.72 over the five years
- Balloon payment: \$77,187.72

Cash Flow Worksheet

Use the Cash Flow worksheet to solve problems with “unequal cash flows”. To solve problems with “equal cash flows”, use the Time-Value-of-Money and Amortization Worksheets.

- To clear Cash Flow Worksheet: **2ND**[CLR WORK]
- To access the Cash Flow worksheet and initial cash flow value (CFo), press **CF**.
- To access the cash flow amount and frequency variables (Cnn/Fnn), press **↓** or **↑**.
- To access the discount rate variable (I), press **NPV**.
- To compute net present value (NPV), net future value (NFV), payback (PB), and discounted payback (DPB), press **↓** or **↑** and **CPT** for each variable.
- To compute the internal rate of return (IRR), press **IRR**.
- To compute the modified internal rate of return (MOD), press **↓** to access the reinvestment rate (RI) variable, key in a value, and press **↓**.

Cash Flow Worksheet Variables

Variable	Key	Display	Variable Type
Initial cash flow	CF	CFo	Enter-only
Amount of <i>n</i> th cash flow	↓	C[01–24/32]	Enter-only
Frequency of <i>n</i> th cash flow	↓	F[01–24/32]	Enter-only
Discounted rate	NPV	I	Enter-only

Net present value	↓ CPT	NPV	Compute-only
Internal rate of return	IRF CPT	IRR	Compute-only

Cash flows consist of an initial cash flow (CFo) and up to 24 (32 for professional BAI+), each of which can have a unique value. You must enter the number of occurrences (up to 9,999), or frequency (F), for each additional cash flow (C01-C24).

eg:p51

Example 10.66. A company pays \$7,000 for a new machine, plans a 20% annual return on the investment, and expects these annual cash flows over the next six years:

Year	Cash Flow Number	Cash Flow Estimate
Purchase	CFo	−\$7,000
1	C01	3,000
2–5	C02	5,000 each year
6	C03	4,000

As the time line shows, the cash flows are a combination of equal and unequal values. As an outflow, the initial cash flow (CFo) appears as a negative value.

Solution.

- Select Cash Flow worksheet: CFo=0.00
- Enter initial cash flow: 7000 CFo=-7,000.00
- Enter cash flow for first year: 3000 C01=3,000.00 F01=1.00
- Enter cash flows for years 2–5: 5000 4 C02=5,000.00 F02=4.00
- Enter cash flow for sixth year: 4000 C03=4,000.00 F03=1.00

Remark:

- To insert or delete cash flow: or .
- Computing NPV: 20 7625.99
- Computing IRR:

□

Bond Worksheet

Price (given yield) with one coupon period or less to redemption:

$$PRI = \frac{RV + \frac{100R}{M}}{1 + \frac{DSR \times Y}{E \times M}} - \frac{A}{E} \times \frac{100R}{M}$$

where

- PRI = dollar price per \$100 par value
- RV = redemption value of the security per \$100 par value (note that $RV=100$ except in those cases where call or put features must be considered)

•

Bond Worksheet Variables

Variable	Key	Display	Variable Type
Settlement date	2nd [BOND]	SDT	Enter only
Annual coupon rate in percent	↓	CPN	Enter only
Redemption date	↓	RDT	Enter only
Redemption value (percentage of par value)	↓	RV	Enter only
Actual/actual day-count method	↓	ACT	Setting
30/360 day-count method	2nd [SET]	360	Setting
Two coupons per year	↓	2/Y	Setting
One coupon per year	2nd [SET]	1/Y	Setting
Yield to redemption	↓	YLD	Enter/compute
Dollar price	↓	PRI	Enter/compute
Accrued interest	↓	AI	Auto-compute

Other Worksheets: Depreciation, Statistics, Date, etc.

To convert between nominal and effective interest rate:

eg:p78

Example 10.67. A bank offers a certificate that pays a nominal interest of 15% with quarterly compounding. What is the annual effective interest rate?

Solution. • Select interest convert worksheet: **2ND** [ICONV]

- Enter nominal interest rate: 15 **ENTER** (NOM = 15.00)
- Enter number of compounding periods per year: **↓ ↓** 4 **ENTER** (C/Y = 4.00)
- Compute annual effective interest rate: **↑** **CPT** $\Rightarrow$ 15.87.

Using formula: $i = (1 + \frac{0.15}{4})^4 - 1 = 0.15865041504$ □

To find the number of days between two dates:

10.5.2 Introduction to Excel Financial Functions

To work on.

10.5.3 Setting loan amortization model on Excel

To work on.

10.5.4 Calculate the value of stock, given the pattern of dividends and yield rates

10.5.5 Calculate net present value, given the investment contributions and investment returns

The syntax for the Npv function is:

Npv(discount_rate, value1, value2, ... value_n)

discount_rate is the discount rate for the period.

value1, value2, ... value_n are the future payments and income for the investment (ie: cash flows). There can be up to 29 values entered.

This first example returns a net present value of \$3,457.19. It assumes that you pay \$7,500 as an initial investment . You then receive the following income for the first four years (EOP): \$3,000, \$5,000, \$1,200, and \$4,000. An annual discount rate of 8% is used.

=NPV(8%, 3000, 5000, 1200, 4000) - 7500

This next example returns a net present value of \$8,660.77. It assumes that you pay \$10,000 as an initial investment. You then receive the following income for the first three years (BOP): \$3,400, \$6,500, and \$10,000. An annual discount rate of 5% is used.

=NPV(5%, 6500, 10000) - 10000 + 3400

Chapter 11

Complex Numbers

Historically, complex numbers are introduced by mathematicians so that the quadratic equations $Ax^2 + Bx + C = 0$ can have solutions even if $B^2 < 4AC$.

Mathematicians found that complex number system (consists of the numbers and the arithmetic operations) is a field with some nice property compare to real numbers. They also discovered that it is possible to study “calculus” over complex-valued functions. With the complex-valued “calculus” tools, mathematicians were able to build a model for hyperbolic geometry and to find a real-valued function $u(x, y)$ that satisfies the two-dimensional Laplace equation, $u_{xx} + u_{yy} = 0$ in the xy -plane for engineering problems describing the behaviour of electrostatic potentials, of equilibrium temperatures, etc.

There are a large number of books on complex function theory such as Churchill and Brown [2008], Saff and Snider [2003], Ahlfors [1979] (a relatively abstract but good book), Rudin [1987] (an extremely abstract book, it is only suitable for a person with highly abstract thinking skill), Freitag and Busam [2005] (theoretical, the exponential function is defined by power series), Marsden [1973], Bak and Newman [2010], Ablowitz and Fokas [2003] (has interesting topics like Painlevé equations, asymptotic evaluations of integrals, and Riemann-Hilbert problems), Conway [1978] (a theoretical but dry book, Conway [1995] is the sequel), Lang [1999] and Berenstein and Gay [1991].

11.1 Algebraic Structure of $\mathbb{C}$

To study complex analysis, we must learn the set of complex numbers and the operations we can do with complex numbers. In this section, we will learn the algebraic operations for complex numbers. In the next section, we will learn the notion of “length” of complex numbers. After that, we will learn the notion of “angles” of complex numbers and its close relation to trigonometry. In the last two sections, we will investigate the “topological” properties of the set of complex numbers and a few basic plane geometrical objects.

Definition 11.1. Let $\mathbb{R}$ be the set of all real numbers (Definition [3.11](#)). A complex number is defined as order pair $z = (x, y)$ of real numbers x and y . The set of all complex numbers is denoted as $\mathbb{C}$. The set of all **non-zero** complex numbers is denoted as $\mathbb{C}^\times$.

Definition 11.2. Let $z = (x, y)$. The *real part* and *imaginary part* of z are

$$x = \Re(z), \quad y = \Im(z)$$

respectively. We denote $(0, 0)$ and $(1, 0)$ as 0 and 1 respectively.

Definition 11.3. Two complex numbers $z_1 = (x_1, y_1)$ and $z_2 = (x_2, y_2)$ are said to be *equal* if $x_1 = x_2$ and $y_1 = y_2$. In other words,

$$z_1 = z_2 \Leftrightarrow \Re(z_1) = \Re(z_2) \wedge \Im(z_1) = \Im(z_2).$$

def:conj

Definition 11.4. The (*complex*) *conjugate* of a complex number $z = (x, y)$ is defined to be

$$\bar{z} = (x, -y).$$

11.1.1 Algebraic Operations of $\mathbb{C}$ and Their Properties

The arithmetic operations of real numbers can be generalised to complex numbers but the calculations are slightly more complicated as follows.

def:complex

Definition 11.5. Let $z = (x, y)$, $z_1 = (x_1, y_1)$ and $z_2 = (x_2, y_2)$ be complex numbers. We define the arithmetic operations of complex numbers as follows:

1. Negation: $-z_1 = (-x, -y)$;
2. Addition: $z_1 + z_2 = (x_1 + x_2, y_1 + y_2)$;
3. Subtraction: $z_1 - z_2 = (x_1 - x_2, y_1 - y_2)$;
4. Multiplication: $z_1 \cdot z_2 = (x_1x_2 - y_1y_2, y_1x_2 + x_1y_2)$;
5. Division: $\frac{z_1}{z_2}$ or $z_1/z_2 = \left(\frac{x_1x_2 + y_1y_2}{x_2^2 + y_2^2}, \frac{x_2y_1 - x_1y_2}{x_2^2 + y_2^2} \right)$.

By the definition and the properties of real numbers, the addition and multiplication are closed.

What concerns us with programming language is that “complex number” calculations are carried out with floating point numbers (Section 40.3.1) and there are round-off errors. So we must realise that “complex number” in any programming language is just an finite precision approximation to the elements of $\mathbb{C}$ rather than the exact representation of $\mathbb{C}$.

By using the definition of the arithmetic operations, we can obtain the basic algebraic properties of complex numbers:

thm:complex-prop

Theorem 11.1. For any $z_1, z_2, z_3, z \in \mathbb{C}$,

1. *Commutative Law for Addition:* $z_1 + z_2 = z_2 + z_1$
2. *Associative Law for Addition:* $(z_1 + z_2) + z_3 = z_1 + (z_2 + z_3)$,
3. *Additive Identity:* $z + 0 = 0 + z = z$;
4. *Existence of additive inverse:* For any z_1 , there is a number z_2 such that $z_1 + z_2 = 0$;
5. *Commutative Law for Multiplication:* $z_1z_2 = z_2z_1$;
6. *Associative Law for Multiplication:* $(z_1z_2)z_3 = z_1(z_2z_3)$;
7. *Multiplicative Identity:* $z \cdot 1 = 1 \cdot z = z$;
8. *Existence of multiplicative inverse:* For any non-zero complex number z_1 , there is a number z_2 such that $z_1 \cdot z_2 = 1$;
9. *Distributive Law:* $z(z_1 + z_2) = zz_1 + zz_2$.

Proof: Let $z_1 = (x_1, y_1)$, $z_2 = (x_2, y_2)$, $z_3 = (x_3, y_3)$ and $z = (x, y)$, where $x_i, y_i, x, y \in \mathbb{R}$.

1. $z_1 + z_2 = (x_1 + x_2, y_1 + y_2) = (x_2 + x_1, y_2 + y_1) = z_2 + z_1$.
2. $(z_1 + z_2) + z_3 = (x_1 + x_2, y_1 + y_2) + (x_3, y_3) = (x_1 + (x_2 + x_3), y_1 + (y_2 + y_3)) = (x_1, y_1) + (x_2 + x_3, y_2 + y_3) = z_1 + (z_2 + z_3)$.
3. $z + 0 = (x, y) + (0, 0) = (x + 0, y + 0) = (x, y) = z$.
4. For any z_1 , let $z_2 = (-x_1, -y_1)$. Then

$$z_1 + z_2 = (x_1, y_1) + (-x_1, -y_1) = (x_1 + (-x_1), y_1 + (-y_1)) = (0, 0) = 0.$$

5. $z_1 \cdot z_2 = (x_1x_2 - y_1y_2, y_1x_2 + x_1y_2) = (x_2x_1 - y_2y_1, x_2y_1 + y_2x_1) = (x_2x_1 - y_2y_1, y_2x_1 + x_2y_1) = z_2 \cdot z_1$.

$$\begin{aligned} 6. \quad (z_1 \cdot z_2) \cdot z_3 &= (x_1x_2 - y_1y_2, y_1x_2 + x_1y_2) \cdot (x_3, y_3) \\ &= \left(\begin{vmatrix} x_1 & y_1 \\ y_2 & x_2 \end{vmatrix}, \begin{vmatrix} y_1 & -x_1 \\ y_2 & x_2 \end{vmatrix} \right) \cdot (x_3, y_3) \\ &= \left(\begin{vmatrix} x_1 & y_1 \\ y_2 & x_2 \end{vmatrix} x_3 - \begin{vmatrix} y_1 & -x_1 \\ y_2 & x_2 \end{vmatrix} y_3, \begin{vmatrix} y_1 & -x_1 \\ y_2 & x_2 \end{vmatrix} x_3 + \begin{vmatrix} x_1 & y_1 \\ y_2 & x_2 \end{vmatrix} y_3 \right) \\ &= (x_1x_2x_3 - y_1y_2x_3 - y_1x_2y_3 - x_1y_2y_3, y_1x_2x_3 + x_1y_2x_3 + x_1x_2y_3 - y_1y_2y_3) \\ &= \left(x_1 \begin{vmatrix} x_2 & y_2 \\ y_3 & x_3 \end{vmatrix} - y_1 \begin{vmatrix} y_2 & -x_2 \\ y_3 & x_3 \end{vmatrix}, y_1 \begin{vmatrix} x_2 & y_3 \\ y_2 & x_3 \end{vmatrix} + x_1 \begin{vmatrix} x_2 & -y_2 \\ x_3 & y_3 \end{vmatrix} \right) \\ &= (x_1, y_1) \cdot \left(\begin{vmatrix} x_2 & y_2 \\ y_3 & x_3 \end{vmatrix}, \begin{vmatrix} y_2 & -x_2 \\ y_3 & x_3 \end{vmatrix} \right) = z_1 \cdot (z_2 \cdot z_3) \end{aligned}$$

7. $z \cdot 1 = (x, y) \cdot (1, 0) = (x \cdot 1 - y \cdot 0, x \cdot 0 + y \cdot 1) = (x, y) = z$.

8. For any non-zero $z_1 = (x_1, y_1)$, $x_1^2 + y_1^2 \neq 0$ and there exists a complex number

$$z_2 = \left(\frac{x_1}{x_1^2 + y_1^2}, \frac{-y_1}{x_1^2 + y_1^2} \right)$$

such that

$$z_1 \cdot z_2 = \left(x_1 \times \frac{x_1}{x_1^2 + y_1^2} - y_1 \times \frac{-y_1}{x_1^2 + y_1^2}, y_1 \times \frac{x_1}{x_1^2 + y_1^2} + x_1 \times \frac{-y_1}{x_1^2 + y_1^2} \right) = (1, 0) = 1.$$

9. $z(z_1 + z_2) = (x, y) \cdot (x_1 + x_2, y_1 + y_2) = (x(x_1 + x_2) - y(y_1 + y_2), y(x_1 + x_2) + x(y_1 + y_2)) = (xx_1 + xx_2 - yy_1 - yy_2, yx_1 + yx_2 + xy_1 + xy_2) = (xx_1 - yy_1, yx_1 + xy_1) + (xx_2 - yy_2, yx_2 + xy_2) = (x, y) \cdot (x_1, y_1) + (x, y) \cdot (x_2, y_2) = zz_1 + zz_2$.

In the language of modern algebra, the first four items are saying that $(\mathbb{C}, +)$ is an *Abelian group* (or *commutative group*). Items 5 to 8 say that $(\mathbb{C}^\times, \cdot)$ is an Abelian group. All the 9 items say that $(\mathbb{C}, +, \cdot)$ is a *field*.

Example 11.2. Find (a) the additive inverse and (b) the multiplicative inverse of $(5, 12)$.

Solution.

(a) additive inverse = $(-5, -12)$.

(b) multiplicative inverse = $\left(\frac{5}{5^2 + 12^2}, \frac{-12}{5^2 + 12^2} \right) = \left(\frac{5}{169}, -\frac{12}{169} \right)$. □

Since $(\mathbb{C}, +, \cdot)$ is a field, we can formally define the concept of power and derive the binomial formula.

Definition 11.6. Given any complex number z and any non-negative integer n , the (n th) power of z is defined recursively as

$$z^0 = 1, \quad z^n = z \cdot z^{n-1}, \quad \text{for } n \geq 1.$$

Theorem 11.3. Let z_1 and z_2 be any complex numbers. Then the **binomial formula** is valid:

$$(z_1 + z_2)^n = \sum_{k=0}^n \binom{n}{k} z_1^k z_2^{n-k}.$$

What makes $\mathbb{C}$ “algebraically” different from $\mathbb{R}$ as a field is that $\mathbb{C}$ has the “conjugate” operation (Definition 11.4) and its properties are characterised by Theorem 11.4.

Theorem 11.4. Let z, z_1 and z_2 be any complex numbers.

1. $\bar{\bar{z}} = z$;
2. $\overline{z_1 \pm z_2} = \bar{z}_1 \pm \bar{z}_2$;
3. $\overline{z_1 \cdot z_2} = \bar{z}_1 \cdot \bar{z}_2$;
4. If $z_2 \neq 0$, $\overline{z_1/z_2} = \bar{z}_1/\bar{z}_2$;

Proof. Let $z = (x, y)$, $z_i = (x_i, y_i)$, $i = 1, 2$.

1. $\bar{\bar{z}} = \overline{(x, -y)} = (x, -(-y)) = (x, y) = z$.
2. $\overline{z_1 \pm z_2} = \overline{(x_1 \pm x_2, y_1 \pm y_2)} = (x_1 \pm x_2, -(y_1 \pm y_2)) = (x_1 \pm x_2, (-y_1) \pm (-y_2)) = (x_1, -y_1) \pm (x_2, -y_2) = \bar{z}_1 \pm \bar{z}_2$
3. $\overline{z_1 \cdot z_2} = \overline{(x_1x_2 - y_1y_2, y_1x_2 + x_1y_2)} = (x_1x_2 - y_1y_2, -(y_1x_2 + x_1y_2)) = (x_1x_2 - (-y_1)(-y_2), (-y_1)x_2 + x_1(-y_2)) = (x_1, -y_1) \cdot (x_2, -y_2) = \bar{z}_1 \cdot \bar{z}_2$.
4. $\overline{z_1/z_2} = \overline{\left(\frac{x_1x_2 + y_1y_2}{x_2^2 + y_2^2}, \frac{x_2y_1 - x_1y_2}{x_2^2 + y_2^2}\right)} = \left(\frac{x_1x_2 + y_1y_2}{x_2^2 + y_2^2}, -\frac{x_2y_1 - x_1y_2}{x_2^2 + y_2^2}\right) = \left(\frac{x_1x_2 + (-y_1)(-y_2)}{x_2^2 + (-y_2)^2}, \frac{x_2(-y_1) - x_1(-y_2)}{x_2^2 + (-y_2)^2}\right) = \frac{(x_1, -y_1)}{(x_2, -y_2)} = \frac{\bar{z}_1}{\bar{z}_2}$

□

11.1.2 The i Notation

With the arithmetic operations in Definition 11.5, we can write

$$(x, y) = (x, 0) + (0, y) = (x, 0) + (0, 1)(y, 0). \quad (11.1)$$

Under the complex arithmetic operation, the set of complex numbers of the form $(x, 0)$ “behaves” just like $\mathbb{R}$. Thus the real number system can be embedded into the complex system and we shall regard the real number system as a subset of the complex number system and the number $(x, 0)$ will be simply denoted as x .

With this in mind, it is possible to prove

Theorem 11.5. For any complex number $z = (x, y)$, $z\bar{z} = x^2 + y^2$.

Proof. $z\bar{z} = (x, y) \cdot (x, -y) = (x^2 - y(-y), x(-y) + yx) = (x^2 + y^2, 0) = x^2 + y^2$. $\square$

Remark 11.6. If we are not allow to write a complex number $(x, 0)$ as a real number, then we will have to write the above equation as $z\bar{z} = (x^2 + y^2, 0)$ which is concise but not cumbersome and it is not demonstrating the special status of complex numbers with zero imaginary part.

By setting $i = (0, 1)$ and consider $(x, 0)$ and $(y, 0)$ as “embedded real numbers”, i.e. we can embed $x \in \mathbb{R}$ into $\mathbb{C}$ through map $\iota: \mathbb{R} \rightarrow \mathbb{C}$, $x \mapsto (x, 0)$. Thus we shall think x as $(x, 0)$ and with the notation of $i = (0, 1)$, it is easy to write $(\text{e:}x+iy)$ as

$$z = (x, y) = x + iy.$$

which was first introduced by Gauss. This is the usual definition use by many complex analysis textbooks and many programming language. We did not start with this definition because it may be confusing mathematically: How can we add a real number x with and imaginary number iy ? With the way we define complex number, we know that $x + iy$ is just a convenient notation for complex number as defined in (II.1).

Notice that the derivation of i is firmly based on real numbers. There is no magic about i contrary to the use of the historical word “imaginary”. Because engineers and scientists normally use i to denote “current”, so they use j to represent the so called imaginary unit $(0, 1)$.

Example 11.7. Write down the output of the Scheme program:

```
1 (define z1 -1+2i)
2 (define z2 3-4i)
3 (define (conj z) (make-rectangular (real-part z1) (- (imag-part z1))))
4 (display (string-append "z1 = " (number->string z1) "\n"))
5 (display (string-append "z2 = " (number->string z2) "\n"))
6 (display (string-append "Re(z1) = " (number->string (real-part z1)) "\n"))
7 (display (string-append "Im(z1) = " (number->string (imag-part z1)) "\n"))
8 (display (string-append "|z1| = " (number->string (magnitude z1)) "\n"))
9 (display (string-append "Arg(z1) = " (number->string (angle z1)) "\n"))
10 (display (string-append "Conj(z1) = " (number->string (conj z1)) "\n"))
11 (display (string-append "z1+z2 = " (number->string (+ z1 z2)) "\n"))
12 (display (string-append "z1-z2 = " (number->string (- z1 z2)) "\n"))
13 (display (string-append "z1*z2 = " (number->string (* z1 z2)) "\n"))
14 (display (string-append "z1/z2 = " (number->string (/ z1 z2)) "\n"))
```

Solution:

```
z1 = -1+2i
z2 = 3-4i
Re(z1) = -1
Im(z1) = 2
|z1| = 2.23606797749979
Arg(z1) = 2.0344439357957027
Conj(z1) = -1-2i
z1+z2 = 2-2i
z1-z2 = -4+6i
z1*z2 = 5+10i
z1/z2 = -11/25+2/25i
```

Example 11.8 (Oon Shea Ming). Express $\frac{(2+3i)(5-2i)}{(1-3i)(3+4i)}$ in the form $a + bi$.

Solution.

$$\begin{aligned}\frac{(2+3i)(5-2i)}{(1-3i)(3+4i)} &= \frac{19+11i}{15-5i} \\ &= \frac{(19+11i)(3+i)}{5(3-i)(3+i)} \\ &= \frac{37}{50} + \frac{49}{50}i\end{aligned}$$

□

By treating a real number as a complex number and by treating a complex number as a pair of real numbers, we are able to find various link between them.

It is possible to related the real and imaginary parts and the complex conjugate as follows.

Theorem 11.9. *Let z be a complex number. Then*

$$\Re(z) = \frac{z + \bar{z}}{2}, \quad \Im(z) = \frac{z - \bar{z}}{2i}.$$

By using Theorem [thm:conj](#) 11.4, it is possible for us to show that a polynomial of real coefficients has complex conjugate roots.

Definition 11.7. A *solution* of an equation $f(z) = 0$ over $\mathbb{C}$ is a value $z_0 \in \mathbb{C}$ such that $f(z_0) = 0$. A *root* of $f(z)$ is a solution of the equation $f(z) = 0$.

Example 11.10. Let $a_i \in \mathbb{R}$ be some constants. If z_0 is a solution of the following equation

$$a_0 z^n + a_1 z^{n-1} + \cdots + a_{n-1} z + a_n = 0. \quad (11.2)$$

[e:polyn](#)

Show that $\bar{z}_0$ is also a solution of the equation.

Solution. Since

$$a_0 z_0^n + a_1 z_0^{n-1} + \cdots + a_{n-1} z_0 + a_n = 0,$$

taking conjugate to LHS and RHS and by applying Theorem [thm:complex-prop](#) 11.1, we obtain

$$a_0 \bar{z}_0^n + a_1 \bar{z}_0^{n-1} + \cdots + a_{n-1} \bar{z}_0 + a_n = 0$$

because $\overline{a_i} = a_i$. This implies $\bar{z}_0$ is a solution of [e:polyn](#) (11.2).

□

Although $\mathbb{R}$ can be considered as a proper subset of $\mathbb{C}$. However, the total ordering property of $\mathbb{R}$ cannot be extended to the complex set $\mathbb{C}$. For any pair of real numbers a and b , the total ordering property means only one of the property below is true:

$$a < b, \quad a = b, \quad \text{or} \quad a > b.$$

This ordering property is compatible with the algebraic properties of $\mathbb{R}$ making $\mathbb{R}$ an *ordered field*.

Definition 11.8. A field $(F, +, \cdot)$ together with a total order $\leq$ on F is an *ordered field* if the order satisfies the following properties:

1. if $a \leq b$ then $a + c \leq b + c$;
2. if $0 \leq a$ and $0 \leq b$ then $0 \leq ab$.

Example 11.11. Show that it is impossible to have a total order for $\mathbb{C}$.

Proof: Suppose it is possible, then either (a) $i = 0$, (b) $0 < i$ or (c) $i < 0$.

(a) It is impossible for $i = 0$ because this implies $(0, 1) = (0, 0) \Rightarrow 1 = 0$. A contradiction.

(b) Suppose $0 < i$, then $i \times 0 < i \times i \Rightarrow 0 < -1$. We cannot say this is a contradiction because this is an ordering in complex number, it may be possible.

So we have $0 < i$, $0 < -1$ and so $0 < -i$ because if $0 < c$, $ac < bc$.

However, in total ordering, $a < b \Rightarrow a + c < b + c$. Therefore, $0 < -i \Rightarrow 0 + i < -i + i \Rightarrow i < 0$. A contradiction because $0 < i$ and $i < 0$ in total ordering.

(c) Suppose $i < 0$, then $i \times i < 0 \times i \Rightarrow -1 < 0$. We may think that this is OK, but let multiply it by i and we obtain $-i < 0$. Adding i to both sides leads to $i + (-i) = 0 < 0 + i = i$. Again, we have $i < 0$ and $0 < i$, a contradiction.

So it is impossible to have a total order on $\mathbb{C}$.

11.1.3 Matrix Representation

A complex number $z = (x, y)$ can also be written in matrix form: $M(z) = \begin{bmatrix} x & -y \\ y & x \end{bmatrix}$. It is easy to verify that the addition and subtraction in this case are similarly to Definition [11.5](#) and have the properties mentioned in Theorem [11.1](#). As for multiplication, it is similar as follows:

$$M(z_1)M(z_2) = \begin{bmatrix} x_1 & -y_1 \\ y_1 & x_1 \end{bmatrix} \begin{bmatrix} x_2 & -y_2 \\ y_2 & x_2 \end{bmatrix} = \begin{bmatrix} x_1x_2 - y_1y_2 & -(x_1y_2 + x_2y_1) \\ x_1y_2 + x_2y_1 & x_1x_2 + y_1y_2 \end{bmatrix} = M(z_1z_2)$$

The conjugate operation corresponds to matrix transpose:

$$M(\bar{z}) = M((x, -y)) = \begin{bmatrix} x & -(-y) \\ -y & x \end{bmatrix} = \begin{bmatrix} x & -y \\ y & x \end{bmatrix}^T = M(z)^T.$$

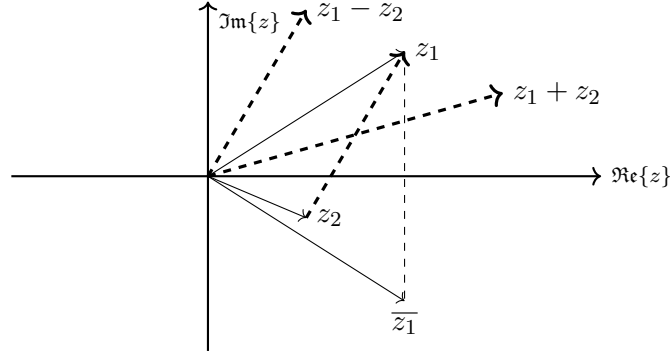
11.2 Geometric Interpretation and Triangle Inequality

Since complex numbers are order pairs, so a complex number $(x, y) \in \mathbb{C}$ can be represented as a point $A(x, y)$ on the Euclidean plane $\mathbb{R}^2$. The plane that represents all the complex numbers will be called *complex plane*, *z-plane* or *Argand diagram*. If we further investigate the “vector space” properties of $\mathbb{C}$, we will find the $\mathbb{C}$ can be viewed as a vector space over $\mathbb{R}$, i.e.

- $(\mathbb{C}, +)$ is a Abelian group;
- For any $z \in \mathbb{C}$, $1 \cdot z = z$;
- For any $\alpha, \beta \in \mathbb{R}$ and any $z \in \mathbb{C}$, $\alpha(\beta z) = (\alpha\beta)z$;
- For any $\alpha, \beta \in \mathbb{R}$ and any $z \in \mathbb{C}$, $(\alpha + \beta)z = \alpha z + \beta z$;
- For any $\alpha \in \mathbb{R}$ and any $z_1 + z_2 \in \mathbb{C}$, $\alpha(z_1 + z_2) = \alpha z_1 + \alpha z_2$.

With this we are able to provide a geometric interpretation for some of the arithmetic operations of $\mathbb{C}$: For any $z_1, z_2 \in \mathbb{C}$,

- $z_1 + z_2$ is the diagonal of parallelogram formed by z_1 and z_2 .
- $z_1 - z_2$ is the third side of triangle formed by z_1 and z_2 .
- The point $\bar{z}_1$ is the reflection of z_1 about the real axis.



A vector space can be enriched into a *normed vector space* by the introduction of the notion of “length”.

Definition 11.9. The *modulus* of a complex number $z = x + iy$ is defined to be

$$|z| = \sqrt{x^2 + y^2}.$$

Geometrically, the modulus $|z|$ is the length of the “vector” z or the distance between the “point” z and the origin in the complex plane.

Example 11.12. Evaluate $|3 - 2i|$.

Solution: $|3 - 2i| = \sqrt{3^2 + (-2)^2} = \sqrt{13}$.

Theorem 11.13 (Properties of Modulus). *For any $z, z_1, z_2 \in \mathbb{C}$,*

1. $\Re(z) \leq |z|, \Im(z) \leq |z|$
2. $|z|^2 = z \cdot \bar{z}$
3. $|z_1 z_2| = |z_1| |z_2|$
4. $|z_1 + z_2| \leq |z_1| + |z_2|$ (*Triangle Inequality*)
5. $||z_1| - |z_2|| \leq |z_1 - z_2|$

Proof: Let $z = x + iy, z_1 = x_1 + iy_1, z_2 = x_2 + iy_2$

1. $\Re(z) = x \leq \sqrt{x^2 + y^2} = |z|; \Im(z) = y \leq \sqrt{x^2 + y^2} = |z|.$
2. $|z|^2 = x^2 + y^2 = (x + iy)(x - iy) = z \cdot \bar{z}$
3. $|z_1 z_2| = |x_1 x_2 - y_1 y_2 + i(x_1 y_2 + x_2 y_1)| = \sqrt{(x_1 x_2 - y_1 y_2)^2 + (x_1 y_2 + x_2 y_1)^2}$
 $= \sqrt{x_1^2 x_2^2 - 2x_1 x_2 y_1 y_2 + y_1^2 y_2^2 + x_1^2 y_2^2 + 2x_1 y_2 x_2 y_1 + x_2^2 y_1^2}$
 $= \sqrt{x_1^2 x_2^2 + y_1^2 y_2^2 + x_1^2 y_2^2 + x_2^2 y_1^2} = \sqrt{(x_1^2 + y_1^2)(x_2^2 + y_2^2)} = |z_1| |z_2|$
4. $|z_1 + z_2|^2 = (x_1 + x_2)^2 + (y_1 + y_2)^2 = x_1^2 + x_2^2 + 2x_1 x_2 + y_1^2 + y_2^2 + 2y_1 y_2$
 $= |z_1|^2 + |z_2|^2 + 2(x_1 x_2 + y_1 y_2)$ (*)

We need to show that

$$x_1 x_2 + y_1 y_2 \leq |z_1| |z_2|. \quad (\dagger)$$

Since $(x_1^2 + y_1^2)(x_2^2 + y_2^2) - (x_1 x_2 + y_1 y_2)^2$
 $= x_1^2 x_2^2 + x_1^2 y_2^2 + y_1^2 x_2^2 + y_1^2 y_2^2 - x_1^2 x_2^2 - y_1^2 y_2^2 - 2x_1 x_2 y_1 y_2$
 $= x_1^2 y_2^2 + x_2^2 y_1^2 - 2x_1 x_2 y_1 y_2 = (x_1 y_2 - x_2 y_1)^2 \geq 0$, therefore, we have $(\dagger)$.

Substituting (†) into (*), we have

$$|z_1 + z_2|^2 \leq |z_1|^2 + |z_2|^2 + 2|z_1||z_2| = (|z_1| + |z_2|)^2 \Rightarrow \text{Triangle inequality.}$$

5. By using triangle inequality, we have

$$\begin{aligned} |z_2| &= |z_1 + (z_2 - z_1)| \leq |z_1| + |z_2 - z_1| \Rightarrow |z_2| - |z_1| \leq |z_2 - z_1|, \\ |z_1| &= |z_2 + (z_1 - z_2)| \leq |z_2| + |z_1 - z_2| \Rightarrow |z_1| - |z_2| \leq |z_1 - z_2|. \end{aligned}$$

Therefore,

$$|z_2| - |z_1| \leq |z_2 - z_1| \Rightarrow -|z_2 - z_1| \leq |z_1| - |z_2| \leq |z_1 - z_2| = |z_2 - z_1|.$$

Use the knowledge you learn in UECM1303 Discrete Mathematics and Applications to prove the following statement.

Example 11.14. Prove that $|z_1 + z_2 + \cdots + z_n| \leq |z_1| + |z_2| + \cdots + |z_n|$ for any positive integer n and any $z_i \in \mathbb{C}$.

Solution. Base case: For any $z_1 \in \mathbb{C}$, $|z_1| \leq |z_1|$.

Assume that for any $z_1, \dots, z_n \in \mathbb{C}$, $|z_1 + z_2 + \cdots + z_n| \leq |z_1| + |z_2| + \cdots + |z_n|$.

For any $z_1, \dots, z_n, z_{n+1} \in \mathbb{C}$, by triangle inequality and induction hypothesis, we have

$$\begin{aligned} |z_1 + z_2 + \cdots + z_n + z_{n+1}| &\leq |z_1 + z_2 + \cdots + z_n| + |z_{n+1}| \\ &\leq |z_1| + |z_2| + \cdots + |z_n| + |z_{n+1}|. \end{aligned}$$

Hence, the statement is true by induction. $\square$

11.3 Polar Form and Roots

For any non-zero complex number, it is possible to assign a unique polar coordinate (r, θ) to a complex number.

def:polar

Definition 11.10. The *polar form* of the complex number $z = (x, y)$ is

$$(r, \theta) : \begin{cases} x = r \cos \theta, \\ y = r \sin \theta. \end{cases}$$

The *trigonometric form* of z is

$$z = r(\cos \theta + i \sin \theta).$$

where $r = |z|$ and the angle $\theta = \arg(z)$ (in radian) is called the *argument* of z which satisfies:

$$(2k - 1)\pi < \arg(z) \leq 2(k + 1)\pi, \quad k = 0, \pm 1, \pm 2, \dots \quad (11.3)$$

e:argument

Note that $\cos \theta + i \sin \theta$ is so basic that Conway [1978] introduced the notion “cis” for it.

Geometrically, $\arg(z)$ represents the angle included between the line segment joining the origin and the point $z = (x, y)$ and the positive x -axis. Because of the “periodicity” of trigonometric functions in Definition 11.10, there are infinitely many values of $\arg(z)$, each differing from the others by multiples of 2π as indicated by (11.3). The “principal value” of $\arg(z)$, denoted by $\text{Arg}(z)$ is the particular value of $\arg(z)$ chosen within the principal interval $(-\pi, \pi]$. Thus we can write

$$\arg(z) = \text{Arg}(z) + 2k\pi, \quad k = 0, \pm 1, \pm 2, \dots$$

Here $\text{Arg}(z)$ is called the *principal value of the argument of z* or simply *principal argument of z* which is defined to be the angle such that

$$\tan(\text{Arg}(z)) = \frac{y}{x}, \quad -\pi < \text{Arg}(z) \leq \pi.$$

It can be calculated as follows:

$$\text{Arg}(z) = \begin{cases} \text{Arctan}(\frac{y}{x}), & x > 0, \\ \text{Arctan}(\frac{y}{x}) + \pi, & x < 0, y \geq 0, \\ \text{Arctan}(\frac{y}{x}) - \pi, & x < 0, y < 0. \end{cases}$$

This was the definition use by the Julia's commands `angle()` and `atan2(x, y)`.

eg:arg1

Example 11.15. Use scientific calculator to calculate the output of the Julia commands: (a) `angle(1-1im)`, (b) `angle(-1)` and (c) `angle(-1-2im)`.

Solution. (a) $\text{angle}(1-1im) = \text{Arctan}(\frac{-1}{1}) = -\frac{\pi}{4} = -0.7853981633974483$

(b) $\text{angle}(-1) = \text{Arctan}(\frac{0}{-1}) + \pi = \pi = 3.141592653589793$

(c) $\text{angle}(-1-2im) = \text{Arctan}(\frac{-2}{-1}) - \pi = -2.0344439357957027$

□

In order to derive a property of the argument, consider the product of any non-zero complex numbers z_1 and z_2 in trigonometric form:

$$\begin{aligned} z_1 \cdot z_2 &= r_1(\cos \theta_1 + i \sin \theta_1) \cdot r_2(\cos \theta_2 + i \sin \theta_2) \\ &= r_1 r_2 (\cos \theta_1 \cos \theta_2 - \sin \theta_1 \sin \theta_2 + i(\cos \theta_1 \sin \theta_2 + \sin \theta_1 \cos \theta_2)) \\ &= r_1 r_2 (\cos(\theta_1 + \theta_2) + i \sin(\theta_1 + \theta_2)). \end{aligned}$$

Note that $\theta_1 = \arg z_1$ and $\theta_2 = \arg z_2$, this implies

$$\arg(z_1 z_2) = \arg(z_1) + \arg(z_2). \quad (11.4) \quad \text{e:arg:mult}$$

With this, it is not difficult to derive

$$\arg\left(\frac{z_1}{z_2}\right) = \arg(z_1) - \arg(z_2)$$

using

$$\arg\left(z_2 \cdot \frac{1}{z_2}\right) = \arg(z_2) + \arg\left(\frac{1}{z_2}\right) \Rightarrow \arg\left(\frac{1}{z_2}\right) = \arg(1) - \arg(z_2) = -\arg(z_2). \quad (11.5) \quad \text{e:arg:inv}$$

Equation (11.4) is sometimes valid when $\arg$ is replaced by Arg . However, this is not always the case.

eg:wwy:1.6.1

Example 11.16. Show that for $z_1 = -1$ and $z_2 = i$, $\text{Arg}(z_1 z_2) \neq \text{Arg}(z_1) + \text{Arg}(z_2)$.

Solution. $\text{Arg}(z_1) = \pi$ according to Example 11.15

$$\text{Arg}(z_2) = \frac{\pi}{2}.$$

$$\text{But } \text{Arg}(z_1 z_2) = \text{Arg}(-i) = -\frac{\pi}{2} \neq \frac{3\pi}{2} = \text{Arg}(z_1) + \text{Arg}(z_2)$$

□

eg:wwy:1.6.2

Example 11.17. Find the principal argument of $\frac{-2}{1 + \sqrt{3}i}$.

Solution.

$$\arg\left(\frac{-2}{1+\sqrt{3}i}\right) = \arg(-2) - \arg(1+\sqrt{3}i) = (\pi + 2k_1\pi) - \left(\frac{\pi}{3} + 2k_2\pi\right) = \frac{2\pi}{3} + 2(k_1 - k_2)\pi$$

$$\text{Hence, } \text{Arg}\left(\frac{-2}{1+\sqrt{3}i}\right) = \frac{2\pi}{3}.$$

□

Example 11.18 (April 2009 Exam, Q1(b)). Determine the real numbers a and b such that

$$a + ib = [3(\cos 40^\circ + i \sin 40^\circ)] [4(\cos 80^\circ + i \sin 80^\circ)] \quad (7 \text{ marks})$$

$$\text{Solution. RHS} = 12(\cos 120^\circ + i \sin 120^\circ) = 12\left(\frac{-1}{2} + \frac{\sqrt{3}}{2}i\right) = -6 + 6\sqrt{3}i$$

□

Example 11.19. Given three distinct points $P_1(z_1)$, $P_2(z_2)$ and $P_3(z_3)$. Derive a formula measuring the angle $\widehat{P_2P_1P_3}$ using the notion of argument.

Solution.

$$\widehat{P_2P_1P_3} = \angle P_3P_1 - \angle P_2P_1 = \arg(z_3 - z_1) - \arg(z_2 - z_1) = \arg \frac{z_3 - z_1}{z_2 - z_1}.$$

□

Equation (11.4) is far more powerful than it looks. We can use it to derive the following theorem.

e:deMoivre

Theorem 11.20 (Abraham de Moivre (1667~1754) Theorem). For any $\theta \in \mathbb{R}$,

$$(\cos \theta + i \sin \theta)^n = \cos n\theta + i \sin n\theta, \quad n \in \mathbb{Z}.$$

Proof. We can prove it by induction for non-positive integer n . When $\boxed{n=0}$, we have

$$(\cos \theta + i \sin \theta)^0 = 1 = \cos 0 + i \sin 0.$$

Suppose that when $n = k$,

$$(\cos \theta + i \sin \theta)^k = \cos k\theta + i \sin k\theta.$$

When $n = k + 1$,

$$\begin{aligned} (\cos \theta + i \sin \theta)^{k+1} &= (\cos \theta + i \sin \theta)^k (\cos \theta + i \sin \theta) \\ &= (\cos k\theta + i \sin k\theta)(\cos \theta + i \sin \theta) \\ &= \cos(k\theta + \theta) + i \sin(k\theta + \theta) \end{aligned}$$

according to (11.4).

de Moivre's formula can be easily derived by using exponent form of complex number. Suppose $z = e^{i\theta} = \cos \theta + i \sin \theta$. Then for any integer n ,

$$z^n = (e^{i\theta})^n = e^{in\theta} \Rightarrow (\cos \theta + i \sin \theta)^n = \cos n\theta + i \sin n\theta.$$

Note that $e^0 = 1 = \cos 0 + i \sin 0$.

To prove the negative case, we first to show that

$$(\cos \theta + i \sin \theta)^{-1} = \cos(-\theta) + i \sin(-\theta).$$

This is because

$$(\cos \theta + i \sin \theta)(\cos(-\theta) + i \sin(-\theta)) = \cos(\theta + (-\theta)) + i \sin(\theta + (-\theta)) = \cos 0 + i \sin 0 = 1.$$

For any positive integer k , using (11.4) and (11.5), we have

$$(\cos \theta + i \sin \theta)^{-k} = \frac{1}{(\cos \theta + i \sin \theta)^k} = \frac{1}{\cos k\theta + i \sin k\theta} = \cos(-k\theta) + i \sin(-k\theta).$$

□

eg:wwy:1.7.1

Example 11.21. Evaluate $(\sqrt{3} + i)^6$.

Solution.

$$\sqrt{3} + i = 2(\cos \frac{\pi}{6} + i \sin \frac{\pi}{6}) \Rightarrow (\sqrt{3} + i)^6 = 2^6(\cos 6(\frac{\pi}{6}) + i \sin 6(\frac{\pi}{6})) = -64.$$

□

From the above example, we learn how to evaluate the n powers of a complex number using de Moivre's formula (11.20). Can we generalise (11.20) to rational numbers? The answer is yes but we have to be careful in the interpretation. First, we know that any rational number can be written as $\frac{n}{m}$ ($m \neq 0, n, m \in \mathbb{Z}$) from *UECM1303 Discrete Mathematics*. Therefore,

$$(\cos \theta + i \sin \theta)^{n/m} = (\cos n\theta + i \sin n\theta)^{1/m} = \cos \phi + i \sin \phi$$

where ϕ is some “angle” we need to determine. Taking power of m , we obtain using (11.4)

$$\cos n\theta + i \sin n\theta = (\cos \phi + i \sin \phi)^m = \cos m\phi + i \sin m\phi \Rightarrow n\theta + 2k\pi = m\phi$$

or

$$\phi = \frac{n\theta + 2k\pi}{m} \Rightarrow (\cos \theta + i \sin \theta)^{n/m} = \cos \frac{n\theta + 2k\pi}{m} + i \sin \frac{n\theta + 2k\pi}{m} \quad (11.6)$$

e:arg:mult2

where $k = 0, \pm 1, \pm 2, \dots$

Let us apply (11.6) to do some calculation and find roots.

Example 11.22. Find $i^{3/4}$.

Solution.

$$i^3 = \cos(\frac{3\pi}{2}) + i \sin(\frac{3\pi}{2}) \Rightarrow i^{3/4} = \cos \frac{3\pi/2 + 2k\pi}{4} + i \sin \frac{3\pi/2 + 2k\pi}{4}, \quad k = 0, 1, 2, 3.$$

So the fourth roots of $-i$ are

$$\cos \frac{3\pi}{8} + i \sin \frac{3\pi}{8}, \quad \cos \frac{7\pi}{8} + i \sin \frac{7\pi}{8}, \quad \cos \frac{11\pi}{8} + i \sin \frac{11\pi}{8}, \quad \cos \frac{15\pi}{8} + i \sin \frac{15\pi}{8}.$$

□

Definition 11.11. Let n be a positive integer. We say that w is an n th root of a nonzero complex number z if

$$w^n = z.$$

To find the root(s) w we proceed as follows:

1. Let $w = r(\cos \theta + i \sin \theta)$ and $z = \rho(\cos \varphi + i \sin \varphi)$ then

$$w^n = z \Rightarrow r^n(\cos n\theta + i \sin n\theta) = \rho(\cos \varphi + i \sin \varphi) \Rightarrow \begin{cases} r^n = \rho \\ n\theta = \varphi + 2k\pi, \quad k = 0, 1, \dots, n-1. \end{cases}$$

2. Thus the n th roots of z are given by

$$w = \sqrt[n]{|z|} \left(\cos \frac{\varphi + 2k\pi}{n} + i \sin \frac{\varphi + 2k\pi}{n} \right), \quad k = 0, 1, \dots, n-1.$$

Definition 11.12. The roots of the equation $w^n = 1$ are called the n th roots of unity.

Geometrically all the n th roots of unity are lying on the unit circle with centre at the origin. Note that one root is at $(1,0)$, i.e, $z = 1$ is one of the n th roots of unity. The other $n - 1$ roots $\omega, \omega^2, \dots, \omega^{n-1}$ are located equidistance on the unity circle and satisfies

$$\begin{aligned} \omega^n = 1 &\Rightarrow \omega^n - 1 = 0 \Rightarrow (\omega - 1)(\omega^{n-1} + \omega^{n-2} + \dots + \omega + 1) = 0 \\ &\Rightarrow \boxed{\omega^{n-1} + \omega^{n-2} + \dots + \omega + 1 = 0}. \end{aligned}$$

eg:1.9.1

Example 11.23. Find the root of the equation $z^3 = 1$.

Solution: Let $z = r(\cos \theta + i \sin \theta)$. Then

$$z^3 = r^3(\cos 3\theta + i \sin 3\theta) = 1 \Rightarrow r = 1 \wedge 3\theta = 2k\pi, \quad k = 0, 1, 2.$$

Therefore,

$$\begin{aligned} z &= 1, \quad \cos \frac{2\pi}{3} + i \sin \frac{2\pi}{3}, \quad \cos \frac{4\pi}{3} + i \sin \frac{4\pi}{3} \\ &= 1, \quad -\frac{1}{2} + i \frac{\sqrt{3}}{2}, \quad -\frac{1}{2} - i \frac{\sqrt{3}}{2}. \end{aligned}$$

Example 11.24 (April 2008 Exam, Q1(a)). Find all five roots of the equation $z^5 - 1 = 0$. (7 marks)

Solution. The five roots are

$$\begin{aligned} z &= \cos \frac{2k\pi}{5} + i \sin \frac{2k\pi}{5}, \quad k = 0, 1, 2, 3, 4 \\ &= 1, \cos \frac{2\pi}{5} + i \sin \frac{2\pi}{5}, \cos \frac{4\pi}{5} + i \sin \frac{4\pi}{5}, \cos \frac{6\pi}{5} + i \sin \frac{6\pi}{5}, \cos \frac{8\pi}{5} + i \sin \frac{8\pi}{5}. \end{aligned}$$

□

Example 11.25 (April 2009 Exam, Q1(c)).

(i) Find the four roots of $z^4 = 4$. (7 marks)

Solution. $z = \sqrt[4]{4}(\cos \frac{2k\pi}{4} + i \sin \frac{2k\pi}{4}), \quad k = 0, 1, 2, 3.$

Therefore, $z = \sqrt{2}, \quad \sqrt{2}i, \quad -\sqrt{2}, \quad -\sqrt{2}i.$

□

(ii) Show that $z^4 + 4$ can be factorised into quadratic factors with real coefficients. (4 marks)

Solution.

$$\begin{aligned} z^4 + 4 &= (z - \sqrt{2}(\cos \frac{\pi}{4} + i \sin \frac{\pi}{4}))(z - \sqrt{2}(\cos \frac{3\pi}{4} + i \sin \frac{3\pi}{4}))(z - \sqrt{2}(\cos \frac{5\pi}{4} + i \sin \frac{5\pi}{4})) \\ &\quad (z - \sqrt{2}(\cos \frac{7\pi}{4} + i \sin \frac{7\pi}{4})) \\ &= (z^2 - 2\sqrt{2} \cos \frac{\pi}{4} z + 2)(z^2 - 2\sqrt{2} \cos \frac{3\pi}{4} z + 2) = (z^2 - 2z + 2)(z^2 + 2z + 2) \end{aligned}$$

□

Example 11.26 (April 2010 Exam, Q5(c)). Find $(-16i)^{1/4}$. (5 marks)

Solution.

$$\begin{aligned}
 (-16i)^{1/4} &= \sqrt[4]{16}(\cos \frac{-\pi/2 + 2k\pi}{4} + i \sin \frac{-\pi/2 + 2k\pi}{4}), \quad k = 0, 1, 2, 3 \\
 &= 2(\cos \frac{-\pi}{8} + i \sin \frac{-\pi}{8}), \quad 2(\cos \frac{3\pi}{8} + i \sin \frac{3\pi}{8}), \quad 2(\cos \frac{7\pi}{8} + i \sin \frac{7\pi}{8}), \\
 &\quad 2(\cos \frac{11\pi}{8} + i \sin \frac{11\pi}{8})
 \end{aligned}$$

□

Example 11.27 (Oon Shea Ming). Find all the roots of the equation $z^6 + 3z^4 + 4z^2 + 2 = 0$.

Solution. We know there are in total six roots.

Put $w = z^2$, it suffices to solve firstly

$$w^3 + 3w^2 + 4w + 2 = 0.$$

It is immediate to remark that -1 is a solution for the above equation. Hence by using Euclidean division. We find that

$$w^3 + 3w^2 + 4w + 2 = (w + 1)(w^2 + 2w + 2).$$

And we find the solutions of the quadratic equation $w^2 + 2w + 2 = 0$ to be $-1 \pm i$.

From $z^2 = -1$, we get

$$z = \pm i$$

From $z^2 = -1 + i = \sqrt{2}e^{3i\pi/4}$, we get $z = \sqrt[4]{2}e^{3i\pi/8}, \sqrt[4]{2}e^{i\pi+3i\pi/8}$.

Using the formula

$$\cos^2 \theta = \frac{1 + \cos \theta}{2}, \quad \sin^2 \theta = \frac{1 - \cos \theta}{2},$$

We find that

$$\cos \frac{3\pi}{8} = +\sqrt{\frac{1 + \cos \frac{3\pi}{4}}{2}} = \sqrt{\frac{1 - 1/\sqrt{2}}{2}}$$

and

$$\sin \frac{3\pi}{8} = +\sqrt{\frac{1 - \cos \frac{3\pi}{4}}{2}} = \sqrt{\frac{1 + 1/\sqrt{2}}{2}}.$$

Hence

$$z = \pm \sqrt[4]{2}e^{3i\pi/8} = \pm \left(\sqrt{\frac{\sqrt{2}-1}{2}} + i\sqrt{\frac{\sqrt{2}+1}{2}} \right).$$

From $z^2 = -1 - i = \sqrt{2}e^{-3i\pi/4}$, we get $z = \sqrt[4]{2}e^{-3i\pi/8}, \sqrt[4]{2}e^{i\pi-3i\pi/8}$.

Similarly,

$$\cos \left(-\frac{3\pi}{8} \right) = +\sqrt{\frac{1 + \cos \left(-\frac{3\pi}{4} \right)}{2}} = \sqrt{\frac{1 - 1/\sqrt{2}}{2}}$$

and

$$\sin \left(-\frac{3\pi}{8} \right) = -\sqrt{\frac{1 - \cos \left(-\frac{3\pi}{4} \right)}{2}} = -\sqrt{\frac{1 + 1/\sqrt{2}}{2}}.$$

Hence

$$z = \pm \sqrt[4]{2}e^{-3i\pi/8} = \pm \left(\sqrt{\frac{\sqrt{2}-1}{2}} - i\sqrt{\frac{\sqrt{2}+1}{2}} \right).$$

Finally, the six roots are

$$\pm i, \pm \sqrt{\frac{\sqrt{2}-1}{2}} \pm i \sqrt{\frac{\sqrt{2}+1}{2}}.$$

□

11.4 Topological Structure of $\mathbb{C}$

The topological structure of $\mathbb{C}$ is studied in Churchill and Brown [2008, Chapter 1, Section 11] under the title “regions of the complex plane”. It is very strange for me to understand why. In that section, what was studied is considered by many to be the concepts related to the “point-set topological structure of $\mathbb{C}$ ”, which are defined below.

The function $d : \mathbb{C} \times \mathbb{C} \rightarrow \mathbb{R}^{\geq 0}$, $d(z, w) := |z - w|$ defines a *metric (or distance)* on $\mathbb{C}$. Since $\mathbb{C}$ is $\mathbb{R}^2$ with field structure, many “geometrical concepts” can be translated from $\mathbb{R}^2$ into complex notation. In particular, $\mathbb{C}$ is a complete metric space, i.e. a metric space where every Cauchy sequence has a limit.

The basic building block for the “topological structure” of $\mathbb{C}$ are the following “ r -neighbourhood of z_0 ” based on the notion of “distance” between two points:

- $\Delta_r(z_0) = \{z \in \mathbb{C} : |z - z_0| < r\}$. It is also denoted as $N(z_0; r)$ or $B(z_0; r)$ in some books.
- $\overline{\Delta_r(z_0)} = \{z \in \mathbb{C} : |z - z_0| \leq r\}$. It is also denoted as $\overline{N(z_0; r)}$ or $\overline{B(z_0; r)}$ in some books.
- $C_r(z_0) = \{z \in \mathbb{C} : |z - z_0| = r\}$.

Definition 11.13. For a subset S of $\mathbb{C}$, a point $z_0 \in S$ is said to be an *interior point* of S if there is some neighbourhood $\Delta_\epsilon(z_0)$ of z_0 (where $\epsilon > 0$) such that $\Delta_\epsilon(z_0) \subset S$. A point z_0 is said to be an *exterior point* of S if there is some neighbourhood $\Delta_\epsilon(z_0)$ of z_0 (where $\epsilon > 0$) such that $\Delta_\epsilon(z_0) \cap S = \emptyset$. If z_0 is neither an interior point nor an exterior point, then z_0 is called a *boundary point* of S .

The set of all the boundary points of S is called the *boundary* of S and is denoted by ∂S .

propn:1

Proposition 11.1. $C_r(z_0)$ is the boundary of $\Delta_r(z_0)$, i.e. $C_r(z_0) = \partial \Delta_r(z_0)$.

Now, we will define the concept of open and close which is fundamental to the topology of $\mathbb{C}$.

Definition 11.14. A set is *open* (in $\mathbb{C}$) if it contains only interior points. A set is said to be *closed* (in $\mathbb{C}$) if it contains all of its boundary points. The *closure* of a set S (in $\mathbb{C}$) is the closed set consisting of all points in S together with the boundary of S .

The topological structure of $\mathbb{C}$ is somewhat “abstract”. So, make sure you try to understand the following example and read the first few chapters of books with the word “topology” in its title.

Example 11.28.

1. $\emptyset$ is open because it contains no points, i.e. it does not violate the requirement that its points must be interior point. $\emptyset$ is also closed because it has no boundary points.
2. $\mathbb{C}$ is open and because every point in it is interior and it does not have boundary points (so it contains the boundary, which is an empty set).
3. $\Delta_r(z_0)$ is open but not closed because it consists of only interior points and not its boundary, which is $C_r(z_0)$ according to Proposition 11.1.

propn:1

4. $\overline{\Delta_r(z_0)}$ is closed but not open because it contains its boundary but at the boundary point z_1 , it is impossible to find ϵ such that $\Delta_\epsilon(z_1) \subset \overline{\Delta_r(z_0)}$.
5. $\{(x, 0) : 0 < x \leq 1\}$ is neither open nor closed. This is because 1 is not an interior point and 0 is a boundary point.

Definition 11.15. A set $D \subset \mathbb{C}$ is said to be *bounded* if there is a number $r \in \mathbb{R}$ such that for every point $z \in D$, $|z| < r$.

Definition 11.16. Given a subset $D \subset \mathbb{C}$ and consider a collection of open subsets S_i of $\mathbb{C}$, $\mathcal{C} = \{S_i \subset \mathbb{C} : i \in I\}$ where I is called the index set. It can be a subset of $\mathbb{N}$ or $\mathbb{R}$ (or some more general set). If $D \subset \bigcup_{i \in I} S_i$, then $\mathcal{C}$ is called the (*open*) *cover* of D .

Definition 11.17. A subset $D \subset \mathbb{C}$ is said to be *compact* if for any open cover $\mathcal{C}$ of D , it is possible to find a finite number of elements $S_1, \dots, S_n$ in $\mathcal{C}$ such that $D \subset S_1 \cup \dots \cup S_n$.

Theorem 11.29 (Heine-Borel Theorem). *A set is compact iff closed and bounded.*

Definition 11.18. Let $S \subset \mathbb{C}$ and $I = [a, b] \subset \mathbb{R}$. Every *continuous* function $\gamma : I \rightarrow S$ is called a *path* in S . The set S is called *path-connected* if every two points in S are in the image of a path in S and S is called *connected* if for any two disjoint open sets $U, V \subset \mathbb{C}$ with $S \subset U \cup V$ one has either $S \subset U$ or $S \subset V$. Any open and connected subset D of the complex plane is called a *region* (or a *domain*).

Theorem 11.30. *Any two points of a region D can be connected by a smooth path.*

Proof. Fix $z_0 \in D$ and consider the sets

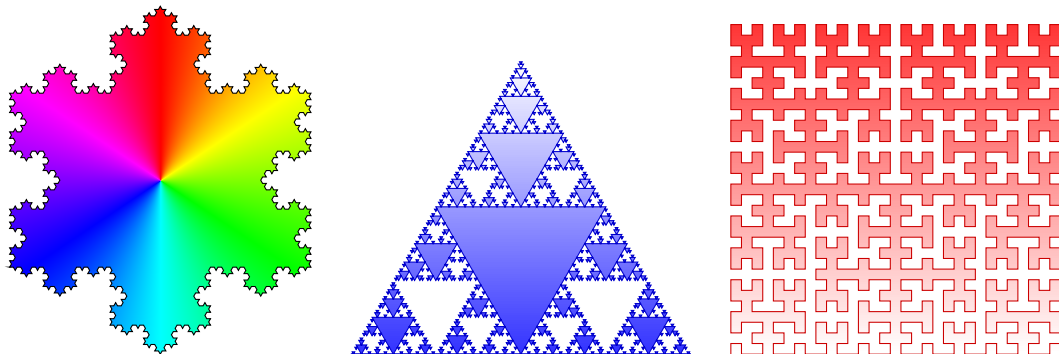
$$A = \{z \in D : z \text{ can be connected to } z_0 \text{ by a smooth path}\}$$

and

$$B = \{z \in D : z \text{ cannot be connected to } z_0 \text{ by a smooth path}\}.$$

Then A, B are open. Clearly, $D = A \cup B$ and $A \cap B = \emptyset$. Since D is connected it follows that either $D \subset A$ or $D \subset B$. Since $z_0 \in D \cap A$ we obtain that $D \subset A$. $\square$

Why do we need all these complicated mathematical concepts of open, closed, compactness, connectivity, etc.? It is because the subsets of $\mathbb{C}$ can be very complicated as demonstrated by the following fractal pictures.



11.5 Basic Geometrical Objects in Complex Plane

There are many interesting propositions about plane geometry which can be found in many books such as Courant et al. [1996]. It is impossible to discuss to many plane geometry problem in such a short course. Instead, we will investigate how complex numbers could be used to describe a few basic geometrical objects in a plane.

First, let us begin with the most fundamental geometric object on a plane. Since $\bar{z}$ is used to denote complex conjugate of z , we will use the notation $\underline{ab}$ to denote the line joining point a and point b in the complex plane. The line and the following theorem describes how to use complex numbers to represent a line.

thm:cplx:line

Theorem 11.31. Suppose a and b are two distinct points in $\mathbb{C}$, the following statements are equivalent:

1. z lies on the line $\underline{ab}$.
2. $\frac{z-a}{b-a} \in \mathbb{R}$.
3. There is a real number t such that $z = (1-t)a + tb$.
4. $\begin{vmatrix} z-a & \bar{z}-\bar{a} \\ b-a & \bar{b}-\bar{a} \end{vmatrix} = 0$.
5. $\begin{vmatrix} z & \bar{z} & 1 \\ a & \bar{a} & 1 \\ b & \bar{b} & 1 \end{vmatrix} = 0$.

Example 11.32. Given points $a = 1 + 3i$ and $b = 4 - 2i$. Write down the line equation in the self-conjugate equation form

$$Az + \overline{A}z + C = 0, \quad C \text{ is a real number.}$$

Solution. From Theorem [11.31](#),

$$\begin{aligned} (z - (1 + 3i))(4 + 2i - (1 - 3i)) &= (4 - 2i - (1 + 3i))(\bar{z} - (1 - 3i)) \\ \Rightarrow (3 + 5i)z - (3 + 5i)(1 + 3i) &= (3 - 5i)\bar{z} - (3 - 5i)(1 - 3i) \\ \Rightarrow (3 + 5i)z - (3 - 5i)\bar{z} - 28i &= 0 \\ \Rightarrow (-5 + 3i)z + (-5 - 3i)\bar{z} + 28 &= 0 \end{aligned}$$

□

Example 11.33. Use the knowledge from linear algebra to find the formula of a triangle in terms its vertices z_1, z_2 and z_3 in a determinant form.

Solution. Since $x_i = \frac{z_i + \bar{z}_i}{2}$ and $y_i = \frac{z_i - \bar{z}_i}{2i}$, for $i = 1, 2, 3$.

$$\begin{aligned} \Delta &= \frac{1}{2} \begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{vmatrix} = \frac{1}{2} \begin{vmatrix} \frac{z_1 + \bar{z}_1}{2} & \frac{z_1 - \bar{z}_1}{2i} & 1 \\ \frac{z_2 + \bar{z}_2}{2} & \frac{z_2 - \bar{z}_2}{2i} & 1 \\ \frac{z_3 + \bar{z}_3}{2} & \frac{z_3 - \bar{z}_3}{2i} & 1 \end{vmatrix} = \frac{1}{8i} \begin{vmatrix} z_1 + \bar{z}_1 & z_1 - \bar{z}_1 & 1 \\ z_2 + \bar{z}_2 & z_2 - \bar{z}_2 & 1 \\ z_3 + \bar{z}_3 & z_3 - \bar{z}_3 & 1 \end{vmatrix} \\ &= \frac{1}{8i} \begin{vmatrix} 2z_1 & z_1 - \bar{z}_1 & 1 \\ 2z_2 & z_2 - \bar{z}_2 & 1 \\ 2z_3 & z_3 - \bar{z}_3 & 1 \end{vmatrix} = \frac{1}{4i} \begin{vmatrix} z_1 & z_1 - \bar{z}_1 & 1 \\ z_2 & z_2 - \bar{z}_2 & 1 \\ z_3 & z_3 - \bar{z}_3 & 1 \end{vmatrix} = \frac{1}{4i} \begin{vmatrix} z_1 & -\bar{z}_1 & 1 \\ z_2 & -\bar{z}_2 & 1 \\ z_3 & -\bar{z}_3 & 1 \end{vmatrix} = \frac{i}{4} \begin{vmatrix} z_1 & \bar{z}_1 & 1 \\ z_2 & \bar{z}_2 & 1 \\ z_3 & \bar{z}_3 & 1 \end{vmatrix} \end{aligned}$$

□

Now, let us investigate fundamental geometric objects related the circle.

If a complex point moves in the complex plane with fixed distance from the point z_0 , it traces a locus of a *circle*. Hence the equation of a circle with radius r and centred at the origin is given by

$$|z - z_0| = r.$$

With a change in the sign, we can obtain a *closed disk*, an *open disk* and an *annulus* below:

$$|z - z_0| \leq r, \quad |z - z_0| < r, \quad r_1 \leq |z - z_0| \leq r_2.$$

eg:1.5.2

Example 11.34. The solution of the equation $|z - 2 - 3i| = 5$ represents the circle. Determine its centre and radius.

Solution. centre: $z_0 = 2 + 3i$; radius, $r = 5$.

□

If we want an *arc* centred at z_0 with a radius r with an angle θ_1 to θ_2 , then the equation is

$$|z - z_0| = r, \quad \theta_1 \leq \text{Arg}(z - z_0) \leq \theta_2.$$

Again, with a change in sign, we can obtain a *sector of circle*

$$|z - z_0| < r, \quad \theta_1 \leq \text{Arg}(z - z_0) \leq \theta_2$$

or just a *sector*

$$\theta_1 \leq \text{Arg } z \leq \theta_2.$$

Some of the more complicated geometrical objects found in analytic geometry can also be expressed with complex number equations.

Definition 11.19. A *hyperbola* is the set of points in a plane defined by the following condition. Given distinct foci a and b , the hyperbola consists of those points for which the absolute difference in distances to these two foci is a real nonzero constant. A defining equation is

$$|z - a| - |z - b| = \pm c. \quad (11.7)$$

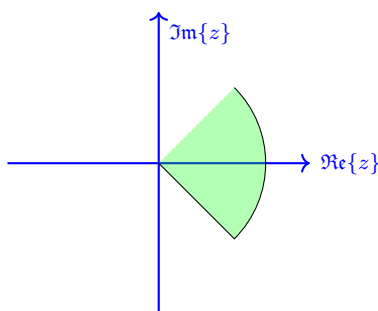
An *ellipse* is the set of those points for which the sum of the distances to foci a and b is a positive constant. A defining equation is

$$|z - a| + |z - b| = c. \quad (11.8)$$

Example 11.35 (April 2008 Exam, Q1(b)). Sketch (or describe) the **closure** of each of the following sets and give an **analytical** proof or a **geometrical** argument:

- (i) $-\frac{\pi}{4} < \arg z < \frac{\pi}{4}$ ($z \neq 0$), (4 marks)

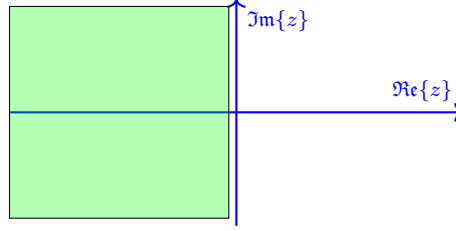
Solution. The closure is $\{z \in \mathbb{C} : |\arg z| \leq \frac{\pi}{4}\}$.



□

- (ii) $|z + 1| < |z - 1|$ (7 marks)

Solution. By calculation: $(x+1)^2 + y^2 < (x-1)^2 + y^2 \Rightarrow x^2 + 2x + 1 < x^2 - 2x + 1 \Rightarrow 4x < 0$. The closure is $\{z \in \mathbb{C} : \Re(z) \leq 0\}$.



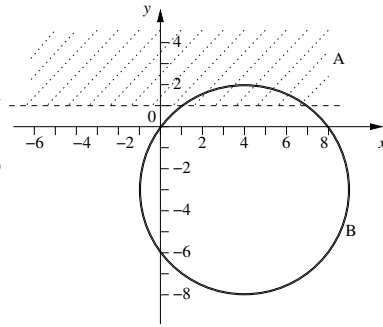
□

Example 11.36 (Dr Oon Shea Ming). Determine the set (a) $A = \{z \in \mathbb{C} \mid \Im(z) > 1\}$ and (b) $B = \{z \in \mathbb{C} \mid |z - 4 + 3i| = 5\}$.

Solution.

(a) $\Im(z) > 1$ is the upper half plane with Cartesian equation $y > 1$.

(b) $|z - (4 - 3i)| = 5$ means that the distance of z to the point $4 - 3i$ is a constant equal to 5. Hence the set is a circle of centre $(4, -3)$ and radius 5.



□

Example 11.37 (Dr Oon Shea Ming). Find the locus of $z \in \mathbb{C}$ that satisfies $\left| \frac{z+2i}{z-i} \right| = 3$.

Solution. When the ratio of distance to two fixed point is a constant different from 1, a reader with geometry background shall know that it is a circle. We can also put $z = x + yi$ and write the algebraic relations:

$$\left| \frac{z+2i}{z-i} \right| = 3 \Leftrightarrow \frac{x^2 + (y+2)^2}{x^2 + (y-1)^2} = 9 \Leftrightarrow x^2 + (y - \frac{11}{8})^2 = \left(\frac{9}{8}\right)^2.$$

Hence, the locus is the circle centred at $\frac{11}{8}i$ with radius $\frac{9}{8}$.

□

11.6 Effective Periods

Kontsevich and Zagier define “(effective) periods” as complex numbers whose real and imaginary parts are values of absolutely convergent integrals

$$p = \int_{\Delta} \frac{f(x_1, \dots, x_n)}{g(x_1, \dots, x_n)} dx_1 \cdots dx_n.$$

Here f and g are polynomials with coefficients in $\mathbb{Q}$ and the integration domain $\Delta \subset \mathbb{R}^n$ is given by polynomial inequalities with rational coefficients.

Example 11.38. $\ln(n) = \int_1^n \frac{dx}{x}$, $\pi = \int_{x^2+y^2 \leq 1} dx dy$, $\zeta(3) = \iiint_{0 < x < y < z < 1} \frac{dx dy dz}{(1-x)yz}$ are periods.

Chapter 12

Complex Functions

ch:cplx:fn

It had taken more than $2\frac{1}{2}$ centuries for mathematicians to come to terms with complex numbers, but the development of the powerful mathematical theory of how to do “calculus” (what we call now “complex analysis”) with functions of complex numbers was astonishingly rapid. Most of the fundamental results were obtained by the mathematicians Cauchy, Dirichlet, Riemann, Weierstrass, and others between 1814 and 1873 — a span of sixty years that changed the face of mathematics forever.

Complex analysis is now a branch of mathematics investigating “complex differentiable” functions defined on some region of the complex plane $\mathbb{C}$. We will discover that “complex differentiability” is much *stronger* than real differentiability and hence complex differentiable functions have much better properties as will be investigated in Topic 4.

12.1 Complex-Valued Functions

To be able to do calculus, we need “complex-valued functions”. We would love to define all the functions in Calculus easily but we will find that it is not simple. In this section, we are able to define rational functions only.

Definition 12.1. Let $D \subset \mathbb{C}$. A *complex(-valued) function* or *complex mapping* defined on S is a rule that assigns to each $z \in D$ a complex number w :

$$f : D \rightarrow \mathbb{C}, \quad z \mapsto w = f(z).$$

The most basic complex-valued function using the complex arithmetic from Topic 1 is the polynomial function and the rational function.

Definition 12.2. The *polynomial function* is a function P over $\mathbb{C}$:

$$P : \mathbb{C} \rightarrow \mathbb{C}, \quad z \mapsto a_0 + a_1z + a_2z^2 + \cdots + a_nz^n.$$

The *rational function* is a function R over $D = \mathbb{C} \setminus \{z_0, \dots, z_n\}$:

$$R : D \rightarrow \mathbb{C}, \quad z \mapsto \frac{a_0 + a_1z + a_2z^2 + \cdots + a_mz^m}{b_0 + b_1z + b_2z^2 + \cdots + b_nz^n}.$$

where $z_i, i = 1, \dots, n$ are the zeros of the polynomial function $b_0 + b_1z + b_2z^2 + \cdots + b_nz^n$.

Example 12.1 (Möbius transform). The Möbius transform $z \mapsto \frac{az + b}{cz + d}$, $ad - bc \neq 0$ is a class of important rational functions which we will investigate in Topic 3.

Example 12.2 (April 2009 Exam, Q1(a)). If $f(z) = z^{24} - 3z^{20} + 4z^{12} - 5z^{16}$, determine real numbers a and b such that

$$f\left(\frac{1+i}{\sqrt{2}}\right) = a + ib. \quad (7 \text{ marks})$$

Solution. $\frac{1+i}{\sqrt{2}} = \cos \frac{\pi}{4} + i \sin \frac{\pi}{4}.$

$$z^{24} = \cos 6\pi + i \sin 6\pi = 1; \quad z^{20} = \cos 5\pi + i \sin 5\pi = -1$$

$$z^{12} = \cos 3\pi + i \sin 3\pi = -1; \quad z^{16} = \cos 4\pi + i \sin 4\pi = 1$$

$$f\left(\frac{1+i}{\sqrt{2}}\right) = 1 - 3(-1) + 4(-1) - 5(1) = 1 + 3 - 4 - 5 = -5.$$

□

Since every complex number can always be written as its real part plus its imaginary part (Topic 1), thus the value of a complex function can be expressed in the (rectangular/Cartesian coordinate) form

$$f(z) = u(x, y) + iv(x, y)$$

where $u = \Re(f(x + iy))$ and $v = \Im(f(x + iy))$ are real-valued functions of two variables.

eg:2.1.2

Example 12.3. Determine the real part and imaginary part of

$$f(z) = |z|^2$$

in rectangular coordinate form.

Solution: In Cartesian coordinates,

$$f(z) = |z|^2 = x^2 + y^2 \Rightarrow u(x, y) = x^2 + y^2 \quad \text{and} \quad v(x, y) = 0$$

eg:2.1.1

Example 12.4. Determine the real part and imaginary part of

$$f(z) = z^2$$

in rectangular coordinate form.

Solution. Solution: Let $z = x + iy$. Then

$$f(z) = (x + iy)^2 = x^2 - y^2 + i(2xy)$$

Hence the real part and the imaginary part of f is

$$u(x, y) = x^2 - y^2, \quad \text{and} \quad v(x, y) = 2xy.$$

□

Graphs are used extensively to investigate the properties of real functions: if $f : \mathbb{R} \rightarrow \mathbb{R}$ is a real-valued function, then the graph of f is a curve in the xy -plane. However, the graph of a complex function $g : \mathbb{C} \rightarrow \mathbb{C}$ lies in $\mathbb{R}^4$ (which is real 4-dimensional space), and so we cannot use graphs to study complex functions. One of the greatest German mathematician Bernhard Riemann developed the concept of complex mapping, where every complex function describes a correspondence between points in two copies of the complex plane. A point in the z -plane (first Argand diagram) is associated with the unique point $w = f(z)$ in the w -plane (second Argand diagram). Hence, the geometric representation of a complex mapping $w = f(z)$ consists of two figures: (a) A subset D of points in the z -plane; and (b) the images of points in D under $w = f(z)$ in the w -plane.

eg:2.2.2

Example 12.5. Consider the mapping $f(z) = z^2$. Find the image of the rectangle bounded by the lines: $x = 0$, $x = a$, $y = 0$, and $y = b$, where a and b are positive numbers.

Solution. From Example 12.4, we have

$$u(x, y) = x^2 - y^2, \quad v(x, y) = 2xy \quad (12.1)$$

First, we determine the image of the vertical line $x = a$ under f . When $x = a$,

$$u = a^2 - y^2 \quad \text{and} \quad v = 2ay \Rightarrow v^2 = 4a^2(a^2 - u)$$

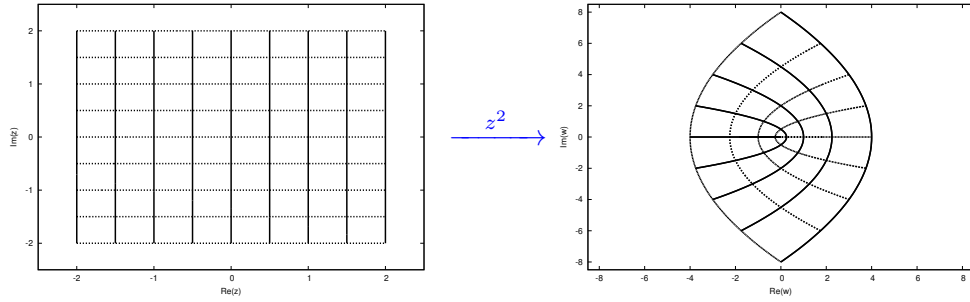
The equation represents a parabola with vertex at a^2 , oriented horizontally, and opening to the left. Since $a > 0$ and $y \geq 0$, this implies that $v \geq 0$. Thus the part of the line $x = a$ lying above the x -axis is mapped to the top half of the parabola.

Next we consider the image of the horizontal line $y = b$. Substituting $y = b$, we have

$$u = x^2 - b^2 \quad \text{and} \quad v = 2bx \Rightarrow v^2 = 4b^2(u + b^2).$$

It represents a parabola with vertex at $-b^2$, oriented horizontally and opening to the right. Since $b > 0$, $x \geq 0$, the part of the line $y = b$ to the right of the y -axis is mapped to the top half of the parabola.

If $x = 0$ then $v = 0$, and $u = -b^2$. The image of this mapping is shown below.



□

Example 12.6. Find the image of a circular arc $\{z \in \mathbb{C} : |z| = a, 0 \leq \text{Arg } z \leq \pi/2\}$ under the mapping $f(z) = z^2$.

Solution. Let the half circular arc be defined by

$$|z| = a, \quad 0 \leq \text{Arg}(z) < \pi/2 \Rightarrow z = a(\cos \theta + i \sin \theta), \quad 0 \leq \theta < \pi/2.$$

Then

$$f(z) = a^2(\cos 2\theta + i \sin 2\theta), \quad 0 \leq \theta < \pi/2.$$

It follows that the image of the half circular arc is a circle centred at the origin and with radius a^2 . □

Example 12.7. Find the curve in $\mathbb{C}$ such that the map $z \mapsto z^2$ is (a) a horizontal line; (b) a vertical line in the w -plane as its image?

Solution. Let C be a real constant

(a) $v = C \Rightarrow 2xy = C$. The hyperbola $y = \frac{C}{2x}$ maps to a horizontal line $v = C$, where

$$u = x^2 - \frac{C^2}{4x^2}, \quad x \in (0, \infty).$$

(b) $u = C$ implies a hyperbola $x^2 - y^2 = C$ is mapped to a vertical line $u = C$ with $v = 2x\sqrt{x^2 - C}$. □

12.2 Limits and Continuity

One of the most fundamental requirement of a function is continuity. If a function is not continuous, it is difficult to “trace” its changes with respect to complex values. To define continuity, we need to concept of limit, which will be defined and studied below.

12.2.1 Limit at $z_0 \in \mathbb{C}$

def:limit

Definition 12.3. Let a function f be defined at all points in some deleted neighbourhood of z_0 (Section 1.5). We say that the *limit* of f as z approaches to z_0 is w_0 if given $\epsilon > 0$, there exists $\delta > 0$ such that

$$0 < |z - z_0| < \delta \Rightarrow |f(z) - w_0| < \epsilon.$$

We write

$$\lim_{z \rightarrow z_0} f(z) = w_0 \quad \text{or} \quad \lim_{z \rightarrow z_0} f(z) = w_0 \quad \text{or} \quad f(z) \rightarrow w_0 \text{ as } z \rightarrow z_0.$$

The definition says that for each disk of radius ϵ , called ϵ -disk and denoted as $N(w_0, \epsilon)$, about w_0 , there is a punctured disk of radius δ about $z = z_0$, denoted as $N^*(z_0, \delta)$ such that the image of each point in $N^*(z_0, \delta)$ lies in $N(w_0, \epsilon)$.

Theorem 12.8. When a limit of a function f exists at a point z_0 , it is unique.

By Definition [12.3](#), in the limit of complex function, z is allowed to approach z_0 from any direction in the complex plane $\mathbb{C}$, that is, along any curve or path through z_0 . Thus a criterion for the nonexistence of a limit is as follows:

Theorem 12.9 (Limit Non-Existence Theorem). If f approaches two complex numbers $w_1 \neq w_2$ for two different curves or paths through z_0 , then $\lim_{z \rightarrow z_0} f(z)$ does not exist.

eg:2.3.1

Example 12.10. Show that the limit $\lim_{z \rightarrow 0} \frac{\bar{z}}{z}$ does not exist.

Proof. We consider two paths leading to the origin: (a) along the real axis; (b) along the imaginary axis.

(a) Along the real axis $z = x$: $\lim_{\substack{x \rightarrow 0 \\ y=0}} \frac{\bar{z}}{z} = \lim_{x \rightarrow 0} \frac{x}{x} = 1$

(b) Along the imaginary axis $z = iy$: $\lim_{\substack{y \rightarrow 0 \\ x=0}} \frac{\bar{z}}{z} = \lim_{y \rightarrow 0} \frac{-iy}{iy} = -1$

By Limit Non-Existence Theorem, $\lim_{z \rightarrow 0} \frac{\bar{z}}{z}$ does not exist. □

eg:2.3.2

Example 12.11. Suppose $f(z) = iz$. Show that $\lim_{z \rightarrow 1} f(z) = i$.

Solution. Observe that

$$|f(z) - i| = |iz - i| = |z - 1|$$

Thus given any $\epsilon > 0$, take $\delta = \epsilon$, then

$$0 < |z - 1| < \delta = \epsilon \Rightarrow |f(z) - i| = |z - 1| < \epsilon.$$

□

The following theorem links the limit of a complex-valued function with two-dimensional real value functions.

thm:2.4.1

Theorem 12.12. Suppose that $f(z) = u(x, y) + iv(x, y)$, $z_0 = x_0 + iy_0$, and $w_0 = u_0 + iv_0$. Then

$$\lim_{z \rightarrow z_0} f(z) = w_0 \quad \Leftrightarrow \quad \lim_{(x,y) \rightarrow (x_0,y_0)} u(x, y) = u_0 \text{ and } \lim_{(x,y) \rightarrow (x_0,y_0)} v(x, y) = v_0.$$

Proof. The result follows immediately from triangle inequality:

- $\Rightarrow$: $|u(x, y) - u_0| \leq |f(z) - w_0|$ and $|v(x, y) - v_0| \leq |f(z) - w_0|$.
- $\Leftarrow$: $|f(z) - w_0| \leq |u(x, y) - u_0| + |v(x, y) - v_0|$.

□

Theorem [12.12](#) is an important theoretical tool that allows us to derive many properties of complex limits from properties of real limits.

thm:2.4.2

Theorem 12.13. Suppose $\lim_{z \rightarrow z_0} f(z)$ and $\lim_{z \rightarrow z_0} g(z)$ exist. Then

1. $\lim_{z \rightarrow z_0} kf(z) = k \lim_{z \rightarrow z_0} f(z)$, where k a complex constant;
2. $\lim_{z \rightarrow z_0} (f(z) \pm g(z)) = \lim_{z \rightarrow z_0} f(z) \pm \lim_{z \rightarrow z_0} g(z)$;
3. $\lim_{z \rightarrow z_0} f(z)g(z) = \left(\lim_{z \rightarrow z_0} f(z) \right) \left(\lim_{z \rightarrow z_0} g(z) \right)$;
4. $\lim_{z \rightarrow z_0} \frac{f(z)}{g(z)} = \frac{\lim_{z \rightarrow z_0} f(z)}{\lim_{z \rightarrow z_0} g(z)}$, provided $\lim_{z \rightarrow z_0} g(z) \neq 0$;

Definition 12.4. A complex function f is said to be *bounded* on region D if there exists a real constant M such that

$$|f(z)| \leq M \quad \text{for all } z.$$

Example 12.14 (April 2009 Exam, Q2(a)). Suppose $f(z)$ is bounded in some neighbourhood of z_0 and $\lim_{z \rightarrow z_0} g(z) = 0$. Show that $\lim_{z \rightarrow z_0} f(z)g(z) = 0$. (5 marks)

Solution. Given that $f(z)$ is bounded by some constant $M > 0$ on some neighbourhood $\delta_1 > 0$ of z_0 .

By definition, $\lim_{z \rightarrow z_0} g(z) = 0$ means: for any $\frac{\epsilon}{M} > 0$, there is a $\delta_2 > 0$ such that

$$0 < |z - z_0| < \delta_2 \Rightarrow |g(z)| < \frac{\epsilon}{M}.$$

For any $\epsilon > 0$, we just need to choose $\delta = \min\{\delta_1, \delta_2\} > 0$, then

$$0 < |z - z_0| < \delta \Rightarrow |f(z)g(z)| < M \frac{\epsilon}{M} = \epsilon.$$

□

By definition, we can determine the limit for polynomial function.

Theorem 12.15. Let $P(z)$ be a polynomial, then $\lim_{z \rightarrow z_0} P(z) = P(z_0)$.

eg:2.4.1

Example 12.16. Find $\lim_{z \rightarrow 1+i} (z^2 + i)$.

Solution. $z^2 + i$ is a polynomial of degree 2. Therefore,

$$\lim_{z \rightarrow 1+i} (z^2 + i) = (1+i)^2 + i = 1 - 1 + 2i + i = 3i.$$

□

However, when we meet with rational function, we need to simplify it in order to find the limit.

eg:2.4.2

Example 12.17. Compute the limit $\lim_{z \rightarrow 1+\sqrt{3}i} \frac{z^2 - 2z + 4}{z - 1 - \sqrt{3}i}$.

Solution. Let $f(z) = z^2 - 2z + 4$ and $g(z) = z - (1 + \sqrt{3}i)$. Notice that $f(1 + \sqrt{3}i) = g(1 + \sqrt{3}i) = 0$.

Using quadratic formula, we find that

$$z^2 - 2z + 4 = (z - 1 + \sqrt{3}i)(z - 1 - \sqrt{3}i).$$

Hence, when $z \neq 1 + \sqrt{3}i$,

$$\frac{f(z)}{g(z)} = z - 1 + \sqrt{3}i.$$

This implies that

$$\lim_{z \rightarrow 1+\sqrt{3}i} \frac{f(z)}{g(z)} = \lim_{z \rightarrow 1+\sqrt{3}i} (z - 1 + \sqrt{3}i) = 2\sqrt{3}i.$$

□

Example 12.18 (April 2010 Exam, Q3(b)). If $\lim_{z \rightarrow z_0} g(z) = B \neq 0$, prove that there exists $\delta > 0$ such that

$$|g(z)| > \frac{1}{2}|B| \quad \text{for} \quad 0 < |z - z_0| < \delta. \quad (7 \text{ marks})$$

Solution. Given for every $\epsilon > 0$, there exists $\delta > 0$ such that

$$0 < |z - z_0| < \delta \Rightarrow |g(z) - B| < \epsilon \Rightarrow ||g(z)| - |B|| \leq |g(z) - B| < \epsilon$$

by using Properties of modulus item 5 (Topic 5). Hence,

$$0 < |z - z_0| < \delta \Rightarrow -\epsilon < |g(z)| - |B| < \epsilon \Rightarrow |B| - \epsilon < |g(z)| \Rightarrow |B| - \frac{1}{2}|B| = \frac{1}{2}|B| < |g(z)|.$$

Note that ϵ is an arbitrary positive number, we can choose $\epsilon < \frac{1}{2}|B|$

□

12.2.2 Limit at ∞

It is sometimes convenient to include the complex plane the point at infinity, denoted by ∞ . It only has extremely limited arithmetic operations defined below:

$$\infty \pm a = a \pm \infty = \infty, \quad \frac{\infty}{a} = \infty, \quad \frac{a}{\infty} = 0, \quad \infty \cdot b = b \cdot \infty = \infty, \quad \frac{b}{0} = \infty$$

where $a \neq \infty$ and $b \neq 0$ (note that b can be ∞). Note that $\infty \pm \infty$, $0 \cdot \infty$, $\frac{\infty}{\infty}$ are, by default, meaningless.

Definition 12.5. For any $\epsilon > 0$, there exists $\delta > 0$ such that

$$|z| > \frac{1}{\delta} \Rightarrow |f(z) - w_\infty| < \epsilon.$$

We denote it as $\lim_{z \rightarrow \infty} f(z) = w_\infty$ or $\lim_{z \rightarrow \infty} f(z) = w_\infty$ or $f(z) \rightarrow w_\infty$ as $z \rightarrow \infty$.

The following theorem gives meaningful statement for limits involving ∞ .

thm:wwy:2.5.1

Theorem 12.19. If z_0 , w_0 and w_∞ are complex numbers, then

- $\lim_{z \rightarrow \infty} f(z) = w_\infty$ if and only if $\lim_{z \rightarrow 0} f(\frac{1}{z}) = w_\infty$;
- $\lim_{z \rightarrow z_0} f(z) = \infty$ if and only if $\lim_{z \rightarrow z_0} \frac{1}{f(z)} = 0$;
- $\lim_{z \rightarrow \infty} f(z) = \infty$ if and only if $\lim_{z \rightarrow 0} \frac{1}{f(\frac{1}{z})} = 0$.

eg:2.5.1

Example 12.20. Find (a) $\lim_{z \rightarrow 1} \frac{iz + 3}{z - 1}$; (b) $\lim_{z \rightarrow \infty} \frac{2z + 1}{z + i}$

Solution. (a) $\lim_{z \rightarrow 1} \frac{z - 1}{iz + 3} = \frac{0}{i + 3} = 0 \Rightarrow \lim_{z \rightarrow 1} \frac{iz + 3}{z - 1} = \infty$
 (b) $\lim_{z \rightarrow \infty} \frac{2z + 1}{z + i} = \lim_{z \rightarrow \infty} \frac{2 + \frac{1}{z}}{1 + \frac{i}{z}} = \frac{2 + 0}{1 + 0} = 2.$ □

12.2.3 Continuity

Now, we can use the notion of limit to define continuity.

Definition 12.6 (Continuity at a Point). A function f is said to be *continuous at z_0* if $\lim_{z \rightarrow z_0} f(z) = f(z_0)$.

Alternatively, we can define continuity without using the notion of limit.

Definition 12.7 (Continuity at a Point, an Equivalent Definition). f is *continuous at z_0* if for any $\epsilon > 0$, there exist $\delta > 0$ such that whenever $|z - z_0| < \delta$,

$$|f(z) - f(z_0)| < \epsilon.$$

Most of the time, it is more convenient to translate the problem of proving the limit for complex function to real function because we are more familiar with real function and the following theorem characterises the link between them.

thm:wwy:2.6.1

Theorem 12.21. Suppose that $f(z) = u(x, y) + iv(x, y)$ and $z_0 = x_0 + iy_0$. Then f is continuous at z_0 if and only if both $u(x, y)$ and $v(x, y)$ are continuous at (x_0, y_0) .

thm:wwy:2.6.2

Theorem 12.22. Suppose f is continuous at z_0 . Let g be continuous at $w_0 = f(z_0)$. Then the composite function $g \circ f$ is continuous at z_0 .

Proof: Given any $\epsilon > 0$, there exists $\epsilon_1 > 0$ such that

$$|w - w_0| < \epsilon_1 \Rightarrow |g(w) - g(w_0)| < \epsilon$$

For $\epsilon_1 > 0$, there exist $\delta > 0$ such that

$$|z - z_0| < \delta \Rightarrow |f(z) - f(z_0)| < \epsilon_1 \Rightarrow |g(f(z)) - g(f(z_0))| < \epsilon.$$

Other algebraic properties of complex limits stated in Theorem [12.13](#) can also be restated in terms of continuity of complex functions.

thm:wwy:2.6.3

Theorem 12.23. *If f and g are continuous at z_0 , then the following functions are continuous at z_0 :*

- $kf(z)$, k a complex constant;
- $f(z) \pm g(z)$;
- $f(z)g(z)$;
- $f(z)/g(z)$ if $g(z_0) \neq 0$.

Continuous complex functions have many important properties that are analogous to properties of continuous real functions.

thm:wwy:2.6.5

Theorem 12.24. *If a complex function f is continuous on a closed and bounded region D , then f is bounded on D .*

Proof: Suppose $f(z) = u(x, y) + iv(x, y)$. Then u and v are continuous for $(x, y) \in D$. It follows that $|f(z)| = \sqrt{u^2 + v^2}$ is continuous in D . Hence f is bounded on D .

Example 12.25 (April 2010 Exam, Q3(a)). Let $f(z) = \begin{cases} \frac{|\Re(z)|^4}{|z|}, & z \neq 0, \\ 0, & z = 0. \end{cases}$

Is $f(z)$ continuous at $z = 0$?

(5 marks)

Solution. Note that when $z \neq 0$,

$$|\Re(z)| \leq |z| \Rightarrow |\Re(z)|^4 \leq |z|^4 \Rightarrow \frac{|\Re(z)|^4}{|z|} \leq |z|^3$$

Therefore, f is continuous at $z = 0$ according to squeeze theorem,

$$0 \leq \lim_{z \rightarrow 0} f(z) = \lim_{z \rightarrow 0} \frac{|\Re(z)|^4}{|z|} \leq \lim_{z \rightarrow 0} |z|^3 = 0 = f(0).$$

□

12.3 Holomorphic Functions and C-R Equations

We know how to define derivative for real functions $f : (a, b) \rightarrow \mathbb{R}$, which is related to the “gradient” of the graph of the function. But how do we make sense of the notion of “derivative” of a complex function? This question is the focus of this section.

12.3.1 Derivatives

Let f be a complex function whose domain D of definition contains a neighbourhood of a point z_0 .

Definition 12.8. The *derivative* of f at z_0 , denoted by $f'(z_0)$ or $\left. \frac{df}{dz} \right|_{z=z_0}$, is defined as

$$f'(z_0) = \lim_{\Delta z \rightarrow 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z} \quad (12.2)$$

provided the limit exists. If it does, we say that the function f is (*complex*) *differentiable* at z_0 .

eg:3.1.1

Example 12.26. Find the derivative of $f(z) = z^n$ at $z_0 \in \mathbb{C}$ where the integer $n \geq 0$.

Solution: By using the binomial formula from Topic 1, we have

$$\begin{aligned} f'(z_0) &= \lim_{\Delta z \rightarrow 0} \frac{(z_0 + \Delta z)^n - z_0^n}{\Delta z} = \lim_{\Delta z \rightarrow 0} \frac{\sum_{k=0}^n \binom{n}{k} z_0^k \Delta z^{n-k} - z_0^n}{\Delta z} \\ &= \lim_{\Delta z \rightarrow 0} \sum_{k=0}^{n-1} \binom{n}{k} z_0^k \Delta z^{n-k-1} = \binom{n}{n-1} z_0^{n-1} \Delta z^0 = n z_0^{n-1}. \end{aligned}$$

eg:3.1.2

Example 12.27. Is $f(z) = |z|^2$ differentiable? Justify your answer.

Solution: At any point z ,

$$\frac{f(z + \Delta z) - f(z)}{\Delta z} = \frac{|z + \Delta z|^2 - |z|^2}{\Delta z} = \frac{z\overline{\Delta z} + \bar{z}\Delta z + \Delta z\overline{\Delta z}}{\Delta z} = \frac{z\overline{\Delta z}}{\Delta z} + \bar{z} + \overline{\Delta z}$$

To show that limit does not exist, we consider two cases:

1. When $\Delta z = \Delta x$,

$$\frac{z\overline{\Delta z}}{\Delta z} + \bar{z} + \overline{\Delta z} = \frac{z\overline{\Delta z}}{\Delta z} + \bar{z} + \overline{\Delta z} = z + \bar{z} + \Delta x \rightarrow z + \bar{z} \quad \text{as } \Delta x \rightarrow 0.$$

2. When $\Delta z = i\Delta y$,

$$\frac{z\overline{\Delta z}}{\Delta z} + \bar{z} + \overline{\Delta z} = -z + \bar{z} - i\Delta y \rightarrow -z + \bar{z} \quad \text{as } \Delta y \rightarrow 0.$$

For the limit to exist, we must have $z = 0$.

At $z = 0$, $\lim_{\Delta z \rightarrow 0} \frac{|\Delta z|^2}{\Delta z} = \overline{\Delta z} \rightarrow 0$ as $\Delta z \rightarrow 0$.

We conclude that the derivative of f exists only at $z = 0$ and its value is 0.

The following question is similar to Example [22.7](#). But the question is imprecise.

Example 12.28 (April 2009 Exam Q3(b), 10 marks; Sept 2012 Exam Q2(b), 7 marks). Show that $f(z) = z^2\bar{z}$ is nowhere differentiable.

Solution. First, we show that f is differentiable at $z = 0$:

$$f'(0) = \lim_{\Delta z \rightarrow 0} \frac{(\Delta z)^2\overline{\Delta z} - 0}{\Delta z} = \lim_{\Delta z \rightarrow 0} \Delta z\overline{\Delta z} = 0.$$

When $\boxed{z \neq 0}$, if f is differentiable at z , then

$$\begin{aligned}
 f'(z) &= \lim_{\Delta z \rightarrow 0} \frac{(z + \Delta z)^2 \overline{z + \Delta z} - z^2 \overline{z}}{\Delta z} = \lim_{\Delta z \rightarrow 0} \frac{(z^2 + 2z\Delta z + (\Delta z)^2)(\overline{z} + \overline{\Delta z}) - z^2 \overline{z}}{\Delta z} \\
 &= \lim_{\Delta z \rightarrow 0} \frac{z^2 \overline{\Delta z} + 2z\overline{z}\Delta z + 2z\Delta z \overline{\Delta z} + \overline{z}(\Delta z)^2 + (\Delta z)^2 \overline{\Delta z}}{\Delta z} \\
 &= \lim_{\Delta z \rightarrow 0} \frac{z^2 \overline{\Delta z}}{\Delta z} + 2z\overline{z} + \lim_{\Delta z \rightarrow 0} 2z\overline{\Delta z} + \lim_{\Delta z \rightarrow 0} \overline{z}\Delta z + \lim_{\Delta z \rightarrow 0} (\Delta z) \overline{\Delta z} \\
 &= \lim_{\Delta z \rightarrow 0} \frac{z^2 \overline{\Delta z}}{\Delta z} + 2z\overline{z}
 \end{aligned}$$

Along x -axis, $\lim_{\Delta z \rightarrow 0} \frac{z^2 \overline{\Delta z}}{\Delta z} = z^2$; But along y -axis, $\lim_{\Delta z \rightarrow 0} \frac{z^2 \overline{\Delta z}}{\Delta z} = -z^2$.
The limit is not unique and f is not differentiable when $z \neq 0$. □

Example 12.29 (April 2008 Exam, Q1(c)). Prove that $\frac{d\overline{z}}{dz}$ does not exist anywhere. (7 marks)

Solution. Let $f(z) = \overline{z}$. Write $z = x + iy$ and $\Delta z = \Delta x + i\Delta y$. If a limit exists, then

$$f'(z) = \lim_{\Delta z \rightarrow 0} \frac{\overline{z + \Delta z} - \overline{z}}{\Delta z} = \lim_{\Delta z \rightarrow 0} \frac{\overline{\Delta z}}{\Delta z}$$

Taking the limit along y -axis and x -axis, we have

$$f'(z) = \lim_{\Delta z = i\Delta y \rightarrow 0} \frac{-i\Delta y}{i\Delta y} = -1; \quad f'(z) = \lim_{\Delta z = \Delta x \rightarrow 0} \frac{\Delta x}{\Delta x} = 1.$$

A contradiction to limit must be unique. Therefore, limit cannot exist and hence function f is nowhere differentiable. □

The function $|f(z)| = |z|^2$ is continuous at each point in the z -plane, yet it is only differentiable at $z = 0$. So continuity of a function at a point does not imply the existence of a derivative there. However, existence of derivative at a point does imply that function is continuous at that point.

thm:3.1.1

Theorem 12.30. If complex function f is differentiable at z , then it is continuous there.

Proof: Given f is differentiable at z , then for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$0 < |z + \Delta z - z| < \delta \Rightarrow \left| \frac{f(z + \Delta z) - f(z)}{\Delta z} \right| < \epsilon$$

Rearranging, we obtain

$$0 < |z + \Delta z - z| < \min\{\delta, 1\} \Rightarrow |f(z + \Delta z) - f(z)| < \epsilon |\Delta z| < \epsilon.$$

Note that when $\Delta z = 0$, $|f(z) - f(z)| = 0 < \epsilon$. Therefore, we have

$$|z + \Delta z - z| < \min\{\delta, 1\} \Rightarrow |f(z + \Delta z) - f(z)| < \epsilon.$$

By definition, f is continuous at point z .

The basic differentiation formulas for complex function are similar to that of real function found in Calculus I. Hence they will be listed here without proof.

thm:diff:prop

Theorem 12.31. If f and g are differentiable at z , then

- $\frac{d}{dz}(cf(z)) = cf'(z)$, where c is a constant;
- $\frac{d}{dz}(f(z) + g(z)) = f'(z) + g'(z)$;
- $\frac{d}{dz}(f(z)g(z)) = f(z)g'(z) + f'(z)g(z)$;
- $\frac{d}{dz}\left(\frac{f(z)}{g(z)}\right) = \frac{g(z)f'(z) - f(z)g'(z)}{(g(z))^2}$, if $g(z) \neq 0$
- $\frac{d}{dz}f(g(z)) = f'(g(z))g'(z)$ (Chain rule)

The L'Hôpital rule can also be generalised to the complex case below.

thm:3.2.1

Theorem 12.32 (L'Hôpital Rule). Suppose that $f(z_0) = g(z_0) = 0$ and that $f'(z_0)$ and $g'(z_0)$ exist, where $g'(z_0) \neq 0$. Then

$$\lim_{z \rightarrow z_0} \frac{f(z)}{g(z)} = \frac{f'(z_0)}{g'(z_0)}.$$

Proof: Let $\Delta = z - z_0$.

$$\lim_{z \rightarrow z_0} \frac{f(z)}{g(z)} = \lim_{z \rightarrow z_0} \frac{\frac{f(z)-f(z_0)}{z-z_0}}{\frac{g(z)-g(z_0)}{z-z_0}} = \lim_{\Delta \rightarrow 0} \frac{\frac{f(z_0+\Delta)-f(z_0)}{\Delta}}{\frac{g(z_0+\Delta)-g(z_0)}{\Delta}} = \frac{\lim_{\Delta \rightarrow 0} \frac{f(z_0+\Delta)-f(z_0)}{\Delta}}{\lim_{\Delta \rightarrow 0} \frac{g(z_0+\Delta)-g(z_0)}{\Delta}} = \frac{f'(z_0)}{g'(z_0)}$$

eg:3.2.1

Example 12.33. Compute $\lim_{z \rightarrow 1+i} \frac{z^2 - 2z + 2}{z - 1 - i}$.

Solution. Let $f(z) = z^2 - 2z + 2$ and $g(z) = z - 1 - i$. Then

$$\begin{aligned} f(1+i) &= (1+i)^2 - 2(1+i) + 2 = 2i - 2 - 2i + 2 = 0 \\ g(1+i) &= 0, \quad g'(1+i) = 1. \end{aligned}$$

Therefore,

$$\lim_{z \rightarrow 1+i} \frac{z^2 - 2z + 2}{z - 1 - i} = \lim_{z \rightarrow 1+i} \frac{2z - 2}{1} = 2(1+i) - 2 = 2i.$$

□

12.3.2 Holomorphic Functions

ec:holomorphic

After we state the formal notion of “derivative”, it is time to define functions which have “derivatives”. We are not interested in studying functions that are not differentiable or are differentiable at only one point.

Definition 12.9. A function $f : D \rightarrow \mathbb{C}$ is said to be a *holomorphic function* (also known as *analytic functions* in many books including our main reference, or *regular functions*, or *conformal maps*) on D if it is differentiable for all $z_0 \in D$.

A function $f : D \rightarrow \mathbb{C}$ is said to be a *holomorphic at a point* z_0 if it is differentiable at z_0 and at every point in some neighbourhood of z_0 .

A holomorphic function whose domain is the whole complex plane is called an *entire function*.

Remark 12.34. According to Theorem [12.31](#), the linear combination of entire functions is also entire.

Example 12.35. Show that $f(z) = z^n$ is an entire function.

Proof. According to Example [12.26](#), f is differentiable at every $z_0 \in \mathbb{C}$ with $f'(z_0) = nz_0^{n-1}$. By definition, f is an entire function. $\square$

In the following, we will learn that a function may be continuous at point z_0 but this does not imply that the function is differentiable at z_0 and if f is not differentiable at z_0 , it is not holomorphic.

Example 12.36 (April 2009 Exam, Q2(b)). Let $f(z) = \begin{cases} \bar{z}^2/z, & z \neq 0, \\ 0, & z = 0. \end{cases}$

1. Does $\lim_{z \rightarrow 0} f(z)$ exists? Explain. (5 marks)

Solution. Let us look at

$$|f(z) - 0| = |\bar{z}^2/z| = \frac{|z|^2}{|z|} = |z| \quad \text{for all } z \neq 0.$$

By definition of limit (Definition [12.3](#)), for any ϵ , there exists $\delta = \epsilon$ such that

$$0 < |z - 0| < \delta \Rightarrow |f(z) - 0| < \epsilon.$$

$\square$

2. Is $f(z)$ continuous at $z = 0$? Explain. (5 marks)

Solution. Since $\lim_{z \rightarrow 0} f(z) = 0 = f(0)$, f is continuous at $z = 0$. $\square$

3. Is $f(z)$ differentiable at $z = 0$? Explain. (5 marks)

Solution. Look at $\lim_{\Delta z \rightarrow 0} \frac{f(\Delta z) - f(0)}{\Delta z} = \lim_{\Delta z \rightarrow 0} \frac{\overline{\Delta z}^2}{\Delta z} = \lim_{\Delta z \rightarrow 0} \frac{\Delta x^2 - \Delta y^2 - 2i\Delta x\Delta y}{\Delta x^2 - \Delta y^2 + 2i\Delta x\Delta y}$

Along y -axis, $\Delta y = 0$, the limit becomes $\lim_{\Delta x \rightarrow 0} \frac{(\Delta x)^2}{(\Delta x)^2} = 1$

However, along $\Delta x = \Delta y$, we have $\lim_{\Delta x \rightarrow 0} \frac{-2i(\Delta x)^2}{2i(\Delta x)^2} = -1$.

The limit is not unique and f is not differentiable at $z = 0$. $\square$

4. Is $f(z)$ “analytic” (holomorphic) at $z = 0$? Explain. (5 marks)

Solution. f is not differentiable at $z = 0$, therefore f is not holomorphic (or “analytic”) at $z = 0$ by definition. $\square$

We have seen in Example [22.7](#) that function $f(z) = |z|^2$ and in Example [12.36](#) that f is only differentiable at $z = 0$ and it not holomorphic at $z = 0$ because there is no neighbourhood of $z = 0$ which f is differentiable.

Definition 12.10. A point in $D \subset \mathbb{C}$ at which f fails to be holomorphic is called a *singular point* of f .

We will discuss the properties of singular points in Topic 6 and we will state the basic properties of holomorphicity as follows.

Theorem 12.37. If f and g are holomorphic in a domain D , so are the functions $f(z) \pm g(z)$ and $f(z)g(z)$. If $g(z) \neq 0$, then $f(z)/g(z)$ is holomorphic.

12.3.3 Cauchy-Riemann (C-R) Equations

In Example [eg:3.1.1](#) 12.26, we showed that computing the derivative of $f(z) = z^n$ is not difficult. However, life is not always so easy. We often encounter complex function written as $f(z) = u(x, y) + iv(x, y)$ and using the definition of derivative to differentiate f may be difficult. Therefore, we shall develop a test for holomorphicity of a complex function f that is based on partial derivatives of its real and imaginary parts u and v .

thm:cr

Theorem 12.38 (Cauchy-Riemann Theorem, Sept 2012 Exam Q2(a), Sept 2011 Exam Q2(a)). Suppose $f(z) = u(x, y) + iv(x, y)$ is differentiable at a point $z = x + iy$. Then at z the first-order partial derivatives of u and v exist and satisfy the Cauchy-Riemann equations:

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}. \quad (12.3)$$

e:cr

(8 marks)

Proof. Write $\Delta z = \Delta x + i\Delta y$. Then

$$\begin{aligned} f'(z) &= \lim_{\Delta z \rightarrow 0} \frac{f(z + \Delta z) - f(z)}{\Delta z} \\ &= \lim_{\Delta x + i\Delta y \rightarrow 0} \frac{u(x + \Delta x, y + \Delta y) + iv(x + \Delta x, y + \Delta y) - u(x, y) - iv(x, y)}{\Delta x + i\Delta y}. \end{aligned}$$

If we choose to let $\Delta z \rightarrow 0$ along a horizontal line, then $\Delta y = 0$ and

$$f'(z) = \lim_{\Delta x \rightarrow 0} \frac{u(x + \Delta x, y) - u(x, y)}{\Delta x} + i \lim_{\Delta x \rightarrow 0} \frac{v(x + \Delta x, y) - v(x, y)}{\Delta x} = \frac{\partial u}{\partial x} + i \frac{\partial v}{\partial x}.$$

On the other hand, if we choose to let $\Delta z \rightarrow 0$ along a vertical line, then $\Delta x = 0$ and

$$f'(z) = \lim_{\Delta y \rightarrow 0} \frac{u(x, y + \Delta y) - u(x, y)}{i\Delta y} + i \lim_{\Delta y \rightarrow 0} \frac{v(x, y + \Delta y) - v(x, y)}{i\Delta y} = -i \frac{\partial u}{\partial y} + \frac{\partial v}{\partial y}.$$

Since the limit is unique, therefore,

$$f'(z) = \frac{\partial u}{\partial x} + i \frac{\partial v}{\partial x} = -i \frac{\partial u}{\partial y} + \frac{\partial v}{\partial y}.$$

By equating the real and imaginary parts, we obtain the Cauchy-Riemann equations. $\square$

Equation [\(12.3\)](#) [e:cr](#) are named Cauchy-Riemann equations in honour of the French mathematician Augustin Louis Cauchy (1789~1857), who discovered and used them, and in honour of the German mathematician Georg Friedrich Riemann (1826~1866), who made them fundamental in his development of the theory of functions of complex variable.

Note that the Cauchy-Riemann equations can be combined into a single equation as

$$\frac{\partial f}{\partial x} = -\frac{\partial f}{\partial y}. \quad (12.4) \quad \text{e:cr:combine}$$

The above theorem states that Cauchy-Riemann equations hold as a necessary condition of f being differentiable at z . The Cauchy-Riemann equations are not sufficient conditions for a function to be differentiable. If the real and imaginary parts of a complex function do not satisfy the Cauchy-Riemann equations at z , then the complex function is not differentiable at z .

eg:wwy:3.3.1

Example 12.39. Consider function $f(z) = z^2 = (x^2 - y^2) + i(2xy)$. Find u , v , u_x , u_y , v_x and v_y . Determine if f satisfies Cauchy-Riemann equations.

Solution:

$$u(x, y) = x^2 - y^2 \quad \text{and} \quad v(x, y) = 2xy.$$

We verify that f satisfies Cauchy-Riemann equations:

$$\frac{\partial u}{\partial x} = 2x = \frac{\partial v}{\partial y}, \quad \frac{\partial u}{\partial y} = -2y = -\frac{\partial v}{\partial x}.$$

eg:wwy:3.3.2

Example 12.40. Consider $f(z) = |z|^2 = x^2 + y^2$. Find u , v , u_x , u_y , v_x and v_y . Determine if f satisfies Cauchy-Riemann equations.

Solution.

$$u(x, y) = x^2 + y^2 \quad \text{and} \quad v(x, y) = 0$$

Thus

$$\frac{\partial u}{\partial x} = 2x, \quad \frac{\partial v}{\partial y} = 0, \quad \frac{\partial u}{\partial y} = 2y, \quad \frac{\partial v}{\partial x} = 0.$$

If Cauchy-Riemann equations hold at (x, y) , then $x = y = 0$. Consequently, f' does not exist for any nonzero z . $\square$

Example 12.41 (Sept 2011 Exam Q2(b); Sept 2010 Exam Q2(b)). Prove that $f(z) = |z|^2$ is nowhere analytic (holomorphic). (8 marks)

Solution. Let $z = x + iy$. Then

$$f(z) = x^2 + y^2$$

Since

$$\frac{\partial u}{\partial x} = 2x \neq \frac{\partial v}{\partial y} = 0$$

(except for $x = y = 0$), f is not holomorphic on $\mathbb{C}$. $\square$

Example 12.42. Determine the analyticity (holomorphicity) of $f(z) = z^2 \bar{z}$. (April 2009 Exam Q3(b), 10 marks; Sept 2009 Exam Q5(c), 5 marks; Sept 2012 Exam Q2(b), 7 marks)

Solution. Let $z = x + iy$. Then

$$f(z) = x^3 + xy^2 + i(x^2y + y^3)$$

Since

$$\frac{\partial u}{\partial x} = 3x^2 + y^2 = \frac{\partial v}{\partial y} = x^2 + 3y^2 \quad \wedge \quad \frac{\partial u}{\partial y} = 2xy = -\frac{\partial v}{\partial x} = -2xy \Rightarrow x = \pm y = 0.$$

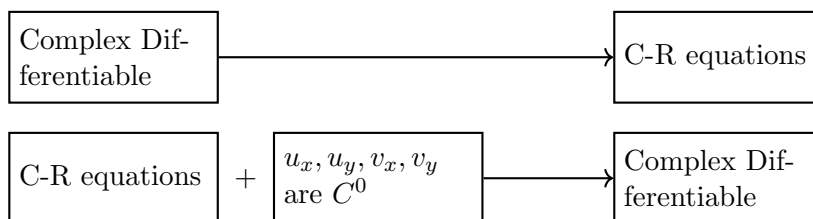
Therefore, f is not holomorphic on $\mathbb{C}$. $\square$

Cauchy-Riemann equations (12.3) do not ensure the existence for $f'(z_0)$. However, together with some additional assumptions imposed on the partial derivatives, Cauchy-Riemann equations give the following criterion for differentiability.

thm:3.3.2

Theorem 12.43 (Sufficient Conditions for Complex Differentiability). *If the real functions $u(x, y)$ and $v(x, y)$ are continuous and have continuous first-order partial derivatives in some neighbourhood of a point z , and if u and v satisfy the Cauchy-Riemann equations at z , then the complex function $f(z) = u(x, y) + iv(x, y)$ is differentiable at z .*

This theorem, together with Cauchy-Riemann Theorem ^{thm:cr} 12.38 imply the following picture.



Proof: For any $\epsilon > 0$, let $\Delta z = \Delta x + i\Delta y$ with $0 < |\Delta z| < \epsilon$. Let

$$\Delta w = f(z + \Delta z) - f(z) = u(x + \Delta x, y + \Delta y) - u(x, y) + i(v(x + \Delta x, y + \Delta y) - v(x, y)) = \Delta u + i\Delta v$$

The assumption that first-order derivatives of u and v are continuous implies that Δu and Δv can be expressed as

$$\begin{aligned}\Delta u &= u_x \Delta x + u_y \Delta y + \epsilon_1 \sqrt{(\Delta x)^2 + (\Delta y)^2} \\ \Delta v &= v_x \Delta x + v_y \Delta y + \epsilon_2 \sqrt{(\Delta x)^2 + (\Delta y)^2}\end{aligned}$$

where $(\epsilon_1, \epsilon_2) \rightarrow (0, 0)$ as $(\Delta x, \Delta y) \rightarrow (0, 0)$.

Thus

$$\Delta w = u_x \Delta x + u_y \Delta y + \epsilon_1 \sqrt{(\Delta x)^2 + (\Delta y)^2} + i \left[v_x \Delta x + v_y \Delta y + \epsilon_2 \sqrt{(\Delta x)^2 + (\Delta y)^2} \right]$$

Replacing u_y by $-v_x$ and v_y by u_x and then dividing Δw by Δz gives

$$\frac{\Delta w}{\Delta z} = u_x + iv_x + (\epsilon_1 + i\epsilon_2) \frac{\sqrt{(\Delta x)^2 + (\Delta y)^2}}{\Delta z} = u_x + iv_x + (\epsilon_1 + i\epsilon_2) \frac{|\Delta z|}{\Delta z}.$$

$$\text{Hence } \lim_{\Delta z \rightarrow 0} \frac{\Delta w}{\Delta z} = u_x + iv_x + \lim_{\Delta z \rightarrow 0} (\epsilon_1 + i\epsilon_2) \frac{|\Delta z|}{\Delta z} = u_x + iv_x.$$

eg:3.3.3

Example 12.44. Show that $f(z) = \frac{x}{x^2 + y^2} - i \frac{y}{x^2 + y^2}$, is complex differentiable at $(x, y) \neq (0, 0)$.

Proof: Given $u = \frac{x}{x^2 + y^2}$ and $v = \frac{-y}{x^2 + y^2}$. We check that u and v satisfy Cauchy-Riemann equations:

$$u_x = \frac{y^2 - x^2}{(x^2 + y^2)^2} = v_y, \quad v_x = \frac{2xy}{(x^2 + y^2)^2} = -u_y.$$

Moreover, all these partial derivatives are continuous except at $(x, y) = (0, 0)$. Sufficient Conditions Theorem ^{thm:3.3.2} 12.43 implies that $f(z)$ is differentiable at $z \neq 0$.

eg:wwy:3.5.2

Example 12.45. Show that $f(z) = \cosh(x) \cos(y) + i \sinh(x) \sin(y)$ is an entire function. Note that $\sinh(x) = \frac{e^x - e^{-x}}{2}$ and $\cosh(x) = \frac{e^x + e^{-x}}{2}$ if they are not taught in Calculus subject.

Solution. Let $u(x, y) = \cosh(x) \cos(y)$ and $v(x, y) = \sinh(x) \sin(y)$. Therefore, the function f satisfies Cauchy-Riemann equations:

$$u_x = \sinh(x) \cos(y) = v_y, \quad u_y = -\cosh(x) \sin(y) = -v_x$$

for all (x, y) .

Since u, v, u_x, u_y, v_x and v_y are all continuous on $\mathbb{R}^2$ because they are the product of continuous functions. According to the sufficient conditions for complex differentiability, f is differentiable at every $z \in \mathbb{C}$ and so f is an entire function. $\square$

Example 12.46 (April 2008 Exam, Q2(a)). Consider the complex function

$$f(z) = |z| + i \operatorname{Arg}(z), \quad z \in \mathbb{C}.$$

- (i) Determine the region where the function f is continuous. (5 marks)

Solution. Since $\operatorname{Arg}(z)$ is only continuous on $\mathbb{C} \setminus (-\infty, 0]$, f is continuous only on $\mathbb{C} \setminus (-\infty, 0]$. $\square$

- (ii) Using Cauchy-Riemann equations, or otherwise, determine the region where the function f is differentiable. (5 marks)

Solution. Given $u = \sqrt{x^2 + y^2}$ and $v = \tan^{-1} \frac{y}{x}$.

$$u_x = \frac{1}{2}(x^2 + y^2)^{-1/2}(2x) = \frac{x}{\sqrt{x^2 + y^2}}$$

and

$$v_y = \frac{1}{1 + (\frac{y}{x})^2}(\frac{1}{x}) = \frac{x}{x^2 + y^2}$$

From $u_x = v_y$ and $(x, y) \neq (0, 0)$, we have $\sqrt{x^2 + y^2} = 1$. It is possible for f to be differentiable at $|z| = 1$. $\square$

- (iii) Where is f analytic (holomorphic)? (3 marks)

Solution. f is nowhere holomorphic because we cannot find a ϵ -neighbourhood of $z \in \mathbb{C}$ in which f is differentiable. $\square$

thm:3.5.1

Theorem 12.47. If $f'(z) = 0$ everywhere in a domain D , then $f(z)$ must be constant throughout D .

Proof: Let $f(z) = u(x, y) + iv(x, y)$. Then

$$0 = f'(z) = u_x + iv_x = v_y - iu_y$$

This implies

$$u_x = u_y = v_x = v_y = 0 \quad \text{in } D$$

Let

- the line segment joining two point P and P' be lying entirely in D ,
- s denote the arc-length function of the line segment measuring from point P ,
- $\mathbf{a}$ denote the unit vector along the line segment.

Then the directional derivative of u in the direction $\mathbf{a}$ is given by

$$\frac{du}{ds} = \nabla u \cdot \mathbf{a} = (u_x, u_y) \cdot (a_1, a_2) = 0.$$

Thus u is constant on the line segment joining P and P' . Since any two points P and Q in D can be joint by finite number of line segments. It follows that u takes the same value

at P and at Q . This implies that u is constant throughout D . Similarly argument leads to the same conclusion for function v .

eg:wwy:3.5.3

Example 12.48 (April 2010 Exam Q4(a); Sept 2011 Exam Q2(c)). Suppose that $f(z)$ and its conjugate $\overline{f(z)}$ are both analytic in a domain D . Show that f must be constant throughout D . (10 marks; 7 marks)

Solution. Let $f(z) = u(x, y) + iv(x, y)$. Then $\overline{f(z)} = u(x, y) + i(-v(x, y)) =: u + iV$.

Since both $f(z)$ and $\overline{f(z)}$ are analytic, we have

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}; \quad \frac{\partial u}{\partial x} = \frac{\partial V}{\partial y} = -\frac{\partial v}{\partial y}, \quad \frac{\partial u}{\partial y} = -\frac{\partial V}{\partial x} = \frac{\partial v}{\partial x}.$$

Hence

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} = -\frac{\partial u}{\partial x} \Rightarrow \frac{\partial u}{\partial x} = 0, \quad \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x} \Rightarrow \frac{\partial v}{\partial x} = 0.$$

and

$$f'(z) = u_x + iv_x = 0 \Rightarrow f(z) = \text{constant}$$

throughout D . □

A variation of the above question can also be asked as follows.

Example 12.49 (Sept 2013 Exam Q2(c)). Let f be analytic (holomorphic) in a domain D . Prove that if f is real-valued for all $z \in D$, then f is constant throughout D .

Solution. Since f is real-valued, we have $f(z) = \overline{f(z)}$, applying the prove in the previous example and we are done. □

12.3.4 Harmonic Functions

If the derivatives in the Cauchy-Riemann equations are sufficiently smooth and if we differentiate in the following way

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \Rightarrow \frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 v}{\partial x \partial y}, \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x} \Rightarrow \frac{\partial^2 u}{\partial y^2} = -\frac{\partial^2 v}{\partial y \partial x}$$

we obtain $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \frac{\partial^2 v}{\partial x \partial y} - \frac{\partial^2 v}{\partial y \partial x} \Rightarrow \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0$. This is the famous Laplace equation.

Definition 12.11. A real-valued function of two variable $h(x, y)$ is said to be *harmonic* if it satisfies the *Laplace equation*

$$h_{xx} + h_{yy} = 0. \tag{12.5}$$

e:laplace

It is not difficult to derive the following theorem relating holomorphic functions and harmonic functions.

Theorem 12.50. If $f(z) = u(x, y) + iv(x, y)$ is holomorphic in a domain D , then u and v are harmonic.

Definition 12.12. If $u(x, y)$ and $v(x, y)$ are harmonic functions defined in domain D and satisfy the Cauchy-Riemann equations throughout D , then we say that v is a *harmonic conjugate* of u .

Theorem 12.51. A function $f(z) = u(x, y) + iv(x, y)$ is holomorphic in a domain D if and only if v is a harmonic conjugate of u .

Example 12.52 (April 2009 Exam, Q3(a)).

- (i) Show that $u(x, y) = 2x(1 - y)$ is harmonic. (3 marks)

Solution. Since $u_{xx} = u_{yy} = 0$, we have $u_{xx} + u_{yy} = 0$. □

- (ii) Find a function $v(x, y)$ such that $f(x + iy) = u(x, y) + iv(x, y)$ is analytic (holomorphic). (5 marks)

Solution. $u_x = 2(1 - y) = v_y \Rightarrow v = 2y - y^2 + f(x)$
 $u_y = -2x = -v_x = -f'(x) \Rightarrow f(x) = x^2$
Hence, $v = x^2 + 2y - y^2$. □

- (iii) Express function f in terms of z . (7 marks)

Solution.

$$f(z) = 2x - 2xy + i(x^2 + 2y - y^2) = 2z + i(x^2 - y^2 + 2ixy) = 2z + iz^2.$$

□

12.3.5 Cauchy-Riemann Equations in Polar Coordinates

Sometimes, if the domain is a disk, it is more convenient to express the variable of the function in trigonometric form, i.e. $z = r(\cos \theta + i \sin \theta) \neq 0$. Then

$$r^2 = x^2 + y^2 \Rightarrow \frac{\partial r}{\partial x} = \frac{x}{r} = \cos \theta, \quad \frac{\partial r}{\partial y} = \frac{y}{r} = \sin \theta$$

and

$$x = r \cos \theta, \quad y = r \sin \theta \Rightarrow \frac{\partial \theta}{\partial x} = -\frac{\sin \theta}{r}, \quad \frac{\partial \theta}{\partial y} = \frac{\cos \theta}{r}.$$

Using chain rule, the first-order partial derivatives, we have

$$\begin{aligned} \frac{\partial u}{\partial x} &= \frac{\partial u}{\partial r} \frac{\partial r}{\partial x} + \frac{\partial u}{\partial \theta} \frac{\partial \theta}{\partial x} = \frac{\partial u}{\partial r} \cos \theta - \frac{1}{r} \frac{\partial u}{\partial \theta} \sin \theta \\ \frac{\partial u}{\partial y} &= \frac{\partial u}{\partial r} \frac{\partial r}{\partial y} + \frac{\partial u}{\partial \theta} \frac{\partial \theta}{\partial y} = \frac{\partial u}{\partial r} \sin \theta + \frac{1}{r} \frac{\partial u}{\partial \theta} \cos \theta \end{aligned}$$

Similarly, we also have

$$\begin{aligned} \frac{\partial v}{\partial x} &= \frac{\partial v}{\partial r} \frac{\partial r}{\partial x} + \frac{\partial v}{\partial \theta} \frac{\partial \theta}{\partial x} = \frac{\partial v}{\partial r} \cos \theta - \frac{1}{r} \frac{\partial v}{\partial \theta} \sin \theta \\ \frac{\partial v}{\partial y} &= \frac{\partial v}{\partial r} \frac{\partial r}{\partial y} + \frac{\partial v}{\partial \theta} \frac{\partial \theta}{\partial y} = \frac{\partial v}{\partial r} \sin \theta + \frac{1}{r} \frac{\partial v}{\partial \theta} \cos \theta. \end{aligned}$$

Cauchy-Riemann equations (12.3) imply that

$$\begin{aligned} \left[\frac{\partial u}{\partial r} - \frac{1}{r} \frac{\partial v}{\partial \theta} \right] \cos \theta - \left[\frac{1}{r} \frac{\partial u}{\partial \theta} + \frac{\partial v}{\partial r} \right] \sin \theta &= 0 \\ \left[\frac{\partial u}{\partial r} - \frac{1}{r} \frac{\partial v}{\partial \theta} \right] \sin \theta + \left[\frac{1}{r} \frac{\partial u}{\partial \theta} + \frac{\partial v}{\partial r} \right] \cos \theta &= 0 \end{aligned}$$

In order that the above two equations are satisfied for all θ , we must have

$$\frac{\partial u}{\partial r} = \frac{1}{r} \frac{\partial v}{\partial \theta} \quad \text{and} \quad \frac{\partial v}{\partial r} = -\frac{1}{r} \frac{\partial u}{\partial \theta}$$

which is the Cauchy-Riemann equations expressed in polar coordinates.

With some calculation, we can find that

$$f'(z) = (\cos \theta - i \sin \theta) \left[\frac{\partial u}{\partial r} + i \frac{\partial v}{\partial r} \right]. \quad (12.6)$$

This formula is useful in computing the derivative when function is given in polar form.

eg:wwy:3.4.1

Example 12.53. Given $f(z) = \frac{1}{z}$, $z \neq 0$. Find $u(r, \theta)$, $v(r, \theta)$, u_r and v_r .

Solution. Since $f(z) = \frac{1}{z} = \frac{1}{r}(\cos \theta - i \sin \theta)$,

$$u(r, \theta) = \frac{\cos \theta}{r}, \quad v(r, \theta) = -\frac{\sin \theta}{r}$$

and

$$u_r(r, \theta) = -\frac{\cos \theta}{r^2}, \quad v_r(r, \theta) = \frac{\sin \theta}{r^2}$$

□

12.4 Elementary Functions

In Calculus, we learned these elementary functions: exp, sin, cos, tan, csc, sec, cot, log, arcsin, arccos, arctan, sinh, cosh, tanh, cosech, sech, coth, etc. They are defined on some subset of $\mathbb{R}$. How can we generalise them to make them sensible on $\mathbb{C}$?

In order to define half of the functions mentioned above, the key is the exponential function. After studying the properties of exponential function, we proceed to trigonometric and hyperbolic functions.

12.4.1 Complex Exponential Function

We learned in Calculus that an exponential function is a mapping

$$\exp : \mathbb{R} \rightarrow \mathbb{R}, \quad x \mapsto e^x$$

where

$$e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n} \right)^n.$$

Can we define exponential function in complex analysis without Calculus as a pre-requisite? The answer is yes and many complex analysis books defined exponential function as

$$\exp : \mathbb{C} \rightarrow \mathbb{C}, \quad z \mapsto e^z = 1 + \sum_{n=1}^{\infty} \frac{z^n}{n!}.$$

This is a nice definition but requires the discussion of “convergence” of series, which we will introduce only in Topic 5. Therefore, I will follow the main reference Churchill and Brown [2008, 89] and Saff and Snider [2003, Section 1.4], a definition which requires the knowledge of exponential, sine and cosine functions from Calculus.

Definition 12.13. Let $z = x + iy$. The (complex) exponential function is defined as a mapping

$$\exp : \mathbb{C} \rightarrow \mathbb{C}, \quad z \mapsto w = \exp(z) = e^z := e^x(\cos y + i \sin y).$$

The following are the basic properties of an exponential function:

1. $e^z \neq 0$ for all $z \in \mathbb{C}$.

This is because $|e^z| = |e^x| |\cos y + i \sin y| = |e^x| > 0$.

2. The *exponential function* is $2\pi i$ -periodic.

This is because

$$e^{2k\pi i} = \cos(2k\pi) + i \sin(2k\pi) = 1, \quad \text{for } k \in \mathbb{Z}.$$

3. e^z is an entire function.

thm:4.1.1

Theorem 12.54. *The exponential function is entire and $\frac{d}{dz}e^z = e^z$.*

Proof: By definition,

$$f(z) = e^z = e^x \cos y + i e^x \sin y$$

The partial derivatives satisfy Cauchy-Riemann equations

$$u_x = e^x \cos y = v_y, \quad u_y = -e^x \sin y = -v_x \quad \text{for all } (x, y) \in \mathbb{R}^2$$

and are continuous. By Sufficient Conditions of Complex Differentiability Theorem [12.43](#), [thm:3.3.2](#), e^z is holomorphic for all $z \in \mathbb{C}$ and hence is entire and

$$\frac{d}{dz}e^z = u_x + i v_x = e^x(\cos y + i \sin y) = e^z.$$

Theorem 12.55 (Algebraic Properties of Exponential Function). *If z, z_1 and z_2 are complex numbers, then*

$$1. e^0 = 1$$

$$2. e^{z_1} e^{z_2} = e^{z_1+z_2}$$

$$3. \frac{e^{z_1}}{e^{z_2}} = e^{z_1-z_2}$$

$$4. (e^z)^n = e^{nz}, \quad n = 0, \pm 1, \pm 2, \dots$$

$$5. (e^z)^{-1} = (e^{-1})^z$$

With the notion of exponential function, it is possible for us to write down the *Euler Formula* (Leonhard Euler, 1707~1783):

$$e^{i\theta} = \cos \theta + i \sin \theta.$$

and any $z = r(\cos \theta + i \sin \theta) \in \mathbb{C}$ can be written in the *exponential form*:

$$z = r e^{i\theta} \quad \text{or} \quad z = |z| e^{i \arg z}.$$

One of the benefit of exponential form is that it allows use to evaluate the product and quotient of complex numbers compactly: Suppose $z_1 = r_1 e^{i\theta_1}$, $z_2 = r_2 e^{i\theta_2}$, then

$$z_1 \cdot z_2 = r_1 r_2 e^{i(\theta_1+\theta_2)}.$$

and the $2\pi i$ -periodicity of exponential function allow us to write the exponential form as

$$z = r e^{i(\theta+2k\pi)}, \quad k = 0, \pm 1, \pm 2, \dots$$

This form is very useful in multi-valued complex function.

eg:wwy:1.8.1

Example 12.56. Express $z = -1 - i$ in exponential form.

Solution. From $|z| = \sqrt{1+1} = \sqrt{2}$ and $\text{Arg } z = -\pi + \tan^{-1} \frac{-1}{-1} = -\frac{3\pi}{4}$, we have

$$z = \sqrt{2}e^{-\frac{3\pi}{4}i}.$$

□

If the complex variable z is expressed in exponential form $z = re^{i\theta}$, the value of the function will also be expressed in terms of the polar system as

$$f(re^{i\theta}) = \tilde{u}(r, \theta) + i\tilde{v}(r, \theta).$$

eg:2.1.2p

Example 12.57. Determine the real part and imaginary part of

$$f(z) = |z|^2$$

in polar coordinate form.

Solution: In polar coordinates,

$$f(re^{i\theta}) = r^2, \quad u(r, \theta) = r^2 \quad \text{and} \quad v(r, \theta) = 0.$$

eg:2.1.1p

Example 12.58. Determine the real part and imaginary part of

$$f(z) = z^2$$

in polar coordinate form.

Solution. *Solution:* When polar coordinates are used,

$$f(re^{i\theta}) = r^2e^{i2\theta} = r^2 \cos(2\theta) + ir^2 \sin(2\theta)$$

Consequently, the real part and imaginary part of f in polar coordinates are, respectively

$$u(r, \theta) = r^2 \cos(2\theta) \quad \text{and} \quad v(r, \theta) = r^2 \sin(2\theta).$$

□

Since $|e^{i\theta}| = 1$ for $0 \leq \theta \leq 2\pi$, the equation of circle with a centre at z_0 and a radius r is given by

$$z = z_0 + re^{i\theta}, \quad 0 \leq \theta \leq 2\pi$$

In Calculus, the “real” exponential function is always positive-valued. However, the value of a complex exponential function can be negative or complex.

eg:4.1.1

Example 12.59. Find complex number z such that $e^z = -1$.

Solution.

$$e^z = -1 = e^{i\pi+2k\pi i} \Rightarrow z = i\pi + 2k\pi i = (2k+1)\pi i, \quad k = 0, \pm 1, \pm 2, \dots$$

□

Example 12.60. Find complex number z such that $e^z = i$.

Solution.

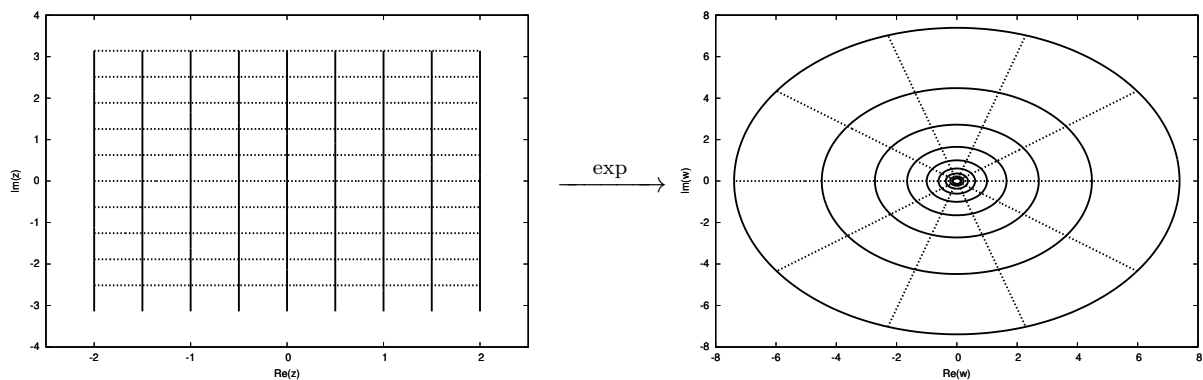
$$e^z = i = e^{i\frac{\pi}{2}+2k\pi i} \Rightarrow z = \left(\frac{1}{2} + 2k\right)\pi i, \quad k = 0, \pm 1, \pm 2, \dots$$

□

With the above characterisation of the exponential function, we learned that complex exponential function is not one-to-one, and all values of e^z are assumed in any infinite horizontal strip of width 2π in the z -plane. The horizontal strip

$$-\infty < x < \infty, \quad -\pi < y \leq \pi$$

is called the *fundamental region* of the complex exponential function. Let us look at how we can plot the mapping of a portion of the fundamental region.



In the above mapping, we find that a point on the vertical line $x = a$ in the fundamental region is of the form (a, y) , $-\pi < y \leq \pi$. This vertical line will be mapped under the exponential function to circle with equation e^{2a} , a circle with radius e^a centred at $(0, 0)$:

$$u = e^a \cos y, \quad v = e^a \sin y \Rightarrow u^2 + v^2 = e^{2a}.$$

Let (x, b) denote a point on the horizontal line $y = b$. The image of (x, b) is a point (u, v) satisfying

$$u = e^x \cos b, \quad v = e^x \sin b \Rightarrow \frac{v}{u} = \frac{\sin b}{\cos b} \Rightarrow v = u \tan b.$$

Thus the image of the horizontal line $y = b$ is a ray $v = u \tan b$.

12.4.2 Complex Trigonometric Functions

The complex trigonometric functions are defined below as a linear combination of “exponential” functions.

Definition 12.14. The *(complex) cosine function* is defined as

$$\cos : \mathbb{C} \rightarrow \mathbb{C}, \quad z \mapsto \cos z = \frac{e^{iz} + e^{-iz}}{2}. \quad (12.7)$$

e:cos

The *(complex) sine function* is defined as

$$\sin : \mathbb{C} \rightarrow \mathbb{C}, \quad z \mapsto \sin z = \frac{e^{iz} - e^{-iz}}{2i}. \quad (12.8)$$

e:sin

The *(complex) tangent function* is defined as

$$\tan : \mathbb{C} \setminus \left\{ \frac{\pi}{2} \pm k\pi : k \in \mathbb{Z} \right\} \rightarrow \mathbb{C}, \quad z \mapsto \tan z = \frac{1}{i} \frac{e^{iz} - e^{-iz}}{e^{iz} + e^{-iz}}. \quad (12.9)$$

e:tan

The *secant*, *cosecant*, and *cotangent* functions are defined as the inverse of cosine, sine and tangent functions respectively as follows:

$$\sec z = \frac{1}{\cos z}, \quad \operatorname{cosec} z = \frac{1}{\sin z}, \quad \cot z = \frac{1}{\tan z}.$$

Consider the case of sine function, it is not difficult to find that when $z = x$ is a real number, then by the definition of exponential function, we have

$$\frac{e^{ix} - e^{-ix}}{2i} = \frac{(\cos x + i \sin x) - (\cos(-x) + i \sin(-x))}{2i} = \frac{i \sin x - i(-\sin x)}{2i} = \sin x.$$

Since exponential function is entire and never take value 0, this implies that sine and cosine functions are entire. It is not difficult to prove the following properties:

- $\sin(-z) = -\sin z$, $\cos(-z) = \cos z$. In other words, sine is odd and cosine is even;
- $\sin^2 z + \cos^2 z = 1$;
- $\sin(z_1 \pm z_2) = \sin z_1 \cos z_2 \pm \cos z_1 \sin z_2$;
- $\cos(z_1 \pm z_2) = \cos z_1 \cos z_2 \mp \sin z_1 \sin z_2$;
- $\sin(z + 2\pi) = \sin z$, $\cos(z + 2\pi) = \cos z$. In other words, sine and cosine functions are 2π -periodic.
- sine and cosine functions are entire.

The derivatives of complex trigonometric function are analogous to that of real trigonometric functions.

Theorem 12.61. For any appropriate z ,

$$\begin{aligned}\frac{d}{dz} \sin z &= \cos z, & \frac{d}{dz} \cos z &= -\sin z, & \frac{d}{dz} \tan z &= \sec^2 z, \\ \frac{d}{dz} \sec z &= \sec z \tan z, & \frac{d}{dz} \operatorname{cosec} z &= \operatorname{cosec} z \cot z, & \frac{d}{dz} \cot z &= -\operatorname{cosec}^2 z.\end{aligned}$$

Now, let us solve some equations with trigonometric functions.

eg:4.3.1

Example 12.62. Solve $\sin z = 0$.

Solution:

$$\sin z = \frac{e^{iz} - e^{-iz}}{2i} = 0 \Rightarrow e^{2iz} - 1 = 0 \Rightarrow 2iz = \log 1 = 2k\pi i, \quad k = 0, \pm 1, \pm 2, \dots$$

Thus the solutions are

$$z = k\pi, \quad k = 0, \pm 1, \pm 2, \dots$$

eg:4.3.2

Example 12.63. Solve $\cos z = 0$

Solution.

$$\cos z = \frac{e^{iz} + e^{-iz}}{2} = 0 \Rightarrow e^{2iz} + 1 = 0 \Rightarrow 2iz = \log(-1) = \pi i + 2k\pi i, \quad k = 0, \pm 1, \pm 2, \dots$$

Thus the solutions are

$$z = \frac{\pi}{2} + k\pi, \quad k = 0, \pm 1, \pm 2, \dots$$

□

The mapping properties of the sine transformation are given below.

eg:4.3.3

Example 12.64. What is the image of the vertical line $x = a$ in the z -plane under the mapping $w = \sin z$?

Solution: The real and imaginary parts of $w = \sin z$ are

$$\begin{aligned}w &= \sin x \cos iy + \cos x \sin iy = \sin x \times \frac{e^y + e^{-y}}{2} + \cos x \times \frac{e^y - e^{-y}}{2i} \\ &= \underbrace{\frac{(e^y + e^{-y}) \sin x}{2}}_{u(x,y)} + i \underbrace{\frac{(e^{-y} + e^y) \cos x}{2}}_{v(x,y)}\end{aligned}$$

Hence the image of the vertical line $z = a + iy$ is the point (u, v) in the w -plane where

$$u = \frac{(e^y + e^{-y}) \sin a}{2}, \quad v = \frac{(e^{-y} + e^y) \cos a}{2}$$

which reduce to

$$\left(\frac{u}{\sin a}\right)^2 - \left(\frac{v}{\cos a}\right)^2 = \left(\frac{e^y + e^{-y}}{2}\right)^2 - \left(\frac{e^y - e^{-y}}{2}\right)^2 = \frac{2 - (-2)}{4} = 1.$$

This represents a *hyperbola* with vertices $(\sin a, 0)$ and $(-\sin a, 0)$ and slant asymptotic $v = \left(\frac{\cos a}{\sin a}\right)u$ and $v = -\left(\frac{\cos a}{\sin a}\right)u$. Since $(a, 0)$ is on the line $x = a$ in the z -plane, the point $(\sin a, 0)$ must be on the image of the line. Therefore the image of the vertical line $x = a$ with $-\pi/2 < a < \pi/2$ and $a \neq 0$ under the sine mapping is the branch of the hyperbola that contains the point $(\sin a, 0)$. Because $\sin(-z) = -\sin z$, it also follows that the image of the line $x = -a$ is the branch of the hyperbola containing the point $(-\sin a, 0)$.

eg: 4.3.4

Example 12.65. Find the image of the horizontal line $y = b$ in the z -plane under the mapping $w = \sin z$.

Solution. The image of the horizontal line $y = b$ under the sine mapping is given by

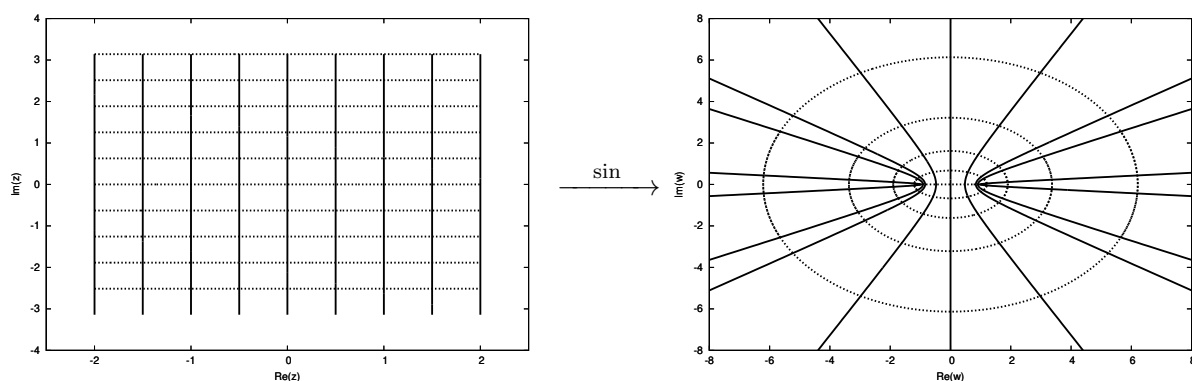
$$u = \frac{(e^b + e^{-b}) \sin x}{2}, \quad v = \frac{(e^{-b} + e^b) \cos x}{2}$$

where $x \in \mathbb{R}$. Rearranging the above equations to eliminate x , we have

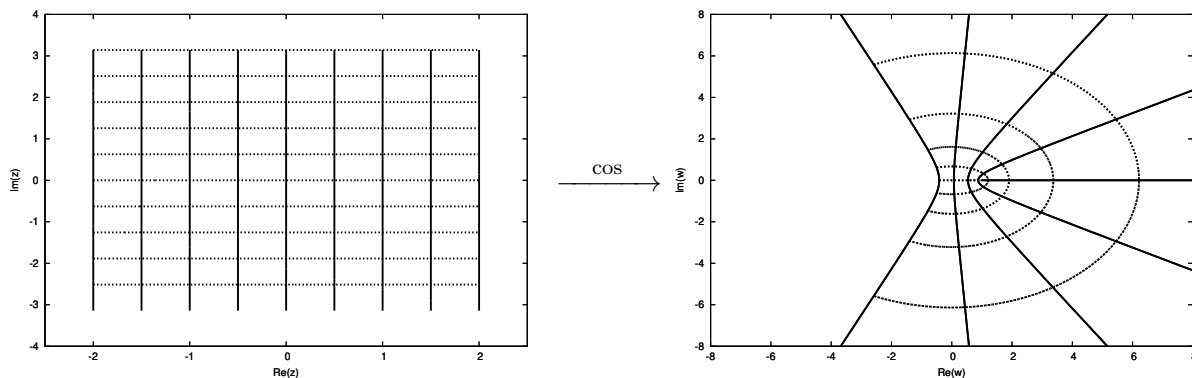
$$\left(\frac{u}{e^b + e^{-b}}\right)^2 = \left(\frac{\sin x}{2}\right)^2, \quad \left(\frac{v}{e^{-b} + e^b}\right)^2 = \left(\frac{\cos x}{2}\right)^2 \Rightarrow \left(\frac{u}{e^b + e^{-b}}\right)^2 + \left(\frac{v}{e^{-b} + e^b}\right)^2 = \frac{1}{4}.$$

This implies that the image of the line $y = b$ is an ellipse. □

Base on the above analysis, we are able to sketch the sine mapping from z -plane to w -plane.



The sketch of the cosine mapping from z -plane (with a region $[-2, 2] \times (-\pi, \pi]$ to w -plane should be similar:



12.4.3 Hyperbolic Functions

The complex hyperbolic functions are defined by replacing the variable x in the exponential function with variable z as follows.

Definition 12.15. The (complex) hyperbolic cosine function is defined as

$$\cosh : \mathbb{C} \rightarrow \mathbb{C}, \quad z \mapsto \cosh z = \frac{e^z + e^{-z}}{2}. \quad (12.10)$$

The (complex) hyperbolic sine function is defined as

$$\sinh : \mathbb{C} \rightarrow \mathbb{C}, \quad z \mapsto \sinh z = \frac{e^z - e^{-z}}{2}. \quad (12.11)$$

The (complex) hyperbolic tangent function is defined as

$$\tanh : \mathbb{C} \setminus \left\{ \frac{\pi i}{2} \pm k\pi i \right\} \rightarrow \mathbb{C}, \quad z \mapsto \tanh z = \frac{\sinh z}{\cosh z}. \quad (12.12)$$

The hyperbolic secant, hyperbolic cosecant, and hyperbolic cotangent functions are defined as the inverse of hyperbolic cosine, hyperbolic sine and hyperbolic tangent functions respectively as follows:

$$\operatorname{sech} z = \frac{1}{\cosh z}, \quad \operatorname{cosech} z = \frac{1}{\sinh z}, \quad \operatorname{coth} z = \frac{1}{\tanh z}.$$

Since the linear combination of entire functions is entire, it follows that $\cosh z$ and $\sinh z$ are entire.

By looking at the definitions of hyperbolic functions and trigonometric functions, we find that they are related by

$$\sin z = -i \sinh(iz), \quad \cos z = \cosh(iz), \quad \tan z = -i \tanh(iz).$$

It is also not difficult for us to obtain the properties of hyperbolic functions based on the properties of trigonometric functions:

- $\sinh(-z) = -\sinh z$, $\cosh(-z) = \cosh z$. In other words, hyperbolic sine function is odd and hyperbolic cosine function is even;
- $\cosh^2 z - \sinh^2 z = 1$;
- $\sinh(z_1 \pm z_2) = \sinh z_1 \cosh z_2 \pm \cosh z_1 \sinh z_2$;
- $\cosh(z_1 \pm z_2) = \cosh z_1 \cosh z_2 \pm \sinh z_1 \sinh z_2$;
- $\sinh$ and $\cosh$ are $2\pi i$ -periodic entire functions.

It is also easy to obtain derivatives of hyperbolic functions by using definition.

Theorem 12.66. For appropriate z ,

$$\begin{aligned} \frac{d}{dz} \sinh z &= \cosh z, & \frac{d}{dz} \cosh z &= \sinh z, & \frac{d}{dz} \tanh z &= \operatorname{sech}^2 z, \\ \frac{d}{dz} \operatorname{sech} z &= -\operatorname{sech} z \tanh z, & \frac{d}{dz} \operatorname{cosech} z &= -\operatorname{cosech} z \cot z, & \frac{d}{dz} \operatorname{coth} z &= -\operatorname{cosech}^2 z. \end{aligned}$$

Example 12.67. Estimate the output of the following Fortran program by using hand calculation with calculator:

```
1 (define z (make-rectangular 1 (/ (* 4.0 (atan 1)) 3.0)))
2 (display (string-append "Given z = " (number->string z) "\n"))
3 (display (string-append "exp(z) = " (number->string (exp z)) "\n"))
4 (display (string-append "sin(z) = " (number->string (sin z)) "\n"))
5 (display (string-append "cos(z) = " (number->string (cos z)) "\n"))
```

```

6 (display (string-append "tan(z) = " (number->string (tan z)) "\n"))
7 (display (string-append "sinh(z) = " (number->string (sinh z)) "\n"))
8 (display (string-append "cosh(z) = " (number->string (cosh z)) "\n"))
9 (display (string-append "tanh(z) = " (number->string (tanh z)) "\n"))

```

Solution. Given $z = (1, \frac{\pi}{3})$. Then

$$\begin{aligned}
e^z &= e^1(\cos \frac{\pi}{3}, \sin \frac{\pi}{3}) = (1.35914091, 2.35410112) \\
e^{iz} &= e^{-\pi/3}(\cos 1, \sin 1) = (0.18960278, 0.29528884) \\
e^{-iz} &= e^{\pi/3}(\cos(-1), \sin(-1)) = (1.53967458, -2.39790108) \\
\sin z &= \frac{e^{iz} - e^{-iz}}{2i} = \frac{(-1.3500718, 2.69318992)}{2i} = (1.34659496, 0.6750359)
\end{aligned}$$

```

Given z = ( 1.0000000000000000 , 1.0471975511965976 )
exp(z) = ( 1.3591409142295230 , 2.3541011180911466 )
sin(z) = ( 1.3465949581259609 , 0.67503589827378041 )
cos(z) = ( 0.86463867926707882 , -1.0513061223909463 )
tan(z) = ( 0.24537876046783191 , 1.0790681863193832 )
sinh(z) = ( 0.58760059682190091 , 1.3363470298378193 )
cosh(z) = ( 0.77154031740762208 , 1.0177540882533274 )
tanh(z) = ( 1.1117850525522133 , 0.26547316157381218 )

```

□

12.5 Inverse Mapping and Multi-Valued Functions

Now, we will proceed to construct the “inverse mapping” for some of the functions mentioned in the previous sections. The simplest may be to consider the inverse mapping of $w = z^n$, which we will try to do in Section 12.5.1. Then we will look into the inverse mapping of exponential function, trigonometric functions and hyperbolic functions.

12.5.1 n th Root Function

Consider the function $z \mapsto z^n = w$. Its “inverse” function is a function where w is given but z is unknown and we need to find z from w . In Topic 1, we learned the z is the n th roots, i.e.

$$w \mapsto z = w^{1/n} = \sqrt[n]{|w|} \left(\cos \frac{\text{Arg } w + 2k\pi}{n} + i \sin \frac{\text{Arg } w + 2k\pi}{n} \right), \quad k = 0, 1, \dots, n-1.$$

We can define the n th root function as a set-valued function below:

Definition 12.16. An n th root function is a set-valued mapping

$$(\)^{1/n} : \mathbb{C} \rightarrow 2^{\mathbb{C}}, \quad z \mapsto \{ \sqrt[n]{|z|} \exp \frac{i \arg z}{n} \}. \quad (12.13)$$

A *principal n th root function* is defined as

$$\sqrt[n]{} : \mathbb{C} \rightarrow \mathbb{C}, \quad z \mapsto \sqrt[n]{z} = \sqrt[n]{|z|} \exp \frac{i \text{Arg}(z)}{n}.$$

Can you see the problem of n th root function? An obvious problem with it is the range is not $\mathbb{C}$ but a power set of $\mathbb{C}$. It is not a function in Calculus! (and Complex Analysis, of course!)

On the other hand, principal n th root function is a function in the sense of Calculus. In the following example, we will see how principal square root function ($n = 2$) maps lines.

Example 12.68. Find the images of the line $x = a$ and $y = b$ under the mapping $z \mapsto \sqrt{z}$.

Solution. Let $w = u + iv = \sqrt{z}$. Then

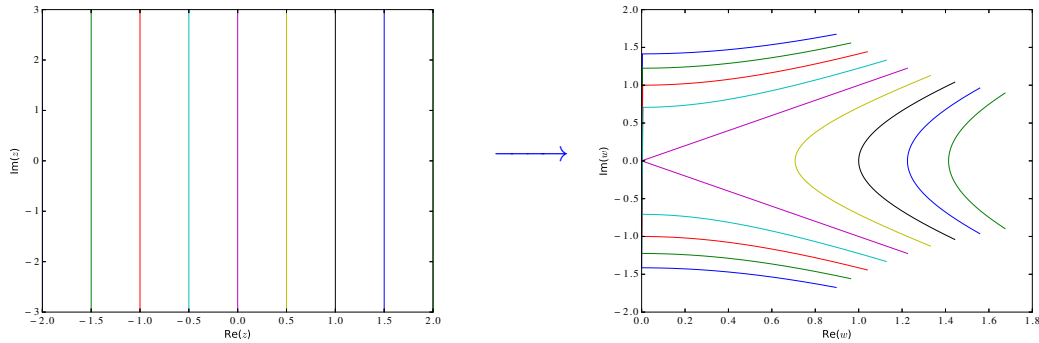
$$z = x + iy = w^2 = u^2 - v^2 + i(2uv) \Rightarrow u^2 - v^2 = x, \quad 2uv = y$$

Vertical line : $x = a$ in the z -plane will be mapping to the curve in w -plane with equation

$$u^2 - v^2 = a \quad (\text{equation of a hyperbola}) \quad (12.14)$$

e:2.5

If $a > 0$, then the function maps the right-half plane given by $\Re(z) > a$ onto the region in the right half-plane satisfying $u^2 - v^2 > a$ and lying to the right of the hyperbola $u^2 - v^2 = a$.

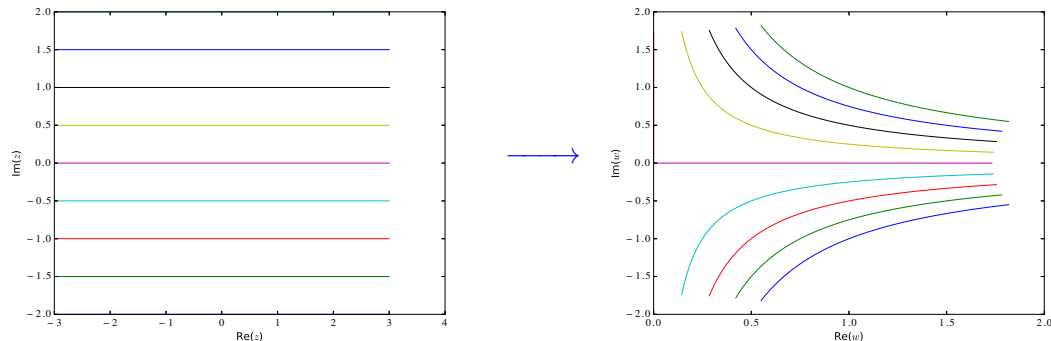


If $b > 0$, the horizontal line $y = b$ will be mapped to the hyperbola in the w -plane with equation

$$uv = \frac{b}{2} \quad (12.15)$$

e:2.6

The upper half plane $\Im(z) > b$ will be mapped onto the region in the first quadrant of the w -plane satisfying (12.15) and lying above the hyperbola with equation (12.15). The following figure shows the mapping:



□

It is possible for us to calculate the derivative of principal n th root function but not the n th root function (because it is not a function!):

$$(\sqrt[n]{z})' = \lim_{\Delta z \rightarrow 0} \frac{\sqrt[n]{z + \Delta z} - \sqrt[n]{z}}{\Delta z} = \lim_{\Delta z \rightarrow 0} \frac{1}{\frac{\Delta z}{\sqrt[n]{z + \Delta z} - \sqrt[n]{z}}} = \lim_{\Delta w \rightarrow 0} \frac{1}{\frac{(w + \Delta w)^n - w^n}{w + \Delta w - w}} = \frac{1}{nw^{n-1}} = \frac{w}{nw^n} = \frac{\sqrt[n]{z}}{nz} \quad (12.16)$$

e:root-n:deriv

where $w + \Delta w = \sqrt[n]{z + \Delta z} \Rightarrow (w + \Delta w)^n = z + \Delta z = w^n + \Delta z$ and $\Delta w \rightarrow 0$ as $\Delta z \rightarrow 0$.

We have to be very careful here because writing something below by copying the formula directly from Calculus does not work:

$$(\sqrt[n]{z})' \neq \frac{1}{n} z^{1/n-1}$$

This is because $z^{1/n-1} = z^{\frac{1-n}{n}}$ should have n values but the left hand side $(\sqrt[n]{z})'$ is single-valued.

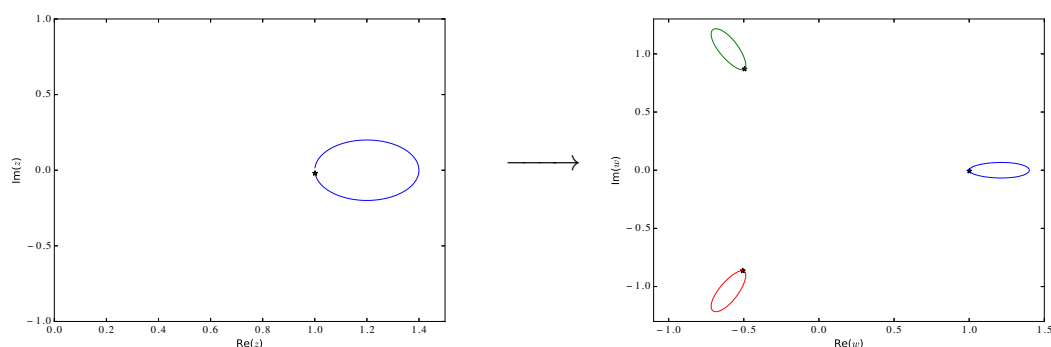
12.5.2 Single-Valued vs Multi-Valued Functions

Before we define “multi-valued function”, we need to recall the meaning of “function” (refer to UECM1303 Discrete Mathematics and Applications), i.e. for every element in domain, there is just one corresponding element in the range. In complex analysis, a function is also called “single-valued function”, just to contrast with a “multi-valued function”. A “multi-valued function” is not a function, however, it can be made into a normal function with the introduction of “Riemann surface”.

To illustrate the “Riemann surface”, we will study the cubic root function

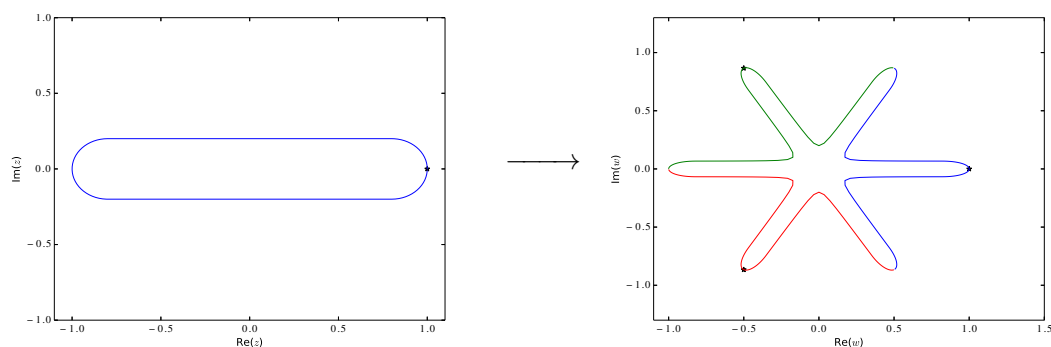
$$z \mapsto z^{1/3} = \sqrt[3]{|z|} \exp \frac{i \arg(z)}{3} \quad (12.17) \quad \text{e:root-3}$$

by looking at the image of the circle $C_1 : |z - 1.2| = 0.2$



The images are “continuous”. This is because the angle “ $\text{Arg}(z)$ ” never “change more than 2π ”.

However, when we pick a curve circling the origin, the angle “ $\text{Arg}(z)$ ” will change by more than 2π and we will have the image starting from “one root”, “jumping” to “another root”:



Let’s say we begin from point 1 and we circle the origin, when we come back to 1, we have

$$1 = \cos 2\pi + i \sin 2\pi \mapsto \exp \frac{i(2\pi)}{3} \neq 1 = \exp 0.$$

So our “mapping” is multi-valued.

Definition 12.17. A “mapping” $f : D \rightarrow \mathbb{C}$ is called a *multi-valued function* if

1. there is a value $z_0 \in D$ such that $f(z_0)$ can have more than one value; and
2. if $f(z_0) = w_1$ and $f(z_0) = w_2$ and $w_1 \neq w_2$, there is a continuous curve ℓ in D such that

$$\ell(0) = \ell(1) = z_0 \bigwedge f(\ell(0)) = w_1 \bigwedge f(\ell(1)) = w_2.$$

For the “multi-valued function” (12.17), ^{e:root-3} the origin is a “branch point” where any curve encircling it will cause the image to “break”, i.e. it will jump from one “root” to another as shown figure above.

Definition 12.18 (Branch Point). Given a multi-valued function f . A point z_0 in the z -plane is called a *branch point* if there is simple closed curve (we will define this in Topic 4) $\ell \subset D$ surrounding z_0 such that $\ell(0) = \ell(1) = z_0$ and $f(\ell(0)) \neq f(\ell(1))$.

Definition 12.19. Given a multi-valued function $f : D \rightarrow \mathbb{C}$. In $G \subset D$, $f|_G : G \rightarrow \mathbb{C}$ is a single-valued continuous function and for all $z \in G$, $f|_G(z) \in f(z)$. Then $f|_G$ is called a *single-valued continuous branch* of f in G .

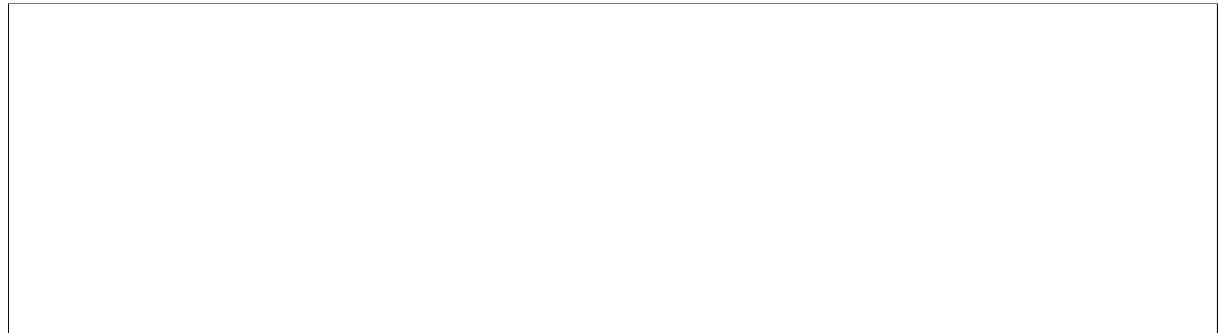
One of the greatest mathematician Georg Friedrich Bernhard Riemann (1826~1866) came up with an approach to turn “multi-valued functions” to “single-valued functions” with “Riemann surface”.

Definition 12.20. A *Riemann surface* is an extension of the ordinary complex plane to a surface that has more than one “sheet”. The multi-valued function will have only one value corresponding to each point on the Riemann surface.

So how can we create a Riemann surface for ^{e:root-3} (12.17)? We need three “sheets”:

$$f : \underbrace{\begin{cases} \{z \in \mathbb{C} : -\pi \leq \arg z < \pi\} \\ \{z \in \mathbb{C} : \pi \leq \arg z < 3\pi\} \\ \{z \in \mathbb{C} : 3\pi \leq \arg z < 5\pi\} \end{cases}}_{\text{“Riemann surface”}} \rightarrow \mathbb{C}, \quad z \mapsto z^{1/3}.$$

Visually, it looks like a figure shown below and f becomes a normal function!



Using the “Riemann sphere” from Section ^{sec:riemann-sphere} 12.6 we can “see” that 0 and ∞ are branch points for the n th root function. On the single-valued branch, the derivative ^{e:root-n:deriv} (12.16) remains true:

$$\frac{d}{dz} z^{1/n} = \frac{1}{n} \left[\frac{z^{1/n}}{z} \right] \quad (12.18)$$

12.5.3 Complex Logarithmic Function

In real analysis, the natural logarithm function $y = \ln(x)$ is often defined as an inverse function of the real exponential function. That is, for a given real number x , $\ln(x)$ is the unique solution to the equation $e^y = x$.

The real exponential function is one-to-one on the real line, there is no ambiguity involved in defining this inverse function. However, complex exponential function is periodic! Given a fixed complex number z , the equation

$$e^w = z$$

has infinitely many solutions! To see why this is the case, suppose that $w = u + iv$ is a solution of $e^w = z$. Then $|e^w| = |z|$. Therefore, we have

$$e^u = |z| \quad \text{and} \quad v = \arg(z) \quad \text{or equivalently,} \quad u = \ln |z|, \quad v = \arg(z).$$

Therefore

$$e^w = z \Leftrightarrow w = \ln |z| + i \arg(z). \quad (12.19)$$

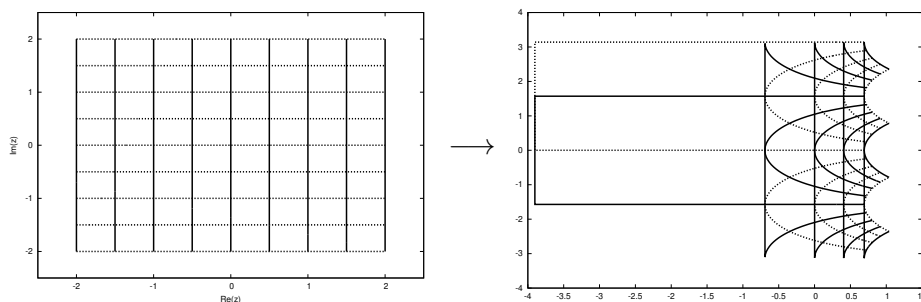
The set of w given in (12.19) is called the complex logarithm of z and denoted by $\log z$.

Definition 12.21. The multiple-valued function $\log z$ is defined by

$$\log : \mathbb{C}^\times \rightarrow \mathbb{C}, \quad z \mapsto \log z = \ln |z| + i \arg(z) \quad (12.20)$$

is called the (complex) logarithm of z .

The logarithmic mapping of the region $[-2, 2] \times [-2, 2]$ to $\mathbb{C}$ is shown below.



Example 12.69. Find the solution to (a) $e^w = i$, (b) $e^w = -2$, (c) $e^w = 5$.

Solution. (a) $w = \ln |i| + i \arg(i) = i(\frac{\pi}{2} + 2k\pi)$, $k = 0, \pm 1, \pm 2, \dots$

(b) $w = \ln |-2| + i \arg(-2) = \ln 2 + i(\pi + 2k\pi)$, $k = 0, \pm 1, \pm 2, \dots$

(c) $w = \ln 5 + i \arg(5) = \ln 5 + 2k\pi i$, $k = 0, \pm 1, \pm 2, \dots$

□

Remark 12.70. The real logarithm of 5, $\ln 5$, is unique in Calculus whereas the complex logarithm of 5, $\log 5$, has infinitely many values.

Example 12.71 (Sept 2012 Exam, Q1(a)(i)). Find all values of $\log(-1 + i\sqrt{3})$. (4 marks)

Solution. $\log(-1 + i\sqrt{3}) = \log(2e^{i2\pi/3 + 2k\pi i}) = \ln 2 + i(\frac{2\pi}{3} + 2k\pi)$, $k = 0, \pm 1, \pm 2, \dots$. □

Theorem 12.72 (Algebraic Properties of Complex Logarithm). Let z, z_1, z_2 be any complex numbers. Then

- $\log(z_1 z_2) = \log z_1 + \log z_2$

- $\log \frac{z_1}{z_2} = \log z_1 - \log z_2$

- $\log z^n = \underbrace{\log z + \dots + \log z}_{n \text{ times}}$

Proof: By using the properties of natural logarithmic function, modulus and argument function, we have

$$\begin{aligned}\log(z_1 z_2) &= \ln |z_1 z_2| + i \arg(z_1 z_2) = \ln |z_1| + \ln |z_2| + i(\arg(z_1) + \arg(z_2)) \\ &= (\ln |z_1| + i \arg(z_1)) + (\ln |z_2| + i \arg(z_2)) = \log z_1 + \log z_2\end{aligned}$$

$$\begin{aligned}\log\left(\frac{z_1}{z_2}\right) &= \ln \left|\frac{z_1}{z_2}\right| + i \arg\left(\frac{z_1}{z_2}\right) = \ln |z_1| - \ln |z_2| + i(\arg(z_1) - \arg(z_2)) \\ &= (\ln |z_1| + i \arg(z_1)) - (\ln |z_2| + i \arg(z_2)) = \log z_1 - \log z_2.\end{aligned}$$

$$\begin{aligned}\log(z^n) &= \ln |z \times \cdots \times z| + i \arg(z \times \cdots \times z) = \ln |z| + \cdots + \ln |z| + i(\arg(z) + \cdots + \arg(z)) \\ &= (\ln |z| + i \arg(z)) + \cdots + (\ln |z| + i \arg(z)) = \log z + \cdots + \log z.\end{aligned}$$

In order to make the logarithmic function “unique”, we need to define the principal value of a complex logarithm using “principal argument” from Topic 1.

Definition 12.22. The *principal value of the complex logarithm* is defined as

$$\text{Log} z = \ln |z| + i \text{Arg}(z). \quad (12.21)$$

e:Log

eg:4.2.2

Example 12.73. Compute the principal value of the complex logarithm for (a) i , (b) $1 + i$, (c) -1

Solution. (a) $\text{Log}(i) = \ln |i| + i \text{Arg}(i) = \pi/2$

$$(b) \text{Log}(1 + i) = \ln |1 + i| + i \text{Arg}(1 + i) = \ln \sqrt{2} + i \frac{\pi}{4}$$

$$(c) \text{Log}(-1) = \ln |-1| + i \text{Arg}(-1) = \pi i$$

□

The principal value of logarithm is not a continuous on $\mathbb{C}$ according to its properties listed below:

- Log is discontinuous at $z = 0$ since $\text{Log} 0$ is undefined.
- Log is discontinuous at every point on the negative real axis, i.e. If $z = x + iy$, then

$$\lim_{y \rightarrow 0^+} \text{Log} z = \ln |x| + \pi i, \quad \lim_{y \rightarrow 0^-} \text{Log} z = \ln |x| - \pi i.$$

Log is continuous on the z -plane excluding the non-positive real axis, i.e.

$$\text{Log} : \mathbb{C} \setminus \{(x, 0) : x \leq 0\} \rightarrow \mathbb{C}, \quad z = r e^{i\theta} \mapsto \underbrace{\ln |r|}_u + i \underbrace{\theta}_v$$

where $-\pi < \theta < \pi$ and $r \neq 0$. Since the first-partial derivatives of u and v are continuous, Log is differentiable.

Since “Log” is a function, we can calculate its derivative:

$$\frac{d}{dz} \text{Log} z = \lim_{\Delta z \rightarrow 0} \frac{\text{Log}(z + \Delta z) - \text{Log} z}{\Delta z} = \lim_{\Delta z \rightarrow 0} \frac{1}{\frac{\Delta z}{\text{Log}(z + \Delta z) - \text{Log} z}} = \frac{1}{\lim_{\Delta z \rightarrow 0} \frac{\Delta z}{\text{Log}(z + \Delta z) - \text{Log} z}}.$$

Since $z = e^w$ and $\Delta w = \text{Log}(z + \Delta z) - \text{Log} z \rightarrow 0$ as $\Delta z \rightarrow 0$, we have

$$e^{\Delta w} = \frac{z + \Delta z}{z} \Rightarrow \Delta z = z e^{\Delta w} - z \Rightarrow \Delta z = e^w (e^{\Delta w} - 1).$$

Therefore,

$$\lim_{\Delta z \rightarrow 0} \frac{\Delta z}{\text{Log}(z + \Delta z) - \text{Log} z} = \lim_{\Delta w \rightarrow 0} \frac{e^w (e^{\Delta w} - 1)}{\Delta w} = e^w = z \Rightarrow \frac{d}{dz} \text{Log} z = \frac{1}{z}.$$

12.5.4 Inverse Trigonometric and Hyperbolic Functions

Definition 12.23. The inverse trigonometric and hyperbolic functions are defined below.

$$\begin{aligned} \bullet \arcsin z &= \frac{1}{i} \log(iz + \sqrt{1 - z^2}); & \operatorname{arcsinh} z &= \log(z + \sqrt{z^2 + 1}); \\ \bullet \arccos z &= \frac{1}{i} \log(z + i\sqrt{1 - z^2}); & \operatorname{arccosh} z &= \log(z + \sqrt{z^2 - 1}); \\ \bullet \arctan z &= \frac{1}{2i} \log \frac{1 + iz}{1 - iz}; & \operatorname{arctanh} z &= \frac{1}{2} \log \frac{1 + z}{1 - z}. \end{aligned}$$

These definitions can be obtained by solving equations, for example,

$$z = \tanh w = \frac{e^w - e^{-w}}{e^w + e^{-w}} = \frac{e^{2w} - 1}{e^{2w} + 1} \Rightarrow (z - 1)e^{2w} + z + 1 = 0 \Rightarrow w = \frac{1}{2} \log \frac{1 + z}{1 - z}.$$

Example 12.74. Find the values of $\arctan(2i)$.

Solution. $\arctan(2i) = \frac{1}{2i} \log\left(-\frac{1}{3}\right) = -\frac{i}{2} \left(\ln \frac{1}{3} + i\pi + 2k\pi i\right) = \frac{\pi}{2} + k\pi + \frac{i \ln 3}{2}, \quad k \in \mathbb{Z}. \quad \square$

Example 12.75. Find the derivative of $\arctan z$ on the single-valued branch.

Solution. At some “ k ”th branch, we have

$$\frac{d}{dz} \arctan z = \frac{d}{dz} \left\{ \frac{1}{2i} [\operatorname{Log}(1 + iz) - \operatorname{Log}(1 - iz) + 2k\pi i] \right\} = \frac{1}{2i} \left[\frac{i}{1 + iz} - \frac{-i}{1 - iz} \right] = \frac{1}{1 + z^2}$$

$\square$

12.5.5 Arbitrary Power Function

Let a be a real number, then the “arbitrary power” x^a is well defined in Calculus. If α is a complex number, is it possible for us to define “arbitrary power” z^α ? The answer is affirmative.

Definition 12.24. Let α be any complex number and $z \neq 0$. The *power function* is defined as the mapping

$$z \mapsto z^\alpha := \exp(\alpha \log z).$$

where $\log$ denotes the multiple-valued complex logarithm defined in Definition [12.21](#). [def:mv-log]

Example 12.76 (April 2010 Exam, Q4(b)). Determine the value of i^i . (5 marks)

Solution.

$$i^i = \exp(i \log i) = \exp(i(\ln |i| + i(\frac{\pi}{2} + 2k\pi))) = \exp\left(-\frac{\pi}{2} - 2k\pi\right)$$

where $k = 0, \pm 1, \pm 2, \dots$. $\square$

The principal value of z^α is given when $k = 0$, i.e. $PV(z^\alpha) = \exp(\alpha \operatorname{Log} z)$. This formula serves to define the principal branch of z^α :

$$\{z \in \mathbb{C} : |z| > 0, -\pi < \operatorname{Arg}(z) < \pi\} = \mathbb{C} \setminus (-\infty, 0].$$

Note that negative real axis and origin are deleted.

Example 12.77. Find the derivative of z^α on the single-valued branch.

Solution. $\frac{d}{dz} z^\alpha = \frac{d}{dz} \left[e^{\alpha(\text{Log} z + 2k\pi i)} \right] = e^{\alpha(\text{Log} z + 2k\pi i)} \frac{d}{dz} [\alpha(\text{Log} z + 2k\pi i)] = z^\alpha \times \alpha \times \frac{1}{z} = \alpha \frac{z^\alpha}{z}.$ $\square$

12.6 Extended Complex Plane and Riemann Sphere

riemann-sphere

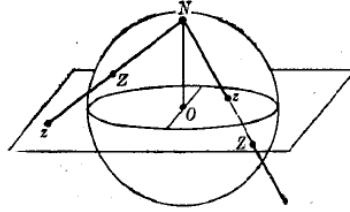
Compare to the set of real numbers $\mathbb{R}$, which needs “two” infinities $+\infty$ and $-\infty$, the set of complex numbers need just one infinity ∞ to make it meaningful.

Definition 12.25. The *extended complex plane* is defined as $\widehat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$. For any $z \in \mathbb{C}$ and any $y \in \widehat{\mathbb{C}} \setminus \{0\}$, we define

$$z + \infty = \infty + z = \infty; \quad y \cdot \infty = \infty \cdot y = \infty.$$

Note that $\infty + \infty$ and $0 \cdot \infty$ are *undefined*!

A geometrical model of the extended complex plane is called the *Riemann sphere*, which can be sketched as follows, where N is the north pole $(0, 0, 1)$ of Riemann sphere, $S^2 = \{(x_1, x_2, x_3) : x_1^2 + x_2^2 + x_3^2 = 1\}$.



It is possible to construct a one-to-one “continuous” mapping between $\widehat{\mathbb{C}}$ and S^2 through *stereographic projection*. Consider a line through the north pole $N(0, 0, 1)$ and a point on the complex plane $(x, y, 0)$: $(x_1, x_2, x_3) = (0, 0, 1) + t(-x, -y, 1)$. Since

$$x_1^2 + x_2^2 + x_3^2 = 1 \Rightarrow (-tx)^2 + (-ty)^2 + (1+t)^2 = 1 \Rightarrow \boxed{t = \frac{-2}{1+x^2+y^2}},$$

we obtain a projection from the extended complex plane to a unit sphere in $\mathbb{R}^3$:

$$z \mapsto \left(\frac{2x}{1+x^2+y^2}, \frac{2y}{1+x^2+y^2}, \frac{1+x^2+y^2-2}{1+x^2+y^2} \right) = \left(\frac{2\Re(z)}{|z|^2+1}, \frac{2\Im(z)}{|z|^2+1}, \frac{|z|^2-1}{|z|^2+1} \right), \quad \infty \mapsto (0, 0, 1).$$

And since $x_1 = -tx$, $x_2 = -ty$ and $x_3 = 1+t$, we can obtain the mapping from the sphere to the complex plane [Saff and Snider, 2003]:

$$(x_1, x_2, x_3) \mapsto (x, y) = \left(\frac{x_1}{-t}, \frac{x_2}{-t} \right) = \left(\frac{x_1}{1-x_3}, \frac{x_2}{1-x_3} \right), \quad x_3 \neq 1, \quad (0, 0, 1) \mapsto \infty.$$

What is the benefit of the stereographic projection? Stereographic projection “preserves” circles and angles. That is, the image of a circle on the sphere is a circle in the plane and the angle between two lines on the sphere is the same as the angle between their images in the plane. A projection that preserves angles is called a *conformal projection*. We shall discuss conformal mapping in Topic 8.

The “holomorphicity” in Section [12.3.2](#) can be generalised to ∞ as follows:

def:infy

Definition 12.26. $f(z)$ is holomorphic at $z = \infty$ if $f(\frac{1}{w})$ is holomorphic at $w = 0$.

On an extended complex plane, it is difficult to “see” that the infinity point is just a “normal” point. But from the sphere, the north pole and south pole are equally “normal”. Therefore, ∞ is treated normally by using the Definition [12.26](#).

Chapter 13

Further Topics in Geometry

<http://www.ams.org/samplings/feature-column/fcarc-circle-limit>

Indra's Pearl

In this topic, we will mainly investigate the *Möbius transform* (or *linear fractional transformation*) mentioned in Example 2.1.3 in Topic 2:

$$T : \hat{\mathbb{C}} \rightarrow \hat{\mathbb{C}}, \quad z \mapsto \frac{az + b}{cz + d} \quad (13.1)$$

e:mobius

where a, b, c , and d are complex numbers with $ad - bc \neq 0$. We are working on the extended complex plane $\hat{\mathbb{C}}$ mentioned in Topic 2.

Möbius transform is important because it is closely related to non-Euclidean geometry and because the set of all Möbius transforms form a group under function composition and has applications in number theory and complex dynamical system. Euclidean geometry used to be very important in the pass (before 19th century?) but it was replaced by Linear Algebra in universities because the later is more general. Some books such as Needham [1998, Chapter 3] (mentioned in Topic 1) and Carathéodory [1954, Chapter 3] bring back the Euclidean geometry and non-Euclidean geometry into complex analysis.

However, due to time constraint, we will be limited to a special non-Euclidean geometry called the “inversive geometry” it is impossible to go in depth into non-Euclidean geometry.

13.1 Translations

In Section [13.5](#), we will learn that a Möbius transform can be factored into reciprocal transform and linear transform. The linear transform can be decomposed into translation, magnification and rotation. Therefore, we shall begin our discussion with translation, followed by magnification and rotation.

Definition 13.1. A *translation* or a *shift* is defined as the mapping (or transform, transformation or map):

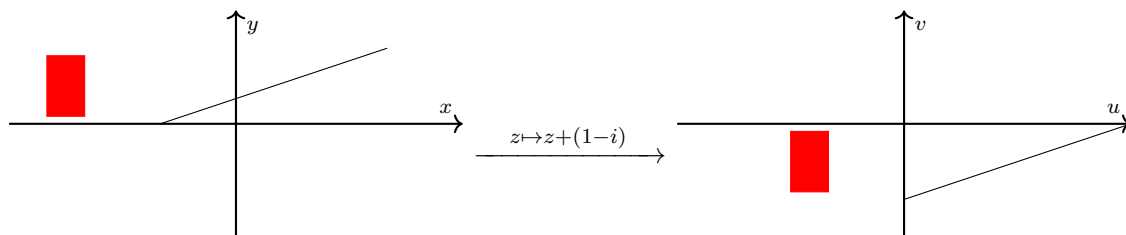
$$T : \hat{\mathbb{C}} \rightarrow \hat{\mathbb{C}}, \quad z \mapsto T(z) = z + b.$$

This is a special case of the Möbius transform ([13.1](#)) with $a = d = 1$ and $c = 0$.

As mentioned in Topic 1 Section 1.5, the basic geometrical objects in a plane are the “lines” and “circular objects” (e.g. circles, arcs). Just as what we did in Topic 2, we will also investigate the image of these geometrical objects.

The image of the point $z = (x, y)$ in the z -plane under T is the point $(x + b_1, y + b_2)$ where $b = (b_1, b_2)$. A line ℓ passing through (p_1, p_2) and (q_1, q_2) will be translated to a line ℓ passing through $(p_1 + b_1, p_2 + b_2)$ and $(q_1 + b_1, q_2 + b_2)$ and a rectangle with vertices z_0, z_1, z_2 , and z_3 will be translated to a rectangle with vertices $z_0 + b, z_1 + b, z_2 + b$ and $z_3 + b$. The shape

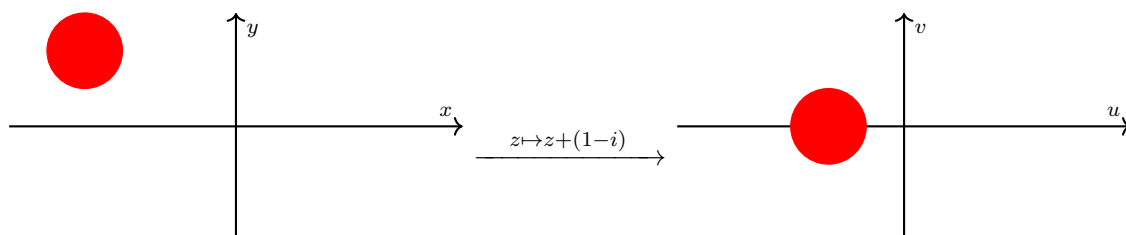
of the original rectangle is preserved under translation. We can observe this in the following illustration.



The image of a circle $C : |z - z_0| = R$ or $z = z_0 + Re^{i\theta}$, $0 \leq \theta < 2\pi$, under a translation is

$$T(C) = \{z_0 + b + Re^{i\theta} : 0 \leq \theta < 2\pi\}.$$

Hence the image of the circle C will be a circle with radius R and centred at $z_0 + b$ as illustrated below.



It is clear that a translation *does not change the shape or size* of a figure in the complex plane. A mapping with this property is sometimes called a *rigid motion*.

Theorem 13.1. *The images of a line and a circle under translation are still a line and a circle respectively.*

13.2 Magnification and Rotation

Definition 13.2. A *magnification* is defined as a mapping

$$T : \widehat{\mathbb{C}} \rightarrow \widehat{\mathbb{C}}, \quad z \mapsto T(z) = az$$

where a is a positive real number, i.e. $a > 0$. This is also a special case of the Möbius transform (I3.1) where $b = c = 0$ and $d = 1$.

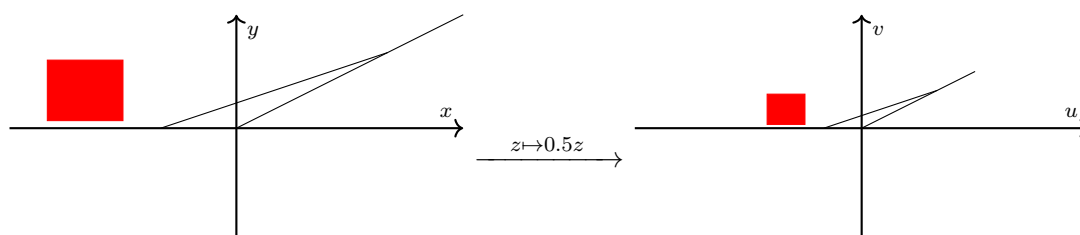
For a given $z \in \mathbb{C}$, if $a > 1$, then $|T(z)| = a|z| > |z|$ implies the point z is being magnified by a . If $a < 1$, then $|T(z)| = a|z| < |z|$ implies the point z is being shrunk by a . For a line $L_0 : \{tp : t \in \mathbb{R}\}$ passing through the origin and point p , the image is the same the original line:

$$T(L_0) = \{atp : t \in \mathbb{R}\}.$$

For a line $L : \{(1-t)p + tq : t \in \mathbb{R}\}$ passing through p and q but passing the origin, will be transformed to a line passing through ap and aq :

$$T(L) = \{(1-t)(ap) + t(aq) : t \in \mathbb{R}\}.$$

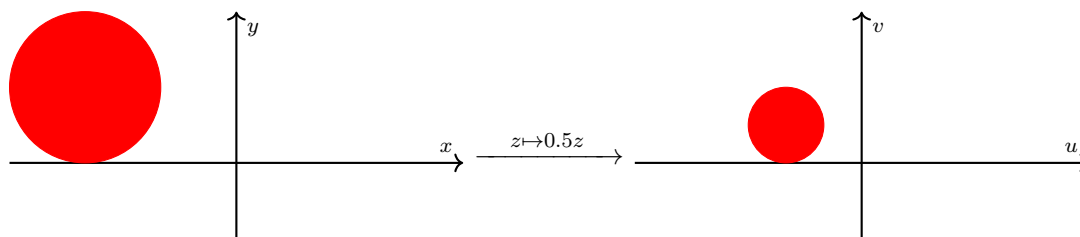
A rectangular will also be “magnified” by a factor of a as illustrated below.



The image of a circle $C = \{z_0 + Re^{i\theta} : 0 \leq \theta < 2\pi\}$ under a magnification is

$$T(C) = \{az_0 + aRe^{i\theta} : 0 \leq \theta < 2\pi\}.$$

The centre and radius are changed from z_0 and R to az_0 and aR respectively.



Theorem 13.2. *The images of a line and a circle under magnification are still a line and a circle respectively.*

We can conclude that a magnification will change the “size” of a geometrical, but it will not change the basic “shape”.

Definition 13.3. A *rotation* is defined as a mapping

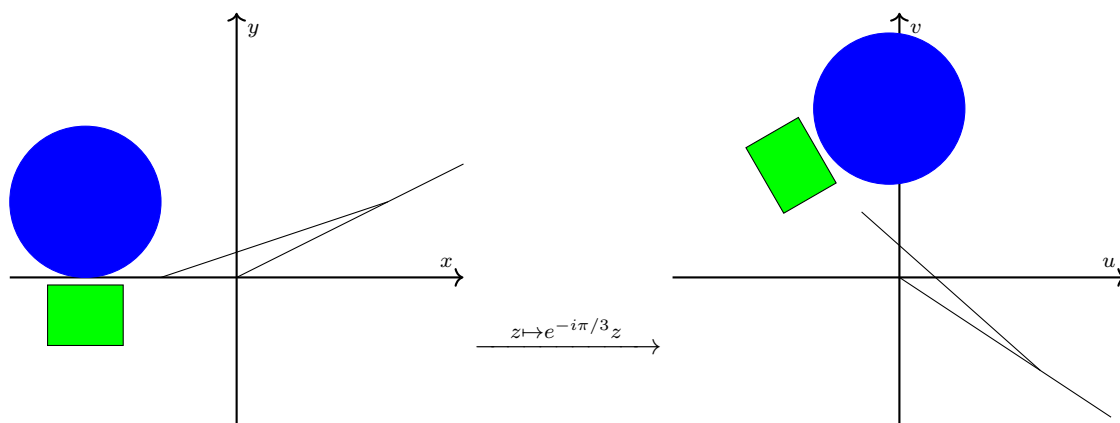
$$T : \widehat{\mathbb{C}} \rightarrow \widehat{\mathbb{C}}, \quad z \mapsto T(z) = e^{i\alpha}z.$$

This is a special case of the Möbius transformation (13.1) with $|a| = 1$, $b = c = 0$ and $d = 1$. le:mobius

Let $z = re^{i\theta}$ be a point in the z -plane. Under the mapping T , z will be “rotated” to $T(z) = re^{i(\theta+\alpha)}$. The linear mapping T can be visualised as a process of rotating a point z *counterclockwise* (respectively *clockwise*) through an angle of α (respectively $-\alpha$) radians about the origin to the point if $\alpha > 0$ (respectively $\alpha < 0$).

Rotations will not change the shape or size of a figure in the complex plane. Thus rotation is also a rigid motion and we have the following theorem and illustrations.

Theorem 13.3. *The images of a line and a circle under a rotation are still a line and a circle respectively.*



13.3 Linear Mapping

Definition 13.4. A *linear map*, *linear mapping*, *linear transform* or *linear transformation* is defined as

$$T : \widehat{\mathbb{C}} \rightarrow \widehat{\mathbb{C}}, \quad z \mapsto T(z) = az + b. \quad (13.2)$$

This is a special case of the Möbius transformation (13.1) where $a \neq 0$, $c = 0$ and $d = 1$. le:mobius

e:linear

It is easy to decompose the linear mapping as the composition of mappings mentioned in the previous sections, i.e. a linear mapping is a composition of rotation followed by a magnification and a translation:

$$T(z) = |a| \left(\frac{a}{|a|} z \right) + b.$$

If we investigate what happens to a point z_0 under the linear mapping T , we will find that z_0 is rotated at the origin by angle $\arg a$ followed by a magnification of factor $|a|$ and finally translated by “vector” b to z_3 :

$$z_0 \mapsto \frac{a}{|a|} z_0 = z_1 \mapsto |a| z_1 = z_2 \mapsto z_2 + b = z_3.$$

Translation, rotation, and magnification all preserve the basic shape of a geometrical object in the complex plane, in particular, lines and circles. Therefore a linear mapping also preserves the basic shape of a geometrical object in the complex plane and we obtain the following theorem.

Theorem 13.4. *The images of a line and a circle under a linear mapping are still a line and a circle respectively.*

eg:11.1.1

Example 13.5 (Images of Two Line Segments). Given the geometrical objects

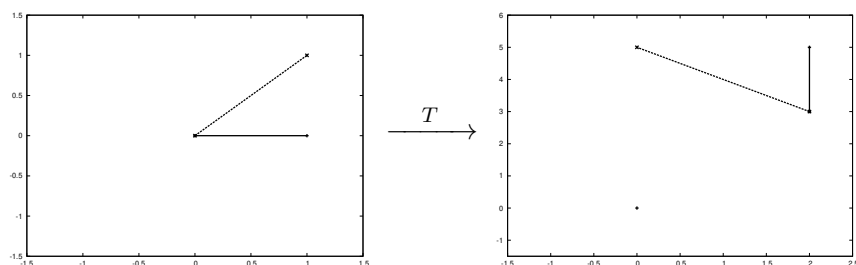
- L_1 : line segment joining 0 and 1;
- L_2 : line segment joining 0 and $1 + i$.

Sketch their images under the linear mapping $T(z) = (2i)z + (2 + 3i)$.

Solution: The images of the end-points of these line segments are:

$$T(L_1) = \text{line segment from } 0 + (2 + 3i) = 2 + 3i \text{ to } (2i)(1 + i0) + (2 + 3i) = 2 + 5i$$

$$T(L_2) = \text{line segment from } 0 + (2 + 3i) = 2 + 3i \text{ to } (2i)(1 + i) + (2 + 3i) = 0 + 5i.$$



Sketching the images of lines and circles under a linear mapping is not very difficult. However, writing down the mathematical equation characterising a line or a circle needs careful calculations as demonstrated by the following example.

mynotes:1.5.2b

Example 13.6. Write down the equations of (a) a line $(-5 + 3i)z + (-5 - 3i)\bar{z} + 28 = 0$ and (b) circle $|z - z_0| = R$ under the linear mapping

$$T : \widehat{\mathbb{C}} \rightarrow \widehat{\mathbb{C}}, \quad z \mapsto \frac{i}{2}z + 3 + 2i.$$

Solution. (a) First, we need to know that a point z on the line in z -plane satisfies $(-5 + 3i)z + (-5 - 3i)\bar{z} + 28 = 0$ and it is mapped to $T(z) = \frac{i}{2}z + 3 + 2i$ in the w -plane. We just need to obtain the inverse image of w , i.e. $z = T^{-1}(w)$:

$$w = \frac{i}{2}z + 3 + 2i \Rightarrow z = \frac{2}{i}(w - 3 - 2i) = -2i(w - 3 - 2i) = -2iw - 4 + 6i,$$

then we can obtain the equation satisfies by w in w -plane:

$$(-5 + 3i)[-2iw - 4 + 6i] + (-5 - 3i)\overline{-2iw - 4 + 6i} + 28 = 0$$

and after simplification, we obtain

$$(6 + 10i)w + 2 - 42i + (6 - 10i)\bar{w} + 2 + 42i + 28 = 0 \Rightarrow \boxed{(6 + 10i)w + (6 - 10i)\bar{w} + 32 = 0}.$$

(b) We can find the solution using the technique in part (a) but it is easier to just map z_0 to $T(z_0)$ and scale R with a factor $|\frac{i}{2}| = \frac{1}{2}$:

$$|w - (\frac{i}{2}z_0 + 3 + 2i)| = \frac{R}{2}.$$

□

13.4 $w = \frac{1}{z}$ Transformation

Definition 13.5. A *reciprocal function*, *reciprocation* or *reciprocal transformation* is defined as

$$T : \widehat{\mathbb{C}} \rightarrow \widehat{\mathbb{C}}, \quad z \mapsto T(z) = \frac{1}{z}.$$

Note that $T(0) = \infty$ and $T(\infty) = 0$.

We would like to know what is the images of lines and circles under the reciprocal function. First, we will investigate the image of a line.

thm:recip:line

Theorem 13.7. *The image of a line under a reciprocal mapping is either a line or a circle.*

Proof: We learned in Fundamentals of Linear Algebra that a line in a plane can be written as

$$\ell : mx + ny + k = 0, \quad m, n, k \in \mathbb{R}.$$

On a complex plane, $x = \frac{z + \bar{z}}{2}$ and $y = \frac{z - \bar{z}}{2i}$ and so the line equation can be written in *self-conjugate form* (Topic 1 E.g. 1.5.2):

$$\frac{mz + m\bar{z}}{2} + \frac{nz - n\bar{z}}{2i} + k = 0 \Rightarrow \frac{m - ni}{2}z + \frac{m + ni}{2}\bar{z} + k = 0 \Rightarrow Az + \bar{A}\bar{z} + k = 0$$

where $A = \frac{m - ni}{2}$.

For $z \in \ell$ is $w = \frac{1}{z}$ and so

$$z = \frac{1}{w} \Rightarrow A\frac{1}{w} + \frac{\bar{A}}{\bar{w}} + k \Rightarrow A\bar{w} + \bar{A}w + kw\bar{w} = 0 \Rightarrow w\bar{w} + \frac{A}{k}\bar{w} + \frac{\bar{A}}{k}w + \frac{A\bar{A}}{k^2} = \frac{A\bar{A}}{k^2}$$

which can be factorised into a circle centre at $-\frac{A}{k}$ of a radius $\frac{|A|}{k}$ when $k \neq 0$:

$$(w + \frac{A}{k})(\bar{w} + \frac{\bar{A}}{k}) = \frac{|A|^2}{k^2} \Rightarrow \left|w + \frac{A}{k}\right|^2 = \left|\frac{A}{k}\right|^2 \Rightarrow \left|w - (-\frac{A}{k})\right| = \left|\frac{A}{k}\right|.$$

and a line $A\bar{w} + \bar{A}w = 0$ when $k = 0$.

eg:2.2.4

Example 13.8 (April 2010 Exam Q2(a)). Find the images of the line (a) $x = a$; (b) $y = b$, where $ab \neq 0$, under the mapping

$$w = f(z) = \frac{1}{z}. \quad (8 \text{ marks})$$

Solution.

$$f(z) = \frac{1}{x+iy} = \frac{x}{x^2+y^2} - i \frac{y}{x^2+y^2}$$

We should expect the images of these circles are vertical and horizontal lines:

(a) When $x = a$,

$$u = \frac{a}{a^2+y^2} \text{ and } v = \frac{-y}{a^2+y^2} \Rightarrow \frac{v}{u} = \frac{-y}{a} \Rightarrow y = \frac{-av}{u}.$$

Hence,

$$u = \frac{a}{a^2 + \left(\frac{-av}{u}\right)^2} \Rightarrow u^2 - \frac{1}{a}u + v^2 = 0 \Rightarrow \left(u - \frac{1}{2a}\right)^2 + v^2 = \left(\frac{1}{2a}\right)^2.$$

This equation represents a circle centred at $\left(\frac{1}{2a}, 0\right)$ with radius $\frac{1}{2|a|}$. Hence all the vertical lines will be mapped to circles centred on the u -axis and pass through the origin.

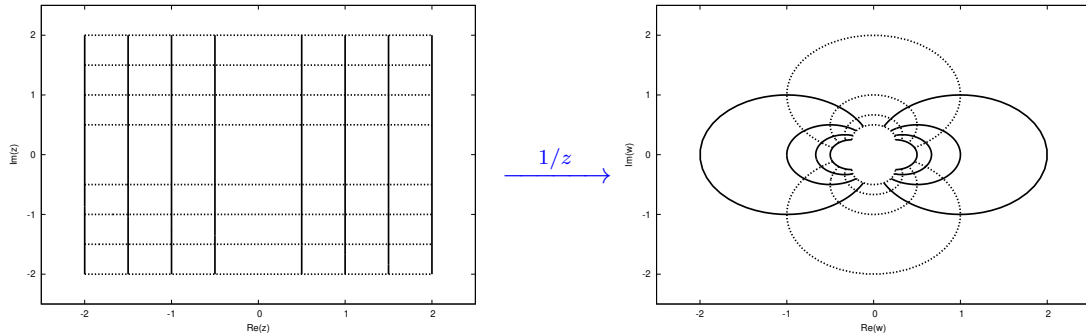
(b) Similarly, one can verify that all the horizontal lines $y = b$ in the z -plane will be mapped to circles centred on the v -axis and pass through the origin.

$$u = \frac{x}{x^2+b^2} \text{ and } v = \frac{-b}{x^2+b^2} \Rightarrow \frac{x}{-b} = \frac{u}{v} \Rightarrow x = \frac{-bu}{v}.$$

Hence,

$$v = \frac{-b}{\left(\frac{-bu}{v}\right)^2 + b^2} \Rightarrow u^2 + \frac{1}{b}v + v^2 = 0 \Rightarrow u^2 + \left(v + \frac{1}{2b}\right)^2 = \left(\frac{1}{2b}\right)^2.$$

This equation represents a circle centred at $\left(0, -\frac{1}{2b}\right)$ with radius $\frac{1}{2|b|}$. Hence all the horizontal lines will be mapped to circles centred on the v -axis and pass through the origin.



□

Note that the line $x = a$ is an unbounded line in z -plane, the ideal point ∞ is not in the complex plane. In order to obtain the entire circle we extend the complex plane by adjoining the complex plane with the ideal point ∞ . On the extended complex plane $\hat{\mathbb{C}}$ the reciprocal mapping will map the line $x = a$ onto the entire circle! In a similar manner, the image of the line $y = b$ under the reciprocal mapping on $\hat{\mathbb{C}}$ is the entire circle.

The image of horizontal and vertical lines are circles, what happens to these circles if we apply reciprocal mapping again? The following example answer this question.

eg:2.2.5

Example 13.9. Find the images of the following circles under the mapping $f(z) = \frac{1}{z}$.

$$(a) \left(x - \frac{1}{2a}\right)^2 + y^2 = \left(\frac{1}{2a}\right)^2, x \neq 0; \quad (b) x^2 + \left(y - \frac{1}{2b}\right)^2 = \left(\frac{1}{2b}\right)^2, y \neq 0.$$

Solution. We should expect the images of these circles are vertical and horizontal lines.

$$(a) \left(x - \frac{1}{2a}\right)^2 + y^2 = \left(\frac{1}{2a}\right)^2 \Rightarrow x^2 + y^2 = \frac{x}{a} \Rightarrow \frac{x}{x^2 + y^2} = u = a$$

$$(b) x^2 + \left(y - \frac{1}{2b}\right)^2 = \left(\frac{1}{2b}\right)^2 \Rightarrow x^2 + y^2 = \frac{y}{b} \Rightarrow \frac{y}{x^2 + y^2} = -v = b \Rightarrow v = -b \quad \square$$

thm:recip:circle

Theorem 13.10. *The image of a circle under a reciprocal mapping is either a line or a circle.*

Proof: A circle centred at P and with a radius R has a general form

$$|z - P| = R \Rightarrow z\bar{z} - P\bar{z} - \bar{P}z + P\bar{P} = R^2$$

Since $z = \frac{1}{w}$, we have

$$\frac{1}{w} \frac{1}{\bar{w}} - \frac{P}{\bar{w}} - \frac{\bar{P}}{w} + P\bar{P} = R^2 \Rightarrow P\bar{P}w\bar{w} - Pw - \bar{P}w + 1 = R^2$$

When $P = 0$, we obtain a line (in self-conjugate form):

$$Pw + \bar{P}w + R^2 - 1 = 0.$$

When $P \neq 0$, we obtain a circle centred at $\frac{1}{\bar{P}}$ with a radius $R/|P|$:

$$(Pw - 1)(\bar{P}w - 1) = R^2 \Rightarrow |Pw - 1| = R \Rightarrow \left|w - \frac{1}{\bar{P}}\right| = \frac{R}{|P|}.$$

The reciprocal mapping can be decomposed as a “circle inversion” followed by a complex conjugation:

$$T(z) = \frac{\bar{z}}{z\bar{z}} = \frac{\bar{z}}{|z|^2}.$$

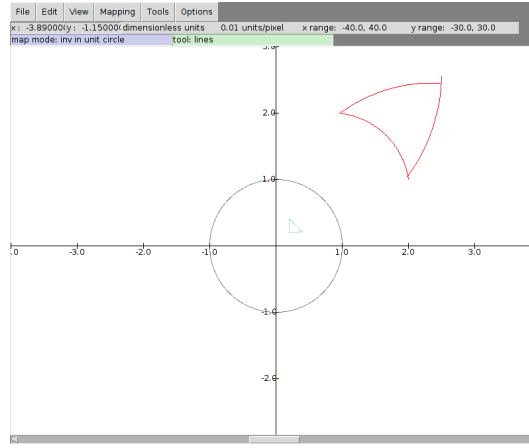
Definition 13.6. A (circle) inversion or a reflection with respect to a unit circle is defined as

$$S : \widehat{\mathbb{C}} \rightarrow \widehat{\mathbb{C}}, \quad z \mapsto S(z) = \frac{z}{|z|^2}.$$

The mapping of inversion in the unit circle will send a point lying outside the unit circle along the ray emanating from zero and containing (r, θ) to the point $(1/r, \theta)$ which lies inside the unit circle according to

$$S(re^{i\theta}) = \frac{re^{i\theta}}{r^2} = \frac{1}{r}e^{i\theta}.$$

Conversely, if $r < 1$, the point (r, θ) will be lying inside the unit circle and its image under the mapping of inversion of the unit circle will be lying outside the unit circle. There is a software called Zmapper which allows us to play around with circle inversion.



Circle inversion is the basic idea in *inversive geometry* (http://en.wikipedia.org/wiki/Inversive_geometry, Schwerdtfeger [1979] and Carathéodory [1954]). Since the complex conjugation just mirror a geometrical object with respect to the real axis, by Theorem 13.10, we obtain the following theorem.

Theorem 13.11. *If a circle C passes through the origin, its image under circle inversion $z \mapsto \frac{z}{|z|^2}$ is a line. If the circle C does not pass through the origin, its image under circle inversion is still a circle.*

13.5 Möbius Transformation on $\hat{\mathbb{C}}$

sec:mobius

The Möbius transformation (13.1) can either be reduced to a linear mapping or be decomposed into a composition of linear mappings and a reciprocal mapping:

- When $c = 0$, $T(z) = \frac{a}{d}z + \frac{b}{d}$, $z \in \mathbb{C}$; $T(\infty) = \infty$.
- When $c \neq 0$, $T(z) = \frac{a}{c} + \frac{bc - ad}{c^2} \left[\frac{1}{z + \frac{d}{c}} \right]$, $z \in \mathbb{C} \setminus \{-\frac{d}{c}\}$; $T(-\frac{d}{c}) = \infty$; $T(\infty) = \frac{a}{c}$.

Geometrically, a Möbius transformation (13.1) can be obtained by first performing stereographic projection (Topic 1) from the plane $\mathbb{C}$ to the Riemann Sphere S^2 , rotating and moving S^2 to a new location and orientation in space, and then performing stereographic projection (from the new position of the sphere) to the plane. These transformations preserve angles, map every straight line to a line or circle, and map every circle to a line or circle.

A short video by Douglas Arnold and Jonathan Rogness called “Möbius Transformations Revealed” (<http://www.ima.umn.edu/~arnold/moebius/>), which is found on YouTube (<https://www.youtube.com/watch?v=0z1fIsUNh04>) illustrates the above paragraph nicely. More interesting geometric theory of Möbius transform can be found in Yaglom [1968, Chapter III], Rademacher [1983, Chapter 9] and Ford [1929, Chapter 1].

The following theorems characterise the basic properties of Möbius transformations.

Theorem 13.12. *A Möbius transformation is one-to-one (bijective) holomorphic function from $\hat{\mathbb{C}}$ to $\hat{\mathbb{C}}$.*

Theorem 13.13. *The composition of any two Möbius transformations is also a Möbius transformation. The inverse of the Möbius transformation (13.1) is the Möbius transformation*

$$T^{-1}(z) = \frac{dz - b}{-cz + a}. \quad (13.3)$$

e:mobius:inv

According to Theorem 13.13, the set of all Möbius transformations forms a group under composition, called the *Möbius group*, and denoted by $\text{Aut}(\widehat{\mathbb{C}})$. This group can be given the structure of a complex manifold in such a way that composition and inversion are holomorphic maps. The Möbius group is then equivalent to the projective linear group $PGL(\mathbb{C}^2) = GL(\mathbb{C}^2)/Z(\mathbb{C}^2)$, which is a special *complex Lie group*. Here, $GL(V)$ is the general linear group of V and $Z(V)$ is the subgroup of all nonzero scalar transformations of V , where $V = \mathbb{C}^2$.

Theorem 13.14 (Fixed Points). *A Möbius transformation that is not the identity map has one or two fixed points.*

Proof: ∞ is a fixed point iff $c = 0$.

The finite fixed points are the solutions in z of the equation $cz^2 + (d - a)z - b = 0$. If $c \neq 0$, this equation has either two distinct roots or one repeated root. If $c = 0$ then there is one root for $d \neq a$ and zero roots for $d = a$. In both cases, the transformation has one or two fixed points.

We need “three image points” to determine a Möbius transformation as stated below.

Theorem 13.15 (Three-fold Transitivity). *If z_1, z_2, z_3 are three distinct points of $\widehat{\mathbb{C}}$ and w_1, w_2, w_3 are three distinct points of $\widehat{\mathbb{C}}$, then there is a unique Möbius transformation T such that $T(z_i) = w_i$ for $i = 1, 2, 3$. If $w = T(z)$, then T is determined by the cross-ratio:*

$$\frac{(w - w_1)(w_2 - w_3)}{(w - w_3)(w_2 - w_1)} = [w, w_1, w_2, w_3] = [z, z_1, z_2, z_3] = \frac{(z - z_1)(z_2 - z_3)}{(z - z_3)(z_2 - z_1)}.$$

Möbius transformations map generalised circles onto generalised circles as stated below.

Theorem 13.16 (Circle-Preserving Property). *If C is a circle in the z -plane and if F is a Möbius transformation T given by (13.1), then the image of C under T is either a circle or a line in the extended w -plane. The image is a line if and only if $c \neq 0$ and the pole $z = -d/c$ is on the circle.*

Proof:

When $c = 0$, T is a linear transformation, hence T maps circles onto circle.

When $c \neq 0$, T is a composition of mappings

$$T(z) = f \circ g \circ h(z)$$

where

$$f(z) = \frac{bc - ad}{c^2}z + \frac{a}{c^2}, \quad g(z) = \frac{1}{z}, \quad h(z) = z + \frac{d}{c}.$$

Observe that h is a translation, g is an inversion and f is a linear transformation. We have shown that they map “generalised circles” (a circle or a line) to “generalised circles”, therefore, the composition of f , g and h will map “generalised circles” to “generalised circles”.

An alternative proof is given in Ford [1929, 8].

Mumford et al. [2002, 91] gave a program to calculate the image of a circle of centre P and radius r mapping to a circle at centre Q and s as follows.

```
1 (define (mobiuscir a b c d P realr)
2   (define Q nil)
3   (define (T z) (/ (+ (* a z) b) (+ (* c z) d))))
```

```

4 (cond
5 ((not (zero? (+ (/ d c) P)))
6 (set! Q (T (- P (/ (* realr realr) (conj (+ (/ d c) P))))))
7 ((positive? realr)

```

However, in the following examples, we will find that for less complicated Möbius transform, it is more convenient to calculate by hand than using formula.

eg:11.5.1

Example 13.17. Find the image of the unit circle $|z| = 1$ under the Möbius transformation

$$T(z) = \frac{z+2}{z-1}$$

What is the image of the disk $|z| < 1$?

Solution: Since $T(1) = \infty$ and so the image of $|z| = 1$ should be a “circle” passing through ∞ , i.e. a line. Thus it suffices to find two image points in order to determine the line:

$$T(-1) = -\frac{1}{2}, \quad T(i) = \frac{-1-3i}{2}.$$

Hence the image of the unit circle under T is the line $u = -\frac{1}{2}$.

Notice that Möbius transformation is a rational function, and so it is continuous on its domain. This implies that the image of the interior of the unit circle is the half plane either on the left or right of $u = -\frac{1}{2}$. The image $T(0) = -2$ is on the left side of the line $u = -\frac{1}{2}$.

So the image of $|z| < 1$ is the half plane described by $w = u + iv$, $u < -\frac{1}{2}$.

Example 13.18 (April 2010 Exam, Q5(a)). Find the image of the interior of the circle $C : |z-2| = 2$ under the Möbius transformation $w = \frac{z}{2z-8}$. (11 marks)

[**Remark:** This question is taken from Saff and Snider [2003, 390]]

Solution. The point $z = 4$ is on C and $T(4) = \frac{4}{8-8} = \infty$. Hence, $T(C)$ is a line.

Picking two points from C : $z_1 = 0$ and $z_2 = 2 + 2i$ and we have

$$T(0) = 0, \quad T(2+2i) = \frac{1+i}{-2+2i} = \frac{-4i}{4+4} = \frac{-i}{2}.$$

This implies the image $T(C)$ is the line $\Re(w) = 0$.

The point 2 is in $|z-2| < 2$ and $T(2) = \frac{2}{-4-8} = -\frac{1}{6}$. Hence the image of $|z-2| < 2$ is $\Re(z) < 0$. □

Example 13.19 (Variation of April 2009 Exam, Q6(b)). Find the images of the circles $C_1 : |z| = 1$ and $C_2 : |z| = 2$ under the transformation $w = \frac{z+2}{z-1}$. What is the image of the disk $|z| < 2$? (15 marks)

Solution. (i) $T(1) = \infty$ and therefore $T(C_1)$ is a line. Take two points i and $-i$ from C_1 and calculate

$$T(i) = \frac{i+2}{i-1} = \frac{(2+i)(-1-i)}{1+1} = \frac{-1-3i}{2}, \quad T(-i) = \frac{-i+2}{-i-1} = \frac{-1+3i}{2}$$

$T(C_1)$ is a line $x = -\frac{1}{2}$.

$$(ii) \quad w = \frac{z+2}{z-1} \Rightarrow w-1 = \frac{3}{z-1} \Rightarrow \frac{3}{w-1} + 1 = z \Rightarrow \left| \frac{3}{w-1} + 1 \right| = 2$$

which expands to

$$\begin{aligned} \frac{3}{w-1} \frac{3}{\bar{w}-1} + \frac{3}{w-1} + \frac{3}{\bar{w}-1} + 1 &= 4 \Rightarrow \frac{3}{(w-1)(\bar{w}-1)} + \frac{1}{w-1} + \frac{1}{\bar{w}-1} = 1 \\ \Rightarrow 3 + \bar{w} - 1 + w - 1 &= (w-1)(\bar{w}-1) = w\bar{w} - w - \bar{w} + 1 \\ \Rightarrow 0 &= w\bar{w} - 2w - 2\bar{w} \Rightarrow (w-2)(\bar{w}-2) - 4 = 0 \Rightarrow |w-2| = 2. \end{aligned}$$

(ii) **Alternatively**, T does not map the points on C_2 to ∞ . Hence the image of C is a “normal” circle C' . Notice that if z is on the circle C , then $\bar{z}$ is also on the circle C . Moreover

$$T(\bar{z}) = \overline{T(z)}.$$

Thus the image circle C' is symmetric about the u -axis.

Furthermore, $T(2) = 4$ and $T(-2) = 0$ are on C' . These two points are endpoints of a diameter. This implies that C' is the circle

$$|w-2| = 2.$$

(iii) To find the image of $|z| < 2$, we check the value $T(0) = -2$, which is outside the circle C' . Thus the image of $|z| < 2$ is the region

$$|w-2| > 2.$$

□

In the above examples, we are given a mapping and asked to sketch the images. However, if we are given two geometrical objects and were asked to construct the “Möbius transform”, then it is more difficult. In fact, we have the following theorem.

Theorem 13.20. *There exist infinitely many Möbius transforms which transform a given circle into a second given circle.*

To construct a particular Möbius transformation T , which maps a given circle in the extended complex plane to a given circle, it suffices to choose T so that it maps three chosen points on one circle to three chosen points on the other. Alternatively, we can pick the centre of the circle and two points on the circle to the centre and two points on the other circle, then by three-fold transitivity theorem, we can determine a Möbius transform.

If a linear transformation which maps a particular circle to a particular circle in the z -plane is required, then a linear transformation suffices.

Example 13.21. Find a linear transformation that maps the circle $|z-i| = 1$ to circle $|w| = 2$.

Solution: Observe that the Möbius transformation that we are looking for will translate the centre from i to 0 and magnify the radius from 1 to 2. Hence the linear transformation

$$T(z) = 2(z-i)$$

will do the job.

Example 13.22. ([Saff and Snider, 2003, 387]) Find a Möbius transformation that maps the circle $C : |z-1| = 1$ to the circle $C' : |w - \frac{3i}{2}| = 2$.

Solution. Pick the centre of C and two points on C : $z = 1$, $z = 0$ and $z = 2$. The images of these three points lie on the line $\Im m(z) = \frac{3}{2}$ passing through the centre of C' :

$$T(1) = -\frac{3i}{2}, \quad T(0) = -2 - \frac{3i}{2}, \quad T(2) = 2 - \frac{3i}{2},$$

Since, there is no rotation, we have real number a such that $T(z) = az + b$, i.e.

$$T(0) = b = -2 - \frac{3i}{2} \Rightarrow T(1) = a - 2 - \frac{3i}{2} = -\frac{3i}{2} \Rightarrow a = 2.$$

Hence, $T(z) = 2z - 2 - \frac{3i}{2}$. □

However, when we were to map circle to “line”, then we need a Möbius transformation with $c \neq 0$.

Example 13.23. ([Saff and Snider, 2003, 391]) Find a Möbius transform of the unit disk $|z| < 1$ onto the right half-plane $\Re(w) > 0$.

Solution. The Möbius transform must map $C : |z| = 1$ to a line $C' : \Re(w) = 0$. Therefore, $c \neq 0$ and we can choose $c = 1$ and pick three points on C such that

$$T(-1) = \frac{-a+b}{-1+d} = \infty, \quad T(-i) = \frac{-ai+b}{-i+d} = 0, \quad T(1) = \frac{a+b}{1+d} = -i.$$

Therefore $d = 1$ and

$$\begin{cases} -ai + b = 0, \\ a + b = -2i \end{cases} \Rightarrow b = ai \Rightarrow a(1+i) = -2i \Rightarrow \begin{cases} a = \frac{-2i}{1+i} = -1-i \\ b = 1-i \end{cases}$$

Hence,

$$T(z) = \frac{(-1-i)z + 1-i}{z+1} = -(1+i) \frac{z+i}{z+1}$$

Note that $w = T(0) = 1-i$ satisfies $\Re(w) > 0$. Therefore, T is one possible solution. □

The following theorem states the general form of Möbius transform which maps a unit disk $|z| \leq 1$ to a unit disk $|w| \leq 1$.

Theorem 13.24. Let the complex number a satisfy $|a| < 1$, $\lambda \in \mathbb{R}$ and the Möbius transformation be

$$T_a(z) = e^{i\lambda} \frac{z-a}{1-\bar{a}z}. \tag{13.4}$$

e:mobius:cir

Then

1. T_a is holomorphic on $\Delta_1(0)$ (recall Topic 1).
2. T_a maps $\Delta_1(0)$ one-to-one onto $\Delta_1(0)$.
3. T_a maps the unit circle onto the unit circle.

Chapter 14

Complex Integration

ch:cplx:int

After the discussion of differentiation in Section [12](#), it is now time to define “integration” for complex functions. But how can we do that?

There are two approaches to the definition of “integration”, one approach is to restrict ourselves to just piecewise smooth curves (called contours) and use the line integrals from Calculus as foundation. Churchill and Brown [2008], Ahlfors [1979], Lang [1999], Freitag and Busam [2005], Marsden [1973], Bak and Newman [2010], Ullrich [2008], Hille [1959], Antimirov et al. [1998] and Titchmarsh [1939] took this approach. This is probably a more popular approach because nearly all curves in complex analysis, scientific and engineering problems are piecewise smooth Ahlfors [1979].

The other approach is to define “integration” as a limiting process of a summation. There are fewer books which took this approach, such as Conway [1978], Carathéodory [1954], Nevanlinna and Paatero [1964], Markushevich [1965], Henrici [1974] and Saff and Snider [2003], which is considered too theoretical.

In Section [14.2](#), we will mention the second approach and just state the properties of contour integrals and demonstrate the calculations of integrals using the knowledge of line integral from Calculus III or Advanced Calculus. We then relate the “integrals” with the anti-derivative in Section [14.3](#).

After that, we will investigate the most important theorems in complex analysis — Cauchy Theorem, its variations and Cauchy Integral Formula in Sections [14.4](#) and [14.5](#). Finally, we will derive some consequences of holomorphicity using the Cauchy Theorem in Sections [14.6](#) and [14.7](#).

14.1 Arcs, Curves and Contours

sec:contour

In the single-variable Calculus, there is only one “path” to integrate from point a to b . However, on a complex plane, when we want to “integrate” from a point z_0 to z_1 , there are many “ways” to travel from z_0 to z_1 , called arcs or paths.

Definition 14.1. An *arc*, or a *path*, or a (*parameterized*) *curve* C is a continuous mapping of the closed interval $[a, b] \subset \mathbb{R}$ to the complex plane $\mathbb{C}$, i.e.

$$C : [a, b] \rightarrow \mathbb{C}, \quad t \mapsto x(t) + iy(t) =: z(t) \quad (14.1)$$

e:param

where $x(t)$ and $y(t)$ are continuous real functions.

Remark 14.1. A *curve* is an equivalent classes of arcs (or paths). In this section, we will assume that a curve is just a path and for the next section, assume that an integral is “parameterization” independent. Read Conway [1978, Chapter 4] for detail explanation.

Definition 14.2. For a curve C , the values $z(a)$ and $z(b)$ are called the *initial* and *terminal* (or *end*) points of C respectively. Any curve with an initial point z_0 and a terminal point z_1 is called an *arc/a path/a curve from z_0 to z_1* .

Definition 14.3. A curve C is said to be *smooth* (or *regular*) if $x'(t)$ and $y'(t)$ are continuous on $[a, b]$ and not simultaneously zero on the open interval (a, b) . In other words, $|z'(t)| \neq 0$ for $t \in (a, b)$.

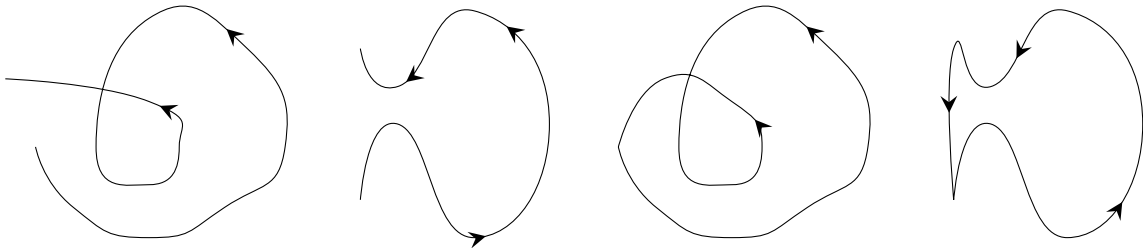
Definition 14.4. A curve C is said to be *piecewise smooth* if it consists of a finite number of smooth curves $C_1, C_2, \dots, C_n$ joined end to end, that is, the terminal point of curve C_k coinciding with the initial point of C_{k+1} for $k = 1, \dots, n - 1$.

Definition 14.5. A piecewise smooth curve on $\mathbb{C}$ is called a *contour*.

Definition 14.6. A curve C is called a *simple curve* if C does not cross itself except possibly at $t = a$ and $t = b$. A curve C is called a *closed curve* if $z(a) = z(b)$. A curve C is called a *simple closed curve* or a *Jordan curve* if C is simple and closed.

Example 14.2. Given an example for (a) a curve which is not simple, (b) a simple curve, (c) a closed curve which is not simple, (d) a Jordan curve.

Solution:



A Jordan curve is important because of the following famous theorem from topology.

Theorem 14.3 (Jordan Curve Theorem). *Any Jordan curve separates the plane into two domains, each having the curve as its boundary. One domain, called the interior, is bounded, the other domain, called exterior, is unbounded.*

Out of many Jordan curves, the circle plays an important role in complex integration. A circle centred at z_0 with radius r can be described by

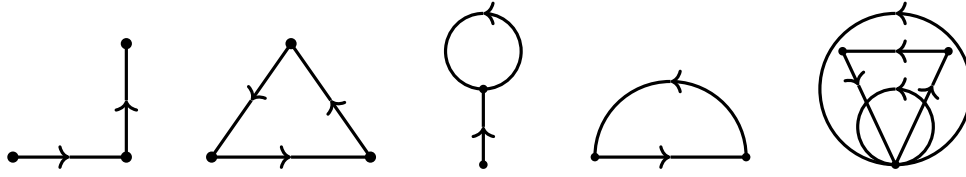
$$z(t) = z_0 + re^{it}, \quad 0 \leq t \leq 2\pi \quad (14.2) \quad \text{e:5.2.1}$$

Definition 14.7. A circle is said to be *positively oriented* if its parametric equation is given as (14.2). A *negatively oriented circle* has parametric equation:

$$z(t) = z_0 + re^{-it}, \quad 0 \leq t \leq 2\pi. \quad (14.3) \quad \text{e:5.2.2}$$

Sometimes a circle imprecisely characterised as $|z - z_0| = r$. In this case, we will **always assume positive orientation** unless otherwise specified.

Example 14.4. ([Saff and Snider, 2003, 169]) Given the following contours, identify which are simple curves, closed curves and simple closed curves.



14.2 Integrals and Contour Integrals

Let C be an arc with parametric representation $z = z(t)$, $a \leq t \leq b$. Let f be a complex-valued function defined on C .

Definition 14.8. A *partition* of the arc is a finite sequence of numbers (t_i) such that

$$P : a = t_0 < t_1 < t_2 < \cdots < t_{n-1} < t_n = b.$$

Definition 14.9. ([Henrici, 1974, 102]) A *system of pivotal points* T corresponding to the partition P is a set of numbers τ_k that satisfies $t_{k-1} \leq \tau_k \leq t_k$, $k = 1, \dots, n$. The *Riemann sum* is defined as

$$S(P, T) = \sum_{n=1}^{\infty} f(z(\tau_k)) [z(t_k) - z(t_{k-1})]. \quad (14.4)$$

Definition 14.10. If there is a complex number A such that, given any $\epsilon > 0$, there is a number $\delta = \delta(\epsilon)$ such that for all partitions P and all corresponding systems of pivotal points T , $|S(P, T) - A| < \epsilon$ then A is called the *integral of f along C* and we denote it as $A = \int_C f(z) dz$.

Definition 14.11. Consider a curve $\gamma : [a, b] \rightarrow \mathbb{C}$. For any partition (t_i) , if $\sum_{i=0}^{n-1} |\gamma(t_{i+1}) - \gamma(t_i)| < \infty$ and a limit exists as $n \rightarrow \infty$. Then the curve γ is said to be *rectifiable*.

Theorem 14.5. A contour C is rectifiable.

Theorem 14.6. If C is rectifiable and f is continuous over C , the integral of f along C exists. It is called a complex (line or contour) integral.

Hence, we know that a continuous function is “integrable” over a contour and it is not difficult to implement the midpoint method using (14.4):

```
1  ;; C : complex function for the curve
2  ;; a, b : real numbers, starting and ending points of C
3  ;; f : complex function
```

```

4 (define (contourint C a b f N)
5   (define h (/ (- b a) N))
6   (define (pts t) (* (f (C (+ t (* 0.5 h)))) (- (C (+ t h)) (C t))))
7   (apply + (map (lambda (i) (pts (+ a (* i h)))) (iota N))))
8
9 (define pi (* 4 (atan 1.0)))
10 (define (C t) (exp (* 0+1i t)))
11 (define (f z) (/ 1 z))
12
13 (display (contourint C 0.0 (* 2 pi) f 1000))
14 (newline)

```

Theoretically, if C is a smooth curve parametrized by (14.1), the complex integral $\int_C f(z)dz$ can be computed using the following formula

$$\int_C f(z)dz = \int_a^b f(z(t))z'(t)dt \quad (14.5) \quad \text{e:int:smooth}$$

and it is related to the real line integral from Calculus II:

$$\int_C f(z)dz = \int_C (u + iv)(dx + idy) = \int_C udx - vdy + i \int_C vdx + udy.$$

eg:5.3.2

Example 14.7. Evaluate the contour integral $\int_C \bar{z}dz$ where C is the curve given by

$$C : z(t) = 3t + it^2, \quad 0 \leq t \leq 2.$$

Solution:

$$\begin{aligned} \int_C \bar{z}dz &= \int_0^2 (3t - it^2)(3 + 2it)dt = \int_0^2 (9t + 2t^3)dt + i \int_0^2 (6t^2 - 3t^2)dt \\ &= \left[\frac{9}{2}t^2 + \frac{1}{2}t^4 \right]_0^2 + i \left[t^3 \right]_0^2 = 26 + 8i \end{aligned}$$

eg:wwy:5.5.2

Example 14.8. Evaluate $\int_C \bar{z}^2 dz$ from 0 to $1 + 2i$ along the curve $C : y = 2x$.

Solution. On C , $z = x + iy = x + i(2x)$, $0 \leq x \leq 1$.

$$\begin{aligned} \int_C \bar{z}^2 dz &= \int_C (x - iy)^2 (dx + idy) = \int_0^1 (x - 2ix)^2 (1 + 2i)dx \\ &= (1 - 2i)^2 (1 + 2i) \int_0^1 x^2 dx = 5(1 - 2i) \times \frac{1}{3} = \frac{5}{3} - \frac{10}{3}i. \end{aligned}$$

□

Since a smooth curve can be computed using (14.5), it is not difficult to obtain the following properties of complex contour integral.

thm:integral:prop

Theorem 14.9. Suppose f and g be continuous in a region D , and C is a smooth curve lying entirely inside D . Then

1. $\int_C kf(z)dz = k \int_C f(z)dz$
2. $\int_C (\alpha f(z) + \beta g(z))dz = \alpha \int_C f(z)dz + \beta \int_C g(z)dz$

3. $\int_{-C} f(z)dz = - \int_C f(z)dz$, where $-C$ denotes the curve having the opposite orientation of C .
4. $\int_C f(z)dz = \int_{C_1} f(z)dz + \int_{C_2} f(z)dz$, where C consists of the smooth curves C_1 and C_2 joined end to end.
5. (ML-Inequality) If f is continuous on a smooth curve C and $|f(z)| \leq M$ for all $z \in C$, then $\left| \int_C f(z)dz \right| \leq ML$, where L is the length of C .

Properties 1 to 4 are commonly used when we try to integrate a linear combination of functions over piecewise smooth curves.

Example 14.10. Given the points $O(0,0)$, $A(1,0)$ and $B(1,1)$.

- (a) Evaluate the integral (i) $\int_{\underline{OAB}} z dz$ and (ii) $\int_{\underline{OB}} z dz$, where $\underline{OAB}$ is a piecewise linear curve joining the line segments $\underline{OA}$ and $\underline{AB}$.
- (b) Are the integrals (i) and (ii) the same?

Solution: (a)(i) Note that $\underline{OA}$ is parametrized by $z = t$, $\underline{AB}$ is parametrized by $z = 1 + it$, $0 \leq t \leq 1$. According to Theorem 14.9,

$$\int_{\underline{OAB}} z dz = \int_{\underline{OA}} z dz + \int_{\underline{AB}} z dz = \int_0^1 t dt + \int_0^1 (1 + it)(idt) = \frac{1}{2} + i - \frac{1}{2} = i.$$

- (a)(ii) Note that $\underline{OB}$ is parametrized by $z = t + it$, $0 \leq t \leq 1$.

$$\int_{\underline{OB}} z dz = \int_0^1 (t + it)(1 + i)dt = \frac{(1 + i)^2}{2} = i.$$

- (b) They are the same and the “integral” independent of the paths.

Example 14.11. Given the points $O(0,0)$, $A(1,0)$ and $B(1,1)$.

- (a) Evaluate the integral (i) $\int_{\underline{OAB}} \Re(3z)dz$ and (ii) $\int_{\underline{OB}} \Re(3z)dz$, where $\underline{OAB}$ is a piecewise linear curve joining the line segments $\underline{OA}$ and $\underline{AB}$.
- (b) Are the integrals (i) and (ii) the same?

Solution. (a) (i) Note that $\underline{OA}$ is parametrized by $z = t$, $\underline{AB}$ is parametrized by $z = 1 + it$, $0 \leq t \leq 1$. According to Theorem 14.9 Properties 1 and 4,

$$\begin{aligned} \int_{\underline{OAB}} \Re(3z)dz &= \int_{\underline{OAB}} 3\Re(z)dz = 3 \int_{\underline{OAB}} \Re(z)dz = 3 \left[\int_{\underline{OA}} \Re(z)dz + \int_{\underline{AB}} \Re(z)dz \right] \\ &= 3 \left[\int_0^1 \Re(t)dt + \int_0^1 \Re(1 + it)(idt) \right] = \frac{3}{2} + 3i. \end{aligned}$$

(a)(ii) Note that $\underline{OB}$ is parametrized by $z = t + it$, $0 \leq t \leq 1$. According to Theorem 14.9 Properties 1,

$$\int_{\underline{OB}} \Re(3z)dz = 3 \int_{\underline{OB}} \Re(z)dz = 3 \int_0^1 \Re(t + it)(1 + i)dt = 3(1 + i) \int_0^1 t dt = \frac{3}{2}(1 + i).$$

(b) They are different due to different paths. □

eg:5.5.5

Example 14.12. Evaluate $\int_C \frac{dz}{z}$ where C represent the path from $z = a$ to $z = -a$ along the upper half of the circle with radius a and centred at the origin.

Solution. The curve C is parametrized by $z = ae^{it}$, $0 \leq t \leq \pi$.

$$\int_C \frac{dz}{z} = \int_0^\pi \frac{iae^{it}dt}{ae^{it}} = i(\pi - 0) = \pi i.$$

□

When C is a closed curve, we shall use the notation $\oint_C f(z)dz$ to represent the complex integral along C in the positive (counter-clockwise) orientation instead of $\int_C f(z)dz$. $\oint_C f(z)dz$ is commonly called *contour integral*.

eg:non-holom

Example 14.13 (April 2009 Exam Q4(a)). Evaluate $\oint_C \bar{z}^2 dz$, where C is the positively oriented circle (i) $C : |z - 1| = 2$. (5 marks) (ii) $C : |z| = 2$. (5 marks)

Solution:

(i) Let $z(t) = 1 + 2e^{it}$, $0 \leq t < 2\pi$, then

$$\begin{aligned} \oint_C \bar{z}^2 dz &= \int_0^{2\pi} (1 + 2e^{-it})^2 (2ie^{it}) dt = 2i \int_0^{2\pi} (1 + 4e^{-it} + 4e^{-2it}) e^{it} dt \\ &= 2i \int_0^{2\pi} e^{it} + 4 + 4e^{-it} dt = 8i(2\pi) = 16\pi i \end{aligned}$$

(ii) Let $z(t) = 2e^{it}$, $0 \leq t < 2\pi$, then

$$\oint_C \bar{z}^2 dz = \int_0^{2\pi} (2e^{-it})^2 (2ie^{it}) dt = 2i \int_0^{2\pi} (4e^{-2it}) e^{it} dt = 8i \int_0^{2\pi} e^{-it} dt = 0$$

eg:5.3.1

Example 14.14. Evaluate the contour integral $\oint_C \frac{1}{z} dz$ where C is the circle $x = r \cos t$, $y = r \sin t$, $0 \leq t \leq 2\pi$.

Solution. The points z on a circle C can be expressed as

$$z(t) = r(\cos t + i \sin t) = re^{it}, \quad 0 \leq t \leq 2\pi$$

and therefore

$$\oint_C \frac{1}{z} dz = \int_0^{2\pi} \frac{1}{re^{it}} (ire^{it}) dt = i \int_0^{2\pi} dt = 2\pi.$$

□

eg:5.5.6

Example 14.15. Evaluate $\oint_C \frac{1}{(z-a)^n} dz$ where C is a circle with radius r and centred at a .

Solution. Circle C is parametrized by $z = a + re^{it}$, $0 \leq t \leq 2\pi$.

$$\oint_C \frac{1}{(z-a)^n} dz = \int_0^{2\pi} \frac{ire^{it} dt}{r^n e^{int}} = \begin{cases} i \left[\frac{1}{r^{1-n}} \right]_0^{2\pi} = 2\pi i, & n = 1, \\ i \left[\frac{r^{1-n}}{i(1-n)} e^{i(1-n)t} \right]_0^{2\pi} = 0, & n \neq 1. \end{cases}$$

□

A striking thing about the results above is that the answers do not depend on the radius of the circle.

Property 5. in Theorem [14.9](#) is used often in the theory of complex integration and is referred to as *ML-inequality* and hence it deserves a proof.

Proof of property (5): Note that $\int_a^b |\gamma'(t)| dt = \int_a^b \sqrt{x'(t)^2 + y'(t)^2} dt = L$ is the length of the smooth curve $C : \gamma(t) = x(t) + iy(t)$. Then

$$\left| \int_C f(z) dz \right| = \left| \int_a^b f(\gamma(t)) \gamma'(t) dt \right| \leq \int_a^b |f(\gamma(t)) \gamma'(t)| dt \leq \int_a^b M |\gamma'(t)| dt = ML.$$

eg:5.4.1

Example 14.16. Suppose C is the circle $|z| = r$. Show that

$$\left| \oint_C e^{\Re z} dz \right| \leq 2\pi r e^r.$$

Solution: Since $z = re^{it}$, $0 \leq t \leq 2\pi$ on C , from Theorem [14.9](#),

$$|e^{\Re z}| = |e^{r \cos t}| \leq e^r \Rightarrow \left| \oint_C e^{\Re z} dz \right| \leq e^r \underbrace{(2\pi r)}_{\text{length of } C}.$$

0101:2014:b7ii

Example 14.17. Use the ML-inequality to prove the following inequality

$$\left| \int_{|z|=4} \frac{z+1}{z^2-3z+2} dz \right| \leq 20\pi.$$

Solution. To carry out the prove, we first estimate the numerator and denominator of $|f(z)|$ on $|z| = 4$:

$$|z+1| \leq |z| + 1 = 5$$

and since $z^2 - 3z + 2 = (z-1)(z-2)$, on $|z| = 4$, we have

$$\min. \text{ distance}(1, z : |z| = 4) = 3 \leq |z-1| \bigwedge \min. \text{ distance}(2, z : |z| = 4) = 2 \leq |z-2|$$

Therefore,

$$\left| \int_{|z|=4} \frac{z+1}{z^2-3z+2} dz \right| \leq \left| \frac{z+1}{z^2-3z+2} \right| (2\pi \times 4) \leq \frac{5}{3 \times 2} \times 8\pi = \frac{20}{3}\pi \leq 20\pi.$$

□

14.3 Anti-Derivatives

sec:anti-deriv

One of the most important theorem in Calculus is the Fundamental Theorem of Calculus, which says that if f is integrable and has an anti-derivative F , then $\int_a^b f(x) dx = F(b) - F(a)$. So it is natural for us to ask if it is possible to generalise the fundamental theorem to complex analysis?

When a complex function $f : D \rightarrow \mathbb{C}$ is given, it is possible to find a function $F : D \rightarrow \mathbb{C}$ such that for $z \in D$,

$$\frac{d}{dz}F(z) = f(z)?$$

We will have to be extremely careful because from Topic 2, we learned that F must be “holomorphic”, otherwise, it is impossible to have a meaningful “differentiation”!

If it is possible for us to find such a holomorphic function F such that $F'(z) = f(z)$, then F is an anti-derivative of f .

Definition 14.12. Let D be a region in $\mathbb{C}$ and $F : D \rightarrow \mathbb{C}$ be a complex function such that $F'(z) = f(z)$ in D . F is called the *anti-derivative* of f .

A necessary and sufficient condition for the existence of anti-derivative is that the integral of a function must be “path-independent”.

Definition 14.13. If an integral depends **only** on the initial and terminal points of the curve C and is independent on the curve C . We say the integral is *path independent*.

path-independent

Theorem 14.18. Suppose $f : D \rightarrow \mathbb{C}$ is continuous, where D is an open connected region. Then f has an antiderivative if and only if the integral between any two points of D is path independent.

Theorem 14.19 (Fundamental Theorem). Let D be an open connected region, C be a contour in D with an initial point z_0 and terminal point z_1 . If $f : D \rightarrow \mathbb{C}$ is continuous and f has an antiderivative $F : D \rightarrow \mathbb{C}$, then

$$\int_C f(z)dz = F(z_1) - F(z_0). \quad (14.6)$$

e:fundamental

A detail proof is available in Conway [1978, Chapter 4].

Although Theorem 14.18 characterises the existence of anti-derivative, it is not very useful. In calculation, we need to find F instead of knowing its existence! Let us practise with some examples below.

Example 14.20. Prove that $\int_0^{\pi i} z^2 \sin z dz = (\pi^2 + 2) \cosh \pi - 2\pi \sinh \pi - 2$.

Solution: By apply the knowledge from Calculus we know can find the anti-derivative using integration by parts:

$$\int z^2 \sin z dz = z^2(-\cos z) - \int (-\cos z) \cdot 2z dz = -z^2 \cos z + 2z \sin z + 2 \cos z + C$$

We can check that $F(z) = -z^2 \cos z + 2z \sin z + 2 \cos z$ is entire (because it is the product and sum of entire functions) and that $F'(z) = f(z)$. Hence,

$$\begin{aligned} \int_0^{\pi i} z^2 \sin z dz &= F(\pi i) - F(0) = (-(\pi i)^2 \cos \pi i + 2\pi i \sin \pi i + 2 \cos \pi i) - (2 \cos 0) \\ &= (\pi^2 + 2) \cosh \pi - 2\pi \sinh \pi - 2. \end{aligned}$$

since $\sin iz = i \sinh z$ and $\cos iz = \cosh iz$ (Topic 2).

Example 14.21. Evaluate $\int_C z^2 dz$ where C is

eg:wwy:5.6.1

1. the curve $y = 1/x^2$ from point $z = 1 + i$ to point $z = 3 + i/9$.

Solution. Note that z^2 has an anti-derivative $F(z) = z^3/3$.

$$\text{Therefore, } \int_C z^2 dz = \left[\frac{z^3}{3} \right]_{1+i}^{3+i/9} = \frac{242}{27} + \frac{2186}{2187}i - \left(-\frac{2}{3} + \frac{2}{3}i \right) = \frac{260}{27} + \frac{728}{2187}i \quad \square$$

2. [April 2010 Exam, Q2(b), 6 marks] the curve $y = \frac{1}{x}$ from point $z = 1 + i$ to point $z = 2 + \frac{i}{2}$.

$$\text{Solution. } \int_C z^2 dz = \left[\frac{z^3}{3} \right]_{1+i}^{2+i/2} = \frac{13}{6} + \frac{47}{24}i - \left(-\frac{2}{3} + \frac{2}{3}i \right) = \frac{17}{6} + \frac{31}{24}i \quad \square$$

14.4 Cauchy Theorems

sec:cauchy-thm

In the previous section, we state the fundamental theorem for contour integral. It is easy to obtain the following result.

Corollary 14.22. *Let D be an open connected region, C be a closed contour in D . If $f : D \rightarrow \mathbb{C}$ is continuous and f has an antiderivative $F : D \rightarrow \mathbb{C}$, then*

$$\int_C f(z) dz = 0. \quad (14.7)$$

e:cauchy

In this section, we will investigate a generalisation of the above results, first stated by the french mathematician Augustin Louis Cauchy (1789~1857). First, we need to classify the “connected domains” based on how many “holes” they have.

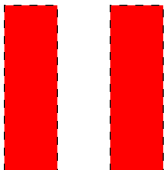
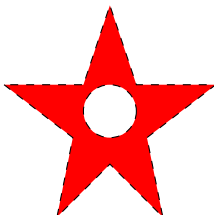
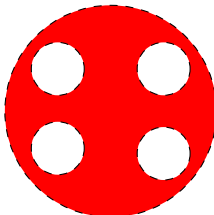
After that, we will study the contour integrals over simple closed curves with a positive orientation. In future, all contour will assume to have positive (or anti-clockwise) orientation!

14.4.1 Simply and Multiply Connected Domains

Definition 14.14 (Simply Connected Domain). A region or a domain (recall Definition 1.4.9 from Topic 1) D is said to be *simply connected* if every simple closed contour C lying entirely in D can be shrunk to a point without leaving D .

Definition 14.15. A region that is not simply connected is called a *multiply connected region*. When there is a “hole” in the multiply connected region, we call it *doubly connected region*.

In lay man term, a simply connected region has no “holes” in it and a multiply connected region is a region with “holes” in it as illustrated in the following figures.

Disconnected	Simply connected	Doubly/Multiply connected	Multiply connected
			

14.4.2 Cauchy Theorem

In 1883, Cauchy proved one of the most important theorem in complex analysis.

Theorem 14.23 (Cauchy's Theorem). *Suppose that complex function f is holomorphic in a simply connected region D and f' is continuous in D . Then for every simple closed contour C in D ,*

$$\int_C f(z)dz = 0.$$

Proof: Let $f(z) = u(x, y) + iv(x, y)$ and assume that u and v and their first partial derivatives are continuous in D . By expressing the contour integral as line integrals, we obtain

$$\oint_C f(z)dz = \oint_C udx - vdy + i \oint_C vdx + udy. \quad (14.8)$$

e:6.1.1

By applying *Green theorem* (Calculus III or Advanced Calculus) below

$$\int_C Pdx + Qdy = \iint_D \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA$$

to the right hand side of (14.8), we have

$$\oint_C f(z)dz = \iint_D \left(\frac{\partial(-v)}{\partial x} - \frac{\partial u}{\partial y} \right) dA + i \iint_D \left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} \right) dA.$$

By applying the Cauchy-Riemann equations, we have $\frac{\partial(-v)}{\partial x} - \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x} - \left(-\frac{\partial v}{\partial x}\right) = 0$ and $\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} = 0$. Hence, $\oint_C f(z)dz = \iint_D 0dA + \iint_D 0dA = 0$.

This argument is good enough for many settings in which Cauchy's theorem is used. However, it is not an optimal argument because it makes the additional assumption about the continuity of f' , i.e. "the continuity of the partial derivatives of u and v ".

Example 14.24. Prove that $\int_{|z|=1} z^2 dz = 0$.

Solution: From Topic 2, we know that $f(z) = z^2$ is an entire function. It's derivative $f'(z) = 2z$ is continuous in $\mathbb{C}$. For the circle $|z| = 1$, which is a simple closed contour in $\mathbb{C}$, by Cauchy Theorem, we have

$$\int_{|z|=1} z^2 dz = 0.$$

14.4.3 Cauchy-Goursat Theorem

The French mathematician Edouard Goursat was he first to prove that the condition of continuity on f' can be omitted [Churchill and Brown, 2008].

Theorem 14.25 ((Cauchy-)Goursat Theorem). *Suppose a function f is holomorphic in a simply connected region D . Then for every simple closed contour C in D ,*

$$\oint_C f(z)dz = 0.$$

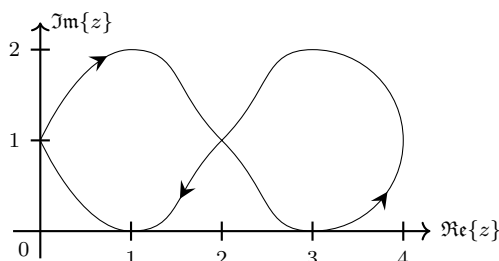
A proof can be found in Churchill and Brown [2008, Chapter 4, Section 47].

The Cauchy-Goursat theorem can be extended to general closed contour as follows.

Theorem 14.26 ((Cauchy-)Goursat Theorem, General Contour). *Suppose a function f is holomorphic in a simply connected region D . Then for every closed contour C in D ,*

$$\oint_C f(z)dz = 0.$$

Example 14.27. Calculate $\oint_C \exp(z^2 + 3z + 5)dz$ given that C defined below.



Solution. First, $f(z) = \exp(z^2 + 3z + 5)$ is a composition of entire functions and is entire, i.e. $D = \mathbb{C}$. For the closed contour C , $\oint_C \exp(z^2 + 3z + 5)dz = 0$ by Cauchy-Goursat Theorem. □

Corollary 14.28. *A function f which is holomorphic throughout a simply connected region D must have an anti-derivative in D .*

This corollary follows from Theorem [14.18](#) and Goursat Theorem. It tells us that entire functions always have possess anti-derivatives.

The Cauchy-Goursat Theorem can be further extended to multiply-connected regions.

Theorem 14.29 (Cauchy-Goursat Theorem, Multiply-Connected Regions). *Let C and C_k , $k = 1, 2, \dots, n$ be the simple closed contours in counter-clockwise direction such that C_k are interior to C and whose interiors have no points in common. If a function f is holomorphic throughout the closed region consisting of all points within and on C except for points interior to any C_k , then*

$$\oint_C f(z)dz + \sum_{k=1}^n \oint_{C_k} f(z)dz = 0. \quad (14.9)$$

The following corollary is an important consequence of the above theorem.

Corollary 14.30 (Principle of Deformation of Path). *Let C_1 and C_2 denote positively oriented simple closed contours where C_2 is interior to C_1 . If a function f is holomorphic in the closed region consisting of those contours and all points between them, then*

$$\oint_{C_1} f(z)dz = \oint_{C_2} f(z)dz.$$

14.5 General Cauchy Integral Formula

In this section, we study a remarkable formula, due to Cauchy, which shows that the values of an holomorphic function at the interior point of a simple closed curve are determined by the values of the function on the curve.

14.5.1 Cauchy Integral Formula on a Disk

Theorem 14.31 (Cauchy Integral Formula on a Disk). *Let f be holomorphic on a closed disk $\overline{D}$ of positive radius and let γ be a circle surrounding the disk. Then*

$$f(z) = \frac{1}{2\pi i} \oint_{\gamma} \frac{f(\zeta)}{\zeta - z} d\zeta \quad (14.10)$$

for all z in the open disk D .

Proof: For a fixed point $z \in D$, define the function

$$g(t) = \oint_C \frac{f(z + t(\zeta - z))}{\zeta - z} d\zeta.$$

Its derivative is

$$g'(t) = \oint_C f'(z + t(\zeta - z)) d\zeta = \left[\frac{f(z + t(\zeta - z))}{t} \right]_{C(0)}^{C(1)} = 0.$$

Therefore, g is a constant function and when $t = 0$ and $t = 1$, we have

$$g(1) = g(0) \Rightarrow \oint_C \frac{f(\zeta)}{\zeta - z} d\zeta = \oint_C \frac{f(z)}{\zeta - z} d\zeta = 2\pi i f(z) \Rightarrow (14.10).$$

14.5.2 Cauchy Integral Formula for Simply Connected Domain

Let D be a simply connected region and let z_0 be a fixed point in D . If f is holomorphic in D , the function $\frac{f(z)}{z - z_0}$ will fail to be holomorphic at z_0 . Hence the integral $\oint_C \frac{f(z)}{z - z_0} dz$ will in general not be zero. However, this integral will have the same value on all contour C about z_0 . This fundamental result is the following well-known theorem.

Theorem 14.32 (Cauchy Integral Formula). *Let $f(z)$ be holomorphic in a domain D . Let C be a simple closed curve in D , within which $f(z)$ is holomorphic, and let z_0 be inside C . Then*

$$f(z_0) = \frac{1}{2\pi i} \oint_C \frac{f(z)}{z - z_0} dz. \quad (14.11)$$

(Sept 2014 Exam Q4(a), 10 marks; Sept 2012 Exam Q4(a), 15 marks)

Note that the domain D is not required to be simply connected, but since f is holomorphic within C , the theorem concerns only a simply connected part of D .

Proof of the Cauchy's Integral Formula Theorem:

Let ρ be a positive number small enough so that the circle $C_\rho : |z - z_0| = \rho$ is interior to C . It follows from the Principle of Deformation of Path (Corollary 14.30) that

$$\oint_C \frac{f(z)}{z - z_0} dz = \oint_{C_\rho} \frac{f(z)}{z - z_0} dz.$$

It remains to show the integral on the right is indeed $f(z_0)2\pi i$. Now since $f(z_0)$ is constant,

$$\oint_{C_\rho} \frac{f(z_0)}{z - z_0} dz = f(z_0) \cdot 2\pi i$$

and

$$\left| \oint_{C_\rho} \frac{f(z)}{z - z_0} dz - f(z_0) \cdot 2\pi i \right| = \left| \oint_{C_\rho} \frac{f(z) - f(z_0)}{z - z_0} dz \right|.$$

Since f is continuous at z_0 , for any $\epsilon > 0$, there exists $\delta > 0$, such that

$$|z - z_0| < \delta \Rightarrow |f(z) - f(z_0)| < \epsilon.$$

Now choose $\rho < \delta$. Then C_ρ is parametrized by $\gamma(t) = z_0 + \rho e^{it}$ and

$$\begin{aligned} \left| \oint_{C_\rho} \frac{f(z) - f(z_0)}{z - z_0} dz \right| &\leq \int_0^{2\pi} \left| \frac{f(\gamma(t)) - f(z_0)}{\gamma(t) - z_0} \right| |\gamma'(t)| dt \\ &< \int_0^{2\pi} \frac{\epsilon}{|\rho e^{it}|} |\rho e^{it}| dt = \epsilon(2\pi - 0) = 2\pi\epsilon. \end{aligned}$$

Since ϵ is arbitrary,

$$\oint_{C_\rho} \frac{f(z) - f(z_0)}{z - z_0} dz = 0 \Rightarrow \text{(14.11)}$$

Let us look at some examples below to learn how to apply Equation (14.11). When the denominator of the integrand is a polynomial the technique of partial fraction is required.

Example 14.33. Calculate $\int_{|z|=4} \frac{z+1}{z^2-3z+2} dz$ and compare your result to Example 14.17.

Solution: We will mostly always use (14.11) of the form

$$\oint_C \frac{f(z)}{z - z_0} dz = 2\pi i f(z_0).$$

The region $|z| < 4$ contains the “poles” 2 and -1 according to the partial fraction:

$$\int_{|z|=4} \frac{z+1}{z^2-3z+2} dz = \int_{|z|=4} \left[\frac{3}{z-2} - \frac{2}{z-1} \right] dz = 2\pi i(3) - 2\pi i(2) = 2\pi i.$$

The estimate 20π is too large in Example 14.17.

Example 14.34 (April 2010 Exam, Q1(a)). Evaluate $\oint_{|z|=5} \frac{6z-10}{z^2-2z-3} dz$. (7 marks)

Solution. The region $|z| < 5$ contains the points 3 and -1 according to the following partial fraction.

$$\begin{aligned} \oint_{|z|=5} \frac{6z-10}{z^2-2z-3} dz &= \oint_{|z|=5} \frac{6z-10}{(z-3)(z+1)} dz = \oint_{|z|=5} \left[\frac{4}{z+1} + \frac{2}{z-3} \right] dz \\ &= 2\pi i(4) + 2\pi i(2) = 12\pi i \end{aligned}$$

□

Example 14.35 (April 2009 Exam, Q4(b)(ii)). Evaluate $\oint_C \frac{2z+1}{z^2+2z-3} dz$ where C is the circle $|z-2|=2$ oriented positively. (5 marks)

Solution. The region $|z-2| < 2$ only contains the point 1 and not -3 .

$$\begin{aligned} \oint_C \frac{2z+1}{z^2+2z-3} dz &= \oint_{|z-2|=2} \frac{2z+1}{(z+3)(z-1)} dz = \oint_{|z-2|=2} \frac{\frac{2z+1}{z+3}}{z-1} dz \\ &= 2\pi i \left[\frac{2(1)+1}{1+3} \right] = \frac{3\pi i}{2}. \end{aligned}$$

□

Example 14.36 (April 2010 Exam, Q3(c)). Find the value of the integral $\oint_C \frac{3z^2 + 2}{(z-1)(z^2+4)} dz$ taken counterclockwise around the circle $|z| = 3$. (13 marks)

Solution. The region $|z| < 3$ contains the points 1 and $\pm 2i$ according to the following partial fraction:

$$\begin{aligned}\oint_C \frac{3z^2 + 2}{(z-1)(z^2+4)} dz &= \oint_C \left[\frac{\frac{3(1)^2+2}{1^2+4}}{z-1} + \frac{\frac{3(-2i)^2+2}{(-2i-1)(-4i)}}{z+2i} + \frac{\frac{3(2i)^2+2}{(2i-1)(4i)}}{z-2i} \right] dz \\ &= \oint_C \left[\frac{1}{z-1} + \frac{1+\frac{1}{2}i}{z+2i} + \frac{1-\frac{1}{2}i}{z-2i} \right] dz \\ &= 2\pi i \left(1 + 1 + \frac{1}{2}i + 1 - \frac{1}{2}i \right) = 6\pi i.\end{aligned}$$

□

eg:7.1.1

Example 14.37. Let C denote the boundary of the square with vertices at $(2, 2)$, $(-2, -2)$, $(2, -2)$, and $(-2, 2)$. Find $\oint_C \frac{e^{-z}}{z - \pi i/2} dz$.

Solution. Note that $\pi i/2$ is in the square surrounded by C and e^{-z} is an entire function, therefore

$$\oint_C \frac{e^{-z}}{z - \pi i/2} dz = 2\pi i (e^{-\pi i/2}) = 2\pi i (-i) = 2\pi.$$

□

14.5.3 Cauchy Integral Formula for Multiply Connected Domain

The Cauchy's integral formula can be extended to multiply connected region.

hy-integral:multi

Theorem 14.38. Suppose that

- C is a simple closed contour and $C_1, C_2, \dots, C_n$ are non-overlapping simple closed contour interior to C ;
- f is holomorphic on all points inside C and exterior to each C_k .

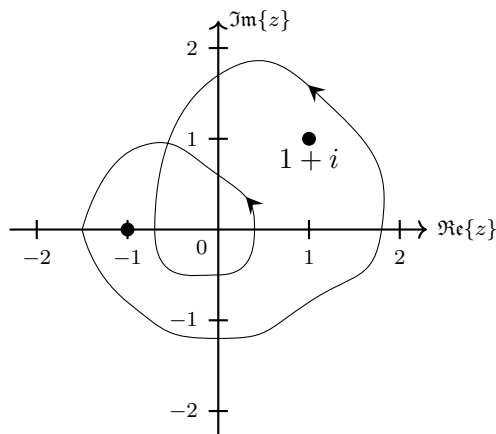
If z_0 is any point inside C bounded by C_1 (the outer boundary) $C_2, \dots, C_n$, then

$$f(z_0) = \frac{1}{2\pi i} \oint_C \frac{f(z)}{z - z_0} dz = \sum_{k=1}^n \frac{1}{2\pi i} \oint_{C_k} \frac{f(z)}{z - z_0} dz. \quad (14.12)$$

e:cauchy-integra

The following example is complicated because the curve is not simple. Therefore, we need to apply Theorem [14.9](#) to decompose the curve into a combination of simple curves and then apply Theorem [14.38](#).

Example 14.39 (Sept 2012 Exam, Q3(b)). Evaluate $\oint_C \frac{\cos z}{(z-1-i)(z^2-4z-5)} dz$, where C is the closed path below. [Do not simplify your answer.] (9 marks)



Solution. Let $f(z) = \frac{\cos z}{(z-1-i)(z^2-4z-5)} = \frac{\cos z}{(z-1-i)(z-5)(z+1)}$. Let C_1 be the small curve encircled $(0,0)$, C_2 be the large curve encircled $(1,1)$.

$$\begin{aligned}\oint_C f(z)dz &= \oint_{C_1} f(z)dz + \oint_{C_2} f(z)dz \\ &= 0 + \left[\oint_{|z+1|=\epsilon} \frac{\frac{\cos z}{(z-1-i)(z-5)}}{z+1} dz + \oint_{|z-1-i|=\epsilon} \frac{\frac{\cos z}{(z-5)(z+1)}}{z-1-i} dz \right] \\ &= 2\pi i \left[\frac{\cos(-1)}{((-1)-1-i)((-1)-5)} + \frac{\cos(1+i)}{(1+i-5)(1+i+1)} \right] \\ &= 2\pi i \left[\frac{\cos 1}{12+6i} + \frac{\cos(1+i)}{(i-4)(2+i)} \right] = 2\pi i \left[\frac{\cos 1}{12+6i} - \frac{\cos(1+i)}{9+2i} \right]\end{aligned}$$

□

An important consequence of Cauchy's integral theorem is that we can show that an holomorphic function possess derivatives of all orders. This is rather remarkable result, and much nicer than in real analysis.

thm:7.1.2

Theorem 14.40. Suppose that f is holomorphic everywhere inside and on a simple closed curve C . If z is any point interior to C , then

$$f'(z) = \frac{1}{2\pi i} \oint_C \frac{f(\zeta)}{(\zeta-z)^2} d\zeta. \quad (14.13)$$

e:cauchy-integra

The derivative f' is holomorphic in D .

In general, for every positive integer n , the n -th derivative $f^{(n)}$ is holomorphic in D

$$f^{(n)}(z) = \frac{n!}{2\pi i} \oint_C \frac{f(\zeta)}{(\zeta-z)^{n+1}} d\zeta, \quad n = 1, 2, \dots. \quad (14.14)$$

e:cauchy-integra

Proof: Let d denote the smallest distance from z to points on C . Let $0 < |\Delta z| < d$.

$$\frac{f(z+\Delta z) - f(z)}{\Delta z} = \frac{1}{2\pi i} \oint_C \left(\frac{1}{\zeta - z - \Delta z} - \frac{1}{\zeta - z} \right) \frac{f(\zeta)}{\Delta z} d\zeta = \frac{1}{2\pi i} \oint_C \frac{f(\zeta)}{(\zeta - z - \Delta z)(\zeta - z)} d\zeta.$$

Then

$$\frac{f(z+\Delta z) - f(z)}{\Delta z} - \frac{1}{2\pi i} \oint_C \frac{f(\zeta)}{(\zeta - z)^2} d\zeta = \frac{1}{2\pi i} \oint_C \frac{f(\zeta)\Delta z}{(\zeta - z - \Delta z)(\zeta - z)^2} d\zeta. \quad (*)$$

Let M denote the maximum value of $|f(z)|$ on C and observe that $|\zeta - z| > d$ and $|\Delta z| < d$. Therefore

$$|\zeta - z - \Delta z| \geq |\zeta - z| - |\Delta z| \geq d - |\Delta z| > 0.$$

Thus, if L is the length of C , we have

$$\left| \oint_C \frac{f(\zeta)\Delta z}{(\zeta - z - \Delta z)(\zeta - z)^2} d\zeta \right| \leq \frac{M|\Delta z|}{(d - |\Delta z|)^2} L \rightarrow 0 \Rightarrow \text{(14.13)} \quad \text{e:cauchy-integral:diff1}$$

according to (*) as $\Delta z \rightarrow 0$.

In the above, we have proven the base case $n = 1$. Assume that

$$f^{(k)}(z) = \frac{k!}{2\pi i} \oint_C \frac{f(\zeta)}{(\zeta - z)^{k+1}} d\zeta, \quad k \geq 1.$$

We now consider $n = k + 1$:

$$\begin{aligned} \frac{f^{(k)}(z + \Delta z) - f^{(k)}(z)}{\Delta z} &= \frac{k!}{2\pi i} \oint_C \left(\frac{1}{(\zeta - z - \Delta z)^{k+1}} - \frac{1}{(\zeta - z)^{k+1}} \right) \frac{f(\zeta)}{\Delta z} d\zeta \\ &= \frac{k!}{2\pi i} \oint_C \frac{f(\zeta)[(k+1)(\zeta - z)^k + \frac{(k+1)k}{2}(\zeta - z)^{k-1}\Delta z + O((\Delta z)^2)]}{(\zeta - z - \Delta z)^{k+1}(\zeta - z)^{k+1}} d\zeta. \end{aligned}$$

Then

$$\begin{aligned} &\frac{f^{(k)}(z + \Delta z) - f^{(k)}(z)}{\Delta z} - \frac{(k+1)!}{2\pi i} \oint_C \frac{f(\zeta)}{(\zeta - z)^{k+2}} d\zeta \\ &= \frac{k!}{2\pi i} \oint_C \left[\frac{f(\zeta)[(k+1)(\zeta - z)^k + O(\Delta z)]}{(\zeta - z - \Delta z)^{k+1}(\zeta - z)^{k+1}} - \frac{(k+1)f(\zeta)}{(\zeta - z)^{k+2}} \right] d\zeta \\ &= \frac{k!}{2\pi i} \oint_C f(\zeta) \left[\frac{[(k+1)(\zeta - z)^{k+1} + O(\Delta z)] - (k+1)(\zeta - z - \Delta z)^{k+1}}{(\zeta - z - \Delta z)^{k+1}(\zeta - z)^{k+2}} \right] d\zeta \\ &= \frac{k!}{2\pi i} \oint_C f(\zeta) \left[\frac{[\frac{(k+1)k}{2}(\zeta - z)^k \Delta z + O((\Delta z)^2)] - (k+1)^2(\zeta - z)^k(-\Delta z) + O((\Delta z)^2)}{(\zeta - z - \Delta z)^{k+1}(\zeta - z)^{k+2}} \right] d\zeta \\ &= \frac{k!}{2\pi i} \oint_C f(\zeta) \left[\frac{[(k+1)(\frac{3}{2}k + 1)(\zeta - z)^k \Delta z + O((\Delta z)^2)]}{(\zeta - z - \Delta z)^{k+1}(\zeta - z)^{k+2}} \right] d\zeta = RHS. \end{aligned}$$

The right hand side is controlled by

$$|RHS| \leq \frac{k!M}{2\pi} \left[\frac{[(k+1)(\frac{3}{2}k + 1)d^k]}{(d - |\Delta z|)^{k+1}d^{k+2}} \right] L|\Delta z| + O(|\Delta z|^2) \rightarrow 0$$

as $\Delta z \rightarrow 0$.

The conclusion follows by induction. Hence, $f^{(n+1)}$ is holomorphic in D .

cor:7.1.1

Corollary 14.41. If $f(z) = u(x, y) + iv(x, y)$ is defined and holomorphic at $z = x + iy$, then $u(x, y)$ and $v(x, y)$ have continuous partial derivatives of all orders at that point.

When (14.14) be written in the form

$$\oint_C \frac{f(z)}{(z - z_0)^n} dz = \frac{2\pi i}{(n-1)!} f^{(n-1)}(z_0), \quad (14.15) \quad \text{e:cauchy-integ}$$

it can be useful in evaluating certain integrals.

Example 14.42 (April 2009 Exam, Q4(b)(i)). Evaluate $\oint_C \frac{1}{(z - z_0)^n} dz$ where n is an integer and C is a simple closed contour oriented positively and z_0 is any point interior to C . (10 marks)

Solution: This question is a generalisation of Example [eg:5.5.6](#) 14.15.

When $n \leq 0$, $\frac{1}{(z - z_0)^n}$ is holomorphic on $\mathbb{C}$. Therefore $\oint_C \frac{1}{(z - z_0)^n} dz = 0$.

Note that $f(z) = 1$ and $D = \mathbb{C}$.

When $n = 1$, $\oint_C \frac{1}{z - z_0} dz = 2\pi i$ by Cauchy Integral Formula (Theorem [thm:cauchy-integral](#) 14.32)

When $n > 1$, $\oint_C \frac{1}{(z - z_0)^n} dz = \frac{2\pi i}{(n - 1)!} \frac{d^{n-1}}{dz^{n-1}} 1 = 0$ by Equation ([e:cauchy-integral:diffn2](#) 14.15).

[eg:7.1.2](#)

Example 14.43. Evaluate $\oint_{|z|=1} \frac{z+1}{z^4+2iz^3} dz$.

Solution. Notice that $\frac{z+1}{z^4+2iz^3} = \frac{\frac{z+1}{z+2i}}{z^3}$. The function $f(z) = \frac{z+1}{z+2i}$ is holomorphic inside and on $|z| = 1$ and

$$f'(z) = -\frac{1-2i}{(z+2i)^2}, \quad f''(z) = \frac{2(1-2i)}{(z+2i)^3}.$$

Hence

$$\oint_{|z|=1} \frac{z+1}{z^4+2iz^3} dz = \frac{2\pi i}{2!} f''(0) = -\frac{\pi}{4} + \frac{\pi}{2}i.$$

□

[eg:7.1.3](#)

Example 14.44. Evaluate $\oint_C \frac{\exp(z^3)}{(z-i)^3} dz$ with C any path enclosing $z = i$ as an interior point.

Solution. Let $f(z) = \exp(z^3)$. Then

$$f'(z) = 3z^2 \exp(z^3), \quad f''(z) = 6z \exp(z^3) + 9z^4 \exp(z^3) = (6z + 9z^4) \exp(z^3).$$

Hence,

$$\oint_C \frac{\exp(z^3)}{(z-i)^3} dz = \frac{2\pi i}{2!} f''(i) = \pi i(6i+9)e^{-i} = \pi(9i-6)e^{-i}.$$

□

Example 14.45 (April 2010 Exam, Q1(b)). Let C be any simple closed contour, described in the positive sense in the z -plane and let

$$f(z) = \oint_C \frac{\zeta^3 + 2\zeta}{(\zeta - z)^3} d\zeta.$$

Show that $f(z) = 6\pi iz$ when z is inside C and that $f(z) = 0$ when z is outside. (8 marks)

Solution. When z is outside,

$$F(\zeta) = \frac{\zeta^3 + 2\zeta}{(\zeta - z)^3} \quad \text{is holomorphic in the domain enclosed by } C.$$

By Cauchy Theorem,

$$\oint_C F(\zeta) d\zeta = \oint_C \frac{\zeta^3 + 2\zeta}{(\zeta - z)^3} d\zeta = 0.$$

When z is inside the region enclosed by C , by generalised Cauchy Integral Formula,

$$\oint_C \frac{\zeta^3 + 2\zeta}{(\zeta - z)^3} d\zeta = \frac{2\pi i}{2!} \frac{d^2}{d\zeta^2} (\zeta^3 + 2\zeta) \Big|_{\zeta=z} = \pi i \frac{d}{d\zeta} (3\zeta^2 + 2) \Big|_{\zeta=z} = 6\pi iz.$$

□

A converse of Cauchy's Theorem is the Morera's Theorem, stated below.

Theorem 14.46 (Morera's Theorem). *If f is continuous in a simply connected domain D and if*

$$\oint_C f(z)dz = 0$$

for every closed contour C in D , then f is holomorphic in D .

Proof: Since f is continuous and $\oint_C f(z)dz = 0$ for every closed contour C in D , we conclude that the line integral $\int_C f(z)dz = 0$ is independent of the path. Let

$$F(z) = \int_{z_0}^z f(\zeta)d\zeta$$

where z_0 is a fixed point in D and z represents any point in D . We know that F is an anti-derivative of f , i.e.,

$$F'(z) = f(z).$$

This implies that F is holomorphic in D and so $f(z) = F'(z)$ is holomorphic.

14.6 Maximum Modulus Principle

ec:max-mod-prin

In this section, we shall derive some results involving maximum values of the moduli of holomorphic functions.

thm:7.3.1

Theorem 14.47 (Local Maximum Modulus Principle). *Suppose that $|f(z)| \leq |f(z_0)|$ at each point z in some neighbourhood $|z - z_0| < \epsilon$ in which f is holomorphic. Then f has the constant value $f(z_0)$ throughout that neighbourhood.*

Proof: Let z_1 be any point other than z_0 in the neighbourhood. Let $\rho = |z_1 - z_0|$. Let C_ρ denotes the circle $|z - z_0| = \rho$. By Cauchy's integral formula,

$$f(z_0) = \frac{1}{2\pi i} \oint_{C_\rho} \frac{f(z)}{z - z_0} dz = \frac{1}{2\pi} \int_0^{2\pi} f(z_0 + \rho e^{i\theta}) d\theta.$$

This means

$$|f(z_0)| \leq \frac{1}{2\pi} \int_0^{2\pi} |f(z_0 + \rho e^{i\theta})| d\theta.$$

On the other hand, $|f(z_0 + \rho e^{i\theta})| \leq |f(z_0)|$ for all θ by assumption and so

$$|f(z_0)| \leq \frac{1}{2\pi} \int_0^{2\pi} |f(z_0 + \rho e^{i\theta})| d\theta \leq \frac{1}{2\pi} \int_0^{2\pi} |f(z_0)| d\theta = |f(z_0)|$$

Hence

$$|f(z_0)| = \frac{1}{2\pi} \int_0^{2\pi} |f(z_0 + \rho e^{i\theta})| d\theta \Rightarrow \frac{1}{2\pi} \int_0^{2\pi} [|f(z_0)| - |f(z_0 + \rho e^{i\theta})|] d\theta = 0$$

which implies

$$|f(z_0)| - |f(z_0 + \rho e^{i\theta})| = 0, \quad 0 \leq \theta \leq 2\pi.$$

because integrating $|f(z_0)| - |f(z_0 + \rho e^{i\theta})| > 0$ must be positive.

That is, $|f(z)| = |f(z_0)|$ for all points on $|z - z_0| = \rho$. Since z_1 is arbitrarily chosen from the disc $|z - z_0| < \epsilon$, $|f(z)| = |f(z_0)|$. Therefore $f(z)$ itself is constant, i.e., $f(z) = f(z_0)$.

We shall state without proof of the following theorem.

Theorem 14.48. *Let f be differentiable on a domain D . Suppose $\{z_n\}$ is a sequence of distinct zeros of f in D , converging to a point of G . Then $f(z) = 0$ on D .*

This is a very strong result, for which there is no analogue for real function. For example, consider the real function

$$h(x) = \begin{cases} 0, & x \leq 0, \\ x^2, & x > 0. \end{cases}$$

This real function is differentiable for all real x and is identically zero on half of the line, but is not identically zero over the entire line.

In the Advanced Calculus course, we learned that a real function which is continuous on a closed interval will be bounded. On the real line, closed intervals are compact.

An analogous result also holds for complex function.

thm:7.4.1

Theorem 14.49. *Suppose f is continuous on a compact subset S of the complex plane. Then f is bounded.*

Note that a set S in $\mathbb{C}$ is compact if and only if it is closed and bounded according to Heine-Borel Theorem (Topic 1).

Proof: Suppose f is not bounded. Then the disk $D_n : |z| \leq n$ cannot contain the image set $f(S)$. This means there is some $z_n \in S$ such that $|f(z_n)| > n$. Now the sequence $\{z_n\}$ is in S . Sequence $\{z_n\}$ is bounded, and hence has a convergent subsequence $\{z_{n_j}\}$ which converges to $z^* \in S$.

Because f is continuous, $f(z_{n_j}) \rightarrow f(z^*)$. Then for some integer m , we can make each $f(z_{n_j})$ lie in the open disk of radius 1 about by choosing $n_j > m$. This contradicts to the fact that $|f(z_{n_j})| > n$.

This result can be improved as follows.

thm:7.4.2

Theorem 14.50. *Let the complex function f be continuous on the compact set. Then there exists z_1 and z_2 in S such that*

$$|f(z_1)| \leq |f(z)| \leq |f(z_2)| \quad \text{for all } z \in S.$$

Proof: By Theorem [thm:7.4.1](#) 14.49, $\{|f(z)| : z \in S\}$ is a bounded set of the real numbers. Let L and U denote respectively the greatest lower bounded and the least upper bound of the set $\{|f(z)| : z \in S\}$.

We will show that there are numbers z_1 and z_2 in S such that $f(z_1) = L$ and $f(z_2) = U$.

Let n be any positive integer. There exists α_n in S such that

$$U - \frac{1}{n} < |f(\alpha_n)| \leq U.$$

This means $|f(\alpha_n)| \rightarrow U$ as $n \rightarrow \infty$. But $\{\alpha_n\}$ is a bounded sequence and therefore has a convergent subsequence $\{\alpha_{n_j}\}$, say $\alpha_{n_j} \rightarrow z_2$. Because f is continuous, $f(\alpha_{n_j}) \rightarrow f(z_2) = U$.

Similarly one can prove that there exists $z_1 \in S$ such that $f(z_1) = L$.

What we have shown so far is that a continuous function defined on a compact set S assumes its maximum value on S . But this does not give any information where the maximum value of $|f(z)|$ is achieved. We will now prove that the maximum value of $|f(z)|$ achieves on the boundary of S .

First we want to show that a differentiable function that is not constant on an open disk cannot have its maximum modulus at the centre of the disk.

thm:7.4.3

Theorem 14.51. *Let f be differentiable and not constant on an open disk D centred at z_0 . Then, for some z in this disk,*

$$|f(z)| > |f(z_0)|.$$

Proof: Suppose $|f(z_0)| \geq |f(z)|$ for all $z \in D$. Let $z(t) = z_0 + \rho e^{i\theta}$, $0 \leq \theta \leq 2\pi$, denote the circle C with small ρ so that this circle is contained in D . Then by Cauchy's integral formula

$$f(z_0) = \frac{1}{2\pi i} \oint_{C_\rho} \frac{f(z)}{z - z_0} dz = \frac{1}{2\pi} \int_0^{2\pi} f(z_0 + \rho e^{i\theta}) d\theta.$$

This means

$$|f(z_0)| \leq \frac{1}{2\pi} \int_0^{2\pi} |f(z_0 + \rho e^{i\theta})| d\theta.$$

Note that $|f(z_0 + \rho e^{i\theta})| \leq |f(z_0)|$ for all θ by assumption and so

$$|f(z_0)| \leq \frac{1}{2\pi} \int_0^{2\pi} |f(z_0 + \rho e^{i\theta})| d\theta \leq \frac{1}{2\pi} \int_0^{2\pi} |f(z_0)| d\theta = |f(z_0)|$$

Hence

$$|f(z_0)| = \frac{1}{2\pi} \int_0^{2\pi} |f(z_0 + \rho e^{i\theta})| d\theta \Rightarrow \frac{1}{2\pi} \int_0^{2\pi} [|f(z_0)| - |f(z_0 + \rho e^{i\theta})|] d\theta = 0.$$

Since the integrand is continuous and nonnegative for $0 \leq \theta \leq 2\pi$. If $|f(z_0)| > |f(z_0 + \rho e^{i\theta})|$ for some θ , then

$$\int_0^{2\pi} (|f(z_0)| - |f(z_0 + \rho e^{i\theta})|) d\theta > 0.$$

Therefore

$$|f(z_0 + \rho e^{i\theta})| = |f(z_0)| \quad \text{for } 0 \leq \theta \leq 2\pi.$$

This means $|f(z)|$ has constant value $|f(z_0)|$ on every circle about z_0 and contained in D . But every point in D is on some circle about z_0 . This implies that $|f(z)| = |f(z_0)|$ for all $z \in D$ ($|f(z)|$ is constant on D). This contradicts to the assumption that f is not constant.

Theorem 14.52 ((Global) Maximum Modulus Principle). *Suppose that f is holomorphic and non-constant on a closed region D bounded by a simply closed curve C . Then the modulus $|f(z)|$ attains its maximum on C .*

Schwarz's lemma is an easy, but very useful, consequence of the maximum modulus principle. The most difficult part is to remember that the German analyst Herman Schwarz (actually Karl Herman Amandus Schwarz, 1843–1921), inventor of the “Schwarz inequality”, “Schwarz lemma”, and “Schwarz alternating method”, spells his name without a “t”, while the French analyst, Laurent Schwartz, 1915–????, inventor of the theory of distributions, uses a “t”.

Suppose that an holomorphic map f maps a disk of radius about a into the disk of radius about $f(a)$. We can use this to estimate the derivative $f'(a)$ and the stretching $|f(b) - f(a)|/|b - a|$. To keep the statement simple, we translate in the domain and range space so that $a = f(a) = 0$ and dilate in the domain and range so that $\epsilon = \delta = 1$. We then get:

Theorem 14.53 (Schwarz's Lemma). *Suppose f maps the open unit disk into itself leaving the origin fixed. Then $|f(z)| \leq |z|$ for all z in the disk, and if equality holds for any nonzero point, then $f(z) = cz$ for some c of modulus 1. Also $|f'(0)| \leq 1$ and if equality holds, then $f(z) = cz$ for some c of modulus 1.*

Proof: The function $g(z) = f(z)/z$ has a removable singularity at the origin, and extends analytically (Topic 5) to the disk with $g(0) = f'(0)$. It satisfies $|g(z)| \leq 1/r$ on the circle of radius $r < 1$, so by the maximum modulus theorem it satisfies this condition on the disk of radius r . Letting r tend to one gives the result. Equality implies g is constant.

14.7 Liouville's Theorem

sec:liouville

The following theorem is a consequence of the Cauchy integral formula ([e:cauchy-integral:diffn](#) (14.14)).

Theorem 14.54 (Cauchy's Inequality). *Suppose that f is holomorphic in a simply connected region D and C is a circle defined by $|z - z_0| = r$ that lies entirely in D . If $|f(z)| \leq M$ for all z on C , then*

$$|f^{(n)}(z_0)| \leq \frac{n!M}{r^n}.$$

Proof: From hypothesis, for $z \in C$,

$$\left| \frac{f(z)}{(z - z_0)^{n+1}} \right| = \frac{|f(z)|}{|z - z_0|^{n+1}} \leq \frac{M}{r^{n+1}}.$$

Thus, applying ML-inequality, we have

$$|f^{(n)}(z_0)| = \frac{n!}{2\pi} \left| \oint_C \frac{f(z)}{(z - z_0)^{n+1}} dz \right| \leq \frac{n!}{2\pi} \frac{M}{r^{n+1}} (2\pi r) = \frac{n!M}{r^n}.$$

Joseph Liouville (1809~1882) is best known for his work on the existence of a transcendental number in 1844 when he constructed an infinite class of such numbers. The following important theorem is one of his contribution to complex analysis. According to Needham [1998, 360], this means that a holomorphic mapping cannot compress the entire complex plane into a region lying inside a disc of finite radius without crushing all the way down to a point.

Theorem 14.55 (Liouville's Theorem). *If f is entire and bounded, then f is a constant function.*

Proof. Suppose f is entire and is bounded, say, $|f(z)| \leq M$ for all z . Then for any point z_0 ,

$$|f'(z_0)| \leq \frac{M}{r}$$

for arbitrarily large r . This means $f'(z_0) = 0$. So f must be constant. $\square$

The following important theorem in algebra can be proved using Liouville's Theorem.

Theorem 14.56 (Fundamental Theorem of Algebra). *If $P(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$ is a non-constant polynomial, then the equation $P(z) = 0$ has at least one root.*

Proof. The proof here is by contradiction. Suppose that $P(z)$ is not zero for $z \in \mathbb{C}$. Then $f(z) = \frac{1}{P(z)}$ is an entire function and bounded in $\mathbb{C}$.

To show that f is bounded, we write

$$P(z) = z^n \left(a_n + \frac{a_{n-1}}{z} + \cdots + \frac{a_1}{z^{n-1}} + \frac{a_0}{z^n} \right).$$

It is possible to choose a sufficiently large R such that

$$\begin{aligned} \left| \frac{a_{n-1}}{z} + \cdots + \frac{a_1}{z^{n-1}} + \frac{a_0}{z^n} \right| &\leq \frac{|a_{n-1}|}{R} + \cdots + \frac{|a_1|}{R^{n-1}} + \frac{|a_0|}{R^n} \\ &\leq \max\{|a_i|\} \frac{1 - R^{-n}}{R - 1} \leq \frac{\max\{|a_i|\}}{R - 1} \leq \frac{|a_n|}{2}. \end{aligned}$$

So when $|z| > R$, we have

$$|P(z)| \geq |z|^n \left(|a_n| - \left| \frac{a_{n-1}}{z} + \cdots + \frac{a_1}{z^{n-1}} + \frac{a_0}{z^n} \right| \right) > R^n \left(|a_n| - \frac{|a_n|}{2} \right) = \frac{R^n |a_n|}{2}.$$

This means

$$|f(z)| < \frac{1}{|P(z)|} < \frac{2}{R^n |a_n|}.$$

So $f(z)$ is bounded for $|z| > R$ and it is continuous in the closed disk $|z| \leq R$, which means f is bounded in $|z| \leq R$ as well.

Hence $f(z)$ is bounded in $\mathbb{C}$. By Liouville's Theorem, f is constant, and therefore $P(z)$ is constant. But $P(z)$ is not a constant. A contradiction. $\square$

Chapter 15

Analyticity

u:cplx:analytic

One of the amazing results in complex analysis is that the existence of the derivative of f at each point of an open set implies that f is infinitely differentiable and is locally represented by a convergent power series [Ullrich, 2008], which we will discuss in this topic. The study of power series $\sum_{n=0}^{\infty} a_n(z - z_0)^n$ and of the characteristic properties of those functions f which can be represented locally as such a power series is called the study of *analyticity*.

Karl Theodor Wilhelm Weierstrass (1815~1897) is best known for his construction of the theory of complex functions by means of power series. Much of the theory of complex sequences and series is analogous to that encountered in Calculus II. According to Weierstrass [Remmert, 1998], complex analysis (or *function theory*) is the theory of convergent Laurent series and integrals are not permitted. However, Weierstrass' approach is limited and many modern complex analysis books have adopted Riemann and Cauchy's geometrical approach.

15.1 Number and Function Sequence

sec:wwy:8.1

In this section, we will define the notion of a sequence of complex numbers and a sequence of functions are the notion of convergence as well as stating useful theorems to determine the convergence of series. The proofs can be found in Knopp [1944]. With the notion of sequence, we can rigorous define series in the next section.

Definition 15.1. A (*number*) *sequence* $\{z_n\}$ is a function whose domain in the set of positive integers and whose range is a subset of complex numbers.

Definition 15.2. We say that sequence $\{z_n\}$ *converges* to L , denoted as $\lim_{n \rightarrow \infty} z_n = L$, if given $\epsilon > 0$, there exists an integer N such that if $n > N$,

$$|z_n - L| < \epsilon.$$

If the sequence has no limit, it is said to be *divergent*.

When a sequence is convergent, it is unique as stated by the following theorem.

Theorem 15.1. *A convergent sequence has a unique limit.*

Example 15.2. Determine whether these sequences (a) $\{\frac{i^n}{n}\}$, (b) $\{e^{in\pi}\}$ converges.

Solution: (a) $\lim_{n \rightarrow \infty} \frac{i^n}{n} = 0$. Therefore, the sequence converges to 0.

(b) When n is even, $\lim_{n \rightarrow \infty} e^{in\pi} = 1$ but when n is odd, $\lim_{n \rightarrow \infty} e^{in\pi} = -1$. Therefore, this sequence diverges.

The following theorem relates “complex number sequence” to two sequences of real numbers.

thm:8.1.1

Theorem 15.3. Suppose $z_n = x_n + iy_n$ and $z = x + iy$. Then

$$\lim_{n \rightarrow \infty} z_n = z \quad \text{if and only if} \quad \lim_{n \rightarrow \infty} x_n = x \quad \text{and} \quad \lim_{n \rightarrow \infty} y_n = y.$$

Proof: The result follows immediately from the fact that

$$|z_n - z|^2 = (x_n - x)^2 + (y_n - y)^2.$$

Example 15.4. Show that $\{\frac{1}{n^2} + i\}$ converges.

Solution: Let the sequence be $\{z_n\}$. Since $x_n = \frac{1}{n^2}$, $y_n = i$ and

$$\lim_{n \rightarrow \infty} x_n = 0, \quad \lim_{n \rightarrow \infty} y_n = i.$$

Therefore, z_n has a limit.

The following theorem is often used in the proof of the convergence or divergence of a sequence.

p:main-criterion2

Theorem 15.5. An arbitrary sequence $\{x_n\}$ is convergent if and only if, given any $\epsilon > 0$, there is an integer N such that for any two integer $m, n > N$ ($m > n$), $|x_m - x_n| < \epsilon$.

Example 15.6. Show that $\{x_n = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots + \frac{(-1)^{n-1}}{2n-1}\}_{n=1}^{\infty}$ converges.

Proof: For any $\epsilon > 0$, there is an $N = \lceil \frac{1}{2\epsilon} \rceil$ such that for any $m, n > N$,

$$|x_m - x_n| = |(-1)^n| \left| \frac{1}{2n+1} - \cdots + \frac{(-1)^{m-n}}{2m+1} \right| < \frac{1}{2n+1} < \epsilon.$$

In many proofs, a variation of Theorem [15.5](#) is used.

p:main-criterion2b

Theorem 15.7. An arbitrary sequence $\{x_n\}$ is convergent if and only if, given any $\epsilon > 0$, there is an integer N such that for any integer $n > N$ and every integer $k \geq 1$, $|x_{n+k} - x_n| < \epsilon$.

thm:8.1.2

Theorem 15.8. Let f be continuous at z^* in a domain D . Then $\lim_{n \rightarrow \infty} f(z_n) = f(z^*)$ for every sequence $\{z_n\}$ converging to z^* .

Now, let us generalise the idea of a number sequence to a function sequence.

Definition 15.3. A function sequence (on domain D) $\{f_n\}$ is a function whose domain is in the set of positive integers and whose range is the set of functions over D .

Function sequence are more complicated than number sequence because the convergence will be dependent on a point in the domain D .

def:8.2.1

Definition 15.4. Suppose $f, f_1, f_2, \dots$ are defined in the same domain D and $z_0 \in D$. We say that *function sequence f_n converges (point-wise) to f at z_0* if

$$\lim_{n \rightarrow \infty} f_n(z_0) = f(z_0).$$

It is useful to have a sequence of function that convergences without depending on a point but rather a region and this is characterised by the idea of uniform convergence below.

def:8.2.1b

Definition 15.5. A function sequence f_n is said to be *convergent uniformly* to f in D if to each $\epsilon > 0$, there exist N , such that for all $z \in D$,

$$|f_n(z) - f(z)| < \epsilon.$$

Being uniformly converging has many benefits. One of them is the preservation of continuity.

Theorem 15.9. Let $\{f_n\}$ be a sequence of continuous functions on domain D . If $\{f_n\}$ converges uniformly, the limit function of f is also continuous.

15.2 Number and Function Series

In mathematics, a number sequence is usually of the form of a series defined below.

Definition 15.6. An (*infinite number*) *series* is a sequence of complex numbers (called the *partial sums*)

$$\{S_n\} := \{z_1, z_1 + z_2, \dots, z_1 + z_2 + \dots + z_n = \sum_{i=1}^n z_i, \dots\}.$$

It is normally more convenient to denote the above sequence by

$$\sum_{n=1}^{\infty} z_n = z_1 + z_2 + \dots + z_n + \dots.$$

Sometimes, our index many start from 0 or other values, such as

$$\sum_{n=0}^{\infty} z_n = z_0 + z_1 + z_2 + \dots + z_n + \dots,$$

it is always possible to rearrange the index to meet the definition above.

Definition 15.7. A series is said to *converges to the sum S* if the partial sum S_n (defined above) converges to S . We then write $\sum_{i=1}^{\infty} z_n = S$ or $\sum_{i=1}^{\infty} z_n = S$ or just $\sum z_n$.

Definition 15.8. A series $\sum_{n=1}^{\infty} z_n$ is said to be *absolutely convergent* if the series $\sum_{n=1}^{\infty} |z_n|$ is convergent.

eg:8.2.2

Example 15.10. Show that $\sum_{n=0}^{\infty} z^n$ is absolutely converging whenever $|z| < 1$ by definition.

Proof: For any $z \in \mathbb{C}$ such that $|z| < 1$ and any $\epsilon > 0$, there exists $N = \frac{\log(\epsilon|1-z|)}{\log|z|}$, for any $n > N$, we have

$$S_n = 1 + z + z^2 + \cdots + z^{n-1} = \frac{1 - z^n}{1 - z}$$

$$\Rightarrow \left| S_n - \frac{1}{1 - z} \right| = \left| \frac{z^n}{1 - z} \right| = \frac{|z|^n}{|1 - z|} < \frac{|z|^N}{|1 - z|} = \frac{\epsilon|1 - z|}{|1 - z|} = \epsilon.$$

In general, it is very inconvenient to prove the convergence of a series by definition. Mathematicians have proved some criteria for verifying whether a series diverges or converges.

Theorem 15.11 (Necessary Condition for Convergence). *If a series $\sum_{n=1}^{\infty} z_n$ is convergent, then the terms $z_n \rightarrow 0$, in other words, $\lim_{n \rightarrow \infty} z_n = 0$.*

Theorem 15.12 (Test for Divergence). *If $\lim_{n \rightarrow \infty} z_n \neq 0$, then $\sum_{n=1}^{\infty} z_n$ is divergent.*

eg: 8.2.2b

Example 15.13. Show that $\sum_{n=0}^{\infty} z^n$ is diverging when $|z| = 1$.

Proof: When $|z| = 1$, $z = e^{i\theta}$ for some θ and $z_n = e^{in\theta} \not\rightarrow 0$ as $n \rightarrow \infty$. Hence, $\sum_{n=0}^{\infty} z^n$ the series diverges when $|z| = 1$.

The following two tests are the complex versions of the ratio and root tests that are encountered in Calculus II.

Theorem 15.14 ((Cauchy's) Ratio Test). *Suppose $\sum z_n$ is a series of nonzero complex numbers such that*

$$\lim_{n \rightarrow \infty} \left| \frac{z_{n+1}}{z_n} \right| = L.$$

- When $L < 1$, then $\sum z_n$ converges absolutely.
- When $L > 1$, then $\sum z_n$ diverges.
- When $L = 1$, we cannot make any judgement, other tests are required.

Theorem 15.15 ((Cauchy's) Root Test). *Suppose $\sum z_n$ is a series such that*

$$\lim_{n \rightarrow \infty} \sqrt[n]{|z_n|} = L.$$

- When $L < 1$, then $\sum z_n$ converges absolutely.
- When $L > 1$, then $\sum z_n$ diverges.
- When $L = 1$, we cannot make any judgement, other tests are required.

A function sequence defined in the previous section is usually of the form of a series defined below.

Definition 15.9. Let f_n be a sequence of complex functions defined in domain D . A *function series* is a sequence of complex functions

$$f_1(z), f_1(z) + f_2(z), \dots, f_1(z) + f_2(z) + \dots + f_n(z) = \sum_{i=1}^n f_i(z) =: S_n(z), \dots$$

The sum $S_n(z)$ is usually called the *partial sum* of the function series and it is normally more convenient to denote the above sequence just by

$$\sum_{n=1}^{\infty} f_n(z) \quad \text{or} \quad \sum_{n=1}^{\infty} f_n(z) \quad \text{or} \quad \sum f_n(z).$$

Definition 15.10. We say that the series $\sum_{n=1}^{\infty} f_n(z)$ *converges pointwise* in D to function f if for each point z_0 , $\sum_{n=1}^{\infty} f_n(z)$ converges and the limit is $f(z_0)$.

Definition 15.11. We say that the series $\sum_{n=1}^{\infty} f_n(z)$ *converges uniformly* in D to function f if the sequence of partial sums $S_n(z)$ converges uniformly to $f(z)$ in D .

The M-test stated below is an important tool to prove the uniform convergence of a function series.

Theorem 15.16. (Weierstrass M-test or comparison test in Lang [1999, 51]) Let $\sum_{n=1}^{\infty} M_n$ be a convergent series of positive real numbers and let f_n be a sequence of complex functions all defined in a domain D . If

$$|f_n(z)| \leq M_n \quad \text{for all } z \in D$$

then the series $\sum_{n=1}^{\infty} f_n(z)$ is uniformly convergent in D and is absolutely convergent for each z in D .

The major reason for us to introduce the notion of uniform converge is because many nice properties of a function is preserved under uniform convergence.

thm:8.2.2

Theorem 15.17. Let the series $\sum_{n=1}^{\infty} f_n(z)$ be uniformly convergent to $f(z)$ in a domain D and let all the functions f_n be continuous in D . Then

1. f is also continuous in D ;
2. $\int_C f(z) dz = \sum_{n=1}^{\infty} \int_C f_n(z) dz$.

thm:8.2.2b

Theorem 15.18. Let the series $\sum_{n=1}^{\infty} f_n(z)$ be uniformly convergent to $f(z)$ in a domain D and let all the functions f_n be differentiable in D . If $\sum_{n=1}^{\infty} f'_n(z)$ is uniformly converging, then

$$f'(z) = \sum_{n=1}^{\infty} f'_n(z), \quad \text{i.e.} \quad \frac{d}{dz} \sum_{n=1}^{\infty} f_n(z) = \sum_{n=1}^{\infty} \frac{d}{dz} f_n(z).$$

eg:8.2.1

Example 15.19. Consider the geometric series $\sum_{n=0}^{\infty} z^n$. Show that the geometric series converges uniformly and absolutely for $|z| \leq K$ where $0 < K < 1$ and it is continuous on $|z| < 1$.

Proof: Since

$$|z^n| \leq M_n = K^n \quad \bigwedge \quad \sum_{n=1}^{\infty} K^n \text{ converges to } \frac{1}{1-K}.$$

By M-test, the geometric series converges uniformly and absolutely for $|z| \leq K$.

Since z^n are continuous in $|z| < 1$, by Theorem 15.17, the geometric series must be continuous for $|z| < 1$.

15.3 Power Series and Taylor Series

This section is extremely important and many Taylor series need to be memorised, otherwise, it would be extremely difficult to find the Laurent series and calculate the residue of a function at a given point in Topic 6.

First, let us define a commonly used function series below.

Definition 15.12. An infinite series of the form

$$\sum_{n=0}^{\infty} a_n (z - z_0)^n \quad (15.1)$$

where a_n are complex constants is called a *power series* in $z - z_0$. The series is said to be *centred at* z_0 .

Analogous to the concept of an interval of convergence in Calculus II, every complex power series has radius of convergence.

Definition 15.13. If there exists a real number R such that series (15.1) converges absolutely whenever $|z - z_0| < R$ and diverges whenever $|z - z_0| > R$, then R is called the *radius of convergence* of series (15.1). If the power series converges absolutely for all z , then we say that its radius of convergence is ∞ .

The circle $|z - z_0| = R$ is called the *circle of convergence* of series (15.1).

Definition 15.14. If a power series has a non-zero radius of convergence, then it is called a *convergent power series*.

If $R = 0$, then power series (15.1) converges only at a point $z = z_0$. If $0 < R < \infty$, then power series at all points interior to the circle $|z - z_0| = R$. If $R = \infty$, power series (15.1) converges for all z .

Theorem 15.20. If $\sum_{n=0}^{\infty} a_n (z - z_0)^n$ is a given power series that has a radius of convergence R , then

$$R = \lim_{n \rightarrow \infty} \left| \frac{a_n}{a_{n+1}} \right|$$

when the limit on the right exists.

Series converges (absolutely) everywhere inside its circle of convergence. Series cannot converge any point outside that circle. However, a power series may converge at some, all, or none of the points of the circle of convergence.

eg:8.3.2

Example 15.21. Consider the power series $\sum_{n=1}^{\infty} (1+3i)^n (z-i)^n$. Determine the radius of convergence and the domain where the series converges absolutely.

Solution: By root test

$$\lim_{n \rightarrow \infty} \sqrt[n]{(1+3i)^n (z-i)^n} = |1+3i||z-i| < 1.$$

The radius of convergence is $R = \frac{1}{|1+3i|} = \frac{1}{\sqrt{10}}$.
Series converges absolutely for $|z-i| < \frac{1}{\sqrt{10}}$.

eg:8.2.4

Example 15.22. Find values of z at which the series $\sum_{n=1}^{\infty} \frac{(-1)^n}{n!} (z-z_0)^n$ is convergent.

Solution. Let $z_n = \frac{(-1)^n}{n!} (z-z_0)^n$. Notice that this problem can be easily solved using ratio test than using root test:

$$\lim_{n \rightarrow \infty} \left| \frac{z_{n+1}}{z_n} \right| = \lim_{n \rightarrow \infty} \frac{|z-z_0|}{n+1} = 0.$$

Thus the series converges absolutely for all complex number z . $\square$

eg:8.2.3

Example 15.23. Find values of z at which the series $\sum_{n=0}^{\infty} \frac{n+1}{2^n} (z-3i)^n$ is convergent.

Solution. Let $z_n = \frac{n+1}{2^n} (z-3i)^n$.

By ratio test:

$$\lim_{n \rightarrow \infty} \left| \frac{z_{n+1}}{z_n} \right| = \lim_{n \rightarrow \infty} \frac{n+2}{2(n+1)} |z-3i| = \frac{1}{2} |z-3i|.$$

By root test:

$$\lim_{n \rightarrow \infty} \sqrt[n]{|z_n|} = \lim_{n \rightarrow \infty} \frac{\sqrt[n]{n+1}}{2} |z-3i| = \frac{1}{2} |z-3i|.$$

Therefore the series is **convergent** if $\frac{1}{2}|z-3i| < 1$ or $|z-3i| < 2$. The series **diverges** when $|z-3i| > 2$.

To determine the convergence at $z = 3i + 2e^{i\theta}$, $-\pi < \theta \leq \pi$ we need to study

$$\sum_{n=1}^{\infty} (n+1)e^{in\theta},$$

by the necessary condition for convergence, $(n+1)e^{in\theta}$ diverges for a fixed θ as $n \rightarrow \infty$, therefore, the series is only convergent for the domain $|z-3i| < 2$. $\square$

Remark 15.24. It is not easy to prove that $(n+1)e^{in\theta}$ diverges by using the necessary condition for convergence. The proof will be of similar nature to the prove of divergence of the sequence of $(\sin(n))_{n=1}^{\infty}$ found in <http://math.stackexchange.com/questions/238997/prove-the-divergence-of-the-sequence-sinn-n-1-infty>.

eg:8.2.5

Example 15.25. Find values of z at which the series $\sum_{n=1}^{\infty} \left(\frac{4n-1}{2n+1} \right)^n (z-z_0)^n$ converges where $z_0 \in \mathbb{C}$ is a constant.

Solution. By using root test,

$$\lim_{n \rightarrow \infty} \sqrt[n]{|z_n|} = \lim_{n \rightarrow \infty} \frac{4n-1}{2n+1} |z - z_0| = 2|z - z_0|$$

Therefore the series is convergent if $2|z - z_0| < 1$ or $|z - z_0| < \frac{1}{2}$.

When $|z - z_0| = \frac{1}{2}$, we have to analyse

$$\sum_{n=1}^{\infty} \left(\frac{4n-1}{2n+1} \right)^n \frac{1}{2^n} e^{in\theta} \quad \text{or equivalently} \quad \sum_{n=1}^{\infty} \left(\frac{4n-1}{4n+2} \right)^n e^{in\theta}$$

Note that by test of divergence, $\left(\frac{4n-1}{4n+2} \right)^n e^{in\theta}$ diverges, therefore the series is convergent if and only if $|z - z_0| < \frac{1}{2}$. $\square$

eg:8.3.1

Example 15.26. Consider the power series $\sum_{n=1}^{\infty} \frac{z^n}{n}$. Determine the radius of convergence and state whether the power series is converging on the circle of convergence.

Solution. By ratio test

$$\lim_{n \rightarrow \infty} \left| \frac{z^{n+1}/(n+1)}{z^n/n} \right| = \lim_{n \rightarrow \infty} \frac{n}{n+1} |z| = |z|.$$

The radius of convergence $R = 1$.

When $|z| = 1$, we have $z = e^{i\theta}$. If $\theta = 2k\pi$, $z = 1$ and from Calculus II, we know that

$$\sum_{n=1}^{\infty} \frac{1}{n} \text{ diverges}$$

according to Integration test. If $\theta \neq 2k\pi$

$$\sum_{n=1}^{\infty} \frac{z^n}{n} = \sum_{n=1}^{\infty} \frac{e^{in\theta}}{n} \text{ converges}$$

by applying the following Lemma [15.1](#) with $\lambda_n = \frac{1}{n} \rightarrow 0$ as $n \rightarrow \infty$. Hence, the power series converges on $\{z \in \mathbb{C} : |z| \leq 1 \wedge z \neq 1\}$. $\square$

e:ineq:trig

Lemma 15.1. Let λ_n be a decreasing sequence. Then

$$\left| \sum_{n=1}^{\infty} \lambda_n e^{in\theta} \right| \leq \frac{\lambda_1}{|\sin \frac{1}{2}\theta|}.$$

If $\lambda_n \rightarrow 0$ as $n \rightarrow \infty$, then $\sum_{n=1}^{\infty} \lambda_n e^{in\theta}$ converges when $\theta \neq 2k\pi$, $k \in \mathbb{Z}$.

thm:8.6.1

Theorem 15.27. If a power series [\(15.1\)](#) converges at z_1 , $z_1 \neq z_0$, then it is absolutely convergent at each point z in the open disk $|z - z_0| < |z_1 - z_0|$.

Proof: Without loss of generality, may assume that $z_0 = 0$. Since series $\sum_{n=0}^{\infty} a_n z_1^n$ is convergent, $a_n z_1^n$ are bounded by, say, M such that for all n ,

$$|a_n z_1^n| \leq M$$

If $|z| < |z_1|$,

$$|a_n z^n| = |a_n z_1^n| \left| \frac{z}{z_1} \right|^n \leq M \left| \frac{z}{z_1} \right|^n$$

Thus by comparison test, $\sum_{n=0}^{\infty} |a_n z^n| < \infty$.

The following results follow immediately from Theorems [thm:8.2.2](#) [thm:8.2.2b](#) [15.17](#) and [15.18](#).

[thm:8.3.1](#)

Theorem 15.28. *A power series $\sum_{n=0}^{\infty} a_n(z - z_0)^n$ with nonzero radius of convergence represents a continuous function within the circle of convergence.*

[thm:8.3.3](#)

Theorem 15.29. *Suppose the power series $f(z) = \sum_{n=0}^{\infty} a_n(z - z_0)^n$ has radius of convergence $R > 0$, then*

$$f'(z) = \sum_{n=1}^{\infty} a_n n (z - z_0)^{n-1} \quad \text{for } |z - z_0| < R.$$

A proof can be found in Lang [1999, Section II, §5].

A theorem by Brook Taylor (1685~1731) is an important results in complex analysis. It explains how we can obtain the coefficients a_n of the power series representation of a function f through differentiation.

Theorem 15.30 (Taylor's Theorem). *Suppose that f is holomorphic throughout a disc $|z - z_0| < R$, centred at z_0 and with radius R . Then f has the power series representation*

$$f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(z_0)}{n!} (z - z_0)^n \quad \text{for } |z - z_0| < R. \quad (15.2)$$

[e:taylor](#)

That is, the series [\(15.2\)](#) converges to $f(z)$ when $|z - z_0| < R$.

Proof: Let z be a point within the circle $|z| < R$. Let C_0 denote the circle $|z| = r_0$, where $r < r_0 < R$. Then

$$\begin{aligned} f(z) &= \frac{1}{2\pi i} \oint_{C_0} \frac{f(\zeta)}{\zeta - z} d\zeta = \frac{1}{2\pi i} \oint_{C_0} \frac{f(\zeta)}{\zeta} \frac{1}{1 - \frac{z}{\zeta}} d\zeta \\ &= \frac{1}{2\pi i} \oint_{C_0} \frac{f(\zeta)}{\zeta} \left\{ \sum_{n=0}^{N-1} \left(\frac{z}{\zeta}\right)^n + \frac{\left(\frac{z}{\zeta}\right)^N}{1 - \frac{z}{\zeta}} \right\} d\zeta \\ &= \sum_{n=0}^{N-1} \frac{f^{(n)}(0)}{n!} z^n + \frac{z^N}{2\pi i} \oint_{C_0} \frac{f(\zeta)}{\zeta^N(\zeta - z)} d\zeta. \end{aligned}$$

Now if M denotes the maximum value of $|f(z)|$ on C_0 ,

$$\left| \frac{z^N}{2\pi i} \oint_{C_0} \frac{f(\zeta)}{\zeta^N(\zeta - z)} d\zeta \right| \leq \frac{r^N}{2\pi} \frac{M}{(r_0 - r)r_0^N} (2\pi r_0) = \frac{Mr_0}{r_0 - r} \left(\frac{r}{r_0}\right)^N \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

To verify formula [\(15.2\)](#), [e:taylor](#) define $g(z) = f(z + z_0)$, then by the above argument, g can be

represented as

$$g(z) = \sum_{n=0}^{\infty} \frac{g^{(n)}(0)}{n!} z^n \quad (15.3)$$

e:8.3.3

Replace z by $z - z_0$ in (15.3), we obtain the Taylor series expansion (15.2) of f .

Definition 15.15. The function series (15.2) is known as the *Taylor series (expansion / representation) of f about $z = z_0$* . When $z_0 = 0$, the series (15.2) becomes

$$f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} z^n \quad (15.4)$$

e:maclaurin

and this is called the *Maclaurin series (expansion / representation) of f* .

Theorem 15.31. *The Taylor and Maclaurin series representations are unique.*

eg:8.4.4

Example 15.32. Find the Maclaurin series for $\frac{1}{z+1}$.

Solution: In Example 15.10, we have $\frac{1}{1-z} = \sum_{n=0}^{\infty} z^n$ (defining $z^0 = 1$ for all $|z| < 1$).

Therefore,

$$\frac{1}{z+1} = \frac{1}{1-(-z)} = \sum_{n=0}^{\infty} (-z)^n = \sum_{n=0}^{\infty} (-1)^n z^n, \quad |z| < 1.$$

eg:8.4.1

Example 15.33. Find the Maclaurin series representation for the exponential function $z \mapsto e^z$ and the function $z \mapsto z^2 e^{3z}$.

Solution: Let $f(z) = e^z$, then $f^{(n)}(z) = e^z$ and hence $f^{(n)}(0) = 1$. It follows that the Maclaurin series representation (15.4) of e^z is

$$e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}, \quad z \in \mathbb{C}.$$

$$z^2 e^{3z} = z^2 \sum_{n=0}^{\infty} \frac{(3z)^n}{n!} = \sum_{n=0}^{\infty} \frac{3^n}{n!} z^{n+2}, \quad z \in \mathbb{C}.$$

eg:8.4.2

Example 15.34. Find the Maclaurin series of $\sin z$, $\cos z$, $\sinh z$ and $\cosh z$.

Solution. For all z , $e^{iz} = \sum_{n=0}^{\infty} \frac{(iz)^n}{n!}$, $e^{-iz} = \sum_{n=0}^{\infty} \frac{(-iz)^n}{n!}$, therefore

$$\sin z = \frac{e^{iz} - e^{-iz}}{2i} = \frac{1}{2i} \sum_{n=0}^{\infty} \frac{(iz)^{2n+1} - (-iz)^{2n+1}}{(2n+1)!} = \frac{1}{2i} \sum_{n=0}^{\infty} \frac{2i(-1)^n}{(2n+1)!} z^{2n+1} = \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{(2n+1)!}$$

$$\cos z = \frac{e^{iz} + e^{-iz}}{2} = \frac{1}{2} \sum_{n=0}^{\infty} \frac{2(iz)^{2n}}{(2n)!} = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} z^{2n}$$

$$\sinh z = -i \sin(iz) = -i \sum_{n=0}^{\infty} \frac{(-1)^n (iz)^{2n+1}}{(2n+1)!} = -i \sum_{n=0}^{\infty} \frac{(-1)^n \cdot (-1)^n \cdot iz^{2n+1}}{(2n+1)!} = \sum_{n=0}^{\infty} \frac{z^{2n+1}}{(2n+1)!}$$

$$\cosh z = \cos(iz) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} (iz)^{2n} = \sum_{n=0}^{\infty} \frac{(-1)^n \cdot (-1)^n}{(2n)!} z^{2n} = \sum_{n=0}^{\infty} \frac{z^{2n}}{(2n)!}$$

□

Example 15.35. Find the coefficients of the Maclaurin series for $\sec z$ up to order 4.

Solution.

$$\begin{aligned}\sec z &= \frac{1}{\cos z} = \frac{1}{1 - \left(\frac{z^2}{2!} - \frac{z^4}{4!} + \cdots\right)} = 1 + \left(\frac{z^2}{2!} - \frac{z^4}{4!} + \cdots\right) + \left(\frac{z^2}{2!} - \frac{z^4}{4!} + \cdots\right)^2 + \cdots \\ &= 1 + \frac{z^2}{2!} - \frac{z^4}{4!} + O(z^6) + \frac{z^4}{4} + O(z^6) + O(z^6) = 1 + \frac{1}{2}z^2 + \frac{5}{24}z^4 + O(z^6).\end{aligned}$$

□

eg:8.8.2

Example 15.36. Find the Maclaurin series representation of (a) $\frac{1}{(1-z)^2}$ and (b) $\frac{z^3}{(1-z)^2}$.

Solution. For $|z| < 1$, we have the Maclaurin series

$$\frac{1}{1-z} = \sum_{n=0}^{\infty} z^n, \quad |z| < 1.$$

By Theorem [15.29](#), differentiating the above series with respect to z gives

$$(a) \quad \frac{-1}{(1-z)^2}(-1) = \sum_{n=1}^{\infty} n z^{n-1}.$$

Multiplying (a) by z^3 , gives

$$(b) \quad \frac{z^3}{(1-z)^2} = \sum_{n=1}^{\infty} n z^{n+2}.$$

□

eg:8.4.5

Example 15.37. Find the Taylor series for $\frac{1}{1-z}$ about $z = 2i$.

Solution: It is possible to the Taylor series using derivative based on [\(15.2\)](#) but it is far more convenient to find it using

$$\begin{aligned}\frac{1}{1-z} &= \frac{1}{1-2i-(z-2i)} = \frac{1}{1-2i} \frac{1}{1-\frac{z-2i}{1-2i}} \\ &= \frac{1}{1-2i} \sum_{n=0}^{\infty} \left(\frac{z-2i}{1-2i}\right)^n = \sum_{n=0}^{\infty} (1-2i)^{-n-1} (z-2i)^n\end{aligned}$$

for $|z-2i| < |1-2i| = \sqrt{5}$.

eg:8.4.6

Example 15.38. Expand $f(z) = (z-1)e^{-2z}$ about $z = 1$.

Solution.

$$\begin{aligned}f(z) &= (z-1)e^{-2(z-1+1)} = e^{-2}(z-1)e^{-2(z-1)} \\ &= e^{-2}(z-1) \sum_{n=0}^{\infty} \frac{(-2(z-1))^n}{n!} = \sum_{n=0}^{\infty} \frac{(-2)^n}{n!e^2} (z-1)^{n+1}\end{aligned}$$

for all z .

□

There is no Maclaurin series for f in the example below since f is not holomorphic at $z = 0$. However, function f still can be repressed as a “power series” involving negative powers of z and this is the idea of Laurent series that we going to study in the next topic.

eg:8.4.7

Example 15.39. Expand $f(z) = \frac{1}{z^3 + z^5}$ into a series involving *powers* of z .

Solution.

$$f(z) = \frac{1}{z^3} \frac{1}{1+z^2} = \frac{1}{z^3} \sum_{n=0}^{\infty} (-z^2)^n = \sum_{n=0}^{\infty} (-1)^n z^{2n-3}.$$

□

15.4 Holomorphicity and Analyticity

sec:analytic

What makes complex analysis so special compare to Calculus is that in Calculus, a function may be smooth, i.e. infinite differentiable in a domain, but it may not be analytic, i.e. it may not be possible for us to represent the real-valued smooth function locally at a point with power series. However, in this section, we will see that holomorphicity (complex differentiability) implies analyticity!

Definition 15.16. We say that a function f is *analytic at z_0* if f is **locally** given by a convergent power series, i.e. there exists a power series (15.1) and some $\epsilon > 0$ such that the series converges absolutely for $|z - z_0| < \epsilon$ to $f(z)$.

Definition 15.17. f is called an *analytic function* on an open set D if f is analytic at every point of D .

The following two theorems are two basic properties of analytic functions.

Theorem 15.40. *The sums, products, and compositions of analytic functions are analytic.*

Theorem 15.41. *The reciprocal of an analytic function that is nowhere zero is analytic, as is the inverse of an invertible analytic function whose derivative is nowhere zero.*

thm:8.3.2

Theorem 15.42. *A power series can be integrated term-by-term within the circle of convergence: that is, if $R \neq 0$ and*

$$f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n, \quad |z - z_0| < R,$$

then for every path C with initial point z_1 and end point z_2 lying inside the circle of convergence,

$$\int_C f(z) dz = \sum_{n=0}^{\infty} a_n \int_{z_1}^{z_2} (z - z_0)^n dz = \sum_{n=0}^{\infty} a_n \frac{(z - z_0)^{n+1}}{n+1} \Big|_{z=z_1}^{z=z_2}.$$

Proof: By Theorem 15.17, we conclude that term-by-term integration along C is permissible. Since each term $a_n(z - z_0)^n$ is holomorphic, the integrals are independent of path, so that $\int_{z_1}^{z_2} f(z) dz$ is independent of path. Hence

$$F(z) = \int_{z_0}^z f(z) dz$$

is well-defined for $|z - z_0| < R$. According to Morera's Theorem,

$$\frac{d}{dz} F(z) = f(z)$$

so that F is analytic.

One of the main results in complex analysis is that all analytic functions f defined on a region $D \subset \mathbb{C}$ can be represented as a convergent power series $\sum_{n=0}^{\infty} a_n(z - z_0)^n$, where z_0 is the centre of the largest disc $S \subset D$, and $z \in S$. Moreover, not only the convergent power series but all formal power series can be uniquely represented by equivalence classes of analytic functions.

Theorem 15.43 (Holomorphicity-Analyticity Theorem). *f is holomorphic in domain D if and only if f is analytic in domain D .*

Proof by Alonso Delfín (CINVESTAV IPN mathematics department, <http://www.math.cinvestav.mx/en/home>).

Let $f : D \rightarrow \mathbb{C}$ be a holomorphic function, i.e. f is differentiable everywhere in D and satisfies Cauchy-Riemann equations. Also let $a \in D$ and $R > 0$ such that

$$\Delta_R(a) = \{z \in D : |z - a| < R\} \subset D,$$

then f is holomorphic in $\Delta_R(a)$. Now take any r such that $0 < r < R$ and define the path $\gamma(t) = a + re^{it}$ for $t \in [0, 2\pi]$. If $\zeta \in \partial\Delta_r(a) = \gamma([0, 2\pi])$ and $z \in \Delta_r(a)$ then clearly

$$\left| \frac{z - a}{\zeta - a} \right| < 1$$

which gives that

$$\frac{1}{\zeta - z} = \frac{1}{\zeta - a} \left(\frac{1}{1 - \frac{z-a}{\zeta-a}} \right) = \frac{1}{\zeta - a} \sum_{n=0}^{\infty} \left(\frac{z-a}{\zeta-a} \right)^n = \sum_{n=0}^{\infty} \frac{(z-a)^n}{(\zeta-a)^{n+1}} \quad (15.5)$$

Furthermore, if $M = \max\{|f(\zeta)| : |\zeta - a| = r\}$, then since $|z - a|/r < 1$ and

$$|f(\zeta)| \times \frac{|z - a|^n}{|\zeta - a|^{n+1}} \leq \frac{M}{r} \left(\frac{|z - a|}{r} \right)^n$$

the Weierstrass M-test gives that the series (15.5) times $f(\zeta)$ converges uniformly to $f(\zeta)/(\zeta - z)$. That is

$$\frac{f(\zeta)}{\zeta - z} = \lim_{k \rightarrow \infty} f(\zeta) \sum_{n=0}^k \frac{(z-a)^n}{(\zeta-a)^{n+1}} \quad (15.6)$$

where the convergence is uniform.

Now, by the Cauchy Integral Formula on a Disk (Section 4.5.1 from Topic 4) we have that for all $z \in \Delta_r(a)$,

$$\begin{aligned} f(z) &= \frac{1}{2\pi i} \int_{\gamma} \frac{f(\zeta)}{\zeta - z} d\zeta \quad (\text{Section 4.5.1}) \\ &= \frac{1}{2\pi i} \int_{\gamma} \lim_{k \rightarrow \infty} f(\zeta) \sum_{n=0}^k \frac{(z-a)^n}{(\zeta-a)^{n+1}} d\zeta \quad (\text{by Equation (15.6)}) \\ &= \lim_{k \rightarrow \infty} \sum_{n=0}^k \frac{1}{2\pi i} \int_{\gamma} f(\zeta) \frac{(z-a)^n}{(\zeta-a)^{n+1}} d\zeta \quad (\text{since the series converges uniformly}) \\ &= \sum_{n=0}^{\infty} \left(\frac{1}{2\pi i} \int_{\gamma} \frac{f(\zeta)}{(\zeta-a)^{n+1}} d\zeta \right) (z-a)^n =: \sum_{n=0}^{\infty} a_n(\gamma)(z-a)^n \end{aligned}$$

Moreover, since $f(z) = \sum_{n=0}^{\infty} a_n(z-a)^n$ then $a_n = f^{(n)}(a)/n!$, then $a_n(\gamma) = a_n = f^{(n)}(a)/n!$, the coefficients don't depend on γ , thus they are independent of r . Then we now have that indeed f has a power expansion of the form $f(z) = \sum_{n=0}^{\infty} a_n(z-a)^n$, and this gives that f is analytic.

If f is analytic on D , then for each $z \in D$, there is a convergence power series at a ϵ -neighbourhood of z and by Theorem 15.29, we know that the function is holomorphic on the ϵ -neighbourhood. Hence, f is holomorphic on D .

In order to understand the importance of the Holomorphicity-Analyticity Theorem, we will analyse the following example.

Example 15.44 (Ablowitz and Fokas [2003]). Prove that the error function $\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$ is entire.

Proof: By using the Maclaurin series of e^t , we obtain

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \sum_{n=0}^{\infty} \frac{(-t^2)^n}{n!} dt = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \int_0^z \frac{t^{2n}}{n!} dt = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{z^{2n+1}}{(2n+1) \cdot n!}, \quad z \in \mathbb{C}.$$

The radius of convergence remains ∞ according to

$$\lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n} = \lim_{n \rightarrow \infty} \frac{(2n+1) \cdot n!}{(2(n+1)+1) \cdot (n+1)!} = \lim_{n \rightarrow \infty} \frac{2n+1}{(2n+3)(n+1)} = 0.$$

So, erf is analytic on $\mathbb{C}$ and by the Holomorphicity-Analyticity Theorem, erf is entire.

15.5 Isolated Zeroes Principle

A very important nice property of analytic function is that its zeros are isolated.

Theorem 15.45 (Zeroes of Analytic Function are Isolated). *Let $D \subset \mathbb{C}$ be some open set and let $f : D \rightarrow \mathbb{C}$ be analytic. Then either f is a constant function, or the set $\{z \in D : f(z) = 0\}$ is totally disconnected.*

ProofWiki:

Suppose f has no zeroes in D . Then the set described in the theorem is the empty set, and we're done.

So we suppose there is $z_0 \in D$ such that $f(z_0) = 0$.

Since f is analytic, there is a Taylor series for f at z_0 which converges for $|z - z_0| < R$. Since $f(z_0) = 0$, we know $a_0 = 0$. Other a_j may be 0 as well. So let k be the least number such that $a_j = 0$ for $0 \leq j < k$, and $a_k \neq 0$. Then we can write the Taylor series for f about z_0 as:

$$\sum_{n=k}^{\infty} a_n(z - z_0)^n = (z - z_0)^k \underbrace{\sum_{n=0}^{\infty} a_{n+k}(z - z_0)^n}_{g(z)}.$$

Note that g is clearly analytic in $|z - z_0| < R$.

Since g is analytic, it is also holomorphic by Holomorphicity-Analyticity Theorem, it is also continuous. Since $g(z_0) = a_k \neq 0$, there is an $\delta > 0$ so that for every $z \in \mathbb{C}$,

$$|z - z_0| < \delta \Rightarrow |g(z) - a_k| < \frac{|a_k|}{2}.$$

But then $g(z)$ can't possibly be 0 in that disk. Hence the result.

15.6 Analytic Continuation

Analytic continuation is a technique to extend the domain of a function to a larger domain. For complex analysis, it is possible to have a beautiful result — the analytic continuation is unique.

The concept of a universal cover was first developed to define a natural domain for the analytic continuation of an analytic function. The idea of finding the maximal analytic continuation of a function in turn led to the development of the idea of *Riemann surfaces*.

The power series in the previous section is generalised by the idea of a *germ*. The general theory of analytic continuation and its generalisations are known as *sheaf theory*.

Complex analysis can be compared to Calculus of one-variable complex functions. Several complex variables, Calculus of multi-variable complex functions are far more complicated where carrying out analytic continuation since the singularities need not be isolated points, and its investigation was a major reason for the development of *sheaf cohomology*.

Definition 15.18. Suppose f is an analytic function defined on a non-empty open subset $D \subset \mathbb{C}$. If G is a larger open subset of $\mathbb{C}$ such that $D \subset G$, and F is an analytic function defined on G such that

$$F(z) = f(z) \quad \forall z \in D,$$

then F is called an *analytic continuation* of f . In other words, the restriction of F to D is the function f we started with.

Example 15.46. Show that $F(z) = \frac{1}{1-z}$ is an analytic continuation of $f(z) = \sum_{n=0}^{\infty} z^n$.

Proof: Since f is defined only on $D = \{z \in \mathbb{C} : |z| < 1\}$ according to Example 15.10, while F is defined on $\mathbb{C} \setminus \{1\}$ and $F(z) = f(z)$ for $|z| < 1$, F is an analytic continuation of f . leg:8.2.2

A holomorphic function is completely determined by its values on a small neighbourhood in D . This is not true for differentiable functions in Calculus. In comparison, complex-differentiability, is a much more rigid notion.

Theorem 15.47 (Identity Theorem). *Given holomorphic functions f and g on a connected open set D , if $f = g$ on some open non-empty subset of D or along an arc interior to D , then $f = g$ on D .*

Proof (Wikipedia):

The connectedness (Topic 1) assumption on the domain D is necessary and is in fact key to a short proof given here. Under this assumption, since we are given that the set is not empty, topologically the claim amounts to that f and g coincide on a set that is both open and closed.

The closedness is immediate from the continuity of f and g .

Therefore, the main issue is to show that the set on which $f = g$ is an open set.

Because a holomorphic function can be represented by its Taylor series everywhere on its domain since it is analytic, it is sufficient to consider the set

$$S = \{z \in D \mid f^{(k)}(z) = g^{(k)}(z) \text{ for all } k \geq 0\}.$$

Suppose w lies in S . Then, because the Taylor series of f and g at w have non-zero radius of convergence, the open disk $\Delta_r(w)$ also lies in S for some r . In fact, r can be anything less than the distance from w to the boundary of D . This shows S is open and proves the theorem.

With the identity theorem, is it possible for us to establish the uniqueness theorem for analytic continuations.

Theorem 15.48 (Analytic Continuation Theorem). *If G is the connected domain of two analytic functions F_1 and F_2 such that $D \subset G$ and for all z in D*

$$F_1(z) = F_2(z) = f(z),$$

then

$$F_1 = F_2 \quad \text{on all of } G.$$

This is because $F_1 - F_2$ is an analytic function which vanishes on the open, connected domain D of f and hence must vanish on its entire domain according to the identity theorem.

A common way to define functions in complex analysis proceeds by first specifying the function on a “small domain” only, and then extending it by analytic continuation. There are various ways one might show that a holomorphic function can be extended to a larger domain. The *Schwarz Reflection Principle* is an important example: If an open set is symmetric about the real axis, then every function holomorphic in the upper half complex plane can be extended by mirror reflection. Sometimes, one can show that a function satisfies a “functional equation” and then use that to do the extending. Gamma function

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt$$

and *Riemann zeta function* [Riemann, 1859] are two famous examples of the analytic continuation of

$$\zeta(z) = \sum_{n=1}^{\infty} \frac{1}{n^z}.$$

By using the M -test, it is easy to see that $\zeta(z)$ converges for $\Re(z) > 1$. By the analytic continuation theorem, it is possible to extend the domain of $\Gamma(z)$ and $\zeta(z)$ to $\mathbb{C} \setminus \{0, -1, -2, \dots\}$ and $\mathbb{C} \setminus \{1\}$ respectively.

So it is always possible for us to obtain an analytic continuation from a power series f ? It is possible that f cannot be continued beyond the disk of definition if the boundary is full of non-isolated singularities. In this case, we say that the circumference is the natural boundary of the function.

Definition 15.19. For a power series

$$f(z) = \sum_{k=0}^{\infty} a_k z^{n_k}$$

with

$$\liminf_{k \rightarrow \infty} \frac{n_{k+1}}{n_k} > 1$$

the circle of convergence is a *natural boundary* or *natural barrier*. Such a power series is called *lacunary*.

A famous example is stated below.

Example 15.49. Show that $f(z) = \sum_{n=0}^{\infty} z^{2^n}$ has a natural boundary $|z| = 1$.

Proof: By ratio test, it is easy to see that f converges for $|z| < 1$. Hence, $|z| = 1$ is a radius of convergence. To show that it is a natural boundary, we study

$$f(z^2) = \sum_{n=0}^{\infty} (z^2)^{2n} = \sum_{n=0}^{\infty} z^{2^{n+1}} = f(z) - z.$$

By induction $f(z^{2^m}) = f(z) - \sum_{j=0}^{m-1} z^{2^j}$. Hence, the value of the function f at all points z_m on the unit circle satisfying the equation $z^{2^m} = 1$ is infinite:

$$f(z_m) = \sum_{n=0}^{\infty} z_m^{2^n} = \sum_{j=0}^{m-1} z_m^{2^j} + \sum_{n=0}^{\infty} z_m^{2^{m+n}} = \sum_{j=0}^{m-1} z_m^{2^j} + \sum_{n=0}^{\infty} 1 = \infty.$$

No matter how small an arc we take on the unit circle, there exist points z_m on the arc such that $f(z_m) = \infty$. Therefore, analytic continuation for f is impossible beyond $|z| = 1$.

The theoretical proof of analytic continuation theorem is complicated. We will just state an theorem which characterises how the largest domain of extension.

Theorem 15.50 (Monodromy Theorem). *Let D be a simply connected domain and f be analytic in some disk $D_0 \subset D$. If f can be analytically continued along any two distinct smooth contours C_1 and C_2 to a point in D , and if there are no singular points enclosed within C_1 and C_2 then the result of each analytic continuation is the same and the function is single-valued.*

This theorem can be extended to the case where the region enclosed by contours C_1 and C_2 contains at most, isolated singular points.

Chapter 16

Residues and Poles

ch:cplx:res

In Chapters [14](#) and [15](#), we have studied holomorphic functions. In this topic, we will study “nearly” holomorphic functions, i.e. functions having some singularities but are holomorphic when the singularities are removed. In Calculus, singularities are considered bad. But the singularities behaved quite well as in complex analysis as we will see in this topic.

Cauchy theorem ([Chapter 14](#)) states that if a function is holomorphic interior to and on a simple closed curve C , then the contour integral of the function around the curve is zero. However, if the function fails to be holomorphic at finite number of point interior to C , then the value of the contour integral along C is no longer zero. There is a specific number, related to the singularity of a function, called the *residue* at the singular point, contributes to the value of the integral and this topic discusses the techniques to find and to apply residue in calculations.

16.1 Singularities of a Function

We have defined the notion of singularity in Topic 2 Definition 2.3.14. A singular point of the complex function f is a point at which f is not holomorphic.

There are two types of singular points, the isolated and non-isolated singularities defined below.

Definition 16.1. A singular point z_0 (of a complex function f) is said to be *isolated* if there is a deleted neighbourhood (Topic 1) of z_0 throughout which f is holomorphic. Otherwise, z_0 is *non-isolated*.

Isolated singular points are considered “well-behaved” while non-isolated singular points normally “behave badly” from an analytical point of view.

Example 16.1. The function $z \mapsto \tan(1/z)$ has a **non-isolated singularity** at $z = 0$.

Example 16.2. Find the isolated singularity for each of the following function.

eg:9.2.1a

$$1. f(z) = \begin{cases} \sin z/z, & z \neq 0, \\ 0, & z = 0 \end{cases} \text{ has an isolated singularity at } z = 0.$$

eg:9.2.1b

$$2. g(z) = \frac{z^3}{(z-1)^2} \text{ has an isolated singularity at } z = 1.$$

eg:9.2.1c

$$3. h(z) = e^{1/z} \text{ has an isolated singularity at } z = 0.$$

In the following definitions, we shall assume f to have an isolated singularity at $z = z_0$.

Definition 16.2. If there is a function $g(z)$, holomorphic at $z = z_0$ and such that $f(z) = g(z)$ for all z in some deleted neighbourhood of z_0 , we say f has a *removable singularity* at z_0 . This means that if the value of f is “corrected” at the point z_0 , it becomes holomorphic there.

def:zero

Definition 16.3. A number z_0 is a *zero* of f if $f(z_0) = 0$. A zero z_0 of f is *isolated* if there is an open disk about z_0 containing no other zeros of f .

def:pole-order-n

Definition 16.4. If, for $z \neq z_0$, the function f can be written in the form $f(z) = P(z)/Q(z)$ where $P(z)$ and $Q(z)$ are holomorphic at z_0 , $P(z_0) \neq 0$, and $Q(z_0) = 0$, we say f has a *pole* at z_0 . If $Q(z)$ has a zero of order n , we say that f has a *pole of order n* .

A function with a pole of order n at $z = z_0$ is usually of the form $\frac{g(z)}{(z - z_0)^n}$ with g holomorphic at $z = z_0$ and $g(z_0) \neq 0$. Pole of order 1 is called a *simple pole*.

Definition 16.5. If f has neither a removable singularity nor a pole at z_0 , we say f has an *essential singularity* at z_0 .

Example 16.3. $f(z) = \begin{cases} \sin z/z, & z \neq 0, \\ 0, & z = 0 \end{cases}$ is not continuous at $z = 0$. However, if we redefine $f(z)$ at 0 as $f(0) = 1$, then f will be continuous at 0 as observe from the power series of f at 0:

$$\frac{\sin z}{z} = \frac{1}{z} \left(z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \cdots \right) = 1 - \frac{z^2}{3!} + \frac{z^4}{5!} - \frac{z^6}{7!} + \cdots.$$

So function f will be holomorphic after redefinition and 0 is a **removable singularity** of f .

Example 16.4. $g(z) = \frac{z^3}{(z-1)^2} = \frac{P(z)}{Q(z)}$. $Q(z) = (z-1)^2$ has a zero of order 2 at $z = 1$, and $P(1) = 1 \neq 0$. Thus $g(z)$ has a **pole of order 2** at $z = 1$.

eg:9.2.1c2

Example 16.5. The function $f(x) = e^{1/x}$ is holomorphic at every $z \neq 0$. Let us restrict z to be real numbers such that $e^{1/x}$. Then

$$\lim_{x \rightarrow 0^+} e^{1/x} = \infty, \quad \lim_{x \rightarrow 0^-} e^{1/x} = 0$$

so that the singularity is not removable. Furthermore, for any positive integer n ,

$$\lim_{x \rightarrow 0^+} x^n e^{1/x} = \infty, \quad \lim_{x \rightarrow 0^-} x^n e^{1/x} = 0$$

so that the singularity is not a pole of order n . Hence f has an essential singularity at $z = 0$.

Example 16.6 (April 2008 Exam, Q3(b)). Identify and classify the singularities of the following functions:

- (i) $\frac{\sin z}{z - \pi}$. (2 mark)

Solution. Since $\lim_{z \rightarrow \pi} \frac{\sin z}{z - \pi} = 1$, π is a removable singularity. □

- (ii) $e^{\frac{1}{z-1}} \log(1+z)$; (2 mark)

Solution. $z = 1$ is an essential singular point because $\ln(1+z) = 2$ when $z = 1$ and $e^{\frac{1}{z-1}}$ is similar to Example 16.5. □
 $z = -1$ is not an isolated singular point of $\ln(1+z)$ since it is **not well defined** in the deleted ϵ neighbourhood Churchill and Brown [2008, 230].

- (iii) $\frac{\sqrt{z}}{z^2 + 1}$ (2 mark)

Solution. $z = \pm i$ are simple poles (regular singular points). $z = 0$ is a singular branch point. □

(iv) $\frac{1 - \cos z}{z^4}$ (2 mark)

Solution. $z = 0$ is a pole of order 2 (regular singular point) because $1 - \cos z \approx z^2/2$. □

16.2 Laurent Series

The Laurent series was named after and first published by Pierre Alphonse Laurent in 1843. Karl Weierstrass discovered it first in 1841 but did not publish it. Pierre Alphonse Laurent (18 July, 1813 ~ 2 September, 1854) was a French mathematician best known as the discoverer of the Laurent series, an expansion of a function into an infinite power series, generalising the Taylor series expansion. He was born in Paris, France. His result was contained in a memoir submitted for the Grand Prize of the Académie des Sciences in 1843, but his submission was after the due date, and the paper was not published and never considered for the prize. Laurent died at age 41 in Paris. His work was not published until after his death.

If a function fails to be holomorphic at a point z_0 , we cannot represent the function by means of Taylor series (Topic 5). Laurent discovered that a function can still be represented by series

$$f(z) = \sum_{n=-\infty}^{\infty} c_n (z - z_0)^n \quad (16.1) \quad \text{e:laurent}$$

involving both positive and negative powers of $z - z_0$ if the singularity is isolated at z_0 .

Definition 16.6 (Principal and Analytic Parts). In the Laurent series (16.1), the sum of all the negative power terms

$$\sum_{n=1}^{\infty} c_{-n} (z - z_0)^{-n}$$

is called the *principal part* of the series and the sum of all the non-negative power terms

$$\sum_{n=0}^{\infty} c_n (z - z_0)^n$$

is called the *analytic part* of the series.

The following theorem tells us the existence of a Laurent series at a point.

Theorem 16.7 (Laurent's Theorem). Suppose $0 \leq r < R$, and that f is holomorphic in the annulus $r < |z - z_0| < R$. If C is any positive oriented simple closed contour around z_0 and lying in the annulus, then, at each point z in the annulus, f has the Laurent series representation (16.1) in which

$$c_n = \frac{1}{2\pi i} \oint_C \frac{f(\zeta)}{(\zeta - z_0)^{n+1}} d\zeta, \quad n = 0, \pm 1, \pm 2, \dots$$

Proof: Let C_1 and C_2 be concentric circles with centre a and radii r_1 and R_2 , where $r < r_1 < R_2 < R$. Let z be a fixed point in D such that $r_1 < |z - z_0| < R_2$. By introducing a crosscut between C_1 and C_2 , it follows from Cauchy's integral formula that

$$f(z) = \frac{1}{2\pi i} \oint_{C_2} \frac{f(\zeta)}{\zeta - z} d\zeta - \frac{1}{2\pi i} \oint_{C_1} \frac{f(\zeta)}{\zeta - z} d\zeta =: I_2 + I_1.$$

Since f is holomorphic in the annulus bounded between C_1 and C_2 , We can write

$$\frac{1}{2\pi i} \oint_{C_2} \frac{f(\zeta)}{\zeta - z} d\zeta = \sum_{n=0}^{\infty} c_n (z - z_0)^n$$

where

$$c_n = \frac{1}{2\pi i} \oint_C \frac{f(\zeta)}{(\zeta - z_0)^{n+1}} d\zeta, \quad n = 0, 1, 2, \dots$$

As for I_1 ,

$$\begin{aligned} I_1 &= -\frac{1}{2\pi i} \oint_{C_1} \frac{f(\zeta)}{\zeta - z} d\zeta = \frac{1}{2\pi i} \oint_{C_1} \frac{f(\zeta)}{z - z_0 - (\zeta - z_0)} d\zeta = \frac{1}{2\pi i} \oint_{C_1} \frac{f(\zeta)}{z - z_0} \frac{1}{1 - \frac{\zeta - z_0}{z - z_0}} d\zeta \\ &= \frac{1}{2\pi i} \oint_{C_1} \frac{f(\zeta)}{z - z_0} \left[\sum_{k=0}^{N-1} \left(\frac{\zeta - z_0}{z - z_0} \right)^k + \frac{(\zeta - z_0)^N}{(z - \zeta)(z - z_0)^{N-1}} \right] d\zeta = \sum_{k=1}^N \frac{c_{-k}}{(z - z_0)^k} + R_N(z) \end{aligned}$$

where

$$\begin{aligned} c_{-k} &= \frac{1}{2\pi i} \oint_{C_1} \frac{f(\zeta)}{(\zeta - z_0)^{-k+1}} d\zeta, \quad k = 1, 2, \dots; \\ R_N(z) &= \frac{1}{2\pi i (z - z_0)^N} \oint_{C_1} \frac{f(\zeta)(\zeta - z_0)^N}{z - \zeta} d\zeta. \end{aligned}$$

Now let d denote the distance from z to z_0 , i.e., $|z - z_0| = d$ and let M denote the maximum value of $|f(\zeta)|$ on C_1 .

Notice that $|\zeta - z_0| = r_1$ and

$$|z - \zeta| = |z - z_0 + z_0 - \zeta| \geq |z - z_0| - |z_0 - \zeta| = d - r_1,$$

we have

$$|R_N(z)| = \frac{1}{2\pi |z - z_0|^N} \left| \oint_{C_1} \frac{f(\zeta)(\zeta - z_0)^N}{z - \zeta} d\zeta \right| \leq \frac{1}{2\pi d^N} \frac{Mr_1^N}{d - r_1} (2\pi r_1) = \frac{Mr_1}{d - r_1} \left(\frac{r_1}{d} \right)^n \rightarrow 0$$

as $n \rightarrow \infty$. This shows that

$$I_1 = \sum_{n=1}^{\infty} \frac{c_{-n}}{(z - z_0)^n}.$$

In order for a Laurent series to be meaningful, we need it to be convergent.

Definition 16.7. We say that the Laurent series *converges absolutely* on domain A if the principal part and the analytic part converges absolutely.

Theorem 16.8. The Laurent series of a function f that is holomorphic in an annulus $R_1 < |z - z_0| < R_2$ converges uniformly to f for $\rho_1 \leq |z - z_0| \leq \rho_2$, where $R_1 < \rho_1$ and $R_2 > \rho_2$.

This theorem implies that Laurent series may be integrated and differentiated termwise. Elementary operations such as addition, subtraction, multiplication of Laurent series behave just like its Taylor series counterpart in Topic 5.

We normally do not have to go through the methods mentioned in the proof in order to obtain the Laurent series. By applying the Taylor series from Topic 5 appropriately, we will obtain Laurent series for a given holomorphic function around an isolated singularity as demonstrated by the following examples.

Example 16.9. Find the Laurent series representation of $e^{1/z}$ about $z = 0$.

Solution: $e^{1/z}$ has a singularity at $z = 0$, but is fine at $z = \infty$. Expanding at infinity gives the Laurent series representation at $z = 0$ (if you think from the Riemann sphere point of view, you will find that the domain of $z = 0$ and $z = \infty$ complements each other):

$$e^{1/z} = \sum_{n=0}^{\infty} \frac{(1/z)^n}{n!} = \sum_{n=0}^{\infty} \frac{1}{n!} z^{-n}.$$

Example 16.10 (April 2009 Exam Q5(b)).

- (i) Find the Laurent series of $\exp(1/z^2)$ at $z = 0$. (8 marks)

Solution. $e^{1/z^2} = \sum_{n=0}^{\infty} \frac{(1/z^2)^n}{n!} = \sum_{n=0}^{\infty} \frac{1}{n!} z^{-2n}.$ □

- (ii) Evaluate $\oint_C \exp(1/z^2) dz$ where C is the circle $|z - 1| = 2$. (8 marks)

Solution. $\oint_C \exp(1/z^2) dz = 2\pi i c_{-1} = 0.$ □

eg:8.5.3

Example 16.11. Find the Laurent series that represent the function $f(z) = z^2 \sin \frac{1}{z^2}$ in the domain $|z| > 0$.

Solution. $f(z) = z^2 \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} \left(\frac{1}{z^2}\right)^{2n+1} = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} z^{-4n}$ □

Example 16.12. Find the Laurent series of $f(z) = \frac{1}{e^z - 1}$ about $z = 0$ (or for the domain $0 < |z| < 2\pi$). Write the series to at least degree 3 for the analytic part.

Solution.

$$\begin{aligned} f(z) &= \frac{1}{1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!} + \frac{z^5}{5!} + O(z^6) - 1} = \frac{1}{z} \left[\frac{1}{1 + \frac{z}{2} + \frac{z^2}{6} + \frac{z^3}{24} + \frac{z^4}{120} + O(z^5)} \right] \\ &= \frac{1}{z} \left[1 - \left(\frac{z}{2} + \frac{z^2}{6} + \frac{z^3}{24} + \frac{z^4}{120} + O(z^5) \right) + \left(\frac{z}{2} + \frac{z^2}{6} + \frac{z^3}{24} + O(z^4) \right)^2 \right. \\ &\quad \left. - \left(\frac{z}{2} + \frac{z^2}{6} + O(z^3) \right)^3 + \left(\frac{z}{2} + O(z^3) \right)^4 + O(z^5) \right] \\ &= \frac{1}{z} \left[1 - \frac{z}{2} + \left(-\frac{1}{6} + \frac{1}{4} \right) z^2 + \left(-\frac{1}{24} + \frac{1}{6} - \frac{1}{8} \right) z^3 + \left(-\frac{1}{120} + \frac{5}{72} - \frac{3}{24} + \frac{1}{16} \right) z^4 \right. \\ &\quad \left. + O(z^5) \right] \\ &= \frac{1}{z} \left[1 - \frac{z}{2} + \frac{1}{12} z^2 + 0z^3 - \frac{1}{720} z^4 + O(z^5) \right] = \frac{1}{z} - \frac{1}{2} + \frac{1}{12} z - \frac{1}{720} z^3 + O(z^4) \end{aligned}$$
□

In the case of rational function, we can make use of partial fractions. First, we state a characterisation of partial fractions.

Theorem 16.13. If $R_{m,n}(z) = \frac{a_m z^m + \cdots + a_1 z + a_0}{b_n (z - z_1)^{d_1} \cdots (z - z_r)^{d_r}}$ is a rational function whose denominator degree $n = d_1 + \cdots + d_r > m$ then $R_{m,n}(z)$ has a partial fraction decomposition of the form:

$$\begin{aligned} R_{m,n}(z) &= \frac{A_0^{(1)}}{(z - z_1)^{d_1}} + \frac{A_1^{(1)}}{(z - z_1)^{d_1-1}} + \cdots + \frac{A_{d_1-1}^{(1)}}{(z - z_1)^1} + \frac{A_0^{(2)}}{(z - z_2)^{d_2}} + \cdots + \frac{A_{d_2-1}^{(2)}}{(z - z_2)^1} \\ &\quad + \cdots + \frac{A_0^{(r)}}{(z - z_r)^{d_r}} + \cdots + \frac{A_{d_r-1}^{(r)}}{(z - z_r)^1} \end{aligned}$$

where $A_s^{(j)}$ are constants.

Example 16.14. Given $f(z) = \frac{1}{z+z^2}$.

1. [April 2010 Exam Q1(c)(i), 5 marks; Sept 2010 Exam Q5(a)(i), 3 marks] Find the Laurent series for f in $0 < |z| < 1$.

Solution: There are two ways to get the solution, as we will demonstrate below:

Partial fraction approach: $\frac{1}{z+z^2} = \frac{1}{z} - \frac{1}{z+1} = \frac{1}{z} - \sum_{n=0}^{\infty} (-1)^n z^n.$

Alternatively, a faster technique: $\frac{1}{z+z^2} = \frac{1}{z} \cdot \frac{1}{1+z} = \sum_{n=0}^{\infty} (-1)^n z^{n-1}.$

2. [Sept 2010 Exam Q5(a)(ii), 3 marks] Find the Laurent series for f in $1 < |z|$.

Solution. $\frac{1}{z+z^2} = \frac{1}{z^2} \cdot \frac{1}{1+\frac{1}{z}} = \sum_{n=0}^{\infty} \frac{(-1)^n}{z^{n+2}}$ □

3. [Sept 2010 Exam Q5(a)(iii), 3 marks] Find the Laurent series for f in $0 < |z+1| < 1$.

Solution. $\frac{1}{z+z^2} = \frac{1}{z+1} \cdot \frac{1}{z+1-1} = \sum_{n=0}^{\infty} (-1)(z+1)^{n-1}$ □

4. [April 2010 Exam Q1(c)(ii), 5 marks; Sept 2010 Exam Q5(a)(iv), 3 marks] Find the Laurent series for f in $1 < |z+1|$.

Solution. $\frac{1}{z+z^2} = \frac{1}{(z+1)^2} \cdot \frac{1}{1-\frac{1}{z+1}} = \sum_{n=0}^{\infty} (z+1)^{-n-2}.$ □

eg: wwy: 8.5.2

Example 16.15. Given $f(z) = \frac{1}{(z-1)(z-2)}$.

1. [Sept 2009 Exam Q5(a)(i)] Find the Laurent series representation of f for $0 < |z| < 1$

Solution. $f(z) = \frac{1}{z-2} - \frac{1}{z-1} = \frac{1}{1-z} - \frac{1}{2} \frac{1}{1-\frac{z}{2}} = \sum_{n=0}^{\infty} z^n - \frac{1}{2} \sum_{n=0}^{\infty} \left(\frac{z}{2}\right)^n = \sum_{n=0}^{\infty} \left(1 - \frac{1}{2^{n+1}}\right) z^n$ □

2. [Sept 2009 Exam Q5(a)(ii)] Find the Laurent series representation of f for $|z| > 2$

Solution. $|z| > 2 \Rightarrow |2/z| < 1$ and $|1/z| < 1$.
 $f(z) = \frac{1}{z-2} - \frac{1}{z-1} = \frac{1}{z} \left[\frac{1}{1-\frac{2}{z}} - \frac{1}{1-\frac{1}{z}} \right] = \frac{1}{z} \left[\sum_{n=0}^{\infty} \left(\frac{2}{z}\right)^n - \sum_{n=0}^{\infty} \left(\frac{1}{z}\right)^n \right] = \sum_{n=0}^{\infty} (2^n - 1) z^{-n-1}.$ □

3. [April 2009 Exam, Q5(a)(ii)] Expand f in a Laurent series valid for the domain $0 < |z-2| < 1$. (4 marks)

Solution. $f(z) = \frac{1}{z-2} \times \frac{1}{1+z-2} = \frac{1}{z-2} \sum_{n=0}^{\infty} (-(z-2))^n = \sum_{n=-1}^{\infty} (-1)^{n+1} (z-2)^n$ □

4. [April 2009 Exam, Q5(a)(i)] Expand f in a Laurent series valid for the domain $1 \leq |z| < 2$ (4 marks)

Solution. The range implies $|\frac{1}{z}| \leq 1$ and $|\frac{z}{2}| < 1$ and so

$$\begin{aligned} f(z) &= \frac{1}{z-2} - \frac{1}{z-1} = \frac{-\frac{1}{2}}{1-\frac{z}{2}} - \frac{1}{z} \times \frac{1}{1-\frac{1}{z}} = -\frac{1}{2} \sum_{n=0}^{\infty} \frac{1}{2^n} z^n - \frac{1}{z} \sum_{n=0}^{\infty} \frac{1}{z^n} \\ &= -\sum_{n=0}^{\infty} z^{-1-n} + \sum_{n=0}^{\infty} (-2^{-1-n}) z^n \end{aligned}$$

□

Example 16.16 (April 2008 Exam, Q3(a)). Find the Laurent series about $z = 0$ of the function $\frac{1}{z(z-1)(z-2)}$ valid in the following regions:

1. $|z| < 1$; (2 marks)

Solution.

$$\begin{aligned} \frac{1}{z(z-1)(z-2)} &= \frac{1}{2z} - \frac{1}{z-1} + \frac{1}{2(z-2)} = \frac{1}{2z} + \frac{1}{1-z} - \frac{1}{4} \frac{1}{1-z/2} \\ &= \frac{1}{2z} + \sum_{n=0}^{\infty} z^n - \frac{1}{4} \sum_{n=0}^{\infty} \left(\frac{z}{2}\right)^n = \frac{1}{2z} + \sum_{n=0}^{\infty} (1 - 2^{-n-2}) z^n. \end{aligned}$$

where $|z| < 1 \wedge |z| < 2$

□

2. $1 < |z| < 2$; (2 marks)

Solution.

$$\begin{aligned} \frac{1}{z(z-1)(z-2)} &= \frac{1}{2z} - \frac{1}{z-1} + \frac{1}{2(z-2)} = \frac{1}{2z} - \frac{1}{z} \frac{1}{1-\frac{1}{z}} - \frac{1}{4} \frac{1}{2(z-2)} \\ &= \frac{1}{2z} - \sum_{n=0}^{\infty} \frac{1}{z^{n+1}} - \frac{1}{4} \sum_{n=0}^{\infty} \left(\frac{z}{2}\right)^n \end{aligned}$$

where $|z| > 1 \wedge |z| < 2$.

□

3. $|z| > 2$. (2 marks)

Solution.

$$\begin{aligned} \frac{1}{z(z-1)(z-2)} &= \frac{1}{2z} - \frac{1}{z-1} + \frac{1}{2(z-2)} = \frac{1}{2z} - \frac{1}{z} \frac{1}{1-\frac{1}{z}} + \frac{1}{2z} \frac{1}{1-\frac{2}{z}} \\ &= \frac{1}{2z} - \frac{1}{z} \sum_{n=0}^{\infty} \frac{1}{z^n} + \frac{1}{2z} \sum_{n=0}^{\infty} \left(\frac{2}{z}\right)^n = \sum_{n=1}^{\infty} \left(\frac{\delta_{n,1}}{2} - 1 + 2^{n-2}\right) z^n \end{aligned}$$

where $|z| > 1 \wedge |z| > 2$.

□

From the above examples, we learned that Laurent series is closely related to the expansion of series about “singular points”. So it is possible to characterise the “isolated singularity” with Laurent series?

The answer is yes. Suppose z_0 is an isolated singularity of f . Then f can be represented by a Laurent series (I6.1) about z_0 . The following theorems show how the nature of the singularity possessed by a function may be determined by its behaviour in a deleted neighbourhood of the singularity.

Theorem 16.17 (Riemann’s Theorem on Removable Singularities). *If f has an isolated singularity at $z = z_0$ and if $\lim_{z \rightarrow z_0} (z - z_0)f(z) = 0$ then the singularity is removable.*

Proof: Multiplying $z - z_0$ to (16.1) gives

$$(z - z_0)f(z) = \cdots + c_{-2}(z - z_0)^{-1} + c_{-1} + \sum_{n=0}^{\infty} c_n(z - z_0)^{n+1}$$

Thus $\lim_{z \rightarrow z_0} (z - z_0)f(z)$ implies that $c_{-n} = 0$, $n = 1, 2, \dots$. Thus

$$f(z) = \sum_{n=0}^{\infty} c_n(z - z_0)^n \quad \text{for } 0 < |z - z_0| < R.$$

Define $f(z_0) = c_0$. Then f is analytic at z_0 . By the Holomorphic-Analytic Theorem from Topic 5, f holomorphic is holomorphic at z_0 .

By using an analogous argument, one can verify the following theorem.

thm:9.3.2

Theorem 16.18 (Riemann's Theorem on Poles). *If f is holomorphic in a deleted neighbourhood of z_0 and if there exists a positive integer n such that*

$$\lim_{z \rightarrow z_0} (z - z_0)^n f(z) \neq 0 \quad \text{but} \quad \lim_{z \rightarrow z_0} (z - z_0)^{n+1} f(z) = 0$$

then f has a pole of order n at $z = z_0$.

Proof: Multiplying $(z - z_0)^{n+1}$ to (16.1), we have

$$(z - z_0)^{n+1} f(z) = \cdots + c_{-(n+1)}(z - z_0)^0 + c_{-n}(z - z_0)^1 + \cdots + c_{-2}(z - z_0)^{n-1} + c_{-1}(z - z_0)^n + \sum_{n=0}^{\infty} c_n(z - z_0)^{2n+1}$$

If $\lim_{z \rightarrow z_0} (z - z_0)^{n+1} f(z) = 0$, then

$$c_{-(n+1)} = c_{-(n+2)} = \cdots = 0.$$

Therefore the Laurent series expansion of f at z_0 reduces to

$$f(z) = c_{-n}(z - z_0)^{-n} + \cdots + c_{-2}(z - z_0)^{-2} + c_{-1}(z - z_0)^{-1} + \sum_{n=0}^{\infty} c_n(z - z_0)^n$$

in which $\lim_{z \rightarrow z_0} (z - z_0)^n f(z) = c_{-n} \neq 0$. This shows that f has a pole of order n at z_0 .

In summary, isolated singular points of a function can be classified by means of its Laurent expansion as follows:

- If the principal part is zero, then z_0 is a removable singularity.
- If the principal part contains only finite number of nonzero terms, then z_0 is a pole. If the first nonzero term in the Laurent expansion is $c_{-n}(z - z_0)^{-n}$, then z_0 is known as a pole of order n .
- If the principal part contains infinitely many terms, then z_0 is an essential singularity.

The essential singularity is the worst-behaved isolated singularity. So far, the following theorem is the only mathematical proposition characterising its local behaviour.

Theorem 16.19 (Casorati-Weierstrass Theorem). *Let z_0 be an essential singularity of the function f , and let D be a disk centred at z_0 on which f is holomorphic except at z_0 . Then $f(D \setminus \{z_0\})$ is dense in the complex plane $\mathbb{C}$. In other words, the values of f on $D^\times$ come arbitrarily close to any complex number.*

A proof is given in [Lang, 1999, Section V, §3].

Let us investigate some examples of how Laurent series could be applied to identify the type of singularity.

eg:wwy:9.3.1

Example 16.20. Determine the type of singularity for $f(z) = \frac{\sin z}{z^5}$ at $z = 0$.

Solution: By looking at the Laurent series of f :

$$f(z) = \frac{1}{z^5} \left(z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \cdots \right) = \frac{1}{z^4} - \frac{1}{3!z^2} + \frac{1}{5!} - \frac{z^2}{7!} + \cdots$$

By definition f has a pole of order 4 at $z = 0$.

Example 16.21. Use the Laurent series of $f(z) = \frac{1 - \cos z}{z^5}$ at the isolated singularity $z = 0$ to determine the type of singularity.

Solution. The Laurent series is

$$\frac{1 - \cos z}{z^5} = \frac{1}{z^5} \left(\frac{z^2}{2!} - \frac{z^4}{4!} + \frac{z^6}{6!} - \cdots \right) = \frac{1}{2!z^3} - \frac{1}{4!z} + \frac{z}{6!} - \cdots$$

and $z = 0$ is a pole of order 3. □

eg:9.3.3

Example 16.22. Consider function $f(z) = \frac{1}{(z-1)^2(z-2)}$. Use the Laurent series of f to determine the type of singularity at $z = 1$.

Solution. The Laurent series representation of f about $z = 1$ is

$$f(z) = -\frac{1}{(z-1)^2} \frac{1}{1-(z-1)} = \frac{-1}{(z-1)^2} \sum_{n=0}^{\infty} (z-1)^n, \quad 0 < |z-1| < 2.$$

Therefore, $z = 1$ is a pole of order 2. □

16.3 Characterisation of Zero and Pole

In the previous section, we learn from Riemann's Theorem on Poles that poles behaved quite well. We also came across the notion of zero in Definition def:zero 16.3 and the zeros of order n in the previous section.

First, let us state an example of non-isolated zero. In this section, we avoid non-isolated zeros and poles and will study only the properties of isolated zeros and poles because they are useful in the calculations of residue.

Example 16.23. Given an example of a complex function which has a non-isolated zero.

Solution: $f(z) = \begin{cases} \sin(1/z), & z \neq 0, \\ 0, & z = 0. \end{cases}$

The zeros of f are 0 and the numbers $\frac{1}{n\pi}$, $n = \pm 1, \pm 2, \dots$.

The point 0 is not an isolated zero because at any given ϵ , it is possible to find other zeros in the ϵ -neighbourhood of 0.

def:zero-order-n

Definition 16.8. An holomorphic function f is said to have a (*isolated*) zero of order n at z_0 if

$$f^{(k)}(z_0) = 0, \quad k = 0, 1, \dots, n-1, \quad f^{(n)}(z_0) \neq 0.$$

eg:9.4.1

Example 16.24. Determine the types of zeros of $f(z) = \sin z$ and $g(z) = z \sin z$.

Solution: f has an isolated zero of order 1 at 0 while g has an isolated zero of order 2 at 0 because $f(0) = g(0) = 0$ and

$$\begin{aligned} f'(z) &= \cos z, & g'(z) &= \sin z + z \cos z, & g''(z) &= 2 \cos z - z \sin z \\ \Rightarrow f'(0) &= 1, & g'(0) &= 0, & g''(0) &= 2. \end{aligned}$$

f and g has isolated zeros at $k\pi$, $k = \pm 1, \pm 2, \dots$ with order 1 since $f(k\pi) = g(k\pi) = 0$ and

$$f'(k\pi) = \cos(k\pi) = (-1)^k, \quad g'(k\pi) = k\pi \cos(k\pi) = (-1)^k k\pi.$$

thm:9.4.1

Theorem 16.25 (Topic 5. Section 5.5). *Let f be holomorphic on an open set D and let z_0 be a zero of f in D . Then z_0 is either an isolated zero of f or there is an open disk about z_0 on which f is identically zero.*

Proof: Express f in Taylor series expansion about z_0 as

$$f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n$$

in some open disk D^* centred at z_0 .

If all $a_n = 0$, then $f(z) = 0$ throughout D^* .

Suppose that some coefficients are not zero. Let m be the smallest integer such that $a_m \neq 0$. Then $a_0 = a_1 = \dots = a_{m-1} = 0$. For $z \in D^*$, $f(z)$ can be written as

$$f(z) = (z - z_0)^m \sum_{n=0}^{\infty} a_{n+m} (z - z_0)^n = (z - z_0)^m g(z)$$

such that $g(z_0) = a_m \neq 0$. So there is some open disk K about z_0 in which $g(z) \neq 0$. Therefore, for z in K , $f(z) \neq 0$.

thm:9.4.2

Theorem 16.26. *Let f be holomorphic on a domain D . Suppose $\{z_n\}$ is a sequence of distinct zeros of f in D and $\{z_n\}$ converges to a point in D . Then $f(z) = 0$ for all z in D .*

Proof: Suppose $z_n \rightarrow b_0$ in D . Since f is continuous, $f(z_n) \rightarrow f(b_0)$. But $f(z_n) = 0$, so $f(b_0) = 0$. Thus b_0 is a zero of f in D . This means b_0 is not an isolated zero, so there must be an open disk D_0 about b_0 in which $f(z) = 0$.

We want to prove that $f(z) = 0$ for all z in D . To do this let w be any point in D , we will show that $f(w) = 0$.

Since D is connected, there is a path Γ in D connected b_0 with w . Choose $r > 0$ such that points of Γ is at a distance at least $2r$ from the boundary of D and choose r to be less than the radius of D_0 .

Now walk along Γ from b_0 to w , along the way selecting points at distance less than r from each others. This yields point $b_1, b_2, \dots, b_n = w$ on Γ . Let D_k denote the open disk of radius r about b_k . By the choice of r , none of these disks reaches the boundary of D . Each b_{k-1} is in $D_{k-1} \cap D_k \cap D_{k+1}$ for $k = 1, 2, \dots, n-1$. Further, $b_0 \in D_0 \cap D_1$ and $w \in D_{n-1} \cap D_n$.

Since $b_1 \in D_0 \cap D_1$, there is a sequence of points in $D_0 \cap D_1$ converging to b_1 . But f is identically zero in D_0 , so $f(z)$ vanishes on this sequence. Since this sequence is also in D_1 , $f(z) = 0$ for $z \in D_1$.

Now choose $b_2 \in D_1 \cap D_2$. Choose a sequence from $D_1 \cap D_2$ and converging to b_2 . since $f(z) = 0$ on D_1 , then $f(z) = 0$ at each point of this sequence. But this sequence is also in D_2 . This implies that $f(z) = 0$ on D_2 .

Continuing in this way, walk along Γ from b_0 to w . We find that $f(z) = 0$ on each of the disks along the way. Finally $f(z) = 0$ on D_n . But $w \in D_n$, so $f(w) = 0$.

Thus we have shown that if a holomorphic (equivalently, analytic) function vanishes on a convergent sequence of points in a domain, then *the function is identically zero throughout the entire domain*.

There is no analogous result for real functions. For example, the real function

$$f(x) = \begin{cases} 0, & x \leq 0, \\ x^2, & x > 0 \end{cases}$$

vanishes on half the line, but is not identically zero over the entire line.

We can't extend the above example to complex value because ordering is not well-defined in complex numbers.

Theorem leads immediately to the conclusion that two analytic functions agree on a convergent sequence in a domain D must be the same.

thm:9.4.3

Theorem 16.27. *Let f and g be analytic on a domain D . Suppose f and g agree on a convergent sequence in D . Then $f(z) = g(z)$ on D .*

Proof: Apply Theorem [thm:9.4.2](#) [16.26](#) to $f(z) - g(z)$.

Recall that when we define the exponential function and trigonometric functions. These functions were defined in such a way that they agreed with their real counterparts. How many ways can these be done. We know the answer from the above theorem that there can be only one extension of a differentiable function from the real to the complex domain!

This is why, when we extend a real function (such as exponential or trigonometric functions) to the complex plane, we can assume that this extension is unique.

thm:9.4.4

Theorem 16.28 (Characterisation of Zeros of Order n). *A function f that is analytic in some disk $|z - z_0| < R$ has a zero of order n at $z = z_0$ if and only if f can be written*

$$f(z) = (z - z_0)^n g(z)$$

where g is analytic at $z = z_0$ and $g(z_0) \neq 0$.

Proof: Sufficiency is trivial based on the properties of holomorphic functions, i.e. $(z - z_0)^n g(z)$ is a product of two analytic functions analytic in a ϵ neighbourhood of z_0 : $(z - z_0)^n$ and $g(z)$.

To prove the necessity, write down the Taylor series expansion of f at z_0 :

$$f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n, \quad a_n = \frac{f^{(n)}(z_0)}{n!}, \quad k = 0, 1, \dots$$

By Definition [def:zero-order-n](#) [16.8](#), $a_0 = a_1 = \dots = a_{n-1} = 0$ and

$$f(z) = a_n (z - z_0)^n + a_{n+1} (z - z_0)^{n+1} + \dots = (z - z_0)^n \underbrace{[a_n + a_{n+1}(z - z_0) + \dots]}_{g(z)}.$$

where $g(z_0) = a_n \neq 0$.

eg:9.4.2

Example 16.29. Determine the type of zero at $z = 0$ for function $f(z) = z \sin(z^2)$.

Solution.

$$f(z) = z(z^2 - \frac{(z^2)^3}{3!} + \frac{(z^2)^5}{5!} - \dots) = z^3(1 - \frac{z^4}{3!} + \frac{z^8}{5!} - \dots).$$

Thus we conclude that $z = 0$ is a zero of order 3 of f . □

Pole of a given function can also be characterized in the manner analogous to Definition 16.8. def:zero-order-n

thm:9.4.5

Theorem 16.30 (Characterisation of Poles of Order n). *A function f analytic in a punctured disk $0 < |z - z_0| < R$ has a pole of order n at z_0 if and only if f can be written*

$$f(z) = \frac{g(z)}{(z - z_0)^n}$$

where g is analytic at z_0 and $g(z_0) \neq 0$.

Proof: Let the Taylor series expansion of the analytic function g at z_0 be

$$\sum_{n=0}^{\infty} \frac{g^{(n)}(z_0)}{n!} (z - z_0)^n$$

where $g(z_0) \neq 0$. Then

$$f(z) = (z - z_0)^{-n} \sum_{n=0}^{\infty} \frac{g^{(n)}(z_0)}{n!} (z - z_0)^n$$

has a pole of order n at z_0 because $(z - z_0)^k f(z)$ is singular for $k = 0, \dots, n-1$, $\lim_{z \rightarrow z_0} (z - z_0)^n f(z) = g(z_0) \neq 0$ and $\lim_{z \rightarrow z_0} (z - z_0)^{n+1} f(z) = 0$.

The necessity of the assertion can be obtained using the Laurent series of the pole of order n at z_0 .

eg:9.4.3

Example 16.31. Consider function $f(z) = \frac{1 + 4i}{(z + 2)^2(z - i)^5}$. Determine all the poles and their corresponding types.

Solution. There are two poles: $z = -2$ and $z = i$.

(a) Write $f(z) = \frac{g(z)}{(z + 2)^2}$, $g(z) = \frac{1 + 4i}{(z - i)^5}$. g is analytic at $z = -2$ and $g(-2) \neq 0$. So f has a pole of order 2 at $z = -2$.

(b) Write $f(z) = \frac{h(z)}{(z - i)^5}$, $h(z) = \frac{1 + 4i}{(z + 2)^2}$. h is analytic at $z = i$ and $h(i) \neq 0$. Thus f has a pole of order 5 at $z = i$. □

The following result characterises how to construct a function having a pole from a function having a zero.

thm:9.4.6

Theorem 16.32. *If functions g and h are analytic at z_0 and h has a zero of order n at z_0 and $g(z_0) \neq 0$, then the function $f(z) = \frac{g(z)}{h(z)}$ has a pole of order n at z_0 .*

Proof: Write $h(z) = (z - z_0)^n \phi(z)$ where $\phi(z)$ is analytic at z_0 and $\phi(z_0) \neq 0$. Then

$$f(z) = \frac{\tilde{g}(z)}{(z - z_0)^n}, \quad \tilde{g}(z) = \frac{g(z)}{\phi(z)}, \quad \tilde{g}(z_0) = \frac{g(z_0)}{\phi(z_0)} \neq 0.$$

This implies that f has a pole of order n at z_0 by Theorem [16.30](#).

16.4 Residue Theorem and Its Computation

If f has an isolated singularity at z_0 , the Laurent series representation of f is given by [\(16.1\)](#). If we integrate the Laurent series representation [\(16.1\)](#) with a simple positively oriented closed curve containing z_0 , we will find that

$$\int_C f(z) dz = \sum_{n=-\infty}^{\infty} c_n \int_C (z - z_0)^n dz = c_{-1} \int_C (z - z_0)^{-1} dz = 2\pi i c_{-1}.$$

Definition 16.9. The coefficient c_{-1} in the Laurent series [\(16.1\)](#) is called the *residue of the function f* at the isolated singularity z_0 . We use the notation

$$\operatorname{Res}(f; z_0) \quad \text{or} \quad \operatorname{Res}(f(z); z_0) \quad \text{or} \quad \operatorname{Res}_{z=z_0} f(z)$$

to denote the residue of f at z_0 . The last symbol was used in Mitrinović and Kečkić [1984] but I find the first two easier to work with.

[eg:9.5.1](#)

Example 16.33. Find the residue of the function $f(z) = e^{1/z}$, $z \neq 0$ at $z = 0$.

Solution: The Laurent series representation of f is

$$f(z) = 1 + \frac{1}{z} + \frac{1}{2!z^2} + \frac{1}{3!z^3} + \cdots.$$

By definition, $\operatorname{Res}(f(z); 0) = c_{-1} = 1$.

How to find residues when we want to calculate the Laurent series quickly?

For the residue at a simple pole, we have the following theorem.

[thm:9.5.1](#)

Theorem 16.34. If f has a simple pole at $z = z_0$, then

$$\operatorname{Res}(f; z_0) = \lim_{z \rightarrow z_0} (z - z_0) f(z).$$

Proof: Laurent series expansion of f is $f(z) = \frac{c_{-1}}{z - z_0} + c_0 + c_1(z - z_0) + \cdots$.

Therefore,

$$\lim_{z \rightarrow z_0} (z - z_0) f(z) = \lim_{z \rightarrow z_0} \left[c_{-1} + c_0(z - z_0) + c_1(z - z_0)^2 + \cdots \right] = c_{-1} = \operatorname{Res}(f; z_0)$$

[thm:9.5.2](#)

Theorem 16.35. If f has a simple pole at $z = z_0$ and $f(z) = \frac{g(z)}{h(z)}$, where g and h are analytic at $z = z_0$ and $g(z_0) \neq 0$, $h(z_0) = 0$, but $h'(z_0) \neq 0$, then

$$\operatorname{Res}(f; z_0) = \frac{g(z_0)}{h'(z_0)}.$$

Proof: Since the function h has a zero of order 1 at z_0 , this implies that

$$h'(z_0) = \lim_{z \rightarrow z_0} \frac{h(z) - h(z_0)}{z - z_0} = \lim_{z \rightarrow z_0} \frac{h(z)}{z - z_0}$$

and

$$\lim_{z \rightarrow z_0} (z - z_0)f(z) = \lim_{z \rightarrow z_0} (z - z_0) \frac{g(z)}{h(z)} = \frac{\lim_{z \rightarrow z_0} g(z)}{\lim_{z \rightarrow z_0} \frac{h(z)}{z - z_0}} = \frac{g(z_0)}{h'(z_0)}$$

Therefore, for simple poles, it is not necessary to find the Laurent series but use Theorem 16.34 (or Theorem 16.30, or Theorem 16.35 instead).

Example 16.36 (April 2010 Exam, Q2(c)). If $f(z) = \frac{1}{z^3 + 1}$, find $\text{Res}(f; z_i)$ where z_i 's are the poles of f . (11 marks)

Solution. First, the poles are

$$\begin{aligned} z_i &= \sqrt[3]{-1} = \exp \frac{i\pi + 2k\pi i}{3}, \quad k = 0, 1, 2 \\ &= \frac{1}{2} + \frac{\sqrt{3}}{2}i, \quad -1, \quad \frac{1}{2} - \frac{\sqrt{3}}{2}i. \end{aligned}$$

Using Theorem 16.34, the calculations are simpler by using L'Hôpital rule:

$$\begin{aligned} \text{Res}(f; -1) &= \lim_{z \rightarrow -1} \frac{z + 1}{z^3 + 1} = \lim_{z \rightarrow -1} \frac{1}{3z^2} = \frac{1}{3} \\ \text{Res}(f; \frac{1}{2} + \frac{\sqrt{3}}{2}i) &= \lim_{z \rightarrow \frac{1}{2} + \frac{\sqrt{3}}{2}i} \frac{1}{3z^2} = \frac{-1 - \sqrt{3}i}{6} \\ \text{Res}(f; \frac{1}{2} - \frac{\sqrt{3}}{2}i) &= \lim_{z \rightarrow \frac{1}{2} - \frac{\sqrt{3}}{2}i} \frac{1}{3z^2} = \frac{-1 + \sqrt{3}i}{6} = \overline{\text{Res}(f; \frac{1}{2} + \frac{\sqrt{3}}{2}i)} \end{aligned}$$

Using Theorem 16.30, the calculations are more complicated:

$$\begin{aligned} \text{Res}(f; -1) &= \frac{1}{(z - \frac{1}{2} - \frac{\sqrt{3}}{2}i)(z - \frac{1}{2} + \frac{\sqrt{3}}{2}i)} \Big|_{z=-1} = \frac{1}{(-\frac{3}{2})^2 + (\frac{\sqrt{3}}{2})^2} = \frac{1}{3}. \\ \text{Res}(f; \frac{1}{2} + \frac{\sqrt{3}}{2}i) &= \frac{1}{(z + 1)(z - \frac{1}{2} + \frac{\sqrt{3}}{2}i)} \Big|_{z=\frac{1}{2} + \frac{\sqrt{3}}{2}i} = \frac{1}{(\frac{3}{2} + \frac{\sqrt{3}}{2}i)(\sqrt{3}i)} = \frac{-1 - \sqrt{3}i}{6} \\ \text{Res}(f; \frac{1}{2} - \frac{\sqrt{3}}{2}i) &= \frac{1}{(z - \frac{1}{2} - \frac{\sqrt{3}}{2}i)(z + 1)} \Big|_{z=\frac{1}{2} - \frac{\sqrt{3}}{2}i} = \frac{1}{(-\sqrt{3}i)(\frac{3}{2} - \frac{\sqrt{3}}{2}i)} = \frac{-1 + \sqrt{3}i}{6} \end{aligned}$$

□

eg:9.5.3

Example 16.37 (Sept 2011 Exam Q3(c); Sept 2010 Exam Q3(c)). Find the residues of $f(z) = \frac{1}{z^4 + 1}$ and hence evaluate $\oint_C \frac{1}{z^4 + 1} dz$ where C is positively oriented circle $|z| = 2$. (8 marks)

Solution. Poles of f are the zeros of $z^4 + 1$:

$$z_k = \sqrt[4]{-1} = \exp \frac{(2k+1)\pi i}{4}, \quad k = 0, 1, 2, 3.$$

$$\text{Res}(f; z_k) = \lim_{z \rightarrow z_k} \frac{z - z_k}{z^4 + 1} = \lim_{z \rightarrow z_k} \frac{1}{4z^3} = \frac{1}{4} e^{-3(2k+1)\pi i/4} = \begin{cases} \frac{-1-i}{4\sqrt{2}}, & k=0, \\ \frac{1-i}{4\sqrt{2}}, & k=1, \\ \frac{1+i}{4\sqrt{2}}, & k=2, \\ \frac{-1+i}{4\sqrt{2}}, & k=3. \end{cases}$$

Using an application of Cauchy's Integral Formula called the Cauchy's Residue Theo-

rem [thm:residue](#) 16.42 in the next section, we find that

$$\oint_{|z|=2} \frac{1}{z^4 + 1} = 2\pi i \sum_{k=0}^3 \text{Res}(f; z_k) = 0.$$

□

With the following example, we can see how Theorem [thm:9.5.2](#) 16.35 can sometimes be very convenient.

eg:9.5.4

Example 16.38. Find the residue of all the poles of $\tan z$.

Solution. It is not difficult to obtain the poles from

$$\tan z = \frac{\sin z}{\cos z} \Rightarrow \cos z = 0 \Rightarrow z = (n + \frac{1}{2})\pi, \quad n = 0, \pm 1, \pm 2, \dots$$

Since

$$\cos z = \cos(z - (n + \frac{1}{2})\pi) \cos(n + \frac{1}{2})\pi - \sin(z - (n + \frac{1}{2})\pi) \sin(n + \frac{1}{2})\pi = -(-1)^n \sin(z - (n + \frac{1}{2})\pi),$$

it is not difficult to see that the poles are simple.

Therefore,

$$\text{Res}(f; (n + \frac{1}{2})\pi) = \lim_{z \rightarrow (n + \frac{1}{2})\pi} (z - (n + \frac{1}{2})\pi) \tan z = \lim_{z \rightarrow (n + \frac{1}{2})\pi} \frac{(z - (n + \frac{1}{2})\pi) \sin z}{-(-1)^n \sin(z - (n + \frac{1}{2})\pi)} = -1$$

for $n = 0, \pm 1, \pm 2, \dots$.

If we can remember Theorem [thm:9.5.2](#) 16.35, then

$$\text{Res}(\tan; (n + \frac{1}{2})\pi) = \frac{\sin(n + \frac{1}{2})\pi}{-\sin(n + \frac{1}{2})\pi} = -1.$$

since $\frac{d}{dz} \cos z = -\sin z$.

□

For the residue at a pole of order n , we have the following theorem.

thm:9.5.3

Theorem 16.39. If f has a pole of order n at $z = z_0$, then

$$\text{Res}(f; z_0) = \frac{1}{(n-1)!} \lim_{z \rightarrow z_0} \frac{d^{n-1}}{dz^{n-1}} (z - z_0)^n f(z). \quad (16.2)$$

e:res:pole:lim

Proof: Because f has a pole of order n at z_0 , the Laurent series representation of f is of the form

$$f(z) = \frac{c_{-n}}{(z - z_0)^n} + \dots + \frac{c_{-1}}{z - z_0} + c_0 + c_1(z - z_0) + \dots$$

By multiplying the above equation with $(z - z_0)^n$, we have

$$(z - z_0)^n f(z) = c_{-n} + \dots + c_{-1}(z - z_0)^{n-1} + c_0(z - z_0)^n + c_1(z - z_0)^{n+1} + \dots$$

In order to find c_{-1} , we have to differentiate both sides $n-1$ times to obtain

$$\frac{d^{n-1}}{dz^{n-1}} (z - z_0)^n f(z) = (n-1)!c_{-1} + n(n-1) \dots 2c_0(z - z_0) + (n+1)n(n-1) \dots c_1(z - z_0)^2 + \dots$$

By taking limit $z \rightarrow z_0$, we can obtain the expression (16.2) after some simplification.

The formula (16.2) can be very useful in determining the residues for low-order poles. For higher order poles, the calculations can become unmanageable, and series expansion is usually easier. As for essential singularities, residues often must be taken directly from series expansions.

eg:9.5.2

Example 16.40. If $f(z) = \frac{1}{(z-1)^2(z-3)}$, find $\text{Res}(f; 1)$ and $\text{Res}(f; 3)$.

Solution: Function f has a simple pole at $z = 3$ and a pole of order 2 at $z = 1$. Hence

$$\begin{aligned}\text{Res}(f; 3) &= \lim_{z \rightarrow 3} \frac{z-3}{(z-1)^2(z-3)} = \frac{1}{(3-1)^2} = \frac{1}{4} \\ \text{Res}(f; 1) &= \lim_{z \rightarrow 1} \frac{1}{1!} \frac{d}{dz} \frac{(z-1)^2}{(z-1)^2(z-3)} = \lim_{z \rightarrow 1} \left(-\frac{1}{(z-3)^2} \right) = -\frac{1}{4}.\end{aligned}$$

Example 16.41 (Sept 2012 Exam, Q3(c)). Let $f(z) = \frac{\sin z}{z(z-3)(4z^2+4z+1)}$. Find all the residues of all the isolated singularities of $f(z)$. (8 marks)

Solution. There are 3 isolated singularities:

$$\begin{aligned}f(z) &= \frac{\sin z}{z(z-3)(2z+1)^2} \Rightarrow z = 0, 3, -\frac{1}{2} \\ \text{Res}(f; 0) &= \lim_{z \rightarrow 0} \frac{z \sin z}{z(z-3)(2z+1)^2} = 0 \quad [\text{removable singularity}] \\ \text{Res}(f; 3) &= \lim_{z \rightarrow 3} \frac{(z-3) \sin z}{z(z-3)(2z+1)^2} = \frac{\sin 3}{147} \quad [\text{simple pole}] \\ \text{Res}(f; -\frac{1}{2}) &= \lim_{z \rightarrow -\frac{1}{2}} \frac{1}{1!} \frac{d}{dz} \frac{(z+\frac{1}{2})^2 \sin z}{z(z-3)(2z+1)^2} = \frac{(z^2-3z) \cos z + (3-2z) \sin z}{4z^2(z-3)^2} \Big|_{z=-\frac{1}{2}} \\ &= \frac{7 \cos \frac{1}{2} - 16 \sin \frac{1}{2}}{49}\end{aligned}$$

□

16.5 Examples of Applications

We now come to the reason why the residue concept is important. Observe that if C is any positively oriented simple closed curve in $0 < |z - z_0| < r$ containing just one isolated singularity z_0 inside, then

$$\int_C f(z) dz = 2\pi i c_{-1} = 2\pi i \text{Res}(f; z_0).$$

This provides the key to evaluating many complex integrals where one can evaluate contour integral by summing the residues at the isolated singularities of f within the closed contour C .

thm:residue

Theorem 16.42 (Cauchy's Residue Theorem). *Let D be a simply connected region and C a simple closed contour lying entirely within D . If function f is analytic on and inside C , except at a finite number of isolated singular points $z_1, z_2, \dots, z_n$ within C , then*

$$\oint_C f(z) dz = 2\pi i \sum_{j=1}^n \text{Res}(f; z_j).$$

Proof: Let $C_1, C_2, \dots, C_n$ be non-overlapping small circles centred at $z_1, z_2, \dots, z_n$, respectively and all are interior to the contour C . Then by deformation principle

$$\oint_C f(z)dz = \sum_{j=1}^n \oint_{C_j} f(z)dz$$

Since $\oint_{C_j} f(z)dz = 2\pi i \operatorname{Res}(f; z_j)$, hence the result follows.

We shall present some worked examples to illustrate how Residue Theorem can be used to evaluate complex integrals.

eg:9.5.5

Example 16.43. Evaluate $\oint_C \frac{1}{(z-1)^2(z-3)}dz$ where the contour C is given by

(a) a rectangle formed by $x = 0, x = 4, y = -1, y = 1$.

Solution: The rectangle contains poles 1 and 3. Recall Example [16.40](#), we obtain

$$\begin{aligned} \oint_C \frac{1}{(z-1)^2(z-3)}dz &= 2\pi i \left[\lim_{z \rightarrow 1} \frac{1}{1!} \frac{d}{dz} \frac{(z-1)^2}{(z-1)^2(z-3)} + \lim_{z \rightarrow 3} \frac{z-3}{(z-1)^2(z-3)} \right] \\ &= 2\pi i \left[-\frac{1}{(1-3)^2} + \frac{1}{(3-1)^2} \right] = 2\pi i \times 0 = 0. \end{aligned}$$

(b) $|z| = 2$.

Solution: Recall Example [16.40](#), we obtain

$$\oint_{|z|=2} \frac{1}{(z-1)^2(z-3)}dz = 2\pi i \operatorname{Res}(f; 1) = -\frac{\pi i}{2}.$$

eg:9.5.7

Example 16.44. Evaluate $\oint_{|z|=2} \tan z dz$.

Solution. According to Example [16.38](#), $\tan z$ has simple poles at the points where $\cos z = 0$, i.e. $z = (k + \frac{1}{2})\pi, k = 0, \pm 1, \pm 2, \dots$. Two simple poles $\tan z$ lying inside $|z| = 2$ are $z = -\frac{\pi}{2}$ and $z = \frac{\pi}{2}$. Hence

$$\oint_C \tan z dz = 2\pi i \left[\operatorname{Res}(\tan; -\frac{\pi}{2}) + \operatorname{Res}(\tan; \frac{\pi}{2}) \right] = -4\pi i$$

□

eg:9.5.6

Example 16.45. Evaluate $\oint_C \tan z dz$ where C is the rectangle formed by $x = -2, x = 5, y = -1, y = 1$.

Solution. Since only three points $-\frac{\pi}{2}, \frac{\pi}{2}, \frac{3\pi}{2}$ lies in between -2 and 5 (inside the rectangle), therefore

$$\begin{aligned} \oint_C \tan z dz &= 2\pi i \left[\operatorname{Res}(\tan; -\frac{\pi}{2}) + \operatorname{Res}(\tan; \frac{\pi}{2}) + \operatorname{Res}(\tan; \frac{3\pi}{2}) \right] \\ &= 2\pi i \left[-1 + (-1) + (-1) \right] = -6\pi i. \end{aligned}$$

□

The Residual Theorem is not only useful for complex integration, it can also be employed to evaluate real integral problems of the following types:

- Definite integrals involving sine and cosine functions:

$$\int_0^{2\pi} f(\sin \theta, \cos \theta) d\theta$$

where can be extended to $f(z)$ a rational expression in z such that $f(e^{i\theta})$ is defined in $0 \leq \theta \leq 2\pi$.

- Improper integrals of rational functions over $(-\infty, \infty)$ or $(0, \infty)$:

$$\int_{-\infty}^{\infty} \frac{p(x)}{q(x)} dx$$

where $p(x)$ and $q(x)$ are polynomials, $q(x)$ has no real roots and the degree of $q(x)$ exceeds the degree of $p(x)$ by at least 2.

- Improper integrals from Fourier Analysis:

$$\int_{-\infty}^{\infty} \frac{f(x)}{g(x)} \cos(mx) dx \quad \text{or} \quad \int_{-\infty}^{\infty} \frac{f(x)}{g(x)} \sin(mx) dx$$

where f and g are normally polynomials and m is a positive constant.

Recall that if f is a real function, the *improper integral* over $\mathbb{R}$ is defined to be

$$\int_{-\infty}^{\infty} f(x) dx = \lim_{R_1 \rightarrow \infty} \int_{-R_1}^0 f(x) dx + \lim_{R_2 \rightarrow \infty} \int_0^{R_2} f(x) dx \quad (16.3)$$

when both the limits exist. Another value that is assigned to $\int_{-\infty}^{\infty} f(x) dx$ defined below is often useful.

Definition 16.10. The *Cauchy principal value* of the improper integral $\int_{-\infty}^{\infty} f(x) dx$ is defined as

$$P.V. \int_{-\infty}^{\infty} f(x) dx = \lim_{R \rightarrow \infty} \int_{-R}^R f(x) dx \quad (16.4)$$

If integral (16.3) converges, its Cauchy principal value (16.4) exists; and

$$\int_{-\infty}^{\infty} f(x) dx = P.V. \int_{-\infty}^{\infty} f(x) dx$$

However, it is not always true that integral (16.3) converges when its Cauchy principal value exists as demonstrated by the following example.

Example 16.46. Let $f(x) = x$. Determine whether $P.V. \int_{-\infty}^{\infty} f(x) dx$ and $\int_{-\infty}^{\infty} f(x) dx$ are meaningful. Justify your answer.

Solution: The following derivation shows that a limit exists and hence the principal value integral exists:

$$P.V. \int_{-\infty}^{\infty} f(x) dx = \lim_{R \rightarrow \infty} \int_{-R}^R x dx = \lim_{R \rightarrow \infty} \left[\frac{x^2}{2} \right]_{-R}^R = \lim_{R \rightarrow \infty} \left[\frac{R^2}{2} - \frac{R^2}{2} \right] = 0.$$

However, the following derivation indicates that limits do not exist:

$$\int_{-\infty}^{\infty} f(x) dx = \lim_{R_1 \rightarrow \infty} \int_{-R_1}^0 x dx + \lim_{R_2 \rightarrow \infty} \int_0^{R_2} x dx = \lim_{R_1 \rightarrow \infty} \left[-\frac{R_1^2}{2} \right] + \lim_{R_2 \rightarrow \infty} \left[\frac{R_2^2}{2} \right].$$

16.5.1 Evaluation of Definite Integrals Involving Sine and Cosine

Residue theory can be useful in evaluating definite integrals over a full period of the form:

$$\int_0^{2\pi} F(\sin m\theta, \cos n\theta) d\theta \quad \text{or} \quad \int_{-\pi}^{\pi} F(\sin m\theta, \cos m\theta) d\theta \quad (16.5) \quad \text{e:wwy:10.3.1}$$

where F is a rational function and m, n are integers. Given the form of an integrand in (16.5) one can reasonably hope that the integral results from the usual parameterization of the unit circle

$$z = e^{i\theta}, \quad 0 \leq \theta \leq 2\pi.$$

We have

$$dz = ie^{i\theta} d\theta \Rightarrow \boxed{d\theta = \frac{dz}{iz}}, \quad \boxed{\cos n\theta = \frac{z^n + z^{-n}}{2}}, \quad \boxed{\sin m\theta = \frac{z^m - z^{-m}}{2i}}.$$

Putting all of this into (16.5)

$$\int_0^{2\pi} F(\sin m\theta, \cos n\theta) d\theta = \oint_{S^1} F\left(\frac{z^m - z^{-m}}{2i}, \frac{z^n + z^{-n}}{2}\right) \frac{dz}{iz}. \quad (16.6) \quad \text{e:wwy:10.3.2}$$

where S^1 is the unit circle.

Example 16.47. Evaluate $\int_0^{2\pi} \frac{1}{a + b \sin \theta} d\theta$, where $a > |b|$.

Solution: Let $z = e^{i\theta}$. Then

$$I = \int_0^{2\pi} \frac{1}{a + b \sin \theta} d\theta = \oint_{S^1} \frac{1}{a + b \left(\frac{z - z^{-1}}{2i}\right)} \frac{dz}{iz} = \oint_{S^1} \frac{2dz}{2aiz + bz^2 - b}$$

The integrand has two simple poles:

$$z_1, z_2 = \frac{-2ai \pm \sqrt{(2ai)^2 - 4b(-b)}}{2b} = \frac{-ai \pm \sqrt{a^2 - b^2}i}{b}$$

Only z_1 lies inside S^1 . Hence

$$\text{Res}(f; z_1) = \lim_{z \rightarrow z_1} (z - z_1) \frac{2}{2aiz + bz^2 - b} = \lim_{z \rightarrow z_1} \frac{2}{2ai + 2bz} = \frac{2}{2ai + 2bz_1} = \frac{1}{\sqrt{a^2 - b^2}i}.$$

Thus

$$\int_0^{2\pi} \frac{1}{a + b \sin \theta} d\theta = 2\pi i \left[\frac{1}{\sqrt{a^2 - b^2}i} \right] = \frac{2\pi}{\sqrt{a^2 - b^2}}.$$

Example 16.48. Evaluate $\int_0^{2\pi} \frac{1}{(a + b \sin \theta)^2} d\theta$ where $a > |b|$.

Solution. From the previous Example 16.47 we know that

$$\int_0^{2\pi} \frac{1}{a + b \sin \theta} d\theta = \frac{2\pi}{\sqrt{a^2 - b^2}} \quad (*)$$

Differentiating (*) with respect to a , gives

$$\begin{aligned} \frac{d}{da} \int_0^{2\pi} \frac{1}{a + b \sin \theta} d\theta &= \int_0^{2\pi} \frac{d}{da} \left[\frac{1}{a + b \sin \theta} \right] d\theta = \int_0^{2\pi} -\frac{1}{(a + b \sin \theta)^2} d\theta \\ &= \frac{d}{da} \frac{2\pi}{\sqrt{a^2 - b^2}} = 2\pi \left(-\frac{1}{2}\right) (a^2 - b^2)^{-3/2} (2a) \end{aligned}$$

$$\text{Therefore } \int_0^{2\pi} \frac{1}{(a + b \sin \theta)^2} d\theta = \frac{2\pi a}{(a^2 - b^2)^{3/2}}. \quad \square$$

Remark 16.49. We can obtain the value of $\int_0^{2\pi} \frac{d\theta}{(a + b \sin \theta)^n}$ or $\int_0^{2\pi} \frac{d\theta}{(a + b \cos \theta)^n}$ (with $|a| > b$) by differentiating $\int_0^{2\pi} \frac{d\theta}{a + b \sin \theta}$ or $\int_0^{2\pi} \frac{d\theta}{a + b \cos \theta}$ $(n-1)$ times.

Example 16.50. Evaluate $\int_0^{2\pi} \frac{1}{a + b \cos \theta} d\theta$, $|a| > b$

Solution. Let $z = e^{i\theta}$. Then $\cos \theta = \frac{z + z^{-1}}{2}$.

$$\begin{aligned} \int_0^{2\pi} \frac{1}{a + b \cos \theta} d\theta &= \oint_{S^1} \frac{1}{a + b(\frac{z+z^{-1}}{2})} \frac{dz}{iz} = \frac{2}{i} \oint_{S^1} \frac{1}{2az + bz^2 + b} dz \\ &= \frac{2}{ib} \oint_{S^1} \frac{1}{(z + \frac{a}{b})^2 + 1 - (\frac{a}{b})^2} dz \\ &= \frac{2}{ib} \oint_{S^1} \frac{1}{z + \frac{a}{b} + \frac{\sqrt{a^2 - b^2}}{b}} \times \frac{1}{z + \frac{a}{b} - \frac{\sqrt{a^2 - b^2}}{b}} dz \\ &= \frac{2}{ib} \times 2\pi i \times \frac{1}{(-\frac{a}{b} + \frac{\sqrt{a^2 - b^2}}{b}) + \frac{a}{b} + \frac{\sqrt{a^2 - b^2}}{b}} \\ &= \frac{2}{b} \times 2\pi \times \frac{1}{(2 \times \frac{\sqrt{a^2 - b^2}}{b})} = \frac{2\pi}{\sqrt{a^2 - b^2}} \end{aligned}$$

$\square$

According to Barry Simon's Comprehensive Analysis Part 2A, Example 16.50 implies

$$\int_0^{2\pi} \cos^{2n} \theta d\theta = \frac{2\pi(2n)!}{2^{2n}(n!)^2}$$

by taking $a = 1$ and expanding both sides using the generalised Binomial Theorem.

Example 16.51.

1. [Sept 2012 Exam Q4(b)] Evaluate $\int_0^{2\pi} \frac{d\theta}{5 + 3 \cos \theta}$. (10 marks)
2. [Sept 2011 Exam Q4(a)] Prove that $\int_0^{2\pi} \frac{d\theta}{5 + 3 \cos \theta} = \frac{\pi}{2}$. (15 marks)

Solution: Let $z = e^{i\theta}$. Then $\cos \theta = \frac{z+z^{-1}}{2}$.

$$\begin{aligned} \int_0^{2\pi} \frac{1}{5 + 3 \cos \theta} d\theta &= \oint_{S^1} \frac{1}{5 + 3(\frac{z+z^{-1}}{2})} \frac{dz}{iz} = \frac{2}{i} \oint_{S^1} \frac{1}{10z + 3z^2 + 3} dz \\ &= \frac{2}{3i} \oint_{S^1} \frac{1}{z + 3} \times \frac{1}{z + \frac{1}{3}} dz = \frac{2}{3i} \times 2\pi i \left[\frac{1}{(-\frac{1}{3}) + 3} \right] = \frac{4\pi}{3} \times \frac{3}{8} = \frac{\pi}{2} \end{aligned}$$

Example 16.52. Evaluate $\int_0^{2\pi} \frac{d\theta}{(5 - 3 \sin \theta)^2}$.

Solution. In the previous Example 16.48, take $a = 5$ and $b = -3$. It follows that

$$\int_0^{2\pi} \frac{d\theta}{(5 - 3 \sin \theta)^2} = \frac{2\pi \cdot 5}{(5^2 - (-3)^2)^{3/2}} = \frac{10\pi}{16^{3/2}} = \frac{5\pi}{32}.$$

Example 16.53. Evaluate $\int_0^{2\pi} \frac{1}{(2 + \cos \theta)^2} d\theta$.

Solution. Differentiating the integral with respect to a in Example 16.50 and taking negative, gives

$$\int_0^{2\pi} \frac{1}{(a + b \cos \theta)^2} d\theta = \frac{2\pi a}{(a^2 - b^2)^{3/2}}.$$

Letting $a = 2$ and $b = 1$, we obtain that $\int_0^{2\pi} \frac{1}{(2 + \cos \theta)^2} d\theta = \frac{4\pi}{3\sqrt{3}}$. □

Example 16.54. Evaluate $I = \int_0^{2\pi} \frac{\cos 3\theta}{5 - 4 \cos \theta} d\theta$.

Solution. Let $z = e^{i\theta}$. Then $\cos n\theta = \frac{1}{2}(z^n + z^{-n})$ and $d\theta = \frac{dz}{iz}$. Hence

$$\begin{aligned} I &= \int_0^{2\pi} \frac{\cos 3t}{5 - 4 \cos t} dt = \oint_{S^1} \frac{\frac{z^3 + z^{-3}}{2}}{5 - 4(\frac{z + z^{-1}}{2})} \frac{dz}{iz} = \frac{i}{4} \oint_{S^1} \frac{z^6 + 1}{z^3(z - \frac{1}{2})(z - 2)} dz \\ &= \frac{i \times 2\pi i}{4} \left[\text{Res}(f, \frac{1}{2}) + \text{Res}(f, 0) \right], \quad f(z) = \frac{z^6 + 1}{(z - \frac{1}{2})(z - 2)} \end{aligned} \quad [0.5 \text{ mark}]$$

$$= \frac{-\pi}{2} \left[\frac{(\frac{1}{2})^3 + (\frac{1}{2})^{-3}}{\frac{1}{2} - 2} + \lim_{z \rightarrow 0} \frac{1}{2!} \frac{d^2}{dz^2} \left(\frac{z^6 + 1}{(z - \frac{1}{2})(z - 2)} \right) \right] \quad [0.5 \text{ mark}]$$

$$= \frac{-\pi}{2} \left[-\frac{65}{12} + \frac{21}{4} \right] = \frac{-\pi}{2} \times \left(-\frac{1}{6} \right) = \frac{\pi}{12}. \quad [0.5 \text{ mark}]$$

The integrand has a pole of order 3 at $z = 0$ and a simple pole at $z = \frac{1}{2}$ inside the domain circle by S^1 :

$$\begin{aligned} \text{Res}(f; 0) &= \lim_{z \rightarrow 0} \frac{1}{2!} \frac{d^2}{dz^2} \left[z^3 \cdot \frac{z^6 + 1}{z^3(z - \frac{1}{2})(z - 2)} \right] = \frac{21}{4} \\ \text{Res}(f; \frac{1}{2}) &= \lim_{z \rightarrow \frac{1}{2}} \left[(z - \frac{1}{2}) \cdot \frac{z^6 + 1}{z^3(z - \frac{1}{2})(z - 2)} \right] = -\frac{65}{12} \end{aligned}$$

where

$$\begin{aligned} \frac{d^2}{dz^2} \left(\frac{z^6 + 1}{(z - \frac{1}{2})(z - 2)} \right) &= 30z^4(z - \frac{1}{2})^{-1}(z - 2)^{-1} + 2(z^6 + 1)(z - \frac{1}{2})^{-3}(z - 2)^{-1} \\ &\quad + 2(z^6 + 1)(z - \frac{1}{2})^{-1}(z - 2)^{-3} - 2(6z^5)(z - \frac{1}{2})^{-2}(z - 2)^{-1} \\ &\quad - 2(6z^5)(z - \frac{1}{2})^{-1}(z - 2)^{-2} + 2(z^6 + 1)(z - \frac{1}{2})^{-2}(z - 2)^{-2} \\ \lim_{z \rightarrow 0} \frac{1}{2!} \frac{d^2}{dz^2} \left(\frac{z^6 + 1}{(z - \frac{1}{2})(z - 2)} \right) &= \frac{1}{2} \left[0 + 2(-8)(-\frac{1}{2}) + 2(-2)(-\frac{1}{8}) - 0 - 0 + 2(4)(\frac{1}{4}) \right] = \frac{21}{4} \end{aligned} \quad [1.5 \text{ marks}]$$

$$\text{Thus } I = \frac{i}{4} \left[2\pi i \left\{ \frac{21}{4} - \frac{65}{12} \right\} \right] = \frac{-\pi}{2} \left(-\frac{1}{6} \right) = \frac{\pi}{12}.$$

Alternatively, C.I.F. method is simpler with the need to calculate the partial fraction:

$$\frac{1}{(z - \frac{1}{2})z^3} = \frac{A}{z - \frac{1}{2}} + \frac{B}{z} + \frac{C}{z^2} + \frac{D}{z^3} \Rightarrow A = 8, D = -2, C = -4, B = -8.$$

Hence

$$\begin{aligned}\int_0^{2\pi} \frac{\cos 3t}{5-4\cos t} dt &= \frac{i}{4} \left[\frac{-\pi i}{12} + 2\pi i \left(\frac{8}{\frac{1}{2}-2} - \frac{8}{0-2} - \frac{4(-1)}{(0-2)^2} - \frac{2 \times 2}{(0-2)^3} \right) \right] \\ &= \frac{i}{4} \left[\frac{-2\pi i}{12} + \frac{-2\pi i}{12} \right] = \frac{\pi}{12}\end{aligned}$$

□

Example 16.55 (April 2008 Exam, Q3(c)). Evaluate the following real integral using the method of contour integration:

$$\int_0^{2\pi} \frac{1 + \sin \theta}{2 + \cos \theta} d\theta. \quad (11 \text{ marks})$$

Solution. Let $z = e^{i\theta}$. Then

$$\begin{aligned}\int_0^{2\pi} \frac{1 + \sin \theta}{2 + \cos \theta} d\theta &= \int_{S^1} \frac{1 + \frac{z-z^{-1}}{2i}}{2 + \frac{z+z^{-1}}{2}} \frac{dz}{iz} = - \int_{S^1} \frac{2i + z - z^{-1}}{4z + z^2 + 1} dz = - \int_{S^1} \frac{2iz + z^2 - 1}{z(z^2 + 4z + 1)} dz \\ &= - \int_{S^1} \frac{2iz + z^2 - 1}{z(z + 2 - \sqrt{3})(z + 2 + \sqrt{3})} dz \\ &= -2\pi i \left[\frac{2iz + z^2 - 1}{(z + 2 - \sqrt{3})(z + 2 + \sqrt{3})} \Big|_{z=0} + \frac{2iz + z^2 - 1}{z(z + 2 + \sqrt{3})} \Big|_{z=-2+\sqrt{3}} \right] \\ &= 2\pi i \left[1 + \frac{2i(\sqrt{3}-2) + (\sqrt{3}-2)^2 - 1}{(2-\sqrt{3})(2\sqrt{3})} \right] = 2\pi i \left[1 + \frac{i(\sqrt{3}-2) + 3 - 2\sqrt{3}}{(2-\sqrt{3})(\sqrt{3})} \right] \\ &= 2\pi i \left[1 + \frac{i + \sqrt{3}}{-\sqrt{3}} \right] = \frac{2\pi i \times i}{-\sqrt{3}} = \frac{2\pi}{\sqrt{3}}\end{aligned}$$

□

16.5.2 Evaluation of Improper Integral of Rational Function on $\mathbb{R}$

In this section we shall demonstrate a technique in finding the improper integrals

$$\int_0^\infty f(x)dx \quad \text{and} \quad \int_{-\infty}^\infty f(x)dx \quad (16.7)$$

of mostly even¹ rational functions $f(x) = \frac{p(x)}{q(x)}$. Here, $q(z)$ has no real roots and has at least one zero above the real axis.

Theorem 16.56. When f is even and the Cauchy principal value exists, then both improper integrals (16.7) converge and

$$P.V. \int_{-\infty}^\infty f(x)dx = \int_{-\infty}^\infty f(x)dx = 2 \int_0^\infty f(x)dx.$$

Procedure of evaluating the improper integrals in (16.7):

- Step 1 : Identify all the distinct zeros, $z_1, \dots, z_n$, of $q(z)$ that lie above the real axis.
- Step 2 : Apply Cauchy's Residue Theorem to integrate the quotient $f(z) = \frac{p(z)}{q(z)}$ around the positively oriented boundary of the semicircular region D . The region D is encircled by a simple closed contour consists of the segment of the real axis from $z = -R$ to $z = R$

¹Recall from Calculus I, a real function f is said to be *even* if $f(-x) = f(x)$ for all x .

and the **upper half circle** C_R . The radius R is taken large enough so that the points $z_1, z_2, \dots, z_n$ all lie inside the closed path:

$$\boxed{\int_{-R}^R f(x)dx + \int_{C_R} f(z)dz = 2\pi i \sum_{k=1}^n \text{Res}(f; z_k).} \quad (*)$$

- Step 3: Verify that $\lim_{R \rightarrow \infty} \int_{C_R} f(z)dz = 0$. Then $(*)$ becomes $\int_{-R}^R f(x)dx = 2\pi i \sum_{k=1}^n \text{Res}(f; z_k)$.
- Step 4: If f is even, then $\int_0^\infty f(x)dx = \pi i \sum_{k=1}^n \text{Res}(f; z_k)$.

For the above procedure to work, we need $f(z)$ to go quickly to 0 when $|z|$ becomes large, so that the integral over the semicircle tends to 0 as the radius R becomes large.

Theorem 16.57. Suppose that there exists a number $B > 0$ such that for all $|z|$ sufficiently large, we have $|f(z)| \leq B|z|^{-2}$. Then

$$\lim_{R \rightarrow \infty} \int_{C_R} f(z)dz = 0.$$

Let us practise the procedure with examples below.

Example 16.58. Evaluate $\int_{-\infty}^{\infty} \frac{1}{1+x^4} dx$.

Solution: Let $f(z) = \frac{1}{1+z^4}$ and take C to be the closed curve consisting of the interval $[-R, R]$ on the x -axis and C_R , the upper semicircle of $|z| = R$ with $R > 1$.

1. Inside the region enclosed by C , $z^4 + 1 = 0$ (Topic 1, n th roots) has two solutions:

$$z = \exp \left[\frac{i(\pi + 2k\pi)}{4} \right] = \begin{cases} e^{\pi i/4}, & k = 0, \\ e^{3\pi i/4}, & k = 1 \end{cases} = \begin{cases} \frac{1+i}{\sqrt{2}}, \\ \frac{-1+i}{\sqrt{2}}. \end{cases}$$

2. By Residue Theorem,

$$\begin{aligned} \oint_C \frac{1}{1+z^4} dz &= 2\pi i \left\{ \text{Res}_{z=e^{\pi i/4}} \frac{1}{1+z^4} + \text{Res}_{z=e^{3\pi i/4}} \frac{1}{1+z^4} \right\} \\ &= 2\pi i \left\{ \lim_{z \rightarrow e^{\pi i/4}} (z - e^{\pi i/4}) \frac{1}{1+z^4} + \lim_{z \rightarrow e^{3\pi i/4}} (z - e^{3\pi i/4}) \frac{1}{1+z^4} \right\} \\ \text{Note 1: } \frac{1}{1+z^4} &= \frac{1}{z^2+i} \frac{1}{z^2-i} = \frac{1}{z - e^{\pi i/4}} \frac{1}{z - e^{5\pi i/4}} \cdot \frac{1}{z - e^{3\pi i/4}} \frac{1}{z - e^{7\pi i/4}} \\ \text{Note 2: Calculations are easier by applying and L'Hôpital Rule} \\ &= 2\pi i \left\{ \frac{1}{(e^{\pi i/4} - e^{5\pi i/4})((e^{\pi i/4})^2 + i)} + \frac{1}{(e^{3\pi i/4} - e^{7\pi i/4})((e^{3\pi i/4})^2 - i)} \right\} \\ &= 2\pi i \left\{ \frac{1}{(e^{\pi i/4} - e^{5\pi i/4})(i+i)} + \frac{1}{(e^{3\pi i/4} - e^{7\pi i/4})(-i-i)} \right\} \\ &= 2\pi i \left\{ \frac{e^{\pi i/4}}{2i(1-(-1))} - \frac{e^{3\pi i/4}}{2i(1-(-1))} \right\} = \frac{\pi}{2} \left\{ \frac{2}{\sqrt{2}} \right\} = \frac{\pi}{\sqrt{2}} \end{aligned}$$

3. By ML-inequality (Topic 4) and an inequality from Topic 1, $|z_1 - z_2| \geq |z_1| - |z_2| \Rightarrow$

$|1 + z^4| \geq |z^4| - 1$, we have

$$0 \leq \left| \int_{C_R} \frac{1}{1 + z^4} dz \right| \leq \underbrace{\max_{C_R} \left| \frac{1}{1 + z^4} \right|}_M \cdot \underbrace{\pi R}_L \leq \frac{\pi R}{R^4 - 1}.$$

When $R \rightarrow \infty$, the **squeeze theorem** (Calculus I) tells us that

$$\lim_{R \rightarrow \infty} \int_{C_R} \frac{1}{1 + z^4} dz = 0.$$

Therefore, $\int_{-\infty}^{\infty} \frac{1}{1 + x^4} dx = \frac{\pi}{\sqrt{2}}.$

Example 16.59. Evaluate $\int_{-\infty}^{\infty} \frac{1}{(1 + x^2)^2} dx.$

Solution: Let $f(z) = \frac{1}{(1 + z^2)^2}$ and consider the upper half plane with a contour of a semi-half circle $C := [-R, R] \cup C_R$. The function f has no singular points on C .

1. The pole of f in the upper half plane is $z^2 + 1 = 0 \Rightarrow z = i$.
2. According to Residue Theorem and formula (I6.2), we have

$$\begin{aligned} \oint_C \frac{1}{(1 + z^2)^2} dz &= 2\pi i \operatorname{Res}(f; i) = 2\pi i \lim_{z \rightarrow i} \frac{d}{dz} \left[\frac{(z - i)^2}{(1 + z^2)^2} \right] = 2\pi i \lim_{z \rightarrow i} \frac{d}{dz} \left[\frac{1}{(z + i)^2} \right] \quad (*) \\ &= 2\pi i \lim_{z \rightarrow i} -2(z + i)^{-3} = \frac{-4\pi i}{(2i)^3} = \frac{-4\pi i}{-8i} = \frac{\pi}{2} \end{aligned}$$

3. By ML-inequality

$$\left| \int_{C_R} \frac{1}{(1 + z^2)^2} dz \right| \leq \pi R \max_{C_R} \left| \frac{1}{(1 + z^2)^2} \right| \leq \pi R \frac{1}{(R^2 - 1)^2} \rightarrow 0$$

as $R \rightarrow \infty$. Hence (*) becomes $\int_{-\infty}^{\infty} \frac{1}{(1 + x^2)^2} dx = \frac{\pi}{2}.$

Example 16.60 (Sept 2010 Exam Q4(a)). Prove that $\int_{-\infty}^{\infty} \frac{1}{x^2 + a} dx = \frac{\pi}{\sqrt{a}}$ where $a > 0$. Hence deduce that $\int_{-\infty}^{\infty} \frac{dx}{(x^2 + 1)^2} = \frac{\pi}{2}.$ (10 marks)

Solution. (i) Let $f(z) = \frac{1}{z^2 + a}$ and consider the upper half plane with a contour of a semi-half circle $C := [-R, R] \cup C_R$. The function f has no singular points on C .

1. The pole of f is $z = \sqrt{a}i$ which is found in Example I6.59.
2. According to Residue Theorem and formula (I6.2), we have

$$\oint_C \frac{1}{z^2 + a} dz = 2\pi i \operatorname{Res}(f; \sqrt{a}i) = 2\pi i \lim_{z \rightarrow \sqrt{a}i} \frac{z - \sqrt{a}i}{z^2 + a} = 2\pi i \lim_{z \rightarrow \sqrt{a}i} \frac{1}{2z} = \frac{2\pi i}{2\sqrt{a}i} = \frac{\pi}{\sqrt{a}}. \quad (*)$$

3. By ML-inequality

$$\left| \int_{C_R} \frac{1}{z^2 + a} dz \right| \leq \pi R \max_{C_R} \left| \frac{1}{z^2 + a} \right| \leq \frac{\pi R}{R^2 - a} \rightarrow 0$$

as $R \rightarrow \infty$. Hence (*) becomes $\int_{-\infty}^{\infty} \frac{1}{x^2 + a} dx = \frac{\pi}{\sqrt{a}}$.

(ii) Differentiate with respect to “parameter” a :

$$\frac{d}{da} \int_{-\infty}^{\infty} \frac{1}{x^2 + a} dx = \frac{d}{da} \frac{\pi}{\sqrt{a}} \Rightarrow \int_{-\infty}^{\infty} \frac{d}{da} \left[\frac{1}{x^2 + a} \right] dx = \int_{-\infty}^{\infty} \frac{-1}{(x^2 + a)^2} dx = -\frac{\pi}{2} a^{-3/2}.$$

When $a = 1$, we obtain $\int_{-\infty}^{\infty} \frac{dx}{(x^2 + 1)^2} = \frac{\pi}{2}$. □

Remark 16.61. The following equality is due to the fact that the function $\frac{1}{x^2 + a}$ is uniformly continuous and differentiable over $\mathbb{R}$:

$$\frac{d}{da} \int_{-\infty}^{\infty} \frac{1}{x^2 + a} dx = \int_{-\infty}^{\infty} \frac{d}{da} \left[\frac{1}{x^2 + a} \right] dx.$$

For further knowledge, look up for “parametric integration techniques” or refer to Lang [1997].

Example 16.62. Evaluate (a) $\int_{-\infty}^{\infty} \frac{1}{1 + x^6} dx$; and (b) $\int_0^{\infty} \frac{1}{1 + x^6} dx$

Solution. (a) Consider $\oint_C \frac{1}{z^6 + 1} dz$ where C is the closed curve consisting the interval $[-R, R]$ on the x -axis and the upper semicircle of $|z| = R$, $R > 1$.

1. The function $f(z) = \frac{1}{z^6 + 1}$ has six simple poles:

$$z = e^{i\pi/6}, \quad e^{i3\pi/6}, \quad e^{i5\pi/6}, \quad e^{i7\pi/6}, \quad e^{i9\pi/6}, \quad e^{i11\pi/6}.$$

Three of these poles $e^{i\pi/6}$, $e^{i3\pi/6}$ and $e^{i5\pi/6}$ lie in the upper half-plane.

2. By the Residue Theorem and L'Hôpital Rule,

$$\oint_C \frac{1}{z^6 + 1} dz = 2\pi i \{ \text{Res}(f; e^{i\pi/6}) + \text{Res}(f; e^{i3\pi/6}) + \text{Res}(f; e^{i5\pi/6}) \}$$

$$\text{Res}(f; e^{i\pi/6}) = \lim_{z \rightarrow e^{i\pi/6}} \frac{z - e^{i\pi/6}}{z^6 + 1} = \frac{1}{6z^5} \Big|_{z=e^{i\pi/6}} = \frac{1}{6} e^{-5\pi i/6}$$

$$\text{Res}(f; e^{i3\pi/6}) = \lim_{z \rightarrow e^{i3\pi/6}} \frac{z - e^{i3\pi/6}}{z^6 + 1} = \frac{1}{6z^5} \Big|_{z=e^{i3\pi/6}} = \frac{1}{6} e^{-3\pi i/6}$$

$$\text{Res}(f; e^{i5\pi/6}) = \lim_{z \rightarrow e^{i5\pi/6}} \frac{z - e^{i5\pi/6}}{z^6 + 1} = \frac{1}{6z^5} \Big|_{z=e^{i5\pi/6}} = \frac{1}{6} e^{-\pi i/6}$$

Thus

$$\begin{aligned} \oint_C \frac{1}{z^6 + 1} dz &= 2\pi i \left\{ \frac{e^{-5\pi i/6}}{6} + \frac{e^{-3\pi i/6}}{6} + \frac{e^{-\pi i/6}}{6} \right\} = \frac{\pi i}{3} e^{-\pi i/6} \{ e^{-4\pi i/6} + e^{-2\pi i/6} + 1 \} \\ &= \frac{\pi i}{3} \left(\frac{\sqrt{3}}{2} - \frac{1}{2}i \right) (1 - \sqrt{3}i) = \frac{\pi i}{3} (-2i) = \frac{2\pi}{3}. \end{aligned}$$

3. By ML-inequality, we have

$$0 \leq \left| \int_{C_R} \frac{1}{1 + z^6} dz \right| \leq \underbrace{\max_{C_R} \left| \frac{1}{1 + z^6} \right|}_M \cdot \underbrace{\pi R}_L \leq \frac{\pi R}{R^6 - 1} \rightarrow 0 \quad \text{as } R \rightarrow \infty.$$

Taking limit $R \rightarrow \infty$, we have $\int_{-\infty}^{\infty} \frac{1}{1+x^6} dx = \frac{2\pi}{3}$.

(b) Notice that function f is even. Thus

$$\int_0^{\infty} \frac{1}{1+x^6} dx = \frac{1}{2} \int_{-\infty}^{\infty} \frac{1}{1+x^6} dx = \frac{\pi}{3}.$$

□

Example 16.63 (April 2010 Exam Q4(c)). Evaluate $\int_{-\infty}^{\infty} \frac{x^2}{x^4+1} dx$. (10 marks)

Solution. Consider the contour integral $I = \oint_C \frac{z^2}{z^4+1} dz$ where C is the closed curve consisting the interval $[-R, R]$ on the x -axis and $\tilde{C}_R$, the upper half of the circle $|z| = R$.

1. We the poles are $e^{\pi i/4}$ and $e^{3\pi i/4}$ from Example [16.58](#).
2. The residue of f at the poles inside C are (applying Theorem [16.35](#), i.e. If $f(z) = \frac{z^2}{z^4+1} = \frac{g(z)}{h(z)}$, then $\frac{g(z)}{h'(z)}$ can be use in finding residue)

$$\begin{aligned} \text{Res}(f; e^{\pi i/4}) &= \lim_{z \rightarrow e^{\pi i/4}} \frac{z^2}{4z^3} = \frac{1}{4z} \Big|_{e^{\pi i/4}} = \frac{e^{-\pi i/4}}{4} \\ \text{Res}(f; e^{3\pi i/4}) &= \frac{1}{4z} \Big|_{e^{3\pi i/4}} = \frac{e^{-3\pi i/4}}{4} \end{aligned}$$

Thus

$$I = \int_{-R}^R \frac{x^2}{1+x^4} dx + \int_{\tilde{C}_R} \frac{z^2}{z^4+1} dz = 2\pi i \left[\frac{\frac{e^{-\pi i/4}}{4} + e^{-3\pi i/4}}{4} \right] = \frac{\pi}{\sqrt{2}} \quad (*)$$

3. By ML-inequality

$$\left| \int_{\tilde{C}_R} \frac{z^2}{z^4+1} dz \right| \leq \pi R \max_{\tilde{C}_R} \left| \frac{z^2}{z^4+1} \right| \leq \pi R \frac{R^2}{R^4-1} \rightarrow 0 \quad \text{as } R \rightarrow \infty.$$

Hence, (*) becomes $\int_{-\infty}^{\infty} \frac{x^2}{x^4+1} dx = \frac{\pi}{\sqrt{2}}.$

□

Up to this point you might wonder the integral would take different value if the curve $\tilde{C}$ is formed by taking the line segment from R to $-R$ and $\tilde{C}_R$, the lower half of the circle $|z| = R$. Now

$$\begin{aligned} \oint_C \frac{z^2}{z^4+1} dz &= 2\pi i \left[\text{Res}\left(\frac{z^2}{z^4+1}; e^{5\pi i/4}\right) + \text{Res}\left(\frac{z^2}{z^4+1}; e^{7\pi i/4}\right) \right] \\ &= 2\pi i \left[\lim_{z \rightarrow e^{5\pi i/4}} \frac{1}{4z} + \lim_{z \rightarrow e^{7\pi i/4}} \frac{1}{4z} \right] = -\frac{\pi}{\sqrt{2}} \end{aligned}$$

This implies

$$\oint_C \frac{z^2}{z^4+1} dz = \int_R^{-R} \frac{x^2}{x^4+1} dx + \int_{\tilde{C}_R} \frac{z^2}{z^4+1} dz = -\frac{\pi}{\sqrt{2}}$$

Taking limit $R \rightarrow \infty$, we will arrive to the same conclusion that

$$\int_{-\infty}^{\infty} \frac{x^2}{x^4+1} dx = \frac{\pi}{\sqrt{2}}.$$

Example 16.64. Evaluate the integral $\int_0^\infty \frac{x^2}{x^6+1} dx$. [Remark: Using Calculus' u -transform is easier]

Solution. Consider the complex function $f(z) = \frac{z^2}{z^6+1}$. Function f has isolated singularities at the zeros of z^6+1 . These singularities are

$$z_k = \exp i \frac{\pi + 2k\pi}{6}, \quad k = 0, 1, 2, 3, 4, 5.$$

1. The first three singularities, namely

$$z_0 = e^{i\pi/6}, \quad z_1 = e^{i3\pi/6} = i, \quad z_2 = e^{i5\pi/6}$$

lie in the upper half plane. These singularities lie in the interior of the semicircular region bounded by the segment $z = x$ ($-R \leq x \leq R$, R large) of the real axis and the upper half C_R of the circle $|z| = R$.

2. By Residue Theorem and Theorem [16.35](#), we have

$$\begin{aligned} \int_{-R}^R \frac{x^2}{x^6+1} + \int_{C_R} f(z) dz &= 2\pi i \sum_{k=0}^2 \text{Res}(f; z_k) \\ &= 2\pi i \left[\lim_{z \rightarrow e^{i\pi/6}} \frac{z^2}{\frac{d}{dz}(z^6+1)} + \lim_{z \rightarrow e^{i3\pi/6}} \frac{z^2}{\frac{d}{dz}(z^6+1)} + \lim_{z \rightarrow e^{i5\pi/6}} \frac{z^2}{\frac{d}{dz}(z^6+1)} \right] \quad (*) \\ &= 2\pi i \left[\frac{(e^{i\pi/6})^{-3}}{6} + \frac{(e^{i3\pi/6})^{-3}}{6} + \frac{(e^{i5\pi/6})^{-3}}{6} \right] = 2\pi i \left[\frac{-i}{6} + \frac{i}{6} + \frac{-i}{6} \right] = \frac{\pi}{3}. \end{aligned}$$

3. By ML-inequality

$$\left| \int_{C_R} \frac{z^2}{z^6+1} dz \right| \leq \pi R \max_{C_R} \left| \frac{z^2}{z^6+1} \right| \leq \frac{\pi R \cdot R^2}{R^6-1} \rightarrow 0$$

as $R \rightarrow \infty$. Hence (*) becomes

$$\int_{-\infty}^{\infty} \frac{x^2}{x^6+1} dx = \frac{\pi}{3} \Rightarrow 2 \int_0^{\infty} \frac{x^2}{x^6+1} dx = \frac{\pi}{3} \Rightarrow \int_0^{\infty} \frac{x^2}{x^6+1} dx = \frac{\pi}{6}$$

□

Example 16.65 (Sept 2010 Exam Q4(a)). Prove that $\int_{-\infty}^{\infty} \frac{dx}{x^2+x+1} = \frac{2\pi}{\sqrt{3}}$. [Remark: Using Calculus' u -transform is easier]

Solution. Let $f(z) = \frac{1}{z^2+z+1}$ and $C = C_R \cup [-R, R]$.

1. The pole(s) of f in upper half plane is $z = \frac{-1 \pm \sqrt{1^2-4}}{2} = \frac{-1}{2} + \frac{\sqrt{3}}{2}i$.
2. By Residue Theorem, we have

$$\begin{aligned} \oint_C f(z) dz &= 2\pi i \text{Res}(f; \frac{-1}{2} + \frac{\sqrt{3}}{2}i) = 2\pi i \lim_{z \rightarrow \frac{-1}{2} + \frac{\sqrt{3}}{2}i} \frac{z - (\frac{-1}{2} + \frac{\sqrt{3}}{2}i)}{z^2 + z + 1} \\ &= 2\pi i \lim_{z \rightarrow \frac{-1}{2} + \frac{\sqrt{3}}{2}i} \frac{1}{2z+1} = \frac{2\pi i}{2(\frac{-1}{2} + \frac{\sqrt{3}}{2}i) + 1} = \frac{2\pi}{\sqrt{3}}. \end{aligned}$$

3. By ML-inequality

$$\left| \int_{C_R} \frac{1}{z^2 + z + 1} dz \right| \leq \pi R \max_{C_R} \left| \frac{1}{z^2 + z + 1} \right| \leq \frac{\pi R}{R^2 - \max_{C_R} |z + 1|} \leq \frac{\pi R}{R^2 - (R + 1)} \rightarrow 0$$

$$\text{as } R \rightarrow \infty. \text{ Hence, } \oint_C f(z) dz = \int_{-\infty}^{\infty} \frac{dx}{x^2 + x + 1} = \frac{2\pi}{\sqrt{3}}.$$

□

Example 16.66 (Sept 2011 Exam, Q4(b)). Prove that $\int_0^{\infty} \frac{1}{(x^2 + a^2)(x^2 + b^2)} dx = \frac{\pi}{2ab(a + b)}$ where $b > a > 0$. [**Remark:** Using Calculus' u -transform is easier] (10 marks)

Solution. Let $f(z) = \frac{1}{(z^2 + a^2)(z^2 + b^2)}$ and $C = [-R, R] \cup C_R$.

1. Since $b > a > 0$, the poles are $z = ai$ and $z = bi$.

2. By Residue Theorem,

$$\begin{aligned} \oint_C f(z) dz &= 2\pi i [\text{Res}(f; ai) + \text{Res}(f; bi)] \\ &= 2\pi i \left[\lim_{z \rightarrow ai} \frac{z - ai}{(z^2 + a^2)(z^2 + b^2)} + \lim_{z \rightarrow bi} \frac{z - bi}{(z^2 + a^2)(z^2 + b^2)} \right] \\ &= 2\pi i \left[\frac{1}{2z(z^2 + b^2) + 2z(z^2 + a^2)} \Big|_{z=ai} + \frac{1}{2z(z^2 + b^2) + 2z(z^2 + a^2)} \Big|_{z=bi} \right] \\ &= 2\pi i \left[\frac{1}{2ai(-a^2 + b^2)} + \frac{1}{2bi(-b^2 + a^2)} \right] = 2\pi i \left[\frac{\frac{1}{2ai} - \frac{1}{2bi}}{b^2 - a^2} \right] = \pi \left[\frac{b - a}{ab(b^2 - a^2)} \right] \end{aligned}$$

3. By ML-inequality,

$$\begin{aligned} \left| \int_{C_R} \frac{1}{(z^2 + a^2)(z^2 + b^2)} dz \right| &= \frac{1}{b^2 - a^2} \left| \int_{C_R} \left[\frac{1}{z^2 + a^2} - \frac{1}{z^2 + b^2} \right] dz \right| \\ &\leq \frac{1}{b^2 - a^2} \left[\left| \int_{C_R} \frac{1}{z^2 + a^2} dz \right| + \left| \int_{C_R} \frac{1}{z^2 + b^2} dz \right| \right] \\ &\leq \frac{1}{b^2 - a^2} \left[\frac{\pi R}{R^2 - a^2} + \frac{\pi R}{R^2 - b^2} \right] \rightarrow 0 \end{aligned}$$

$$\text{as } R \rightarrow \infty. \text{ Hence, } \int_{-\infty}^{\infty} \frac{1}{(x^2 + a^2)(x^2 + b^2)} dx = \frac{\pi}{ab(a + b)}$$

$$4. f \text{ is even, therefore, } \int_0^{\infty} \frac{1}{(x^2 + a^2)(x^2 + b^2)} dx = \frac{\pi}{2ab(a + b)}$$

□

16.5.3 Evaluation of Improper Integrals from Fourier Analysis

Residue theory can be useful in evaluating some improper integrals from Fourier Analysis of the form

$$\int_{-\infty}^{\infty} f(x) \sin mx dx \quad \text{or} \quad \int_{-\infty}^{\infty} f(x) \cos mx dx$$

where m denotes a positive integer.

Example 16.67 (Sept 2009 Exam, Q3(b)). Let R be a positive number. Let C be the positively oriented closed curve consisting of the portion of real axis from $-R$ to R and the upper half of the circle $|z| = R$. By integrating $f(z) = \frac{e^{i3z}}{(z^2 + 1)^2}$ over C , prove that

$$\int_{-\infty}^{\infty} \frac{\cos 3x}{(x^2 + 1)^2} dx = \frac{2\pi}{e^3}. \quad (15 \text{ marks})$$

Solution: A solution can be found the textbook Churchill and Brown [2008, 270].

1. There is only one pole in the upper half plane, i.e. $z = i$
2. By Residue Theorem,

$$\begin{aligned} \oint_C \frac{e^{i3z}}{(z^2 + 1)^2} dz &= 2\pi i \operatorname{Res}(f; i) = 2\pi i \lim_{z \rightarrow i} \frac{d}{dz} (z - i)^2 \frac{e^{i3z}}{(z^2 + 1)^2} = 2\pi i \lim_{z \rightarrow i} \frac{d}{dz} \frac{e^{i3z}}{(z + i)^2} \\ &= 2\pi i \lim_{z \rightarrow i} \left[3ie^{i3z}(z + i)^{-2} - 2e^{i3z}(z + i)^{-3} \right] = 2\pi i \left[3ie^{-3}(2i)^{-2} - 2e^{-3}(2i)^{-3} \right] \\ &= \frac{2\pi i}{e^3} \left[-\frac{3i}{4} - \frac{i}{4} \right] = \frac{2\pi}{e^3} \quad (*) \end{aligned}$$

3. By ML-inequality and $z = x + iy$, $y > 0$, $|e^{i3z}| = e^{-3y} < 1$, we have

$$\left| \int_{C_R} \frac{e^{i3z}}{(z^2 + 1)^2} dz \right| \leq \pi R \max_{C_R} \left| \frac{e^{i3z}}{(z^2 + 1)^2} \right| \leq \frac{\pi R e^{-3y}}{(R^2 - 1)^2} \rightarrow 0$$

as $R \rightarrow \infty$. Hence, (*) becomes $\int_{-\infty}^{\infty} \frac{\cos 3x}{(x^2 + 1)^2} dx = \frac{2\pi}{e^3}$.

The integral $\int_{-\infty}^{\infty} \frac{e^{itx}}{x^2 + 1} dx$ arises in probability theory as a scalar multiple of the characteristic function of the Cauchy distribution, which can be evaluated using the technique mentioned above.

Example 16.68. Evaluate $\int_{-\infty}^{\infty} \frac{\cos mx}{x^2 + 1} dx$ and $\int_{-\infty}^{\infty} \frac{\sin mx}{x^2 + 1} dx$.

Solution. Consider the function $f(z) = \frac{e^{imz}}{z^2 + 1}$ and let C denote the contour consisting $[-R, R]$ on the x -axis and C_R , the upper half of the circle $|z| = R$.

1. f has just a simple pole at $z = i$ lying inside C .
2. By Residue Theorem,

$$\int_C f(z) dz = 2\pi i \operatorname{Res}(f; i) = 2\pi i \lim_{z \rightarrow i} (z - i) \frac{e^{imz}}{(z - i)(z + i)} = 2\pi i \left[\frac{e^{-m}}{i + i} \right] = \pi e^{-m}.$$

Therefore

$$\int_C f(z) dz = \int_{-R}^R \frac{e^{imx}}{x^2 + 1} dx + \int_{C_R} \frac{e^{imz}}{z^2 + 1} dz = \pi e^{-m}$$

3. On the upper semicircle C_R , $z = x + iy$ with $y > 0$. By ML-Inequality,

$$\begin{aligned} \left| \int_{C_R} \frac{e^{imz}}{z^2 + 1} dz \right| &\leq \pi R \max_{C_R} \left| \frac{e^{imz}}{z^2 + 1} \right| \leq \frac{\pi R \max_{C_R} |e^{imx} \cdot e^{-my}|}{\min_{C_R} |z^2 + 1|} \\ &\leq \frac{\pi R e^{-my}}{R^2 - 1} \leq \frac{\pi R}{R^2 - 1} \rightarrow 0 \quad \text{as } R \rightarrow \infty. \end{aligned}$$

4. As $R \rightarrow \infty$, we obtain that

$$\lim_{R \rightarrow \infty} \int_{-R}^R \frac{e^{imx}}{x^2 + 1} dx + \lim_{R \rightarrow \infty} \int_{C_R} \frac{e^{imz}}{z^2 + 1} dz = \pi e^{-m} \Rightarrow \int_{-\infty}^{\infty} \frac{e^{imx}}{x^2 + 1} dx = \pi e^{-m}$$

Since $e^{imx} = \cos mx + i \sin mx$, therefore extracting real and imaginary parts gives us

$$\int_{-\infty}^{\infty} \frac{\cos mx}{x^2 + 1} dx = \pi e^{-m}, \quad \int_{-\infty}^{\infty} \frac{\sin mx}{x^2 + 1} dx = 0.$$

□

Remark 16.69. Note that $f(x) = \frac{\cos mx}{x^2 + 1}$ is a symmetric function. Hence

$$\int_0^{\infty} \frac{\cos mx}{x^2 + 1} dx = \frac{\pi e^{-m}}{2}.$$

eg: wwy:10.4.2

Example 16.70. Evaluate $\int_0^{\infty} \frac{x \sin x}{x^2 + 4} dx$.

Solution. Note that the integrand is even, we can write

$$\int_0^{\infty} \frac{x \sin x}{x^2 + 4} dx = \frac{1}{2} \int_{-\infty}^{\infty} \frac{x \sin x}{x^2 + 4} dx = \frac{1}{2} I.$$

Consider the contour integral where the contour $C = [-R, R] \cup C_R$ and choose $f(z) = \frac{ze^{iz}}{z^2 + 4}$ such that $f(x) = \frac{x(\cos x + i \sin x)}{x^2 + 4}$ when $x \in \mathbb{R}$.

1. f has just a pole $z = 2i$ on the upper half plane.

2. By Residue Theorem,

$$\begin{aligned} \oint_C \frac{ze^{iz}}{z^2 + 4} dz &= \int_{C_R} \frac{ze^{iz}}{z^2 + 4} dz + \int_{-R}^R \frac{xe^{ix}}{x^2 + 4} dx = 2\pi i \operatorname{Res}\left(\frac{ze^{iz}}{z^2 + 4}; 2i\right) \\ &= 2\pi i \lim_{z \rightarrow 2i} \frac{ze^{iz}}{z + 2i} = 2\pi i \frac{(2i)e^{-2}}{4i} = e^{-2}\pi i. \end{aligned} \quad (*)$$

3. By the inequality $|\int_C f(z) dz| \leq \int_C |f(z)| |dz|$ we have

$$\left| \int_{C_R} \frac{ze^{iz}}{z^2 + 4} dz \right| \leq \int_0^{\pi} \left| \frac{Re^{i\theta} e^{iRe^{i\theta}}}{R^2 e^{i2\theta} + 4} \right| R d\theta \leq \int_0^{\pi} \frac{e^{-R \sin \theta}}{1 - (\frac{2}{R})^2} d\theta = \frac{1}{1 - (\frac{2}{R})^2} \int_0^{\pi} e^{-R \sin \theta} d\theta$$

So by symmetry of sine function with respect to $x = \pi/2$, we have

$$\begin{aligned}\int_0^\pi e^{-R \sin \theta} d\theta &= \int_0^\delta e^{-R \sin \theta} d\theta + \int_\delta^{\pi-\delta} e^{-R \sin \theta} d\theta + \int_{\pi-\delta}^\pi e^{-R \sin \theta} d\theta \\ &= 2 \int_0^\delta \boxed{e^{-R \sin \theta}} d\theta + 2 \int_\delta^{\pi/2} e^{-R \sin \theta} d\theta \\ &\leq 2 \int_0^\delta 1 d\theta + 2 \int_\delta^{\pi/2} e^{-R \sin \delta} d\theta = \delta + 2e^{-R \sin \delta}(\pi - 2\delta)\end{aligned}$$

where $\delta > 0$ is an arbitrary small number, $\boxed{e^{-R \sin \delta} < 1}$ and $e^{-R \sin \theta} < e^{-R \sin \delta}$ when $\delta < \theta \leq \frac{\pi}{2}$.

When $R \rightarrow \infty$, $\sin \delta > 0$ and

$$e^{-R \sin \delta}(\pi - 2\delta) \rightarrow 0$$

Now, note that $\delta > 0$ is arbitrary, if $\int_0^\pi e^{-R \sin \theta} d\theta$ is not zero, then we can choose δ smaller than it, a contradiction. So it has to be zero and hence, (*) becomes

$$\Im\left(\int_{-\infty}^\infty \frac{x e^{ix}}{x^2 + 4} dx\right) = \Im\left(\frac{\pi}{e^2} i\right) \Rightarrow I = \frac{\pi}{e^2} \Rightarrow \int_0^\infty \frac{x \sin x}{x^2 + 4} dx = \frac{\pi}{2e^2}.$$

□

In the above example, we find that the estimate of $\int_0^\pi e^{-R \sin \theta} d\theta$ is important. Our proof is very complicated. In fact, it can be simplified if employed the following inequality.

Theorem 16.71 (Jordan's Inequality). For $x \in [0, \frac{\pi}{2}]$,

$$\frac{2x}{\pi} \leq \sin x \leq x. \quad (16.8)$$

e:ineq:jordan

Since e^x is an increasing function, therefore, for $R > 1$ and any $0 \leq x \leq \frac{\pi}{2}$, $\boxed{e^{-R \sin x} \leq e^{-\frac{2xR}{\pi}}}$ (16.8) $\Rightarrow$
 $-R \sin x \leq -\frac{2Rx}{\pi} \Rightarrow e^{-R \sin x} \leq e^{-\frac{2xR}{\pi}}$, which implies

$$\int_0^{\frac{\pi}{2}} e^{-R \sin x} dx \leq -\frac{\pi}{2R} \left[e^{-\frac{2Rx}{\pi}} \right]_0^{\frac{\pi}{2}} = \frac{\pi}{2R} (1 - e^{-R}) \leq \frac{\pi}{2R} \quad (16.9)$$

e:ineq:jordan2

If you are wondering what happens when we are integrating something with a singularity, the following is an example with a “removable singularity”. Even the “singularity” is removable, in order to use Residue Theorem (which is based on Cauchy Integral Formula from Topic 4), we need to “avoid” the singularity because the contour is not allowed to pass through the singularity.

Example 16.72. Show that $\int_0^\infty \frac{\sin x}{x} dx = \frac{\pi}{2}$.

Solution. Consider a contour C consists of

- two upper semicircles C_ϵ and C_R with radius $0 < \epsilon < R$ in positive direction; and
- line segments on the axes $[-R, -\epsilon]$ and $[\epsilon, R]$.

Since $z = 0$ is the only singularity of $f(z) = e^{iz}/z$ and is lying outside the contour C . Hence,

$$\oint_C \frac{e^{iz}}{z} dz = \int_{-R}^{-\epsilon} \frac{e^{ix}}{x} dx + \int_{-C_\epsilon} \frac{e^{iz}}{z} dz + \int_\epsilon^R \frac{e^{ix}}{x} dx + \int_{C_R} \frac{e^{iz}}{z} dz = 0. \quad (*)$$

Notice that by replacing x by $-x$ in the first integral in the middle becomes

$$\int_{-R}^{-\epsilon} \frac{e^{ix}}{x} dx = - \int_{\epsilon}^R \frac{e^{-ix}}{x} dx \Rightarrow \int_{-R}^{-\epsilon} \frac{e^{ix}}{x} dx + \int_{\epsilon}^R \frac{e^{ix}}{x} dx = 2i \int_{\epsilon}^R \frac{\sin x}{x} dx.$$

We can calculate the integral over C_{ϵ} with $z = \epsilon e^{i\theta}$, $0 \leq \theta \leq \pi$:

$$\int_{-C_{\epsilon}} \frac{e^{iz}}{z} dz = - \int_{C_{\epsilon}} \frac{e^{iz}}{z} dz = - \int_0^{\pi} \frac{e^{i\epsilon e^{i\theta}}}{\epsilon e^{i\theta}} i\epsilon e^{i\theta} d\theta = -i \int_0^{\pi} e^{i\epsilon e^{i\theta}} d\theta.$$

Since ϵ is suppose to arbitrary small, we want to estimate

$$\lim_{\epsilon \rightarrow 0} \int_0^{\pi} e^{i\epsilon e^{i\theta}} d\theta \stackrel{?}{=} \int_0^{\pi} e^{i0} d\theta = \pi.$$

To prove that the above equal is true, we look at

$$\left| \int_0^{\pi} e^{i\epsilon e^{i\theta}} d\theta - \pi \right| = \left| \int_0^{\pi} e^{i\epsilon e^{i\theta}} d\theta - \int_0^{\pi} 1 d\theta \right| = \left| \int_0^{\pi} (e^{i\epsilon e^{i\theta}} - 1) d\theta \right| \leq \max_{C_{\epsilon}} |e^{i\epsilon e^{i\theta}} - 1| \pi \quad (\dagger)$$

by ML-inequality.

Let us note that

$$0 \leq \theta \leq \pi \Rightarrow \epsilon \sin \theta \geq 0 \Rightarrow e^{-\epsilon \sin \theta} \leq 1 \Rightarrow 0 \leq 1 - e^{-\epsilon \sin \theta} \leq \epsilon \sin \theta \quad (\Delta)$$

where $f(x) = e^{-x} + x \Rightarrow f'(x) = 1 - e^{-x} \geq 0$ for $x \geq 0$, f is an increasing function with $f(0) = 0$, which means $f(x) \geq 1$. We will now estimate $(\dagger)$:

$$\begin{aligned} |e^{i\epsilon e^{i\theta}} - 1| &= |e^{i\epsilon \cos \theta - \epsilon \sin \theta} - e^{-\epsilon \sin \theta} + e^{-\epsilon \sin \theta} - 1| \\ &\leq |e^{i\epsilon \cos \theta - \epsilon \sin \theta} - e^{-\epsilon \sin \theta}| + |e^{-\epsilon \sin \theta} - 1| = e^{-\epsilon \sin \theta} |e^{i\epsilon \cos \theta} - 1| + |1 - e^{-\epsilon \sin \theta}| \\ &\leq |\cos(\epsilon \cos \theta) + i \sin(\epsilon \cos \theta) - 1| + \epsilon \sin \theta \\ &= \sqrt{(\cos(\epsilon \cos \theta) - 1)^2 + \sin^2(\epsilon \cos \theta)} + \epsilon \sin \theta = \sqrt{2 - 2 \cos(\epsilon \cos \theta)} + \epsilon \sin \theta \\ &\leq |\epsilon \cos \theta| + \epsilon \sin \theta \rightarrow 0 \quad \text{as } \epsilon \rightarrow 0 \end{aligned}$$

because $2 - 2 \cos x = 2(1 - \cos x) \leq 2(\frac{x^2}{2}) = x^2$. Again, $g(x) = \frac{x^2}{2} + \cos x - 1 \geq 0$ because $g(0) = 0$ and $g'(x) = x - \sin x > 0$ when $0 < x \leq \pi$.

By now, we confirm that $\lim_{\epsilon \rightarrow 0} \int_0^{\pi} e^{i\epsilon e^{i\theta}} d\theta = \pi$.

On C_R , $z = R e^{i\theta}$, $0 \leq \theta \leq \pi$. By applying the “Jordan’s inequality” ([le:ineq:jordan2](#) (16.9), we have

$$\begin{aligned} \left| \int_{C_R} \frac{e^{iz}}{z} dz \right| &= \left| \int_0^{\pi} \frac{e^{iR e^{i\theta}}}{R e^{i\theta}} iR e^{i\theta} d\theta \right| \leq \int_0^{\pi} |e^{iR \cos \theta - R \sin \theta}| d\theta = \int_0^{\pi} |e^{-R \sin \theta}| d\theta \\ &= 2 \int_0^{\pi/2} e^{-R \sin \theta} d\theta \leq \frac{\pi}{eR} \rightarrow 0 \quad \text{as } R \rightarrow \infty. \end{aligned}$$

So, as $\epsilon \rightarrow 0$ and $R \rightarrow \infty$, $(*)$ becomes $\oint_C \frac{e^{iz}}{z} dz = 2 \int_0^{\infty} \frac{i \sin x}{x} dx + (-i\pi) + 0 = 0 \Rightarrow \int_0^{\infty} \frac{\sin x}{x} dx = \frac{\pi i}{2i} = \frac{\pi}{2}$. □

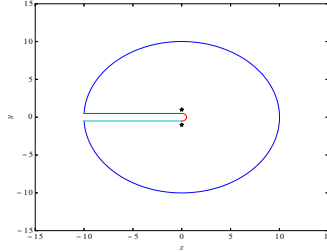
16.5.4 Evaluating Integrals with “Multi-Valued” Functions

Due to time constraint, it is impossible for us to discuss more complicated integrals with integrands involving n th root function, logarithmic function, x^{α} , etc. These examples, which

involve branch cut and carefully chosen contour, such as $\int_0^\infty \frac{x^{-a}}{x+1} dx$ ($0 < a < 1$) can be found in Churchill and Brown [2008, 280]. An example is given below to demonstrate the complexity of performing integration.

Example 16.73 (Wikipedia, Methods of Contour Integration). Evaluate $\int_0^\infty \frac{\ln x}{(1+x^2)^2} dx$.

Solution: Consider $f(z) = \left(\frac{\text{Log} z}{1+z^2}\right)^2$. In Topic 2, we learned that $\text{Log} z$ has a domain $\mathbb{C} \setminus (-\infty, 0]$ (due to the **branch cut**). So we will need to consider a contour $C = C_R \cup L_a \cup C_\epsilon \cup L_b$ which looked like a Pac-Man facing left:



where $C_R = \{\sqrt{R^2 + \epsilon^2}e^{it} : -\pi + \arctan \frac{\epsilon}{R} < t < \pi - \arctan \frac{\epsilon}{R}\}$,
 $L_a = \{-R(1-t) + \epsilon i : 0 \leq t \leq 1\}$,
 $C_\epsilon = \{\epsilon e^{it} : \pi/2 \geq t \geq -\pi/2\}$,
 $L_b = \{-Rt - \epsilon i : 1 \geq t \geq 0\}$.

1. The poles inside the domain are $z = \pm i$.

2. By Residue Theorem,

$$\begin{aligned} \oint_C f(z) dz &= 2\pi i \left\{ \text{Res}(f; i) + \text{Res}(f; -i) \right\} \\ &= 2\pi i \left\{ \lim_{z \rightarrow i} \frac{d}{dz} \left[(z-i)^2 \left(\frac{\text{Log} z}{1+z^2} \right)^2 \right] + \lim_{z \rightarrow -i} \frac{d}{dz} \left[(z+i)^2 \left(\frac{\text{Log} z}{1+z^2} \right)^2 \right] \right\} \\ &= 2\pi i \left\{ \left(\frac{2\text{Log} z}{z(z+i)^2} - \frac{2(\text{Log} z)^2}{(z+i)^3} \right) \Big|_{z=i} + \left(\frac{2\text{Log} z}{z(z-i)^2} - \frac{2(\text{Log} z)^2}{(z-i)^3} \right) \Big|_{z=-i} \right\} \\ &= 2\pi i \left\{ -\frac{\pi}{4} + \frac{\pi^2}{16}i + \left(-\frac{\pi}{4} - \frac{\pi}{16}i\right) \right\} = 2\pi i \times \left(-\frac{\pi}{2}\right) = -\pi^2 i \end{aligned}$$

3. By ML-inequality,

$$\begin{aligned} \left| \int_{C_R} \left(\frac{\text{Log} z}{1+z^2} \right)^2 dz \right| &\leq 2\pi \sqrt{R^2 + \epsilon^2} \left(\frac{\max_{C_R} |\text{Log} z|}{R^2 + \epsilon^2 - 1} \right)^2, \quad z = \sqrt{R^2 + \epsilon^2}e^{it} \\ &\leq 2\pi \sqrt{R^2 + \epsilon^2} \left[\frac{(\ln \sqrt{R^2 + \epsilon^2})^2 + \pi^2}{(R^2 + \epsilon^2 - 1)^2} \right] \rightarrow 0 \quad \text{as } R \rightarrow \infty \end{aligned}$$

$$\text{and } \left| \int_{C_\epsilon} \left(\frac{\text{Log} z}{1+z^2} \right)^2 dz \right| \leq \pi \epsilon \frac{(\ln \epsilon)^2 + \pi^2}{(1 - \epsilon^2)^2} \rightarrow 0 \quad \text{as } \epsilon \rightarrow 0.$$

Hence,

$$\begin{aligned}
\oint_C f(z)dz &= \lim_{\epsilon \rightarrow 0} \lim_{R \rightarrow \infty} \int_{-R}^0 \left[\frac{\text{Log}(x + i\epsilon)}{1 + (x + i\epsilon)^2} \right]^2 dx + \lim_{\epsilon \rightarrow 0} \lim_{R \rightarrow \infty} \int_0^{-R} \left(\frac{\text{Log}(x - i\epsilon)}{1 + (x - i\epsilon)^2} \right)^2 dx \\
&= \lim_{\epsilon \rightarrow 0} \lim_{R \rightarrow \infty} \left\{ - \int_R^0 \left[\frac{\text{Log}(-x + i\epsilon)}{1 + (-x + i\epsilon)^2} \right]^2 dx \right\} \\
&\quad + \lim_{\epsilon \rightarrow 0} \lim_{R \rightarrow \infty} \left\{ - \int_0^R \left[\frac{\text{Log}(-x - i\epsilon)}{1 + (-x - i\epsilon)^2} \right]^2 dx \right\} \\
&= \lim_{\epsilon \rightarrow 0} \int_0^\infty \left(\frac{\ln \sqrt{x^2 + \epsilon^2} + i\pi}{1 + (-x + i\epsilon)^2} \right)^2 dx - \lim_{\epsilon \rightarrow 0} \int_0^\infty \left(\frac{\ln \sqrt{x^2 + \epsilon^2} - i\pi}{1 + (-x - i\epsilon)^2} \right)^2 dx \\
&= \int_0^\infty \left(\frac{\ln \sqrt{x^2} + i\pi}{1 + x^2} \right)^2 dx - \int_0^\infty \left(\frac{\ln \sqrt{x^2} - i\pi}{1 + x^2} \right)^2 dx \\
&= \int_0^\infty \frac{(\ln |x|)^2 + 2\pi i \ln |x| - \pi^2}{(1 + x^2)^2} dx - \int_0^\infty \frac{(\ln |x|)^2 - 2\pi i \ln |x| - \pi^2}{(1 + x^2)^2} dx \\
&= \int_0^\infty \frac{(\ln |x|)^2 + 2\pi i \ln |x| - \pi^2 - (\ln |x|)^2 + 2\pi i \ln |x| + \pi^2}{(1 + x^2)^2} dx \\
&= \int_0^\infty \frac{4\pi i \ln |x|}{(1 + x^2)^2} dx = -\pi^2 i \Rightarrow \int_0^\infty \frac{\ln x}{(1 + x^2)^2} dx = \frac{-\pi^2 i}{4\pi i} = -\frac{\pi}{4}.
\end{aligned}$$

16.6 Argument Principle and Rouché Theorem

Finally, Cauchy's residue theorem is used to derive two important theorems. Before the derivation, we will define a special class of functions.

Definition 16.11. A function f is said to be *meromorphic* in a domain D if for every $z \in D$, z is either holomorphic or has a pole.

By definition, an entire function is a special case of meromorphic function. The rational functions mentioned in Topic 2 are meromorphic functions on $\mathbb{C}$.

Theorem 16.74 (Argument Principle). *If f is holomorphic and nonzero at each point of a simple closed positively oriented contour C and is meromorphic inside C , then*

$$\frac{1}{2\pi i} \oint_C \frac{f'(z)}{f(z)} dz = N_0(f) - N_p(f) \quad (16.10)$$

where $N_0(f)$ and $N_p(f)$ are, respectively, the number of zeros and poles of f inside C (multiplicity included).

Proof [Saff and Snider, 2003, 372]:

Let $G(z) = f'(z)/f(z)$. For each point $z \in C$, G is holomorphic because f is assumed to be non-zero on C .

Inside C , the singularities of G are the points where f has a zero or a pole. Consider a point z_0 inside C that is a zero of order m (Definition 16.8), then we know that $f(z) = (z - z_0)^m h(z)$, where h is analytic and $h(z_0) \neq 0$. Hence, in some deleted neighbourhood of z_0 , we have

$$G(z) = \frac{m}{z - z_0} + \frac{h'(z)}{h(z)}.$$

e:argument-princ

def:zero-order-n

Since the function h'/h is analytic at z_0 , the above formula shows that G has a simple pole at z_0 with residue equal to m .

Consider a point z_p inside C that is a pole of order k (Definition [16.4](#)), then we know that

$$f(z) = \frac{H(z)}{(z - z_0)^k}$$

where H is analytic at z_p and $H(z_p) \neq 0$. Hence, in some deleted neighbourhood of z_p , we have

$$G(z) = \frac{-k}{z - z_p} + \frac{H'(z)}{H(z)}.$$

Since the function H'/H is analytic at z_p , we find that z_p is a simple pole of G .

By residue theorem,

$$\oint_C G(z)dz = \sum_{z \text{ is a pole or a zero of } f} 2\pi i \times (\text{order of } z) \Rightarrow \oint_C G(z)dz = 2\pi i[N_0(f) - N_p(f)].$$

Now, suppose that a function f is holomorphic on C and meromorphic inside C . With the Argument Principle, we know how many times $f(C)$ winds around the origin $w = 0$. If f gets “minor changes” to $g(z) = f(z) + h(z)$, where h is also holomorphic. We would like to know how small the “changes” h be to guarantee that $g(C)$ winds around the origin the same number of times as $f(C)$. This was answered by the following theorem.

Theorem 16.75 (Rouché Theorem). *If f and h are each functions that are holomorphic inside and on a simple closed contour C and if the strict inequality*

$$|h(z)| < |f(z)|$$

holds at each point on C , then f and $f + h$ must have the same total number of zeros (counting multiplicities) inside C .

Example 16.76. (Saff and Snider [2003]) Prove that all five zeros of $g(z) = z^5 + 3z + 1$ lie in the disk $|z| < 2$.

Solution. First, we know that $f(z) = z^5$ has 5 roots in the disk $|z| < 2$.

So let's try to study $h(z) = 3z + 1$ to see if $|h(z)| < |f(z)|$ at each point on $|z| = 2$:

$$|3z + 1| \leq 3|z| + 1 = 3(2) + 1 = 7 < |z^5| = |z|^5 = 32.$$

The inequality is satisfied and by Rouché Theorem, all the roots of g are all inside the disk $|z| < 2$. □

Refer to the following for more examples:

HenriciP-AppliedAndComputationalComplexAnalysis-Vol3-Wiley1986
MitrinovicDS-KeckicJD-TheCauchyMethodOfResiduesTheoryAndAppl1-1984

Chapter 17

Harmonic Function

A real-valued C^2 function on an open subset $D \subset \mathbb{C}$ is called *harmonic* if it satisfies the Laplace equation (Topic 2 Section 2.3.4). We have seen in Topic 2 that both the real and imaginary parts of a holomorphic function are harmonic. In this topic we will discuss more examples on how to find the harmonic conjugates as well as mentioning that harmonic functions share similar properties to holomorphic functions such as the mean value property. We also explained how complex functions could be used to solve very simple Dirichlet problem.

The study of harmonic functions leads to the development of potential theory, which can be found in Ransford [1995], Evans [1927], Berenstein and Gay [1991], Helms [1969] and Doob [2001]. Potential theory has applications in the theory of partial differential equations (UECM3003) and some special type of harmonic functions have been applied in the *morphing* of two- and three-dimensional *geometric objects*.

17.1 $\frac{\partial}{\partial z}$ and $\frac{\partial}{\partial \bar{z}}$

In this section we shall present another representation of Cauchy-Riemann equations and Laplace equation. Let $z = x + iy$. The “complex variable” pair $(z, \bar{z})$ can be treated as two independent variables. They are related to the “real variable” pair x and y by $x = \frac{z + \bar{z}}{2}$ and $y = \frac{z - \bar{z}}{2i}$. To see how a function change with respect to the complex variable pair and real variable pair, we applying the transformation rules from Calculus:

$$\begin{cases} \frac{\partial f}{\partial x} = \frac{\partial f}{\partial z} \frac{\partial z}{\partial x} + \frac{\partial f}{\partial \bar{z}} \frac{\partial \bar{z}}{\partial x} = \frac{\partial f}{\partial z} + \frac{\partial f}{\partial \bar{z}} \\ \frac{\partial f}{\partial y} = \frac{\partial f}{\partial z} \frac{\partial z}{\partial y} + \frac{\partial f}{\partial \bar{z}} \frac{\partial \bar{z}}{\partial y} = i \frac{\partial f}{\partial z} - i \frac{\partial f}{\partial \bar{z}} \end{cases} \Rightarrow \begin{bmatrix} \frac{\partial f}{\partial z} \\ \frac{\partial f}{\partial \bar{z}} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & -i \\ 1 & i \end{bmatrix} \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}$$

Here, $\frac{\partial}{\partial z}$ and $\frac{\partial}{\partial \bar{z}}$ are considered as the symbolic derivatives with respect to z and $\bar{z}$, respectively. Hence, we define the “complex variable” derivative operators formally below.

Definition 17.1. $\frac{\partial}{\partial z} = \frac{1}{2} \left[\frac{\partial}{\partial x} - i \frac{\partial}{\partial y} \right]$, $\frac{\partial}{\partial \bar{z}} = \frac{1}{2} \left[\frac{\partial}{\partial x} + i \frac{\partial}{\partial y} \right]$, $dz = dx + idy$, $d\bar{z} = dx - idy$. The operator $\frac{\partial}{\partial \bar{z}}$, abbreviated as $\bar{\partial}$ is called the *Cauchy-Riemann operator*.

By applying the Cauchy-Riemann equations, we can obtain the *homogeneous Cauchy-Riemann equation* [Berenstein and Gay, 1991]:

$$\frac{\partial f}{\partial \bar{z}} = \frac{1}{2} \left[\frac{\partial}{\partial x}(u + iv) + i \frac{\partial}{\partial y}(u + iv) \right] = \frac{1}{2} \left[\frac{\partial u}{\partial x} + i \frac{\partial v}{\partial x} + i \left(\frac{\partial u}{\partial y} - \frac{\partial v}{\partial y} \right) \right] = \frac{1}{2}(0 + 0i) \Rightarrow \boxed{\frac{\partial f}{\partial \bar{z}} = 0}.$$

This implies that a holomorphic function is independent of variable $\bar{z}$ and this is a reason why holomorphic functions are so special compare to two-variable real functions [Ahlfors, 1979].

Using this formulation, it is easy to see why $f(z) = |z|^2$ is not holomorphic:

$$\frac{\partial f}{\partial \bar{z}} = \frac{\partial}{\partial \bar{z}}(z\bar{z}) = z \neq 0$$

except for $z = 0$. Even if f is complex differentiable at $z = 0$, there is no neighbourhood for it to be complex differentiable, so f is nowhere holomorphic. Berenstein and Gay [1991] uses $\bar{\partial}$ to develop the whole complex analysis. However, the languages of algebraic topology and differential calculus were used which made this book too abstract for undergraduate study.

By calculation, it is possible to show that $4\frac{\partial}{\partial z}\frac{\partial}{\partial \bar{z}} = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$. Therefore, with the above notions, it is possible to provide an alternative definition for harmonic functions.

Definition 17.2. A two-variable function h is called *harmonic* if it satisfies the Laplace equation

$$\frac{\partial}{\partial z}\frac{\partial}{\partial \bar{z}}h(z, \bar{z}) = 0.$$

17.2 Harmonic Conjugate

We know that the real and imaginary parts of an holomorphic function are harmonic. But if we are given a harmonic function $h(x, y)$, is it possible to find a harmonic conjugate $v(x, y)$ so that $h(x, y) + iv(x, y)$ becomes a holomorphic function? In the section 17.2.1, we will learn that if the domain D is simply-connected, every real-valued harmonic function h admits a harmonic conjugate, i.e., a real-valued function v , for which $h + iv$ is holomorphic. In section 17.2.2, we will practise with some examples to find harmonic conjugate.

17.2.1 Relation with Holomorphic Functions

In Topic 2, we learned that the real and imaginary parts of a holomorphic function are harmonic. But is every harmonic function either the real or imaginary part of a holomorphic function? A confirmation is available below when the domain is “nice”.

Theorem 17.1. Let h be a function harmonic on a simply connected domain D . Then there is a holomorphic function f such that $h = \Re(f)$ on D .

For the harmonic function h on a region D , how can we find another harmonic function v on D such that $f(x + iy) = h(x, y) + iv(x, y)$ becomes a holomorphic function on D ? If such a function v exists, it is called a “harmonic conjugate” of h . It is defined in Topic 2, we will recall definition below.

Definition 17.3. If $u(x, y)$ and $v(x, y)$ are harmonic functions defined in domain D and satisfy the Cauchy-Riemann equations throughout D , then we say that v is a *harmonic conjugate* of u .

Example 17.2. Show that $\text{Arg}(z)$ is a harmonic conjugate of $\ln|z|$ on $\mathbb{C} \setminus (-\infty, 0]$.

Proof: We learned in Topic 2 that $\text{Log}z = \ln|z| + i\text{Arg}(z)$ is holomorphic in the domain $D = \mathbb{C} \setminus (-\infty, 0]$. Therefore, both $\ln|z|$ and $\text{Arg}(z)$ are harmonic in D . Since $u = \ln|z|$ and $v = \text{Arg}(z)$ satisfied the C-R equations according to Topic 2, $\text{Arg}(z)$ is a harmonic conjugate of $\ln|z|$.

It is impossible to find harmonic conjugate for any harmonic function. For example, the harmonic function $h = \ln|z|$ in Example 17.2 is harmonic on $D = \mathbb{C}^\times$ but h **does not have** a

harmonic conjugate on $\mathbb{C}^\times$ (we learned that $\text{Arg}(z)$ is not continuous on $(-\infty, 0]$ in Topic 1). However, mathematicians found that it is possible only for a special type of domain according to the following important theorem from the “Calculus of Manifolds”.

Theorem 17.3 (Harmonic Conjugate Existence Theorem). *Suppose that D is an open and connected region. Then every harmonic function u on D has a harmonic conjugate v on D if and only if D is simply connected.*

Is the harmonic conjugate v unique? Mathematicians found that on any given simply connected region, v is unique up to an additive constant.

If v is a harmonic conjugate of h , is h a harmonic conjugate of v ? The answer is negative. That means even if $f(z) = h + iv$ is holomorphic, $g(z) = v + ih$ may not be holomorphic as demonstrated by the following example.

eg:wwy:3.6.2

Example 17.4. Consider $f(z) = z^2 = (x^2 - y^2) + i(2xy) =: u + iv$. Show that v is a harmonic conjugate of u but u is not a harmonic conjugate of v .

Solution. (a) We know that $f(z) = z^2$ is holomorphic from Topic 2. Therefore u and v satisfy the Cauchy-Riemann equations and so v is a harmonic conjugate of u .

(b) But for $g(z) = U + iV = 2xy + i(x^2 - y^2)$ is no longer holomorphic because

$$U_x = 2y, \quad V_y = -2y, \quad U_y = 2x, \quad V_x = 2x$$

Thus Cauchy-Riemann equations are satisfied only when $x = y = 0$. □

So if v is the harmonic conjugate of h , what is the harmonic conjugate of v ? It is not difficult to confirm this: $-h$ is the harmonic conjugate of v . Furthermore, h and v are linked geometrically as stated by the following theorem.

Theorem 17.5. *If h is harmonic and v is a harmonic conjugate of h , then the level curves $h(x, y) = C$ and $v(x, y) = \tilde{C}$ are mutually orthogonal, where C and $\tilde{C}$ are some constants.*

17.2.2 Finding Harmonic Conjugate

sec:examples

How can we find a harmonic conjugate for a given harmonic function h ? Let v be a harmonic conjugate of h . Then h and v should satisfy the Cauchy-Riemann equations and so we find v by following these steps:

1. $v = \int h_x dy + f(x)$ (or $v = -\int h_y dx + g(y)$, depending which is easier to integrate);
2. $v_x = \frac{d}{dx} \int h_x dy + f'(x) = -h_y$;
3. Integrate $f'(x)$ to obtain $f(x)$ and $v = \int h_x dy + f(x)$.

Example 17.6. Show that $h(x, y) = x^2 - y^2 + x$ is harmonic on $\mathbb{C}$ and find a harmonic conjugate.

Solution: (a) Since $h_x = 2x + 1 \Rightarrow h_{xx} = 2$, $h_y = -2y \Rightarrow h_{yy} = -2$ and $h_{xx} + h_{yy} = 0$ on $\mathbb{C}$, h is harmonic on $\mathbb{C}$.

(b) By using C-R equations, we have

$$\begin{aligned} h_x = 2x + 1 = v_y &\Rightarrow v = (2x + 1)y + f(x) \\ h_y = -2y = -v_x &= -(2y + f'(x)) \Rightarrow -f'(x) = 0 \Rightarrow f(x) = C. \end{aligned}$$

Therefore, the harmonic conjugate of h is $v = 2xy + y + C$.

asmar2005:p654

Example 17.7 (April 2010 Exam, Q5(b)). Let $u(x, y) = x^3 - 3xy^2$.

- (i) Verify that the function u is harmonic (on $\mathbb{C}$). (3 marks)

Solution: $u_{xx} = 6x$ and $u_{yy} = -6x$. Then $u_{xx} + u_{yy} = 0$ and u is harmonic.

- (ii) Find a function v such that $f(z) = u + iv$ is analytic. (6 marks)

Solution:

$$u_x = 3x^2 - 3y^2 = v_y \Rightarrow v = 3x^2y - y^3 + f(x)$$

$$u_y = -6xy = -v_x = -(6xy + f'(x)) \Rightarrow f'(x) = C$$

Therefore $v = 3x^2y - y^3 + C$.

Example 17.8 (Sept 2010 Exam, Q3(a)). Let $u(x, y) = y^3 - 3x^2y$, where x, y are real variables. Show that u is harmonic and find a real valued function $v(x, y)$ such that $f(x + iy) = u(x, y) + iv(x, y)$ is analytic (holomorphic). (10 marks)

Solution. (a) $u_x = -6xy \Rightarrow u_{xx} = -6y$.

$$u_y = 3y^2 - 3x^2 \Rightarrow u_{yy} = 6y.$$

Since $u_{xx} + u_{yy} = 0$, u is harmonic.

$$(b) u_x = -6xy = v_y \Rightarrow v = -3xy^2 + f(x)$$

$$u_y = 3y^2 - 3x^2 = -v_x = -(-3y^2 + f'(x)) \Rightarrow f'(x) = 3x^2 \Rightarrow f(x) = x^3 + C.$$

The harmonic conjugate $v(x, y) = -3xy^2 + x^3 + C$. □

Example 17.9 (April 2008 Exam, Q2(b)).

- (i) Prove that the function $\varphi(x, y) = e^{-x}(x \cos y + y \sin y)$ is harmonic. (4 marks)

Solution.

$$\frac{\partial \varphi}{\partial x} = (-e^{-x}x + e^{-x}) \cos y - e^{-x}y \sin y \Rightarrow \frac{\partial^2 \varphi}{\partial x^2} = (e^{-x}x - 2e^{-x}) \cos y + e^{-x}y \sin y$$

$$\frac{\partial \varphi}{\partial y} = -e^{-x}x \sin y + e^{-x}(\sin y + y \cos y) \Rightarrow \frac{\partial^2 \varphi}{\partial y^2} = e^{-x}(-y \sin y - x \cos y + 2 \cos y)$$

Hence $\frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} = 0$ and so φ is harmonic. □

- (ii) Find an analytic (holomorphic) function f of $z = x + iy$ such that the real part of f is $\varphi(x, y)$. (6 marks)

Solution. By definition, $\varphi(x, y) = \Re(f(x + iy)) =: u$. We need to find a harmonic conjugate $v(x, y)$ and $f(z)$ such that $f(x + iy) = \varphi(x, y) + iv(x, y)$.

$$u_x = (-e^{-x}x + e^{-x}) \cos y - e^{-x}y \sin y = v_y$$

$$\Rightarrow v = (-e^{-x}x + e^{-x}) \sin y - e^{-x} \left[\sin y - y \cos y \right] + f(x)$$

$$u_y = -e^{-x}x \sin y + e^{-x}(\sin y + y \cos y) = (-e^{-x}x + e^{-x}) \sin y + e^{-x}y \cos y$$

$$= -v_x = - \left[(e^{-x}x - 2e^{-x}) \sin y + e^{-x} \left[\sin y - y \cos y \right] + f'(x) \right] \Rightarrow f'(x) = 0$$

Hence, $v = e^{-x}y \cos y - xe^{-x} \sin y + C$ and

$$f(z) = e^{-x}x \cos y + e^{-x}y \sin y + ie^{-x}y \cos y - ix e^{-x} \sin y + C$$

$$= e^{-x}((x + iy) \cos y - i(x + iy) \sin y) = ze^{-x}(\cos y - i \sin y) = ze^{-x}e^{-iy} = ze^{-z}.$$

17.3 Mean Value Property

Harmonic functions share many properties of holomorphic functions (Topic 4), especially the mean value property and the maximum value principle.

For $(x, y) \in \mathbb{R}^2 \setminus \{(x_0, y_0)\}$, define

$$v(x, y) = \frac{1}{2} \ln((x - x_0)^2 + (y - y_0)^2).$$

thm:log:normal

Proposition 17.1 (Normal Derivative of the Logarithm). *Recall the notion of $C_r(z_0) = \{(x, y) \in \mathbb{C} : (x - x_0)^2 + (y - y_0)^2 = r^2\}$. For $(x, y) \in C_r((x_0, y_0))$, we have*

$$\frac{\partial v}{\partial n}(x, y) = \frac{1}{r}.$$

Proof: Any point on $C_r((x_0, y_0))$ has the form $(x, y) = (x_0 + r \cos t, y_0 + r \sin t)$, $0 \leq t \leq 2\pi$.

From Calculus 3, the unit normal vector on C_r is given by

$$\nabla \sqrt{(x - x_0)^2 + (y - y_0)^2} = \left(\frac{\frac{1}{2} \times 2(x - x_0)}{\sqrt{(x - x_0)^2 + (y - y_0)^2}}, \frac{\frac{1}{2} \times 2(y - y_0)}{\sqrt{(x - x_0)^2 + (y - y_0)^2}} \right)$$

and

$$\begin{aligned} \frac{\partial v}{\partial n} \Big|_{(x,y) \in C_r} &= \nabla v \cdot \nabla \sqrt{(x - x_0)^2 + (y - y_0)^2} \\ &= \frac{(x - x_0)^2}{[(x - x_0)^2 + (y - y_0)^2]^{3/2}} + \frac{(y - y_0)^2}{[(x - x_0)^2 + (y - y_0)^2]^{3/2}} \\ &= \frac{1}{\sqrt{(x - x_0)^2 + (y - y_0)^2}} = \frac{1}{r} \quad \text{since } (x - x_0)^2 + (y - y_0)^2 = r^2 \text{ on } C_r((x_0, y_0)) \end{aligned}$$

where

$$\nabla v = \frac{1}{2} \left[\frac{2(x - x_0)}{(x - x_0)^2 + (y - y_0)^2} + \frac{2(y - y_0)}{(x - x_0)^2 + (y - y_0)^2} \right]$$

Proposition 17.2 (Harmonicity of the Logarithm). *The function $v(x, y)$ is harmonic at all points $(x, y) \neq (x_0, y_0)$.*

Proof: Note that in Example [17.2](#), $\ln |z|$ is only defined on $\mathbb{C} \setminus (-\infty, 0]$. Here, v is defined on $\mathbb{C}^\times$:

$$\begin{aligned} \frac{\partial v}{\partial x} &= \frac{x - x_0}{(x - x_0)^2 + (y - y_0)^2} \Rightarrow \frac{\partial^2 v}{\partial x^2} = \frac{(x - x_0)^2 + (y - y_0)^2 - (x - x_0)[2(x - x_0)]}{[(x - x_0)^2 + (y - y_0)^2]^2} \\ &= \frac{(y - y_0)^2 - (x - x_0)^2}{[(x - x_0)^2 + (y - y_0)^2]^2}. \end{aligned}$$

Similarly,

$$\frac{\partial v}{\partial y} = \frac{y - y_0}{(x - x_0)^2 + (y - y_0)^2} \Rightarrow \frac{\partial^2 v}{\partial y^2} = \frac{(x - x_0)^2 - (y - y_0)^2}{[(x - x_0)^2 + (y - y_0)^2]^2}$$

Therefore, $v_{xx} + v_{yy} = 0$ on $\mathbb{C}^\times$.

Suppose h is harmonic on D and $z_0 \in D$ is arbitrary. If $r > 0$ is such that $\overline{\Delta_r(z_0)} \subset D$, then we can choose a disk B with $\overline{\Delta_r(z_0)} \subset B \subset D$, and since B is simply-connected $h = \Re(f)$ on B for some holomorphic function f on B . Cauchy's integral formula then gives

$$f(z_0) = \frac{1}{2\pi i} \oint_{|z - z_0| = r} \frac{f(z)}{z - z_0} dz = \frac{1}{2\pi} \int_0^{2\pi} f(z_0 + re^{i\theta}) d\theta.$$

Taking real parts we get the Mean Value Theorem.

Theorem 17.10 (Mean Value Theorem). *If $h : D \rightarrow \mathbb{R}$ is harmonic and $\overline{\Delta_r(z_0)} \subset D$, then*

$$h(x, y) = \frac{1}{2\pi} \int_0^{2\pi} h(x + r \cos \theta, y + r \sin \theta) d\theta. \quad (17.1)$$

e:mvp

Proof: Denote the right side as $\phi(r)$ where $0 \leq r \leq a$ where a is the maximum distance between (x, y) and ∂D .

When $r = 0$, $\phi(0) = \frac{1}{2\pi} \int_0^{2\pi} h(x, y) d\theta = h(x, y)$.

Since ϕ is an integral of the continuous function h , ϕ is a continuous function of r on $[0, a]$. If we can show that ϕ is constant over $[0, a]$, then $\phi(r) = \phi(0) = h(x, y)$.

First, we show that $\phi(r_1) = \phi(r_2)$ for all $0 < r_1 < r_2 \leq a$.

$$\begin{aligned} \frac{1}{r_2} \oint_{C_{r_2}} h ds - \frac{1}{r_1} \oint_{C_{r_1}} h ds &= \oint_{C_{r_2}} h \frac{\partial v}{\partial n} ds - \oint_{C_{r_1}} h \frac{\partial v}{\partial n} ds && [\text{Theorem 17.1}] \\ &= \oint_{\Gamma_{r_1, r_2}} h \frac{\partial v}{\partial n} ds = \oint_{\Gamma_{r_1, r_2}} v \frac{\partial h}{\partial n} ds && [\text{Green2}] \\ &= \oint_{C_{r_2}} v \frac{\partial h}{\partial n} ds - \oint_{C_{r_1}} v \frac{\partial h}{\partial n} ds = \ln r_2 \oint_{C_{r_2}} \frac{\partial h}{\partial n} ds - \ln r_1 \oint_{C_{r_1}} \frac{\partial h}{\partial n} ds \\ &= \ln r_2 \times 0 - \ln r_1 \times 0 = 0 && \left[\frac{\partial h}{\partial n} \text{ is conservative} \right] \end{aligned}$$

where Γ_{r_1, r_2} is the boundary of an annulus of radius r_1 and r_2 . The direction points towards the annulus. The label [Green2] denotes the Green's Second Identity Theorem from Calculus 3:

$$\iint_{\Omega} (u \Delta v - v \Delta u) dx dy = \oint_{\Gamma} \left(u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \right) ds$$

where u and v are functions on the domain Ω . Note that when $u = h$ and v are both harmonic, $\Delta u = 0$ and $\Delta v = 0$, the left hand side of Green's Second Identity Theorem becomes zero.

Remark 17.11. Using the notion of complex analysis, (17.1) can be written in a “shorter form”:

$$h(z) = \frac{1}{2\pi} \int_0^{2\pi} h(z + re^{i\theta}) d\theta. \quad (17.2)$$

From this theorem, it is possible to derive some integration formula which is difficult to obtain using the techniques we learned in Calculus 1 to 3 as demonstrated by the following example.

Example 17.12. Show that for any $a, b \in \mathbb{R}$,

$$\int_0^{2\pi} e^{b+\sin \theta} \cos(a + \cos \theta) d\theta = 2\pi e^b \cos a.$$

Solution. It is easy to show that $h(x, y) = e^y \cos x$ is harmonic for $(x, y) \in \mathbb{R}^2$. By the mean value theorem, we have

$$h(a, b) = e^b \cos a = \frac{1}{2\pi} \int_0^{2\pi} h(a + r \cos \theta, b + r \sin \theta) d\theta$$

which simplifies to the equation below

$$e^b \cos a = \frac{1}{2\pi} \int_0^{2\pi} e^{b+r \sin \theta} \cos(a + r \cos \theta) d\theta.$$

Taking $r = 1$, we obtain our result. □

The mean value property characterises harmonic functions.

Theorem 17.13. *If $u : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ is a continuous function satisfying the mean value property, then it is harmonic.*

A harmonic function satisfies the “maximum value principle”: it does not assume its maximum on a region unless it is constant there. Indeed the conclusion is true for any continuous function with the mean-value property, or even just satisfying the inequality

$$h(x, y) \leq \frac{1}{2\pi} \int_0^{2\pi} h(x + r \cos \theta, y + r \sin \theta) d\theta.$$

The reverse inequality similarly shows that a harmonic function satisfies the “minimum value principle”.

thm:max-min

Theorem 17.14 (Maximum-Minimum (Value) Principle). *Let D be a domain of a harmonic function h .*

1. *If h attains a maximum or a minimum inside D , then h is a constant on D .*
2. *If h extends continuously to $\overline{D}$ and $h \leq 0$ on ∂D , then $h \leq 0$ on D .*

Proof: It is enough to prove the statement concerning the maximum of h . If we let $v = -h$, we have the v is harmonic and a minimum value of h corresponds to the maximum value of v .

Suppose that h attains a maximum value $(x_0, y_0) \in D$.

Step 1: Let C_r be any positively oriented circle in D with a centre at (x_0, y_0) . By assumption, $M = h(x_0, y_0) \geq h(x, y)$ for all $(x, y) \in D$. By Mean Value Theorem, we have

$$M = h(x_0, y_0) = \frac{1}{2\pi} \int_0^{2\pi} h(x_0 + r \cos \theta, y_0 + r \sin \theta) d\theta.$$

Consider the continuous function $f(\theta) = h(x_0 + r \cos \theta, y_0 + r \sin \theta)$ for $\theta \in [0, 2\pi]$. Then

$$M \geq h(x, y) \Rightarrow f(\theta) \leq M \quad \text{for } 0 \leq \theta \leq 2\pi.$$

The average of $f(\theta)$ is equal to M implies that $f(\theta) = M$ (otherwise, since $f(\theta) \leq M$, the average should be less than M). Thus h is constant and equal to $h(x_0, y_0)$ on C_r . Since r is arbitrary, it follows that h is constant and equal to $h(x_0, y_0)$ on any circle that we can fit around (x_0, y_0) in D .

Step 2: Let (x, y) be any point in D . Join (x_0, y_0) to (x, y) using a polygonal line. Cover the polygonal line by finitely many overlapping disks $S_0, S_1, \dots, S_n$ with S_0 centred at (x_0, y_0) and S_j centred at (x_j, y_j) , the point of intersection of ∂S_{j-1} and the polygonal line.

Remark: It is possible to cover the polygonal line with finitely many disk because polygonal lines are closed and bounded and Heine-Borel Theorem (Topic 1) says that finite covers exists.

The function h is constant and equals the maximum value M on S_0 . By continuity of h , it follows that $h(x_1, y_1) = M$. We learn from step 1 that h is constant in the largest disk we can fit around (x_1, y_1) . Hence h is constant and equals M in S_1 . Continue in this manner, we conclude that h is constant and equals M in S_2 to S_n . Thus $h(x, y) = M$. Since (x, y) is arbitrary it follows that h is constant in D .

If the domain D is not bounded, the function h need not attain a maximum or minimum anywhere inside D or on its boundary. For example, $h(x, y) = xy$ is harmonic in $D = \mathbb{R} \times [0, \infty)$.

The function h can be arbitrary large and arbitrary small in the interior of D and so there is not maximum or minimum value.

However, if D is bounded, we have the following corollary.

Corollary 17.15. *Let D be a bounded domain and h a harmonic function on D such that h is continuous on ∂D , the boundary of D . Then h attains its maximum value M and its minimum value m on ∂D .*

From a lecturer's point view, this is a nice result because a lecturer just need to construct a harmonic function and an exam question for calculus class is available.

Example 17.16. Find the maximum and minimum values of $h(x, y) = x^2 - y^2$ for $(x, y) \in D = [0, 2] \times [0, 1]$.

Solution: Of course, we learned in Calculus 2 that to find maximum or minimum, we need to solve $h_x = h_y = 0$ for critical points and then compare them with the boundary values to find the maximum value and minimum value.

First, it is easy to show that $h_{xx} + h_{yy} = 2 + (-2) = 0$ and so h is harmonic. We don't need to solve $h_x = h_y = 0$ for critical points. We just need to find maximum values and minimum values from the boundary:

- When $x = 0$, $h_{\min} = -1$, $h_{\max} = 0$;
- When $x = 2$, $h_{\min} = 3$, $h_{\max} = 4$;
- When $y = 0$, $h_{\min} = 0$, $h_{\max} = 4$;
- When $y = 1$, $h_{\min} = -1$, $h_{\max} = 3$.

Hence, $-1 < h < 4$ in the interior of D and the minimum value is -1 and the maximum value is 4 .

17.4 Application to Simple Dirichlet Problem

The Dirichlet problem (which we will discuss in Topic 9) is a problem of finding a harmonic function on a given domain with prescribed boundary values. For “nice” domains, a solution is unique as shown below.

Theorem 17.17 (Uniquess Theorem). *Let D be a domain of $\mathbb{C}$ and $\phi : \partial D \rightarrow \mathbb{R}$ be a continuous function. The Dirichlet problem is to find a harmonic function h on D such that $\lim_{z \rightarrow z_0} h(z) = \phi(z_0)$ for all $z_0 \in \partial D$. There is at most one solution h to the Dirichlet problem.*

Proof: If the Dirichlet problem has two solutions h_1 and h_2 . Then $h_1 - h_2$ is harmonic on D and extends continuously to $\overline{D}$ and is zero on ∂D . By applying the Maximum Principle (Theorem 17.14, item 2.), since $h_1 - h_2$ is zero on the boundary, $h_1 - h_2 = 0$ in D .

If the domain is not only nice, but also “simple” and related to the elementary functions we discussed in Topic 2 such as the logarithmic function, then we can obtain the solution.

Example 17.18. (Saff and Snider [2003]) Find a harmonic function $h(x, y)$ in the upper half-plane and takes the following values on the x -axis: $h(x, 0) = 0$ for $x \in (-\infty, -1) \cup (1, \infty)$ and $h(x, 0) = 1$ for $x \in (-1, 1)$.

Solution: The harmonic function should have the following form:

$$h(x, y) = A_1 \operatorname{Arg}(z - (-1)) + A_2 \operatorname{Arg}(z - 1) + B.$$

[Question: How do we know that the above function is harmonic?]

On the x -axis (the boundary), we have

$$h(x > 1, 0) = A_1 \times 0 + A_2 \times 0 + B = 0 \Rightarrow B = 0.$$

$$h(-1 < x < 1, 0) = A_1 \times 0 + A_2 \times \pi = 1 \Rightarrow A_2 = \frac{1}{\pi}.$$

$$h(x < -1, 0) = A_1 \times \pi + \frac{1}{\pi} \times \pi = 0 \Rightarrow A_1 = -\frac{1}{\pi}.$$

Therefore, the harmonic solution is $h(x, y) = \frac{\text{Arg}(z - 1) - \text{Arg}(z + 1)}{\pi}$.

Example 17.19. ([Kwok, 2010, Exercise 3.25, p143]) Solve the Dirichlet problem of h on the domain $D = \{z \in \mathbb{C} : \Im(z) > 0\}$ with the boundary condition

$$h(x, 0) = \begin{cases} U_1, & x < x_1, \\ U_2, & x_1 < x < x_2, \\ \vdots & \vdots \\ U_n, & x_{n-1} < x < x_n, \\ U_{n+1}, & x_n < x. \end{cases}$$

Solution. According to Saff and Snider [2003, 141], the harmonic function should have the following form:

$$h(x, y) = A_1 \text{Arg}(z - x_1) + A_2 \text{Arg}(z - x_2) + \cdots + A_n \text{Arg}(z - x_n) + B.$$

Therefore, on the boundary, we have

$$h(x > x_n, 0) = B = U_{n+1}$$

$$h(x_{n-1} < x < x_n, 0) = A_n \pi + U_{n+1} = U_n \Rightarrow A_n = \frac{U_n - U_{n+1}}{\pi}$$

$$h(x_{n-2} < x < x_{n-1}, 0) = A_{n-1} \pi + \frac{U_n - U_{n+1}}{\pi} \pi + U_{n+1} = U_{n-1} \Rightarrow A_{n-1} = \frac{U_{n-1} - U_n}{\pi}$$

$\vdots$

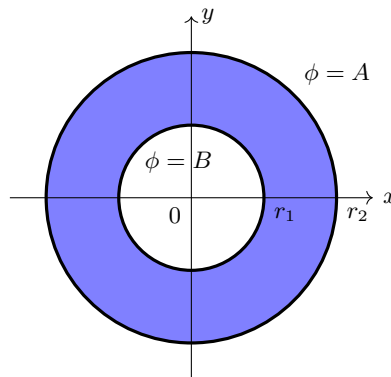
$$h(x < x_1) = A_1 \pi + U_2 = U_1 \Rightarrow A_1 = \frac{U_2 - U_1}{\pi}$$

Hence, the solution is

$$h(x, y) = \frac{U_2 - U_1}{\pi} \text{Arg}(z - x_1) + \cdots + \frac{U_{n+1} - U_n}{\pi} \text{Arg}(z - x_n) + U_{n+1}.$$

□

Example 17.20. ([Saff and Snider, 2003, Exercise 3.4, Question 5]) Find a general formula for a solution $\phi(x, y)$ to the Dirichlet problem inside the washer depicted in the following figure.



Solution. To find a harmonic function which is constant on a circle, we will think of $\ln |z|$, which is the real part of $\text{Log}(z)$. Therefore, we should have a general form

$$\phi(z) = a \ln |z| + b.$$

From the boundary conditions

$$\phi(r_1 e^{i\theta}) = a \ln r_1 + b = B, \quad \phi(r_2 e^{i\theta}) = a \ln r_2 + b = A.$$

By solving the linear algebra problem above, we have

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{\ln r_1 - \ln r_2} \begin{bmatrix} 1 & -1 \\ -\ln r_2 & \ln r_1 \end{bmatrix} \begin{bmatrix} B \\ A \end{bmatrix} \Rightarrow \phi(x, y) = \frac{(B - A) \ln \sqrt{x^2 + y^2} + A \ln r_1 - B \ln r_2}{\ln r_1 - \ln r_2}.$$

□

Chapter 18

Conformal Mapping

In 1569 the Flemish cartographer Gerardus Mercator (1512–1594) devised a cylindrical map projection that preserves angles. The Mercator projection is still used today for world maps. Another map projection known to the ancient Greeks is the stereographic projection. It is also locally conformal, i.e., angle preserving. At that time, conformal mapping is used to draw a world atlas and one can make money out of selling atlas.

In this topic, we will discover that holomorphic functions are locally conformal in general. We then proceed to investigate various conformal mappings which translates from one basic geometric object in z -plane to another geometric object in w -plane. Later we will find that although some simply connected domain are too complicated, it is theoretically possible to map it conformally to a unit disk.

Conformal mapping is a very beautiful theory in complex analysis because it has many good properties. Conformal mapping has also been important in certain engineering problems in the past because a physical problem defined in some complicated region or with a complicated boundary can be solved by using a complex function to map the region or boundary to a simpler region or boundary on which the problem may be solved. The inverse function is then used to provide the solution of the original problem.

18.1 Local Properties of Holomorphic Maps

This section is an extension of the study of holomorphic maps from Topic 2 and 3. We find that holomorphic functions preserves angles (and its “shape”) at a point if the derivative is not zero.

18.1.1 Preservation of Angles

In order to study the angle preservation property of a holomorphic function f at a point z_0 , we need to define the “angle” of a curve at the given point z_0 .

Definition 18.1. Let C be a smooth curve in a complex plane parameterized by

$$z = z(t), \quad a \leq t \leq b$$

and passing through the point z_0 at $a < t_0 < b$, i.e. $z(t_0) = z_0$. The *angle* of the curve C of the point z_0 is $\arg(z'(t_0))$.

Let f be a holomorphic function. If the image of a point z in the z -plane under f is denoted by w , then the image of the curve C in the z -plane under mapping f , denoted as Γ is parameterized as

$$w(t) = f(z(t)), \quad a \leq t \leq b.$$

If $f'(z_0) \neq 0$, then by chain rule,

$$w'(t_0) = f'(z(t_0))z'(t_0). \quad (18.1) \quad \text{e:11.1.1}$$

By taking argument, we obtain (Topic 1)

$$\arg(w'(t_0)) = \arg(f'(z(t_0)) + \arg(z'(t_0)) \quad (18.2)$$

e:11.1.2

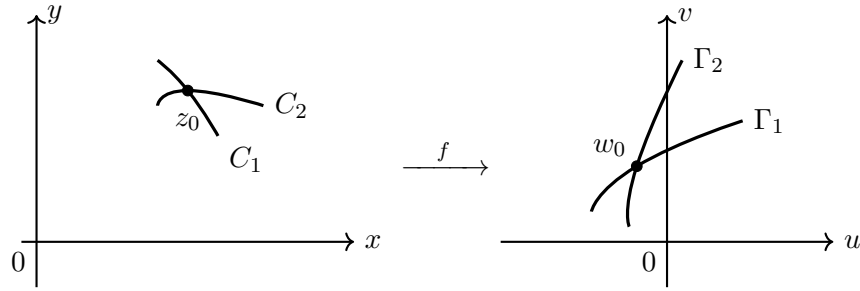
which is useful in describing the relationships of the directions of the tangents of C and Γ at point z_0 and $w_0 = f(z_0)$, respectively. So the effect of the transformation f is to rotate the angle of inclination of the tangent vector at z_0 through the angle $\alpha = \arg(f'(z_0))$ to obtain the angle of inclination of the tangent vector T^* at $w_0 = f(z_0)$.

Let θ_0 be the angle of inclination of a directed tangent line to C at z_0 and φ_0 be the angle of inclination of a directed tangent line to Γ at $w_0 = f(z_0)$. Then (18.2) can be restated as

$$\varphi_0 = \arg(f'(z_0)) + \theta_0. \quad (18.3)$$

e:11.1.3

Now let C_1 and C_2 be two smooth curves intersected at z_0 with angles θ_1 and θ_2 of inclination of directed tangent lines to C_1 and C_2 , respectively. Let Γ_1 and Γ_2 be respectively the image of C_1 and C_2 under the mapping f .



Then the angles of inclination of the tangent line to Γ_1 and Γ_2 at $w_0 = f(z_0)$ are respectively

$$\begin{aligned} \varphi_1 &= \arg(f'(z_0)) + \theta_1, \\ \varphi_2 &= \arg(f'(z_0)) + \theta_2. \end{aligned}$$

It follows that $\varphi_2 - \varphi_1 = \theta_2 - \theta_1$. The angle $\varphi_2 - \varphi_1$ from Γ_1 to Γ_2 is the same as in magnitude and sense as the angle $\theta_2 - \theta_1$ from C_1 to C_2 . In other words, the *angle of intersection of two lines at z_0 is being preserved under the mapping f at z_0 if f is holomorphic at z_0 and $f'(z_0) \neq 0$!*

The above analysis leads to the following theorem.

thm:11.3.1

Theorem 18.1. Let $f : D \rightarrow \tilde{D}$ be holomorphic. If $f'(z) \neq 0$, then f preserves angles in the sense that to each pair of curves in D intersecting at angle α there corresponds a pair of curves in $\tilde{D}$ intersecting at angle α .

Now we can define the local conformal property below.

Definition 18.2. A mapping f is said to be (locally) conformal at a point z_0 if f is holomorphic at z_0 and $f'(z_0) \neq 0$. Alternatively, we define f to be locally conformal at a point z_0 if f is angle-preserving at z_0 (according to Theorem 18.1).

If $f'(z_0) = 0$ at a point z_0 in the z -plane, then $\arg(f'(z_0))$ has no meaning in (18.2) and f is normally not locally conformal.

eg:11.3.2b

Example 18.2. Find all points where the mapping $f(z) = z^2$ is locally conformal.

Solution: From $f'(z) = 2z \neq 0$, f is locally conformal when $z \neq 0$.

When $z = 0$, the real axis $C_1 : z_1(t) = t$ and the imaginary axis $C_2 : z_2(t) = it$ intersected at the origin with angle $\theta = \pi/2$. But the image curves of the axes are

$$\Gamma_1 : w_1(t) = t^2, \quad \Gamma_2 : w_2(t) = -t^2.$$

The angle between Γ_1 and Γ_2 at the intersection in w -plane is $\varphi = \pi$. Thus the function is

not conformal at $z = 0$.

eg:11.3.2a

Example 18.3. Show that the entire function $\exp : \mathbb{C} \rightarrow \mathbb{C}$ is locally conformal at every point in the z -plane.

Proof: Since $\frac{d}{dz} \exp(z) = \exp(z) \neq 0$ for all $z \in \mathbb{C}$, $\exp$ is locally conformal at every point.

eg:11.3.4

Example 18.4. Find all points where the mapping $f(z) = \sin z$ is (locally) conformal.

Solution. Sine function is entire. Hence the mapping is conformal if $f'(z) = \cos z \neq 0$:

$$\cos z = 0 \Leftrightarrow z = \frac{2n+1}{2}\pi, \quad n = 0, \pm 1, \pm 2, \dots$$

Thus the mapping f is locally conformal at point z except those points stated above. $\square$

Definition 18.3. A mapping that preserves the **magnitude of the angle** between two smooth curves is called an *isogonal mapping*.

eg:11.3.1

Example 18.5. Consider the complex mapping $z \rightarrow \bar{z}$. Define two curves in z -plane

$$C_1 : z_1(t) = t + (2t - t^2)i, \quad C_2 : z_2(t) = t + \frac{1}{2}(t^2 + 1)i, \quad 0 \leq t \leq 2.$$

These curves intersect at z_0 . Determine the angle of intersection of C_1 and C_2 at z_0 and the angle of the images at $\bar{z}_0$.

Solution: These curves intersect at $z_0 = 1 + i$ ($t = 1$).

The tangent vectors to the curve C_j at z_0 are $z'_1(1) = 1 + i0$ and $z'_2(1) = 1 + i$. So the angle between C_1 and C_2 at z_0 is

$$\theta = \arg(z'_2(1)) - \arg(z'_1(1)) = \frac{\pi}{4}.$$

The parametric equations of the image curves are

$$\Gamma_1 : w_1(t) = \overline{z_1(t)} = t - (2t - t^2)i, \quad \Gamma_2 : w_2(t) = \overline{z_2(t)} = t - \frac{t^2 + 1}{2}i, \quad 0 \leq t \leq 2.$$

These two image curves intersect in w -plane at $w_0 = 1 - i$ and the tangent vectors to the image curves are $w'_1(1) = 1 + 0i$ and $w'_2(1) = 1 - i$. The angle between Γ_1 and Γ_2 at w_0 is

$$\varphi = \arg(w'_2(1)) - \arg(w'_1(1)) = -\frac{\pi}{4}.$$

Note that the angles θ and φ are equal in magnitude but the angle measured in z -plane is **counterclockwise** but corresponding angle measured in w -plane is clockwise. Thus the two angles are equal in **magnitude** but not equal in sense. Thus the mapping $z \mapsto \bar{z}$ is only an isogonal mapping and is not a conformal mapping.

18.1.2 Local Amplification

We have seen that $\arg(f'(z_0))$ gives the angle of rotation to the tangent vector of C at z_0 under the mapping f . Can we attach a geometrical meaning to the modulus $|f'(z_0)|$?

Let Δs be the arc length between z_0 and z along C , and Δt be the arc length between w_0 and w along Γ . If both z and w are parameterized by the respectively arc lengths, we have

$$\frac{f(z) - f(z_0)}{z - z_0} = \frac{w - w_0}{z - z_0} = \frac{|w - w_0|e^{i\arg(w-w_0)}}{|z - z_0|e^{i\arg(z-z_0)}} = \frac{|w - w_0|}{\Delta t} \frac{\Delta s}{\Delta s |z - z_0|} \exp \left[i \arg \frac{w - w_0}{z - z_0} \right]$$

When $z \rightarrow z_0$ and $w \rightarrow w_0$, we have

$$\lim_{z \rightarrow z_0} \frac{\Delta s}{|z - z_0|} = 1, \quad \lim_{w \rightarrow w_0} \frac{|w - w_0|}{\Delta t} = 1.$$

Hence, taking limit $z \rightarrow z_0$ and then absolute value, we have

$$|f'(z_0)| = \lim_{z \rightarrow z_0} \frac{\Delta t}{\Delta s} = \lim_{z \rightarrow z_0} \frac{|f(z) - f(z_0)|}{|z - z_0|} \Rightarrow |f(z) - f(z_0)| \approx |f'(z_0)||z - z_0|.$$

Therefore $|f'(z_0)|$ is a scale factor of the holomorphic mapping f at z_0 .

Since the angle are preserved according the previous subsection, therefore, the shape of any geometric object in a small neighbourhood of z_0 will just be magnified by the scale $|f'(z_0)|$.

18.1.3 Local Inverse

If a geometrical object in z -plane preserves its angle and is magnified by a factor to w -plane, we would suspect that an “open” disk will map nicely to an open disk. Hence, a holomorphic function should “preserve” openness as well. Recall from Topic 2 that a domain an open connected subset of $\mathbb{C}$. The following theorem informs us that openness is preserved.

Theorem 18.6. *If f is a non-constant holomorphic function in a domain D , then $f(D)$ is a domain.*

Definition 18.4. A holomorphic function f is called *univalent* in D if takes no value more than once in D , i.e. it is one-to-one from D to $f(D)$.

A holomorphic function f that is locally conformal at z_0 has a local inverse as characterised by the following theorem.

thm:11.2.1

Theorem 18.7 (Inverse Function Theorem). *Let f be holomorphic at z_0 with $f'(z_0) \neq 0$. Then f is univalent in the neighbourhood of z_0 , i.e. there is a unique holomorphic inverse function f^{-1} in some open neighbourhood N of $w_0 = f(z_0)$ such that for all $w \in N$, $f(f^{-1}(w)) = w$ and $f'(z)(f^{-1})'(w) = 1$.*

Proof: Let $w_0 = f(z_0)$. We shall show that there exists an open neighbourhood N of w_0 such that we can specify a unique solution to the equation $w = f(z)$.

Define $g(z) = f(z) - w_0$ so that $g(z_0) = 0$ and $g'(z_0) \neq 0$. Then g has a simple zero at z_0 . Since zeros are isolated (since g is analytic, see Topic 5), there exists $\rho > 0$ such that g is holomorphic on the closed disk $|z - z_0| \leq \rho$ and has no zeros in the deleted disk $0 < |z - z_0| \leq \rho$.

Let $C_\rho(z_0)$ denote the circle $|z - z_0| = \rho$. Let $m = \min_{C_\rho(z_0)} |g(z)|$ and γ is any point such that $|\gamma| < m$. It follows by Rouché's theorem that g and $g - \gamma$ have the same number of zeros inside C . But g has only one zero at z_0 . Hence $g(z) = \gamma$ has exactly one solution inside $C_\rho(z_0)$ for $|\gamma| < m$. Let $w = \gamma + w_0$. Then $f(z) - w_0 = w - w_0$, i.e. $f(z) = w$, has exactly one solution inside $C_\rho(z_0)$, as required.

18.1.4 Local Property at Critical Points

When a holomorphic function ceases to be locally conformal, such a point is a critical point defined below.

Definition 18.5. If f is holomorphic at z_0 and $f'(z_0) = 0$, then z_0 is called a *critical point* of f .

Holomorphic functions are not conformal at critical points as characterised by the following theorem.

Theorem 18.8. Let f be holomorphic at the critical point. If $n > 1$ is an integer such that

$$f'(z_0) = f''(z_0) = \cdots = f^{(n-1)}(z_0) = 0 \quad \text{and} \quad f^{(n)}(z_0) \neq 0,$$

then the angle between any two smooth curves intersecting at z_0 is increased by a factor of n by the mapping f . In particular, f is not conformal at z_0 .

Proof: Express the function in Taylor series form

$$f(z) = \sum_{k=0}^{\infty} \frac{f^{(k)}(z_0)}{k!} (z-z_0)^k = f(z_0) + \sum_{k=n}^{\infty} \frac{f^{(k)}(z_0)}{k!} (z-z_0)^k = f(z_0) + \frac{f^{(n)}(z_0)}{n!} (z-z_0)^n [1+g(z)] \quad (18.4)$$

e:11.3.3

$$\text{where } g(z) = \frac{n!(z-z_0)^n}{f^{(n)}(z_0)} \left[\frac{f^{(n+1)}(z_0)}{(n+1)!} + \frac{f^{(n+2)}(z_0)}{(n+2)!} (z-z_0) + \cdots \right].$$

Notice that g is holomorphic and $g(z_0) = 0$.

Let Γ be the image of a smooth curve C under the transformation. Note that the angle of inclination θ_0 and φ_0 are limits of $\arg(z-z_0)$ and $\arg(f(z)-f(z_0))$, respectively, as $z \rightarrow z_0$ along C . By taking argument on (18.4) we obtain

$$\arg(f(z)-f(z_0)) = \varphi_0 = \arg\left(\frac{f^{(n)}(z_0)}{n!}\right) + \arg((z-z_0)^n) = \arg(f^{(n)}(z_0)) + n\theta_0.$$

18.2 Elementary Conformal Mapping

In this section, we shall investigate the “global” conformal mappings between various elementary geometric objects, in which many of them have been introduced in Topic 1. Two basic geometric objects are important and deserve definitions.

Definition 18.6. The *upper half plane* $\mathbb{H}$ is defined as

$$\mathbb{H} = \{z \in \mathbb{C} : \Im(z) > 0\}.$$

Definition 18.7. The *unit disk* Δ is defined as

$$\Delta = \{z \in \mathbb{C} : |z| < 1\}.$$

Recall from Topic 1 that the boundary of a domain D is denoted as ∂D . Using this notation, we have $\partial\mathbb{H} = \mathbb{R}$ and $\partial\Delta = S^1$, a circle of unit radius.

In the following subsections, we will only limit ourselves to just a few conformal mappings characterised by elementary functions (Topic 2), for more examples, consult Kober [1957].

18.2.1 Conformal Mapping

Let D be a domain in $\mathbb{C}$ and $\tilde{D} := f(D)$ is an image of the function f .

Definition 18.8. A mapping $f : D \rightarrow \tilde{D}$ is said to be a *conformal mapping* (or *map* or *transform*) if f is locally conformal at every point $z_0 \in D$ and f is one-to-one from D to $\tilde{D}$.

Definition 18.9. Two domains D_1 and D_2 are said to be *conformal equivalent* if there exists a conformal mapping from D_1 to D_2 .

One of the simplest conformal mapping from $\widehat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ to $\widehat{\mathbb{C}}$ is given by the reciprocation (Topic 3).

Example 18.9. Show that the reciprocation $w = \frac{1}{z}$ is a conformal map from $\widehat{\mathbb{C}}$ to $\widehat{\mathbb{C}}$.

Proof: First, we need to show that $f(z) = \frac{1}{z}$ is one-to-one from $\widehat{\mathbb{C}}$ to $\widehat{\mathbb{C}}$:

- When $z \neq 0$, z and $w = \frac{1}{z}$ has one-to-one correspondence.
- $z = 0$ and $w = \infty$ are one-to-one.
- $z = \infty$ and $w = 0$ are one-to-one.

Now, we need to show that reciprocation is locally conformal at every point in $\widehat{\mathbb{C}}$.

- When $z \in \mathbb{C}$, $f'(z) = -\frac{1}{z^2} \neq 0$, reciprocation is locally conformal. Note that when $z = 0$, $f'(0) = \infty \neq 0$.
- When $z = \infty$, $f'(\infty) = \lim_{z \rightarrow 0} \frac{d}{dz} f\left(\frac{1}{z}\right) = \lim_{z \rightarrow 0} \frac{d}{dz} (z) = 1 \neq 0$.

Hence, the reciprocation is a conformal map from $\widehat{\mathbb{C}}$ to $\widehat{\mathbb{C}}$.

The conformal mapping from $\widehat{\mathbb{C}}$ to $\widehat{\mathbb{C}}$ is just a Möbius transform. This why Möbius transform is so important in complex analysis.

Example 18.10. Show that the Möbius transformation T (Topic 3) is conformal from $\widehat{\mathbb{C}}$ to $\widehat{\mathbb{C}}$.

Solution. Proof: First, we need to show that $T(z) = \frac{az + b}{cz + d}$ is one-to-one from $\widehat{\mathbb{C}}$ to $\widehat{\mathbb{C}}$:

- When $c = 0$, we obtain a linear transform, which is one-to-one from $\mathbb{C}$ to $\mathbb{C}$ and ∞ to ∞ . Note that $a \neq 0$ otherwise it contradicts with the definition $ad - bc = ad \neq 0$.
- When $c \neq 0$,
 - $T(-\frac{d}{c}) = \infty$ and $T(\infty) = \frac{a}{c}$.
 - When $z \neq -\frac{d}{c}$, it is one-to-one from $\mathbb{C} \setminus \{-\frac{d}{c}\}$ to $\mathbb{C} \setminus \{\frac{a}{c}\}$.

Now, we need to show that T is locally conformal at every point in $\widehat{\mathbb{C}}$:

- When $z \neq \infty$, and we note that $ad - bc \neq 0$,

$$T'(z) = \frac{a(cz + d) - c(az + b)}{(cz + d)^2} = \frac{ad - bc}{(cz + d)^2} \neq 0.$$

Note that $T'(-\frac{d}{c}) = \infty \neq 0$.

- When $z = \infty$, $T'(\infty) = \lim_{z \rightarrow 0} \frac{d}{dz} T\left(\frac{1}{z}\right) = \lim_{z \rightarrow 0} \frac{d}{dz} \left[\frac{bz + a}{dz + c} \right] = \frac{bc - da}{c^2} \neq 0$.

Hence, T is conformal from $\widehat{\mathbb{C}}$ to $\widehat{\mathbb{C}}$. □

18.2.2 Conformal Mappings between $\mathbb{H}$ and Δ

By applying the knowledge of Topic 3, both $\mathbb{H}$ and Δ can be viewed as generalised circles in $\widehat{\mathbb{C}}$. This inspires us that the conformal mapping between Δ and Δ , $\mathbb{H}$ and $\mathbb{H}$, $\mathbb{H}$ and Δ are special Möbius transforms.

In Topic 3, we have stated a theorem characterising the conformal mappings between Δ and Δ . Rewrite the theorem slightly gives us a theorem below.

Theorem 18.11. *Let the complex number a satisfy $|a| < 1$, $\alpha \in \mathbb{R}$ and the Möbius transformation be*

$$T_a(z) = e^{i\alpha} \frac{z - a}{az - 1}. \quad (18.5)$$

e:mobius:cir

Then T_a is a conformal mapping from Δ to Δ .

In Tutorial 3 question 6, we can obtain a theorem for the conformal mapping from $\mathbb{H}$ to $\mathbb{H}$

Theorem 18.12. *The conformal map from $\mathbb{H}$ to $\mathbb{H}$ has the following general form*

$$T(z) = \frac{az + b}{cz + d}, \quad a, b, c, d \in \mathbb{R}, \quad ad - bc > 0. \quad (18.6)$$

e:H-isom

Example 18.13. Determine the images of $\mathbb{H}$ under the conformal mappings (a) $w = \frac{1+z}{1-z}$ and (b) $w = \frac{z+1}{z-1}$.

Solution. (a) Since $a, b, c, d \in \mathbb{R}$ and $ad - bc = 1 \times 1 - 1 \times (-1) = 2 > 0$, the image is $\mathbb{H}$.

(b) The two transform differ by a negative sign, so the image is rotated by 180° to become $\Im(w) < 0$. $\square$

Now let us investigate the general conformal mapping from a half-plane $\mathbb{H}$ to a unit circle Δ given by a Möbius transform.

These transforms will map points on the line $\Im(z) = 0$ to the points on the circle $|w| = 1$. Without loss of generality, we select $z = 0$, $z = 1$ and $z = \infty$ on $\partial\mathbb{H}$. Then

$$T(0) = \left| \frac{a(0) + b}{c(0) + d} \right| = \left| \frac{b}{d} \right| = 1, \quad T(1) = \left| \frac{a + b}{c + d} \right| = 1, \quad T(\infty) = \left| \frac{a}{c} \right| = 1.$$

From the second equation above, we can obtain

$$\left| \frac{a}{c} \right| \left| \frac{1 + \frac{b}{a}}{1 + \frac{d}{c}} \right| = 1 \Rightarrow \left| 1 + \frac{b}{a} \right| = \left| 1 + \frac{d}{c} \right| \Rightarrow 1 + \frac{b}{a} + \frac{\bar{b}}{a} + \left| \frac{b}{a} \right|^2 = 1 + \frac{d}{c} + \frac{\bar{d}}{c} + \left| \frac{d}{c} \right|^2.$$

Since $|b| = |d|$ and $|a| = |c|$, we have $\left| \frac{b}{a} \right| = \left| \frac{d}{c} \right|$ and may cancel some terms above to obtain

$$\frac{b}{a} + \frac{\bar{b}}{a} = \frac{d}{c} + \frac{\bar{d}}{c} \Rightarrow \Re\left(\frac{b}{a}\right) = \Re\left(\frac{d}{c}\right).$$

Together with $\left| \frac{b}{a} \right| = \left| \frac{d}{c} \right|$, we have $\frac{b}{a} = \frac{d}{c}$ or $\frac{b}{a} = \frac{\bar{d}}{c}$. Since $ad - bc \neq 0$, therefore $\frac{b}{a} = \frac{\bar{d}}{c}$ and if we let $z_0 = -\frac{b}{a}$, we obtain the following theorem.

Theorem 18.14. *The general conformal transform from the half plane $\mathbb{H}$ to a unit circle Δ is*

$$T(z) = e^{i\alpha} \frac{z - z_0}{z - \bar{z}_0}$$

where $\Im(z_0) > 0$ and $\alpha \in \mathbb{R}$.

thm:H-disk

Remark 18.15. Find the inverse conformal mapping from Δ to $\mathbb{H}$ is going to be easy according to the Tutorial 3 question 5 (which is related to the matrix representation of a Möbius transform):

$$T(z) = e^{-i\alpha} \frac{-\bar{z}_0 z + z_0}{-z + 1} = e^{-i\alpha} \frac{|z_0|(e^{-i\beta} z - e^{i\beta})}{z - 1}, \quad \alpha, \beta = \text{Arg}(z_0) \in \mathbb{R}, \quad \Im(z_0) > 0. \quad (18.7)$$

Although we have the above theorems, it is difficult to describe all possible mappings between “half planes” and circles. The following examples demonstrates this situation.

Example 18.16. Show that the image of the right half-plane $\Re(z) > 0$ under the conformal mapping $w = \frac{z-1}{z+1}$ is the unit disk $|w| < 1$.

Solution: There are two approaches to proving this fact. Both require us to recall what we learned in Topic 3.

Proof using Topic 3: First find the inverse mapping

$$wz + w = z - 1 \Rightarrow z(w - 1) = -1 - w \Rightarrow z = \frac{1+w}{1-w}.$$

Then, we can calculate the image as follows

$$\Re(z) > 0 \Rightarrow z + \bar{z} > 0 \Rightarrow \frac{1+w}{1-w} + \frac{1+\bar{w}}{1-\bar{w}} > 0 \Rightarrow 1+w-\bar{w}-w\bar{w}+1-w+\bar{w}-w\bar{w} > 0$$

because $(1-w)(1-\bar{w}) = 1 - 2\Re(w) + |w|^2 \geq 1 - 2|w| + |w|^2 = (1-|w|)^2 > 0$ due to $\Re(w) \leq |w|$.

Hence, after cancellation of terms, we have the image

$$2 - 2w\bar{w} > 0 \Rightarrow 1 > |w|^2 \Rightarrow |w| < 1.$$

Proof using Theorem [18.14](#): First, we rotate $\Re(z) > 0$ to $\mathbb{H}$ by $T_1(z) = e^{i\pi/2}z = iz$ (we learned the idea of rotation in Topic 3) and we know $T_2(z) = \frac{z-i}{z+i}$ maps $\mathbb{H}$ to Δ according to Theorem [18.14](#). Therefore,

$$(T_2 \circ T_1)(z) = \frac{iz-i}{iz+i} = \frac{z-1}{z+1}$$

is a mapping from $\Re(z) > 0$ to Δ .

Example 18.17. Show that the conformal mapping $w = \frac{1}{z}$ maps the half-plane $\Re(z) > \frac{1}{2}$ to the unit disk $|w-1| < 1$.

Solution. Proof: The image of the mapping is given by

$$\Re(z) = \frac{z+\bar{z}}{2} > \frac{1}{2} \Rightarrow z+\bar{z} > 1 \Rightarrow \frac{1}{w} + \frac{1}{\bar{w}} > 1 \Rightarrow \bar{w} + w > w\bar{w} \Rightarrow w\bar{w} - w - \bar{w} + 1 < 1$$

Which tells the image is $|w-1| < 1$. □

Example 18.18. Determine the image of (a) $|z| < 1$ (b) $\Im(z) > 0$ under the conformal mapping $w = \frac{z-1}{z+1}$.

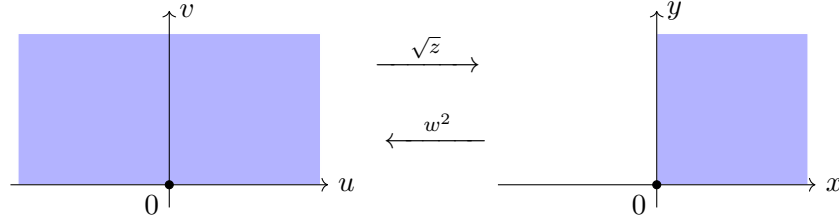
Solution. (a) The boundary $|z| = 1$ has a point -1 which maps to ∞ . Hence, it is mapped to a line. From $T(1) = 0$ and $T(i) = i$, we know that the image is $\Re(w) < 0$.

(b) The boundary $\Im(z) = 0$ has a point -1 which maps to ∞ , so the image is a line. From $T(0) = -1$ and $T(1) = 0$, the image is $\Im(w) > 0$. □

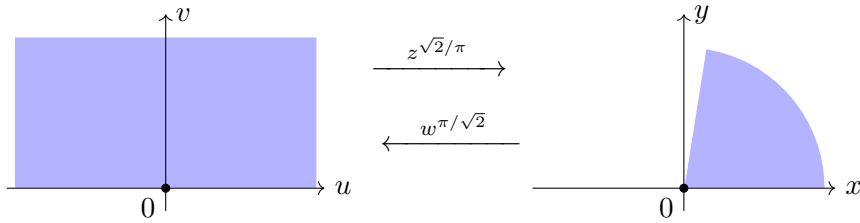
18.2.3 Conformal Mappings for Generalised Rectangles and Circular Shape Objects

As we have mentioned in the previous section, Δ and $\mathbb{H}$ are two fundamental geometrical objects. In this section, we will extend our study to include sectors and annulus.

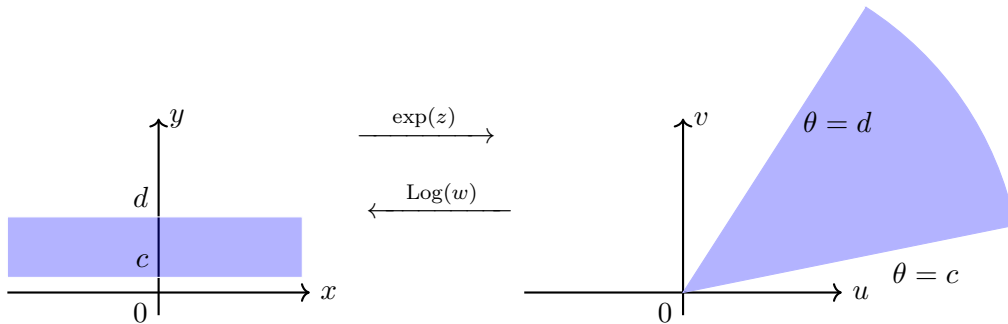
First, we will note that the n -th root function from $\mathbb{H}$ to a sector of angle π/n , i.e. $L = \{z \in \mathbb{C} : 0 < \arg(z) < \pi/n\}$ is a conformal mapping. The inverse mapping, the n -th power function from L to $\mathbb{H}$ is also a conformal mapping. In particular, the square root function $z \mapsto \sqrt{z}$ maps a half-plane $\mathbb{H}$ to quarter sector and the square function $w \mapsto w^2$ maps a quarter sector to $\mathbb{H}$ as shown below.



The principal arbitrary function (Topic 2, Section 2.5.5) will be a conformal mapping from $\mathbb{H}$ to $L = \{z : 0 < \arg(z) < \alpha\}$ when it has the form $f(z) = z^{\alpha/\pi}$ and $0 < \alpha \leq \pi$. For example, when $\alpha = \sqrt{2}$, we have

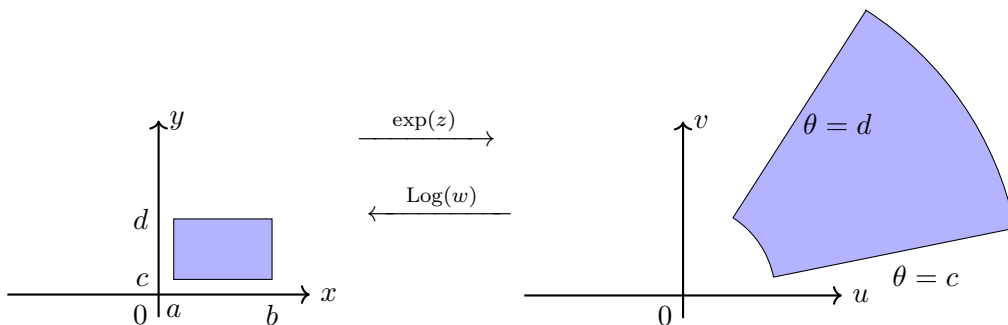


What if we are interested in “rectangles” instead of the upper half plane $\mathbb{H}$? Exponential function and principal logarithmic function can be of great use to us because an exponential function is $2\pi i$ -periodic. So when we limit ourselves to infinite long rectangle $R = \{z \in \mathbb{C} : c < \Im(z) < d\}$, $0 < c < d < 2\pi$, we will have a conformal mapping between R and a sector $\exp(R) = \{w \in \mathbb{C} : c < \arg(w) < d\}$. When $c = 0.1$ and $d = 1$, we have the following picture:

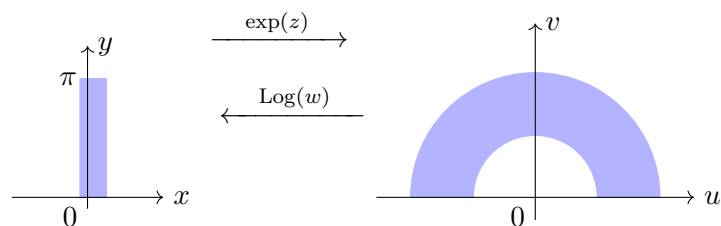


Note that when $c = 0$ and $d = \pi$, we have a conformal mapping between R and $\mathbb{H}$.

If we have a normal rectangle $R = (a, b) \times (c, d)$, where $0 < c < d < 2\pi$, the image $\exp(R)$ is a portion of a sector as illustrated below.



Out of these, the most important is the case where $R = (a, b) \times (0, \pi)$, the image $\exp(R)$ is a semi-annulus as illustrated below.



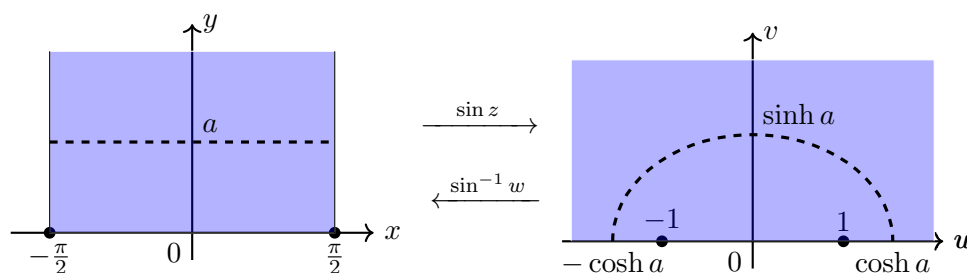
Another important case is when $R = (-\infty, 0) \times (0, \pi)$, $\exp$ is a conformal mapping from R to upper half unit disk $\Delta \cap \mathbb{H}$.

From the above analysis, we know that exponential function can be used as a conformal mapping between an infinite rectangle and the upper plane $\mathbb{H}$. If we consider the sine function $w = \sin z$. Let $z = x + iy$ and $w = u + iv$. From Topic 2, we know that

$$\sin(x + iy) = \sin x \cos iy + \sin iy \cos x = \underbrace{\sin x \cosh y}_u + i \underbrace{\cos x \sinh y}_v. \quad (18.8)$$

e:sine-rule

If we limit ourselves to the infinite upper band $R = (-\pi/2, \pi/2) \times \mathbb{R}$, then sine function is a conformal mapping from R to $\mathbb{H}$. Beware that although we draw the boundaries, the boundaries are not part of the conformal mapping because the boundaries do not preserve angle at $z = \pm \pi/2$.



Example 18.19 (April 2009 Exam, Q6(a)). Let a and b be two positive numbers.

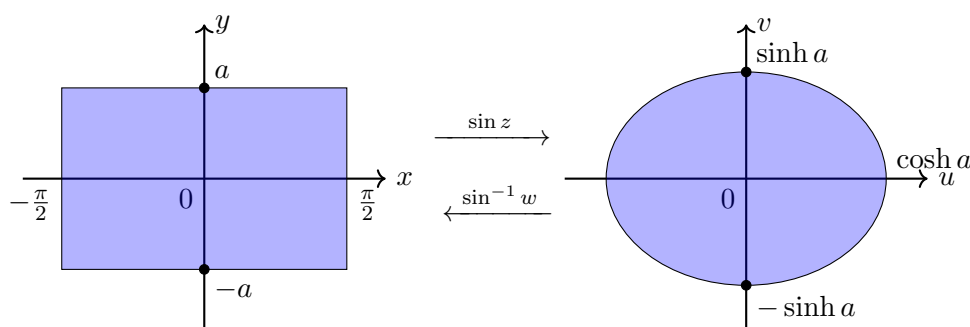
- (i) Find the image of $z = a$ under the mapping $w = \sin z$ (5 marks)

Solution. $-1 \leq w = \sin a \leq 1$ since a is a positive number. □

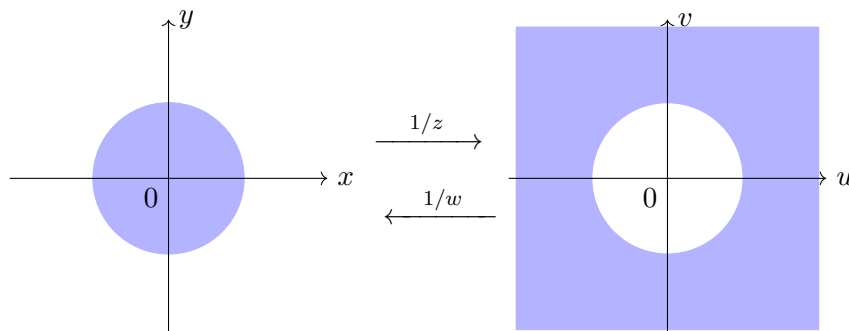
- (ii) Find the image of $z = ib$ under the mapping $w = \sin z$. (5 marks)

Solution. $w = \sin ib = \frac{e^{-b} - e^b}{2i} = i \sinh b$. Since b is a positive number, the image is the upper imaginary axis. □

If the domain is a suitable rectangle $[-\pi/2, \pi/2] \times [-a, a]$, then its image under the conformal mapping $z \mapsto \sin z$ is given below. Note that the upper rectangle $[-\pi/2, \pi/2] \times [0, a]$ maps to the upper ellipse while the lower rectangle $[-\pi/2, \pi/2] \times [-a, 0]$ maps to the lower ellipse.



Most of the above geometric objects D are bounded. If we want to find the conformal mapping for the complement of D , we would need to find a conformal mapping from D to Δ and take reciprocation because the reciprocation is a conformal mapping from Δ to $\hat{\mathbb{C}} \setminus \overline{\Delta}$ as shown below.



We will investigate a few more conformal mappings in the tutorial. But until here, we should be asking, given any two geometrical objects D_1 and D_2 , is it always possible for us to map D_1 to D_2 ?

18.3 Riemann Mapping Theorem

sec:rmt

A partial answer to the above question, which is a significant result, was obtained by Riemann in his 1851 PhD thesis [Needham, 1998, 180]. The result basically says that any two proper subsets of $\mathbb{C}$ which are simply connected are conformally equivalent.

m:riemann-mapping

Theorem 18.20 (Riemann Mapping Theorem). *There exists a conformal mapping f from the unit disk Δ to any simply connected (Topic 4) proper subdomain D of the complex plane $\mathbb{C}$. Let $z_0 \in D$, if $f(0) = z_0$ and $\arg(f'(0)) = \alpha \in [0, 2\pi)$, then f is unique.*

A proof of this theorem can be found in Osgood [1900], Hille [1959, Chapter 17], Freitag and Busam [2009, Chapter IV], Nehari [1952, Chapter V], Bieberbach [1964, §18], Ahlfors [1979, Chapter 6] and Bak and Newman [2010, Chapter 14].

This theorem is important because it says that it is theoretically possible for us to solve the Laplace equation on any simply connected domain because we can always reduced it to the unit disk. We will investigate very simple examples in Topic 9.

However, there is no conformal map that maps $\mathbb{C}$ onto Δ because if such a holomorphic function exists, then $|f(z)| = |w| < 1$ and according to Liouville's Theorem (Topic 4), f has to be constant.

According to Ahlfors [1979], the boundary of the domain mentioned in the theorem may be complicated.

18.4 Numerical Conformal Mapping

Knowing that a conformal mapping exists in principle is not the same as knowing how to solve the numerical problem: given D , to calculate f to a given degree of accuracy. In some simple cases the conformal mapping can be written as a composition of elementary functions, such as Möbius transformations, powers, and roots, trigonometric functions, elliptic integrals, etc. mentioned in the earlier section. This section tries to mention two numerical methods where for a given domain D (which can be approximated by a polygon), it is possible to construct a mapping from either $\mathbb{H}$ or Δ to D .

A map from one domain D_1 to D_2 is often described as a composition of a map from Δ to D_1 composed with the inverse of a map from Δ to D_2 . Depending on the nature of the numerical description, it may be difficult to find the inverse map.

After Riemann put forth his mapping theorem in 1851, Elwin Bruno Christoffel and Herman Amandus Schwarz independently discovered the Schwarz-Christoffel formula, which is based on this idea: If $f_k(z) = (z - a_k)^{\frac{\alpha_k}{\pi} - 1}$, $k = 1, \dots, n$, then

$$\arg(f'(z)) = \arg(f_1(z)) + \dots + \arg(f_n(z))$$

will give us a function f which maps the real axis with points a_k to a polygon with points $f(a_k)$.

Theorem 18.21 (Schwarz-Christoffel Mapping Theorem). *Let P be the interior of a simple polygon whose boundary is piecewise linear having vertices $w_1, \dots, w_n$ and non-zero interior angles $\alpha_1, \dots, \alpha_n$ in counter-clockwise order. Let f be any conformal map from Δ to P . Then*

$$f(z) = A + B \int_{z_0}^z \prod_{k=1}^n \left(1 - \frac{\zeta}{z_k}\right)^{-\beta_k} d\zeta. \quad (18.9)$$

e:sc-map

Here $\beta_k = \frac{\alpha_k}{\pi} - 1$. The parameters A , B and the pre-vertices z_k are unknown a priori satisfies the relation $w_k = f(z_k)$, for $k = 1, \dots, n$.

The main challenge in computing (18.9) is to solve the parameter problem of finding the pre-vertices z_k ! After the pre-vertices are found, we still need to evaluate the integral in (18.9). General numerical integration methods can have problems with integrals of this form for some points, Gauss-Jacobi quadrature (a type of Gauss quadrature, see UECM3033 Numerical Methods) is a better approximation to $\int_{-1}^1 f(x)(1-x)^\alpha(1+x)^\beta dx$, which is closely related to (18.9).

There are quite a few of numerical conformal mapping software implemented based on the Schwarz-Christoffel (SC) mapping idea above [Driscoll and Trefethen, 2002, 56]:

- Schwarz-Christoffel Toolbox for MATLAB: it is the most user friendly numerical software.
- SCPACK (Fortran 77): It only maps from the disk, using the side-length formulation of the parameter problem.
 - Set the values: N = Number of vertices, w(k) = vertices, NPTSQ = number of quadrature points per subinterval, qwork has a dimension NPTSQ*(2N+3), BETAM(k) = external angles divided by $-\pi$.
 - qinit: Computes Gauss-Jacobi quadrature nodes and weights.
 - scsolve: Solve the “parameter problem” for a given polygon.
 - wsc: Find the SC mapping from Δ to P .
 - zsc: Find the inverse SC mapping from P to Δ .
- Cap (Fortran 77): Maps the unit disk onto a circular arc polygonal domain [Papamichael and Stylianopoulos, 2010, 29].

However, there are numerical conformal mapping algorithms which are not based the SC mapping, such as [Driscoll and Trefethen, 2002, 56]:

- Zipper: finds a numerical approximation of the conformal map (and its inverse) of the unit disk onto any Jordan region.
- CONFPACK: A Fortran package by Hough which implements Symm’s equation for simply connected regions (interior or exterior) with piecewise smooth boundaries.

The zipper algorithm assumes that there are an even number of boundary points, $z_0, z_1, \dots, z_{2n+1}$. The first map

$$\varphi_1(z) = \sqrt{\frac{(z - z_2)(z_1 - z_0)}{(z - z_0)(z_1 - z_2)}}$$

maps the complement in the extended plane of the circular arc through z_0, z_1, z_2 onto $\mathbb{H}$ such that $\varphi_1(z_2) = 0$, $\varphi_1(z_3) = \zeta_3^{(1)}$ and $\varphi_1(z_4) = \zeta_4^{(1)}$. Then similar maps φ_i , $i = 2, \dots, n$ map the image of the rest of the points $\zeta_3^{(1)}, \zeta_4^{(1)}$, then $\zeta_5^{(2)}, \zeta_6^{(2)}, \dots$ to $\mathbb{H}$ and we have

$$\varphi_n \circ \dots \circ \varphi_1 : \mathbb{H} \setminus \gamma \rightarrow \mathbb{H}.$$

Here γ is an arc passing through z_0, z_1, z_2 .

Lastly, we have $\varphi_n(\zeta_{2n+1}^{(n-1)}) = \zeta_{2n+1}^{(n)}$ and we define $\zeta_{2n+2}^{(n)}$ to be a circle intersecting real axis at an angle α at 0 passing through $\zeta_{2n+1}^{(n)}$. A final map of the following form

$$\varphi_{n+1}(z) = C \left(\frac{z}{1 - z/\zeta_{2n+2}^{(n)}} \right)^{\pi/\alpha}$$

will turn the original image D approximated by $z_0, \dots, z_{2n+1}$ to $\mathbb{H}$.

The software Zipper works as follows:

1. Create a Jordan curve (approximation) with

```
xcm -grid -gridcolor orange -subgridcolor yellow -o init.dat
```

2. Put more than 200 points on the boundary of the polygon, with at least 5 points per edge with:

```
./zipper/polygon < ./input/polygon.inp > logfile
```

where polygon.inp define the following information:

```
init.dat
200
poly.dat
```

3. Compute the conformal map to the disk with Zipper algorithm:

```
./zipper/zipper # reads poly.dat and generates poly.img
```

4. Compute the pre-images of the data points

```
./zipper/origpre
```

The pre-images can be viewed with

```
xcm -use 2 red blue poly.pre orig.pre
```

Chapter 19

Applications

Complex analysis has important application in signal processing (which is closely related the very abstract mathematics called harmonic analysis, see McRobert [1947], Berenstein and Gay [1991], Groemer [1996]) and is very useful in solving two-dimensional (2D) Laplace equation. *Fourier transform*, *Laplace transform* and *z-transform* are three important *linear transforms* in signal processing. These linear transforms are used to transform systems from “time-domain” to “frequency”-like domain, which made the original problem easier to solve, and are investigated in Section [19.1](#).

Conformal mapping (Topic 8) was used in the past to solve 2D Laplace equation, which are important in 2D potential theory, and will be discussed in Section [19.2](#).

Complex analysis can be applied to study the solution of certain analytic PDEs [Kracht and Kreyszig, 1988], however, it is beyond the current course to study them.

19.1 Integral Transforms

An integral transform is linear and important in the analysis of some ordinary differential equations (UECM2023) and some partial differential equations (UECM3003) which are associated with quantum mechanics, fluid mechanics, electromagnetic theory, elastic theory, etc. [Ablowitz and Fokas, 2003, Chapter 6]

In this section, we will mention two integral transforms, i.e. Fourier transform and Laplace transform. Then we will mention how the can be applied in linear AC circuits and classical control theory with examples.

Definition 19.1. A *linear transform* L of an appropriate set of functions $Func(D)$ over the domain $D \subset \mathbb{R}$ to another set of functions G satisfies the following property:

$$L : Func(D) \rightarrow G, \quad \alpha f + \beta g \in F(D) \mapsto \alpha L\{f\} + \beta L\{g\}$$

for all $\alpha, \beta \in \mathbb{R}$ and all $f, g \in Func(D)$.

Definition 19.2. An *integral transform* of a function $f : D \rightarrow \mathbb{C}$ is given by

$$T\{f(x)\} = \int_D K(x, s) f(s) ds$$

where $K(x, s)$ is called the *kernel* of the transform.

19.1.1 Fourier Transform

Fourier transform is an integral transform which is closely related to complex analysis [Dym and McKean, 1972, Chapter 3], [Roos, 1969, Chapter 2].

Definition 19.3. The *Fourier Transform* of a Lebesgue integrable function (this should be introduced in real analysis) $f : \mathbb{R}^n \rightarrow \mathbb{C}$ is defined as

$$\mathcal{F}\{f(x)\} = \int_{\mathbb{R}^n} e^{-2\pi i \mathbf{x} \cdot \xi} f(\mathbf{x}) d\mathbf{x} = F(\xi) \quad \text{or} \quad \widehat{f}(\xi) \quad (19.1)$$

where the kernel is $K(\mathbf{x}, \xi) = e^{-2\pi i \mathbf{x} \cdot \xi}$.

When $n = 1$, we have

$$\mathcal{F}\{f(x)\} = \int_{-\infty}^{\infty} e^{-2\pi i x \xi} f(x) dx = F(\xi). \quad (19.2)$$

eg:two-exp

Example 19.1. Find the Fourier Transform for $f(t) = e^{-|t|}$.

$$\begin{aligned} \text{Solution: } \mathcal{F}\{e^{-|t|}\} &= \int_{-\infty}^{\infty} e^{-2\pi i \xi t - |t|} dt = \int_{-\infty}^0 e^{-2\pi i \xi t + t} dt + \int_0^{\infty} e^{-2\pi i \xi t - t} dt \\ &= \frac{1}{1 - 2\pi i \xi} + \frac{1}{1 + 2\pi i \xi} = \frac{2}{1 + (2\pi \xi)^2}. \end{aligned}$$

Definition 19.4. The *inverse Fourier transform* of a non-periodic function $F : \mathbb{C} \rightarrow \mathbb{R}$ is defined as

$$\mathcal{F}^{-1}\{F(\xi)\} = \int_{-\infty}^{\infty} F(\xi) e^{2\pi i \xi x} d\xi = f(x).$$

Remark 19.2. There are various definition of Fourier transform, we are trying to following the definition used in signal processing literature and some analysis books, such as Brigham [1988], Kaiser [1994], Stein and Shakarchi [2003].

The Fourier transform is useful in science and engineering because many linear ODEs and a few linear PDEs can be transformed to an algebra problem. After reducing the algebra problem to a simpler form, one can use inverse Fourier transform to obtain the solution either in integral form and many times in analytic form. Examples can be found in signal processing books such as Brandwood [2003], Jennison [1961] or *UECM3003 Partial Differential Equations*.

19.1.2 Laplace Transform

Laplace transform is an integral transform which works on causal functions. A rigorous theoretical investigation can be found in Doetsch [1974].

Definition 19.5. A function $f(t)$ defined on $\mathbb{R}$ is called a *causal function* if $f(t)$ is zero when $t < 0$.

We can construct a causal function $f(t)$ from any function $g(t)$ over $\mathbb{R}$ by multiplying it *Heaviside step function* $H(t)$, i.e. $f(t) = g(t)H(t)$. If $f(t)$ is a causal function, $f(t) = H(t)f(t)$.

def:laplace

Definition 19.6. The *Laplace transform* of a causal function $f(t)$ is defined by the following improper integral:

$$F(s) = \mathcal{L}\{f(t)\} = \int_0^{\infty} f(t) e^{-st} dt = \lim_{L \rightarrow \infty} \lim_{\varepsilon \rightarrow 0+} \int_{-\varepsilon}^L f(t) e^{-st} dt. \quad (19.3)$$

Here the kernel $K(s, t) = e^{-st}$, $s \in \mathbb{C}$ is called the *transformed variable*, its real part larger than a real number, usually 0 if not otherwise stated.

e:ft:n

e:ft

e:ltransform

The behaviour of $f(t)$ for $t < 0$ is usually ignored, thus the inverse Laplace transform will not recover $f(t)$ for $t < 0$, so $\mathcal{L}^{-1}\{F(s)\} = f(t)H(t)$. If one is interested in the region of $t < 0$, a bilateral Laplace transform $F_B(s) = \mathcal{L}_B\{f(t)\} = \int_{-\infty}^{\infty} f(t)e^{-st}dt$ should be used.

Remark 19.3. Fourier transform (48.2) and Laplace transform (32.1) are “formally” related when $s = 2\pi i\xi$ and the function “ f ” is causal.

Not all functions have Laplace transform, only functions that are of *exponential order* (defined below) have Laplace transform.

Definition 19.7. A piecewise-continuous function is said to be of *exponential order* α_f if there exists a real constant α_f such that $|f(t)| \leq Ke^{\alpha_f t}$ and if $c > \alpha_f$ then $\lim_{t \rightarrow \infty} \frac{f(t)}{e^{ct}} = 0$.

Functions that are of exponential order include power functions t^n , polynomial functions, exponential function a^t ($a > 1$), etc.

Theorem 19.4. If a function f is piecewise continuous on the interval $0 \leq t \leq A$ for every positive A and there are real constants $K > 0$, α_f and $M > 0$ such that

$$|f(t)| \leq Ke^{\alpha_f t}$$

then the Laplace transform (32.1) exists for $s > \alpha_f$.

Definition 19.8. Let $F(s)$ be the Laplace transform of a function $f(t)$, $t > 0$. The contour integral (Topic 4)

$$f(t) = \mathcal{L}^{-1}\{F(s)\} = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} F(s)e^{st}ds. \quad (19.4)$$

where $c > \alpha_f$. The function $f(t)$ is called the *inverse Laplace transform* of $F(s)$.

Laplace transform (32.1) and inverse Laplace transform (19.4) are normally used in classical control theory to study the “response” of a “plant” which can be approximated by a constant coefficient ODE system (refer to UECM2023 Ordinary Differential Equation)

$$\dot{\mathbf{x}} = A\mathbf{x} + B\mathbf{u}, \quad \mathbf{x}(0) = \mathbf{x}_0 \quad (19.5)$$

with respect to various types of “input” $\mathbf{u}$.

The study of system response is shown below:

1. Apply Laplace transform (32.1) to (19.5) to obtain a “transfer function”

$$\mathbf{X}(s) = \underbrace{(s\mathbf{I} - A)^{-1}}_{F(s)} [B\mathbf{U}(s) + \mathbf{x}(0)]. \quad (19.6)$$

2. Use algebraic manipulations, especially *partial fractions*, to transform $F(s)$ in (19.6) into a combination of recognisable forms;
3. Utilise the properties of inverse Laplace transform and Laplace transform table to find the analytic expression for $\Phi(t) = \mathcal{L}^{-1}\{F(s)\}$. Otherwise, leave the expression in integral form and solve it numerically:

$$\mathbf{x}(t) = \mathcal{L}^{-1}\{F(s)B\mathbf{U}(s)\} + \mathcal{L}^{-1}\{F(s)\}\mathbf{x}(0) = \int_0^t \Phi(t - \tau)B\mathbf{u}(\tau)d\tau + \Phi(t)\mathbf{x}(0).$$

For examples, please refer to UECM2023 Ordinary Differential Equations, Kwok [2010, Chapter 7] or Mathews and Howell [1997, Chapter 11].

Modern control system are digital instead of analogue. Therefore, the technique of Laplace transform has to be complemented with z -transform to investigate the “frequency response” of digital or computerised control systems [Moudgalya, 2007].

19.1.3 Complex Numbers in AC Circuit Theory

When Fourier transform is applied to the study of AC (alternating current) linear circuits operating at a certain frequency, we will be dealing with complex numbers and its arithmetic which we learned in Topic 1.

Electricity is produced by generators at power stations and then distributed by a vast network of transmission lines (called the National Grid system) to industry and for domestic use. It is easier and cheaper to generate alternating current (AC) and it is more conveniently distributed since its voltage can be readily altered using transformers Bird [2014].

The analysis of basic single-phase AC electric circuits containing impedances and ideal AC supplies can be done with the use of “complex numbers” (“phasors” in engineering literature). In this simplified model, operating at a frequency f , the relation between two physical quantities that characterises a resistor, a capacitor and an inductor:

$$V_R = RI_R, \quad C \frac{dV_C}{dt} = I_C, \quad V_I = L \frac{dI_I}{dt}$$

are going to be simplified to

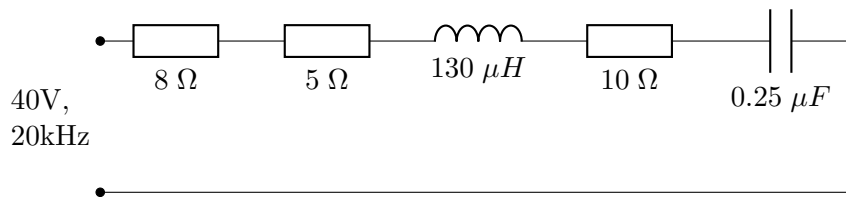
$$V_R = RI_R, \quad V_C = \frac{1}{2\pi i f C} I_C, \quad V_I = 2\pi i f L I_I$$

where constants R , C and L are known as the resistance (unit: Ω), capacitance (unit: F) and inductance (unit: H) respectively. Note that engineers use “ j ” instead of “ i ”.

Definition 19.9. The *impedance* Z of a circuit is the ratio of the voltage against the current. The *admittance* Y is the inverse of impedance Z .

Based on rule that the voltage across two nodes in with linear components in series will be the sum of the voltages of each component and the sum of all currents entering a node is zero, we can solve the problems of single-phase AC electric circuits.

Example 19.5. [Bird, 2014, 225] The following circuit components are connected in series across a 40V, 20kHz supply. Calculate (a) the circuit peak current $|I|$, (b) the circuit phase angle ϕ . For engineers, the circuit current is $I = V/Z$ and the circuit phase angle is $\text{Arg}(Z) \times \frac{180^\circ}{\pi}$. Write your answers to at least 4 decimal places.



Solution: (a) Using what we learned in SPM physics on the resistance of series circuit, we can formulate

$$\begin{aligned} V &= I \left(8 + 5 + 2\pi i (20k)(130\mu) + 10 + \frac{1}{2\pi i (20k)(0.25\mu)} \right) \\ &= I(23 + 16.3363i - 31.8310i) = (23 - 15.4947i)I \end{aligned}$$

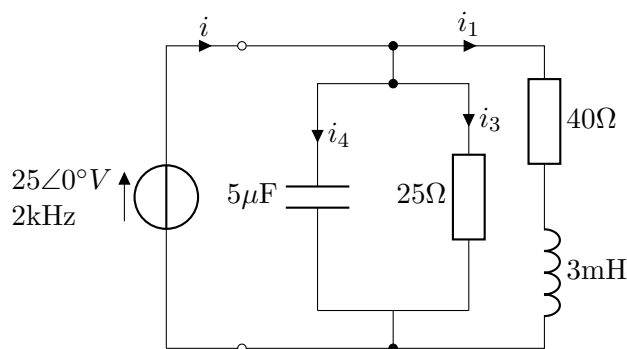
Hence, $I = \frac{40}{23 - 15.4947i} = 1.1962 + 0.8059i \Rightarrow |I| = 1.4423$.

(b) The phase angle is

$$\phi = \text{Arg}(23 - 15.4947i) \times \frac{180^\circ}{\pi} = -\tan^{-1} \frac{15.4947}{23} \frac{180^\circ}{\pi} = -33.9675^\circ.$$

Remark: Engineers called this “33.9675° lagging” (when the angle is positive, it is called “leading”).

Example 19.6. [Bird, 2014, 384] For the network shown below



determine (a) the total network impedance, (b) the total network admittance, (c) the supply current I .

Solution. (a) $f = 2000\text{Hz}$

- $\frac{1}{Z_1} = 2\pi i(2000)(5 \times 10^{-6}) + \frac{1}{25} = 0.04 + 0.06283185i$
- $Z_2 = 40 + 2\pi i(2000)(3 \times 10^{-3}) = 40 + 37.6991i$

Hence, $\frac{1}{Z} = \frac{1}{Z_1} + \frac{1}{Z_2} = 0.04 + 0.06283185i + \frac{1}{40 + 37.6991i} = 0.0532 + 0.0504i \Rightarrow Z \approx 9.9143 - 9.3769i$

Engineers like to write in the form $\approx 13.65\angle -43.40^\circ$

(b) $Y = \frac{1}{Z} \approx 0.0532 + 0.0504i \approx 0.76\angle 43.40^\circ$

(c) $I = \frac{V}{Z} \approx \frac{25\angle 0^\circ}{13.65\angle -43.40^\circ} = 1.83\angle 43.40^\circ$ □

19.1.4 Classical Control Theory

In system engineering, a single input single output linear system is usually described by an ODE:

$$a_n y^{(n)} + a_{n-1} y^{(n-1)} + \cdots + a_1 y' + a_0 y = b_m u^{(m)} + b_{m-1} u^{(m-1)} + \cdots + a_1 u' + a_0 u$$

where $y = y(t)$ is the output and $u = u(t)$ is the input.

To study the system response to input, we normally assume $y(0) = \cdots = y^{(n-1)}(0) = 0$ and $u(0) = \cdots = u^{(m-1)}(0) = 0$ and use Laplace transform (or Fourier Transform) to obtain

$$(a_n s^n + a_{n-1} s^{n-1} + \cdots + a_1 s + a_0)Y(s) = (b_m s^m + b_{m-1} s^{m-1} + \cdots + b_1 s + b_0)U(s),$$

in short,

$$Y(s) = \frac{B(s)}{A(s)}U(s).$$

The ratio $\frac{B(s)}{A(s)}$ is called the *transfer function* of the system. In general, the transfer function (against the original state equation) are related by the table below:

Type	Mathematical Model	State Equation	Transfer Function
analog system	differential equations	$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}, t)$ $\mathbf{y} = \mathbf{g}(\mathbf{x})$	$\mathbf{Y}(s) = \mathbf{G}(\mathbf{X}(s))$ $= \mathbf{F}(\mathbf{X}(s), \mathbf{U}(s), s)$
digital system	difference equations	$\mathbf{x}_k = \mathbf{f}(\mathbf{x}_{k-1}, \mathbf{u}_{k-1}, t_{k-1})$ $\mathbf{y}_k = \mathbf{g}(\mathbf{x}_k)$	$\mathbf{Y}(z) = \mathbf{G}(\mathbf{X}(z))$ $= \mathbf{F}(\mathbf{X}(z), \mathbf{U}(z), z)$

The state equation characterises the system's behaviour in time-domain whereas the transfer function characterises the system's behaviour in frequency-domain. The state equation and transfer function are linked by a linear transform defined below.

Root-locus technique is a useful tool for analysis and design of classical controllers. This technique chooses the controller gains that make the loci pass through given poles that come from the performances [Boukas and AL-Sunni, 2011b, 141].

Assume the transfer function can be factorised as

$$G(s) = k \frac{(s + z_1)(s + z_2) \cdots (s + z_m)}{(s + p_1)(s + p_2) \cdots (s + p_n)}$$

where k , $-z_i$, $-p_i$ are the gain, the zeros and the poles of the system respectively. If we choose a proportional controller with the following transfer function

$$C(s) = K_p$$

then the determination of the gain of the controller, K_p , is as follows:

1. Let $K = kK_p$ and $1 + K_pG(s) = 0$.
2. Draw the root locus for K varying from 0 to ∞ .
3. Determine the intersection between the loci and the line corresponding to the desired damping ratio ξ ($\cos \theta = \xi$) and get the dominant pair of poles. Let s_d be the one with positive imaginary part.
4. Compute the gain K that gives the pole s_d and then

$$K_p = \frac{(s_d + p_1)(s_d + p_2) \cdots (s_d + p_n)}{k(s_d + z_1)(s_d + z_2) \cdots (s_d + z_m)}.$$

Example 19.7. Find K_p for $G(s) = \frac{5}{s(s+1)}$ so that the system is stable with an overshoot $\leq 5\%$ and a steady state error of zero.

Solution: Let $K = 5K_p$, then

$$1 + \frac{K}{s(s+1)} = 0 \Rightarrow s^2 + s + K = 0 \Rightarrow (s + \frac{1}{2})^2 - \frac{1}{4} + K = 0 \Rightarrow s_d = -\frac{1}{2} + \frac{1}{2}i$$

More examples can be found in control system textbooks such as Ogata [2002], Dorf and Bishop [2011] and Boukas and AL-Sunni [2011b].

19.2 Application to Potential Theory

In this section, we will try to explore the connection between complex functions and two-dimensional (2D) Dirichlet problem, which is essential in the understanding of “static state” problem in physics and engineering. In physics, “static state” problem is an important part of potential theory.

Conformal mapping technique from Topic 8 can be applied to transform a general 2D Dirichlet problem to relatively easy to solve Dirichlet problem. However, with the introduction of powerful finite element methods, conformal mapping technique is no longer popular but serves as theoretical tool.

19.2.1 General Dirichlet Problem

Let us state the general Dirichlet problem.

Definition 19.10. If the values of a harmonic function are prescribed along the boundary, the problem is known as the *boundary value problem of the first kind* or (*general*) *Dirichlet problem*. If the values of the normal derivative of a harmonic function are prescribed on the boundary, the problem is known as the *boundary value problem of the second kind* or *Neumann problem*.

Given a two-dimensional bounded domain D and a function $g : \partial D \rightarrow \mathbb{R}$, the *Dirichlet problem* consists of finding a function $h : \bar{D} \rightarrow \mathbb{R}$ such that h is harmonic on D , continuous on $\bar{D}$, and coincident with g on ∂D . The Dirichlet problem can be formulated as

$$\begin{cases} \Delta h = 0, & (x, y) \in D, \\ h = g(x, y), & (x, y) \in \partial D =: C. \end{cases} \quad (19.7)$$

e:dirichlet

An important consequence of the maximum and minimum principle (Topic 6) is **uniqueness** for the Dirichlet problem. If h_1 and h_2 are both harmonic on D , continuous on $\bar{D}$, and agree on ∂D , then $h_1 - h_2$ is harmonic and vanishes on the boundary, but it takes its maximum and minimum on the boundary, so it is identically zero, i.e., $h_1 = h_2$.

The existence of solution is more complicated theoretical problem. In the section [sec:poisson](#) [e:dirichlet](#) 19.2.3, we will learn that if the function g in Equation (19.7) is continuous, then a solution exists.

19.2.2 Transformations of Boundary Conditions

Before trying to “solve” [e:dirichlet](#) (19.7), we need to transform it to something “nice”, for example Δ , so that we can formulate the solution. Theoretically, it is always possible to transform a domain with boundary points more than two values to Δ according to Riemann Mapping Theorem (Topic 8). What we are concerned is this: Is harmonic property conformally invariant?

The answer is positive. Harmonic property is preserved under conformal mapping.

Theorem 19.8 (Invariance of Laplace Equation). Suppose that f is a holomorphic function mapping a region D into a region $\tilde{D}$ and $\tilde{h}$ is a harmonic function on $\tilde{D}$. Then the function $h = \tilde{h} \circ f$ is harmonic on D . Thus, if $\tilde{h}$ satisfies $\Delta \tilde{h} = 0$ on $\tilde{D}$, then h satisfies $\Delta h = 0$ on D .

Proof: Let z_0 be a point in D and $w_0 = f(z_0)$. Topic 7 states that $\tilde{h}$ has a harmonic conjugate $\tilde{v}$ on a disk around w_0 (because a disk is simply connected). Then $\tilde{h} + i\tilde{v}$ is holomorphic on this disk. By the composition of holomorphic function, $(\tilde{h} + i\tilde{v}) \circ f$ is holomorphic on $\tilde{D}$. Therefore, $\Re((\tilde{h} + i\tilde{v}) \circ f) = \Re(\tilde{h} \circ f) = \tilde{h} \circ f$ is harmonic at z_0 . Since z_0 is arbitrary, $\tilde{h} \circ f$ is harmonic on D .

To see how Theorem [thm:laplace:inv](#) 19.8 can be of use, one can consider the Dirichlet problem [e:dirichlet](#) (19.7) on a domain D . We can transform D to a Dirichlet problem over a simpler domain such as $\mathbb{H}$ or Δ or an annulus. The Dirichlet problem over a simpler domain is normally easier to solve. These steps can be formalised as follows:

1. Describe clearly the region $\tilde{D} = f(D)$ and its boundary $\tilde{C} = f(C)$ in the w -plane.
2. Since f is one-to-one, we have an inverse function f^{-1} defined on $\tilde{D}$ and $\tilde{C}$.
3. Solve the transformed Dirichlet problem on $\tilde{D}$:

$$\begin{cases} \Delta \tilde{h}(w) = 0, & w \in \tilde{D}, \\ \tilde{h}(w) = g(f^{-1}(w)), & w \in \tilde{C}. \end{cases}$$

4. Let $h(z) = \tilde{h}(f(z))$ for all $z \in D$. Then h is a solution of our original Dirichlet problem on D (19.7) by the invariance of Laplace equation.

Note that the above procedure can be modified to solve the homogeneous Neumann problem, i.e. $\frac{dh}{dn} = 0$ on ∂D .

We will illustrate the above procedure with two examples below.

Example 19.9. ([Saff and Snider, 2003, p386]) Find a function $h(x, y)$ harmonic inside the unit disk $|z| < 1$ and satisfying the boundary condition $h(x, y) = \begin{cases} 1 & y > 0, \\ -1 & y < 0. \end{cases}$

Solution. We learned how to solve a Dirichlet problem on $\mathbb{H}$ in Topic 7, so we will find a conformal mapping from $D = \Delta$ to $\tilde{D} = \mathbb{H}$:

$$w = f(z) = -i \left[\frac{z+1}{z-1} \right], \quad \text{This is a special case of Topic 8 Equation (8.7)}$$

The transformed Dirichlet problem is

$$\begin{aligned} \Delta \tilde{h}(u, v) &= 0, & w = (u, v) \in \mathbb{H}, \\ \tilde{h}(u, 0) &= \begin{cases} 1, & u < 0, \\ -1, & u > 0. \end{cases} \end{aligned}$$

The transformed Dirichlet problem has a solution of the form $\tilde{h}(u, v) = A \operatorname{Arg}(w) + B$ (Check out Topic 7 Section 7.4). So

$$\begin{aligned} \tilde{h}(u > 0, 0) &= A(0) + B = -1, \\ \tilde{h}(u < 0, 0) &= A(\pi) - 1 = 1 \Rightarrow A = \frac{2}{\pi}. \end{aligned}$$

Hence,

$$\begin{aligned} \tilde{h}(u, v) &= \frac{2}{\pi} \operatorname{Arg}(w) - 1 \\ \Rightarrow h(x, y) &= \frac{2}{\pi} \operatorname{Arg}\left(-i \left[\frac{z+1}{z-1} \right]\right) - 1 = \frac{2}{\pi} \operatorname{Arg}\left(\frac{-i(x^2 + y^2 - 1) - 2y}{x^2 + y^2 - 1}\right) - 1 \\ &= \frac{2}{\pi} \tan^{-1}\left(\frac{x^2 + y^2 - 1}{2y}\right) - 1 = \frac{2}{\pi} \left[\tan^{-1}\left(\frac{x^2 + y^2 - 1}{2y}\right) - \frac{\pi}{2} \right] \\ &= \frac{2}{\pi} \left[-\cot^{-1}\left(\frac{x^2 + y^2 - 1}{2y}\right) \right] = \frac{2}{\pi} \left[\tan^{-1}\left(\frac{2y}{1 - (x^2 + y^2)}\right) \right] \end{aligned}$$

where $\cot^{-1} x = \tan^{-1} \frac{1}{x}$, $\tan^{-1} x + \cot^{-1} x = \frac{\pi}{2}$ and $-\tan^{-1} x = \tan^{-1}(-x)$. □

Example 19.10 (April 2008 Exam, Q4).

- (a) Prove that the transformation

$$f(z) = i \left(\frac{1-z}{1+z} \right)$$

maps the interior of the unit disc $\Delta = \{z \in \mathbb{C} : |z| < 1\}$ conformally to the upper half-plane $\mathbb{H} = \{z \in \mathbb{C} : \Im(z) > 0\}$ and maps the upper and lower unit semicircles C_+ and C_- onto the positive and negative real axis $\mathbb{R}_+$ and $\mathbb{R}_-$ respectively. (12 marks)

Proof: f is a Möbius transform, therefore, it maps the boundary of a Δ , which is a circle to a line if one of the image point is at infinity:

$$f(1) = 0, \quad f(i) = i \frac{1-i}{1+i} = 1, \quad f(-1) = \infty.$$

The upper semicircle C_+ maps to $\mathbb{R}_+$.

Complementary, C_- maps to $\mathbb{R}_-$. Mathematically, we prove it as follows:

$$f(e^{-it}) = i \frac{1 - e^{-it}}{1 + e^{-it}} = i \frac{(1 - e^{-it})(1 + e^{it})}{2 + 2 \cos t} = \frac{-2 \sin t}{2(1 + \cos t)} < 0$$

for $0 < t < \pi$.

(b) Prove that the function

$$\varphi(u, v) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{u}{v}\right)$$

solves the Dirichlet problem in the upper half-plane (**note**: there is a typo mistake in the original question):

$$\frac{\partial^2 \varphi}{\partial u^2} + \frac{\partial^2 \varphi}{\partial v^2} = 0 \quad \text{on } \mathbb{H}, \quad \varphi(u, v) = \begin{cases} 0, & \text{on } \mathbb{R}_-, \\ 1, & \text{on } \mathbb{R}_+. \end{cases} \quad (6 \text{ marks})$$

Proof: $\varphi(u < 0, v = 0) = \frac{1}{2} + \frac{1}{\pi} \arctan(-\infty) = \frac{1}{2} + \frac{1}{\pi} \times \frac{-\pi}{2} = 0;$

$$\varphi(u > 0, v = 0) = \frac{1}{2} + \frac{1}{\pi} \arctan(+\infty) = \frac{1}{2} + \frac{1}{\pi} \times \frac{\pi}{2} = 1.$$

$$\frac{\partial^2 \varphi}{\partial u^2} = \frac{1}{\pi} \frac{\partial}{\partial u} \left[\frac{1}{1 + \left(\frac{u}{v}\right)^2} \times \frac{1}{v} \right] = \frac{1}{\pi} \frac{-v}{(u^2 + v^2)^2} \times 2u$$

$$\frac{\partial^2 \varphi}{\partial v^2} = \frac{1}{\pi} \frac{\partial}{\partial v} \left[\frac{1}{1 + \left(\frac{u}{v}\right)^2} \times \frac{-u}{v^2} \right] = \frac{1}{\pi} \frac{u}{(u^2 + v^2)^2} \times 2v \Rightarrow \frac{\partial^2 \varphi}{\partial u^2} + \frac{\partial^2 \varphi}{\partial v^2} = 0$$

(c) Determine the solution to the Dirichlet problem in the unit disk Δ :

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} = 0 \quad \text{on } \Delta, \quad \varphi(x, y) = \begin{cases} 0, & \text{on } C_-, \\ 1, & \text{on } C_+. \end{cases} \quad (6 \text{ marks})$$

Solution. From the given $\varphi : \mathbb{H} \rightarrow \mathbb{R}$, $\psi : \Delta \rightarrow \mathbb{R}$ and conformal mapping $f : \Delta \rightarrow \mathbb{H}$, we have

$$f(z) = i \frac{1 - 2iy - x^2 - y^2}{1 + 2x + x^2 + y^2} = \underbrace{\frac{2y}{1 + 2x + x^2 + y^2}}_u + i \underbrace{\frac{1 - x^2 - y^2}{1 + 2x + x^2 + y^2}}_v$$

$$\Rightarrow \psi(z) = \varphi(f(z)) = 1 + \frac{1}{\pi} \arctan\left(\frac{2y}{1 - x^2 - y^2}\right).$$

□

19.2.3 Poisson's Integral Formula

sec:poisson

In order to solve a general Dirichlet problem (19.7) on a disk centred at 0 and of radius R , $D = \Delta_R(0)$, we need the Poisson kernel.

Definition 19.11. The *Poisson kernel* is defined as

$$P_r(\theta) = \Re \left(\frac{1 + re^{i\theta}}{1 - re^{i\theta}} \right) = \frac{1 - r^2}{1 - 2r \cos \theta + r^2} = \sum_{n=-\infty}^{\infty} r^{|n|} e^{in\theta}$$

where $0 \leq r < 1$, $\theta \in \mathbb{R}$.

Theorem 19.11. Let the radius $0 < R < +\infty$. If the real-valued functional $f : \partial\Delta_R(0) \rightarrow \mathbb{R}$ is continuous, then the function h defined below

$$\begin{aligned} h(z) &= f(z), & z \in \partial\Delta_R(0), \\ h(re^{i\theta}) &= \frac{1}{2\pi} \int_0^{2\pi} P_{r/R}(\theta - t) f(Re^{it}) dt, & 0 \leq r < R, \quad \theta \in \mathbb{R}. \\ &= \frac{1}{2\pi} \int_0^{2\pi} \frac{R^2 - r^2}{R^2 - 2Rr \cos(\theta - t) + r^2} f(Re^{it}) dt \end{aligned} \quad (19.8)$$

is continuous on $\overline{\Delta_R(0)}$ and harmonic on $\Delta_R(0)$.

The formula (19.8) is known as the *Poisson Integral Formula* and it is a solution to the general Dirichlet problem on a disk of radius R .

Theorem 19.12 (Harnack's Inequalities). If h is non-negative and harmonic on $\Delta_R(a)$, then

$$\frac{R-r}{R+r} h(a) \leq h(a + re^{i\theta}) \leq \frac{R+r}{R-r} h(a)$$

for $0 \leq r < R$.

19.2.4 Heat Problems

In a static heat flow problem, we can think of the temperature $T(x, y)$ as the solution of Dirichlet problem on D (19.7). The level curves $T(x, y) = C$ represent the curves of constant temperature, called the *isotherms*, inside D . The curves of heat flow inside D are the curves that give the direction along which heat is flowing inside D . The “curves of heat flow” are given by the level curves of the harmonic conjugate of T .

Example 19.13. Determine the isotherms and curves of heat flow for the Dirichlet problem

$$\begin{cases} \Delta T = 0, & (x, y) \in \mathbb{H}, \\ T = \begin{cases} 100, & x > 0, \\ 50, & x < 0. \end{cases} \end{cases}$$

Solution: The boundary values are 50 and 100, so the temperature T inside the region will vary between these two values. Note that the following function is a harmonic function on $\mathbb{R} \times (0, \infty)$:

$$T = \Im(A \log z + B) = A \operatorname{Arg}(z) + B.$$

When $y = \Im(z) = 0$, $x = \Re(z) > 0$,

$$T(x, 0) = A \times 0 + B = 100 \Rightarrow B = 100.$$

When $y = \Im(z) = 0$, $x = \Re(z) < 0$,

$$T(x, 0) = A \times \pi + B = 50 \Rightarrow A = \frac{-50}{\pi}.$$

Hence, the isotherms are given by the lines $T = 100 - \frac{50}{\pi} \operatorname{Arg}(z) = C$, $50 \leq C \leq 100$ or in tangent form:

$$\operatorname{Arg}(z) = \tan^{-1} \frac{y}{x} = \frac{(100 - C)\pi}{50} \Rightarrow y = x \tan\left(\frac{(100 - C)\pi}{50}\right)$$

The curves of heat flow is given by the harmonic conjugate (Topic 7) $v = \tilde{C}$ where

$$v = -\Re(100 - \frac{50}{\pi} \text{Log} z) = \frac{50}{\pi} \ln |z| - 100 = \tilde{C} \Rightarrow x^2 + y^2 = C'.$$

Remark 19.14. It is possible for us to have another solution? From Topic 7, we know that the maximum-minimum principle says that the solution is unique. Therefore, the solution we guessed in the above example is the only solution.

Example 19.15. ([Saff and Snider, 2003, Section 3.4]) Find the temperature distribution $T(x, y)$ that is harmonic in the domain $D = \{z \in \mathbb{C} : 1 < |z| < 2\}$ and takes values $T(z) = 20^\circ\text{C}$ when $|z| = 1$ and $T(z) = 30^\circ\text{C}$ when $|z| = 2$.

Solution. Since the boundary is circular and the boundary values are constant. We guess that the temperature distribution has the form

$$T(x, y) = A \text{Log}|z| + B$$

because Log is harmonic on circular domain and has the described feature.

Since $|z| = 1 \Rightarrow T(z) = A \text{Log} 1 + B = 20 \Rightarrow B = 20$ and $|z| = 2 \Rightarrow T(z) = A \text{Log} 2 + B = 30 \Rightarrow A = 10/\text{Log} 2$.

Therefore, $T(x, y) = \frac{10 \text{Log}|z|}{\text{Log} 2} + 20$. □

19.2.5 Electrostatic Field

In electricity and magnetism, the force of attraction or repulsion between charged particles is the gradient of a harmonic function h , known as the *electrostatic potential*. The level curves $h(x, y) = C$, represent the *equi-potential lines*. In this case, the level curves of a harmonic conjugate v of h give the direction of the electric force in the field and are known as the *line of force*.

Example 19.16. [Saff and Snider, 2003, 435] Find the electrostatic potential ϕ inside a capacitor formed by two nested parallel but non-concentric cylindrical conductors in the domain $D = \{z : |z| < 1\} \setminus \{z : |z - 0.3| \leq 0.3\}$ and takes that value 0 on the inner circle $C_1 : \{z : |z - 0.3| = 0.3\}$ and 1 on the outer circle $C_2 : \{z : |z| = 1\}$.

Solution. We learned in Tutorial 8 that the conformal mapping below

$$f(z) = \frac{z - \frac{1}{3}}{z - 3}$$

will allow us to transform the non-concentric “circles” to concentric circles at 0 of radius $|f(1)| = |f(-1)| = \frac{1}{3}$ and $|f(0)| = |f(0.6)| = \frac{1}{9}$.

The transform Dirichlet problem has a general solution

$$\tilde{h}(w) = A \text{Log}|w| + B$$

where

$$\tilde{h}(|w| = \frac{1}{3}) = A \text{Log} \frac{1}{3} + B = 1$$

$$\tilde{h}(|w| = \frac{1}{9}) = A \text{Log} \frac{1}{9} + B = 0 \Rightarrow A = \frac{1}{\ln 3}, B = 2$$

Hence,

$$\begin{aligned} h(z) &= \tilde{h}(f(z)) = \frac{\operatorname{Log} \left| \frac{z - \frac{1}{3}}{z - 3} \right|}{\ln 3} + 2 = \frac{\operatorname{Log} \left| \frac{3z - 1}{z - 3} \right| - \ln 3}{\ln 3} + 2 = \frac{1}{2 \ln 3} \operatorname{Log} \frac{(3x - 1)^2 + (3y)^2}{(x - 3)^2 + y^2} + 1 \\ &= \frac{1}{\ln 3} \left[\ln 3 + \frac{1}{2} \operatorname{Log}((3x - 1)^2 + 9y^2) - \frac{1}{2} \operatorname{Log}((x - 3)^2 + y^2) \right] \end{aligned}$$

□

The domain of the following example is a disk of radius R . We can apply the Poisson integral formula (19.8). However, we find that the integration process is complicated. In comparison, the approach in Example 19.9 is probably more convenient for us to find an analytic solution.

Example 19.17. [Kwok, 2010, Example 7.1.1] An infinitely long metal cylinder of radius R is cut into two halves. The upper half is grounded and the lower half is maintained at a fixed potential K . Find the electrostatic potential at points inside the cylinder.

Solution: The infinite length of the metal cylinder can be nicely approximated by an equivalent the Dirichlet problem in the domain $|z| < R$ with boundary condition

$$h(r, \theta) = \begin{cases} 0, & 0 < \theta = \arg(z) < \pi, \\ K, & \pi < \theta < 2\pi. \end{cases}$$

The solution is

$$h(r, \theta) = \frac{1}{2\pi} \int_{\pi}^{2\pi} \frac{R^2 - r^2}{R^2 - 2Rr \cos(\theta - t) + r^2} \times K dt$$

To calculate evaluate the above integral, we use the “universal substitution” $s = \tan \frac{t - \theta}{2}$ so that $\cos(t - \theta) = \frac{1 - s^2}{1 + s^2}$ and $dt = \frac{2}{1 + s^2} ds$. The integral is then reduced to

$$\begin{aligned} h(r, \theta) &= \frac{K}{2\pi} \int_{\tan(\frac{\pi - \theta}{2})}^{\tan(\frac{2\pi - \theta}{2})} \frac{R^2 - r^2}{R^2 - 2Rr \frac{1 - s^2}{1 + s^2} + r^2} \frac{2}{1 + s^2} ds \\ &= \frac{K(R^2 - r^2)}{\pi} \int_{\cot \frac{\theta}{2}}^{-\tan(\frac{\theta}{2})} \frac{1}{(R^2 + r^2)(1 + s^2) - 2Rr(1 - s^2)} ds \\ &= \frac{K(R^2 - r^2)}{\pi(R - r)^2} \int_{\cot \frac{\theta}{2}}^{-\tan(\frac{\theta}{2})} \frac{1}{1 + (\frac{R+r}{R-r})^2 s^2} ds = \frac{K}{\pi} \int_{\cot \frac{\theta}{2}}^{-\tan(\frac{\theta}{2})} \frac{1}{1 + u^2} du \\ &= \frac{K}{\pi} \left[\tan^{-1} u \right]_{\cot \frac{\theta}{2}}^{-\tan(\frac{\theta}{2})} = \frac{K}{\pi} \left[\tan^{-1} \frac{R+r}{R-r} s \right]_{\cot \frac{\theta}{2}}^{-\tan(\frac{\theta}{2})} \\ &= \frac{K}{\pi} \left[\tan^{-1} \frac{\frac{R+r}{R-r} (-\tan(\frac{\theta}{2}) - \cot(\frac{\theta}{2}))}{1 - (\frac{R+r}{R-r})^2 \tan(\frac{\theta}{2}) \cot(\frac{\theta}{2})} \right] \\ &= \frac{K}{\pi} \left[\tan^{-1} \frac{(R^2 - r^2)(\tan(\frac{\theta}{2}) + \cot(\frac{\theta}{2}))}{4Rr} \right] = \frac{K}{\pi} \left[\tan^{-1} \frac{R^2 - r^2}{2Rr \sin \theta} \right] \end{aligned}$$

where $u = \frac{R+r}{R-r} s$.

Note that we use the following formula from trigonometry and Calculus:

$$\begin{aligned}\tan(A \pm B) &= \frac{\tan A \pm \tan B}{1 \mp \tan A \tan B} \Rightarrow \tan A - \tan B = \tan(A - B)(1 + \tan A \tan B), \\ \tan(\pi - \frac{\theta}{2}) &= -\tan \frac{\theta}{2}, \quad \tan(\frac{\pi}{2} - \frac{\theta}{2}) = \cot \frac{\theta}{2}, \\ \tan^{-1} x - \tan^{-1} y &= \tan^{-1} \frac{x - y}{1 + xy}, \quad \int \frac{1}{1 + x^2} dx = \tan^{-1} x + C, \\ \sin \theta &= \frac{2 \tan \frac{\theta}{2}}{1 + \tan^2 \frac{\theta}{2}} = \frac{2}{\cot \frac{\theta}{2} + \tan \frac{\theta}{2}}.\end{aligned}$$

For more examples, consult Gibbs [1958].

19.2.6 Fluid Flow

One of the most significant applications of conformal mapping is to flows [Fisher, 1999]. For a sourceless and irrotational flow on some domain D , the velocity vector $V = \phi_x(x, y) + i\phi_y(x, y)$ where ϕ is the velocity potential. The function ϕ has a harmonic conjugate $\psi(x, y)$ such that the velocity vector is tangent to a curve $\psi(x, y) = C_2$. The curves $\psi(x, y) = C_2$ are called the *streamlines* of the flow, and the function ψ is called the *stream function*. In particular, a boundary across which fluid cannot flow is a streamline.

The holomorphic function

$$F(z) = \phi(x, y) + i\psi(x, y)$$

is called the *complex potential* of the flow and the velocity is $V = \overline{F'(z)}$.

Steady fluid flow problems are typically of two types. The first is when an impermeable object is placed in a flow of uniform velocity across the plane. The resulting flow *around* the object is to be found, assuming a (virtually) uniform velocity far from the object.

The second type of flow is the flow through a channel or around a corner, or generally in a simply-connected region whose boundary consists of straight lines except for a few irregularities or corners. The boundary is impermeable, and the flow is assumed to be (virtually) uniform “at ∞ ”, i.e. far from the irregular part of the boundary. Both types of flow problem can be handled by consideration of appropriate conformal mappings.

Example 19.18. [Churchill and Brown, 2008, 409] Let a long circular cylinder of unit radius be placed in a large body of fluid flowing with a uniform velocity, the axis of the cylinder being perpendicular to the direction of flow. Determine the complex potential F as well the velocity around the cylinder.

Solution: To determine the steady flow around the cylinder, we represent the cylinder by the circle $|z| = 1$ and let the flow distant from it to be parallel to the x axis and to the right. Symmetry shows that the points on the x axis exterior to the circle may be treated as boundary points, and so we need to consider only the upper half plane, i.e. $\mathbb{H}$.

The boundary of the region D of the flow, consisting of the upper semi-circle and the parts of the x -axis exterior to the circle, is mapped onto the entire u -axis by the Joukowski’s function (Tutorial 8, Q7, $\alpha = 2$):

$$w = z + \frac{1}{z}.$$

The region D is mapped onto the $\mathbb{H}$.

The complex potential for the corresponding uniform flow in $\mathbb{H}$ is

$$\tilde{F}(w) = Aw, \quad A > 0$$

and the complex potential of the original steady state flow in D is

$$F(z) = \tilde{F}(z + \frac{1}{z}) = A(z + \frac{1}{z}), \quad A > 0.$$

The velocity is

$$V = \overline{F'(z)} = \overline{A(1 - \frac{1}{z^2})} = A(1 - \frac{1}{\bar{z}^2}).$$

Part III

Linear Algebra

Chapter 20

Vectors

ch:linalg:vec

In scheme, a vector is a data storage of fixed length n .

In armadillo C++, vectors are encoded as column vectors (default) and row vectors.

```
arma::cx_vec      = arma::cx_colvec  = arma::Col<cx_double> colvec(n); // 64-bit
arma::cx_fvec     = arma::cx_fcolvec = arma::Col<cx_float>  colvec(n); // 32-bit
arma::cx_rowvec   = arma::Row<cx_double> rowvec(n); // 64-bit
arma::cx_frowvec  = arma::Row<cx_float>  rowvec(n); // 32-bit
```

The tools asymptote and povray can be used for creating relatively pretty diagrams for vectors.

Textbooks:

Poole, D. (2015). Linear algebra a modern introduction.(4th ed.). USA: Cengage Learning

Strang, G. (2006). Linear algebra and its applications. (4th ed.). New York: Thomson.

DeFranza, J., & Gagliardi, D. (2008). Introduction to linear algebra with applications. New York: McGraw-Hill.

Anton, H. and Busby, R.C., 2003. Contemporary linear algebra. Hoboken, N. J.: John Wiley & Sons.

Farlow, J, Hall, J, McDill, J, West, B. (2007). Differential equations and linear algebra. (2nd ed.). Harlow: Prentice-Hall/Pearson.

Need to add complex vectors. References:

PaulRHalmos-FiniteDimensionalVectorSpaces-Springer1974;

LinearAlgebra2ndEd-KennethHoffman+RayKunze-1971 (inner product space over complex field)

The primary purpose of linear algebra is to build up the clarity of thought and expression. One thing we can learn from the definitions, theorems and proofs is how to think clearly and express mathematics clearly, to avoid misunderstanding and confusion.

Linear algebra begins as the measure of basic geometric objects such as the measurement of lengths and areas of triangles and rectangles. In the 17th century, it was applied in the study of mechanics (part of physics) and in 20th century, it was applied in the study of statistics.

In this topic, we will begin with two-dimensional vectors and then proceed to the study of three-dimensional vectors, and the related concepts. Finally, we try to generalise various concepts to n -dimensional vectors.

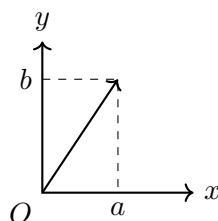
20.1 Two-Dimensional Vectors

Intuitively, a *vector* is a quantity which has *magnitude* and *direction*. Vector quantities are denoted as: $\mathbf{u}$, $\vec{u}$ or $\underline{u}$, $\underline{u}$ or $\overrightarrow{OA}$.

A *two-dimensional vector* (normally called a *2-d vector*) is written in the form of two-tuple (x, y) . For example, $(1, 2)$, $(-2, 3)$. The *components* of a 2-d vector $\mathbf{v} = (x, y)$ are numbers x

and y . Two vectors are *equal* iff they have the same corresponding components.

Example 20.1. When a 2-d coordinate system is given, the vector (a, b) can be sketched in the 2-dimensional plane as follows (the sketched below are for $a, b > 0$).



If the initial point of a vector $\mathbf{v}$ is A and the terminal point is B . We write $\mathbf{v} = \overrightarrow{AB}$. Take any point $A(x_0, y_0)$ as the initial point and $B(x_0 + a, y_0 + b)$ as the terminal point. Then, we write $\overrightarrow{AB} = (a, b)$.

Example 20.2. If $A = (-2, 3)$ and $B = (1, 5)$, what is the vector $\overrightarrow{AB}$?

Solution. $\overrightarrow{AB} = \overrightarrow{OB} - \overrightarrow{OA} = (3, 2)$

□

Example 20.3. Suppose A and B are points in the 2-dimensional plane and vector $\overrightarrow{AB} = (3, -1)$. If $A = (-4, 2)$, what are the coordinates of B ?

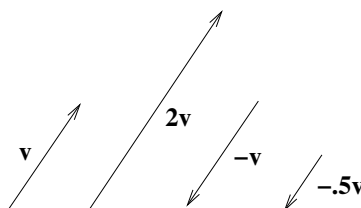
Solution. $\overrightarrow{AB} = (3, -1) = \overrightarrow{OB} - \overrightarrow{OA} = \overrightarrow{OB} - (-4, 2)$.

Therefore, $\overrightarrow{OB} = (-1, 1)$.

□

Definition 20.1 (Scalar Multiplication). A *scalar* is a quantity which has *only* magnitude. Let $\mathbf{v}$ be a nonzero vector and k be a nonzero real number (scalar). Then the *scalar multiple* of $\mathbf{v}$ is the vector $k\mathbf{v}$ which is parallel to $\mathbf{v}$ but with $|k|$ times the magnitude of $\mathbf{v}$. If $k > 0$, then $k\mathbf{v}$ is in the same direction as $\mathbf{v}$. If $k < 0$, then $k\mathbf{v}$ is in the opposite direction to $\mathbf{v}$. $0\mathbf{v} = (0, 0)$ is a vector with zero length. It is called a *zero vector* and is denoted as $\mathbf{0}$.

Example 20.4. If a 2-d vector $\mathbf{v}$ is given, the sketches of the scalar multiples $2\mathbf{v}$, $-\mathbf{v}$ and $-0.5\mathbf{v}$ can be shown below.



Definition 20.2. Two vectors $\mathbf{u}$ and $\mathbf{v}$ are said to be *parallel* if one is a scalar multiple of the other, i.e. if $\mathbf{u} = c\mathbf{v}$ or $\mathbf{v} = k\mathbf{u}$ for some scalars c, k .

Definition 20.3 (Vector Addition). Let $\mathbf{u} = (a, b)$ and $\mathbf{v} = (p, q)$ be two vectors. The *sum* (or *addition*) of the two vectors is defined as

$$\mathbf{u} + \mathbf{v} = (a + p, b + q).$$

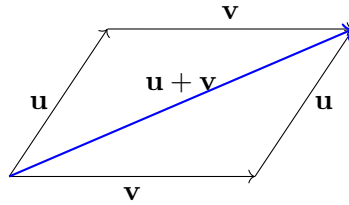
Example 20.5. Let $\mathbf{u} = (1, 2)$ and $\mathbf{v} = (3, 1)$. Then $\mathbf{u} + \mathbf{v} = (4, 3)$

Definition 20.4 (Vector Subtraction). Let $\mathbf{u} = (a, b)$ and $\mathbf{v} = (p, q)$ be two vectors. The *difference* (or *subtraction*) of the two vectors is defined as

$$\mathbf{u} - \mathbf{v} = \mathbf{u} + (-\mathbf{v}) = (a - p, b - q).$$

Example 20.6. $(7, 2) - (5, 3) = (2, -1)$.

Geometrically, the vector addition can be described by the **parallelogram rule**. This means, if the arrows for $\mathbf{u}$ and $\mathbf{v}$ are drawn with the same starting point, $\mathbf{u} + \mathbf{v}$ is the diagonal of the parallelogram whose sides are $\mathbf{u}$ and $\mathbf{v}$.



Another geometrical interpretation of the vector addition would be given by the **triangle rule**, i.e. if the arrow of $\mathbf{v}$ is drawn with its starting point being the terminal point of $\mathbf{u}$, then $\mathbf{u} + \mathbf{v}$ is the third side of the triangle with sides $\mathbf{u}$ and $\mathbf{v}$.

thm:1.1.6

Theorem 20.7 (Properties of Vector Addition and Scalar Multiplication). Let $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ be vectors and let c and d be scalars.

1. (Closed under addition) $\mathbf{u} + \mathbf{v}$ is a vector.
2. (Additive Commutative Law) $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$.
3. (Additive Associative Law) $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$.
4. (Zero) $\mathbf{u} + \mathbf{0} = \mathbf{0} + \mathbf{u} = \mathbf{u}$.
5. (Negatives) There is a vector called the negative of $\mathbf{u}$ and denoted $-\mathbf{u}$ with the property that $\mathbf{u} + (-\mathbf{u}) = (-\mathbf{u}) + \mathbf{u} = \mathbf{0}$.
6. (Closed under scalar multiplication) $c\mathbf{u}$ is a vector.
7. (Scalar Associativity) $c(d\mathbf{u}) = (cd)\mathbf{u}$.
8. (One) $1\mathbf{u} = \mathbf{u}$.
9. (Distributive Law) $c(\mathbf{u} + \mathbf{v}) = c\mathbf{u} + c\mathbf{v}$ and $(c + d)\mathbf{u} = c\mathbf{u} + d\mathbf{u}$.

With the above rules, we can simplify calculations with vectors as demonstrated in the following examples.

Example 20.8. $4(\mathbf{u} - 3\mathbf{v}) + 6(-2\mathbf{u} + 3\mathbf{v}) = -8\mathbf{u} + 6\mathbf{v}$

def:lin:comb

Definition 20.5. A *linear combination* of vectors $\mathbf{u}$ and $\mathbf{v}$ is a vector of the form $a\mathbf{u} + b\mathbf{v}$, where a and b are real numbers. More generally, a linear combination of k vectors $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k$ is a vector of the form

$$c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + \dots + c_k\mathbf{u}_k,$$

where $c_1, c_2, \dots, c_k$ are real numbers.

Example 20.9. Why is $(-5, 9)$ a linear combination of $\mathbf{u} = (-2, 3)$ and $\mathbf{v} = (-1, 1)$?

Solution: Because $(-5, 9) = 4(-2, 3) - 3(-1, 1) = 4\mathbf{u} - 3\mathbf{v}$.

Example 20.10. Is $(3, 0)$ a linear combination of $(-1, 2)$ and $(-3, 3)$?

Solution. Let a and b be two real numbers. $(3, 0) = a(-1, 2) + b(-3, 3) = (-a - 3b, 2a + 3b)$. From here, we can find $a = 3$ and $b = -2$. Therefore the answer is yes since $(3, 0) = 3(-1, 2) + (-2)(-3, 3)$. $\square$

Example 20.11. Is $(1, 1)$ a linear combination of $(1, -1)$ and $(-1, 1)$?

Solution. No. If there exist $a, b \in \mathbb{R}$ such that $(1, 1) = a(1, -1) + b(-1, 1)$. Then $a - b = 1$ and $b - a = 1$ which implies $-1 = 1$. A contradiction. $\square$

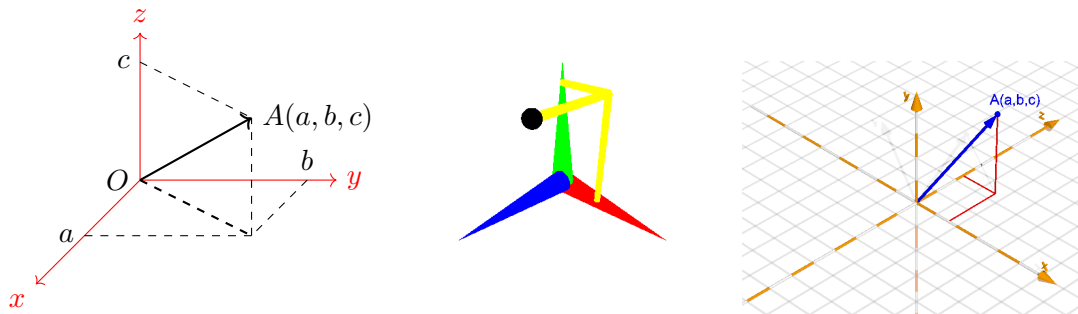
Theorem 20.12. If 2-dimensional nonzero vectors $\mathbf{u}$ and $\mathbf{v}$ are not parallel, then the set of linear combinations of $\mathbf{u}$ and $\mathbf{v}$ is the entire xy -plane.

Definition 20.6. The *standard basis* in the 2-dimensional plane are $\mathbf{i} = (1, 0)$ and $\mathbf{j} = (0, 1)$.

Example 20.13. (a, b) in standard basis form is $a\mathbf{i} + b\mathbf{j}$.

20.2 Three-Dimensional Vectors

In order to “sketch” a three-dimensional (abbreviated as 3d) vectors, we need to specify the coordinate system. However, there are 3 different coordinate systems that are used in practice — the normal right-handed system used in mathematics teaching (often called oblique project in computer graphics), the OpenGL right-handed system and the Renderman (and DirectX) left-handed system. They are sketched below.



Remark 20.14. In order to demonstrate the complexity of OpenGL program, the C program that generated the middle picture is demonstrated below.

```
#include <math.h>
#include <GL/glut.h>

GLint winx=500, winy=400;          // Window Size
GLfloat a=0.8, b=1.7, c=1.2;
GLfloat cPos[]={2.5,2.5,2.5};      // Camera Position. Looking at origin
GLfloat cU[]={0.,1.,0.};           // Camera Up in Z-direction
GLfloat clipping_near=0., clipping_far=-10.;

void init(void);
void keyboard(unsigned char key, int x, int y);
void display(void);
void reshape(int, int);
void renderCylinderP(float, float, float, float, float, float, float, int);

int main(int argc, char **argv)
{
    glutInit(&argc, argv);
```

```

    glutInitDisplayMode(GLUT_DOUBLE | GLUT_RGB);
    glutInitWindowPosition(100, 100);
    glutInitWindowSize(winx, winy);
    glutCreateWindow("OpenGL Coordinate System");

    init();
    glutDisplayFunc(display);
    glutKeyboardFunc(keyboard);
    glutReshapeFunc(reshape);
    glutMainLoop();
    return 0;
}

void keyboard(unsigned char key, int x, int y) {
    switch (key) {
        case 'q': ;
        case 27: exit (0); // ESCAPE key
    }
}

void init(void) {
    glClearColor(1.,1.,1.,0.);
}

void display(void) {
    glClear(GL_COLOR_BUFFER_BIT);
    gluLookAt (cPos[0],cPos[1],cPos[2],0.,0.,0.,cU[0],cU[1],cU[2]);

    glPushMatrix();
    glColor3f(1.,0.,0.);
    glTranslatef(0.,0.,0.);
    // rotate z-vector along y-axis to become x-axis
    glRotatef(90.0f, 0.0f, 1.0f, 0.0f);
    GLUQuadricObj *quadObj = gluNewQuadric();
    gluQuadricDrawStyle(quadObj, GLU_FILL);
    // gluCylinder(quadObj, base_radius, top_radius, height, slices, stacks);
    gluCylinder(quadObj, 0.2, 0.0, 2, 10, 4);
    glPopMatrix();

    glPushMatrix();
    glColor3f(0.,1.,0.);
    glTranslatef(0.,0.,0.);
    // rotate z-vector along x-axis to become y-axis
    glRotatef(90.0f, -1.0f, 0.0f, 0.0f);
    gluCylinder(quadObj, 0.2, 0.0, 2, 10, 4);
    glPopMatrix();

    // z-axis
    glPushMatrix();
    glColor3f(0.,0.,1.);
    glTranslatef(0.,0.,0.);
    gluCylinder(quadObj, 0.2, 0.0, 2, 10, 4);
    glPopMatrix();

    // Coordinate indicator
    renderCylinderP(a,0.,0.,a,b,0.,0.05,10);
    renderCylinderP(0.,b,0.,a,b,0.,0.05,10);
    renderCylinderP(a,b,0.,a,b,c,0.05,10);

    // The Point A(a,b,c)
    glPushMatrix();
    glColor3f(0.,0.,0.);
    glTranslatef(a,b,c);
    gluSphere(quadObj, 0.1, 15, 15);
    glPopMatrix();

    glutSwapBuffers();
}

void reshape (int w, int h) {
    glViewport (0, 0, (GLsizei)w, (GLsizei)h);
    glMatrixMode (GL_PROJECTION);
    glLoadIdentity ();
    //gluPerspective (60, (GLfloat)w / (GLfloat)h, 1.0, 100.0);
    gluPerspective (60, (GLfloat)w / (GLfloat)h, clipping_near, clipping_far);
    glMatrixMode (GL_MODELVIEW);
}

```

```

/* Rendering cylinder between two points */
void renderCylinder(float x1, float y1, float z1,
    float x2, float y2, float z2, float radius,
    int subdivisions, GLUquadricObj *quadric) {
    float vx = x2-x1, vy = y2-y1, vz = z2-z1;

    //handle the degenerate case of z1 == z2 with an approximation
    if(vz == 0) vz = .0001;

    float v = sqrt( vx*vx + vy*vy + vz*vz );
    float ax = 57.2957795*acos( vz/v );
    if ( vz < 0.0 ) ax = -ax;
    float rx = -vy*vz, ry = vx*vz;

    glPushMatrix();
    glColor3f(1.,1.,0.);
    glTranslatef( x1,y1,z1 );
    glRotatef(ax, rx, ry, 0.0);
    gluQuadricOrientation(quadric, GLU_OUTSIDE);
    gluCylinder(quadric, radius, radius, v, subdivisions, 1);

    //draw the first cap
    gluQuadricOrientation(quadric, GLU_INSIDE);
    gluDisk( quadric, 0.0, radius, subdivisions, 1);
    glTranslatef( 0,0,v );
    //draw the second cap
    gluQuadricOrientation(quadric, GLU_OUTSIDE);
    gluDisk( quadric, 0.0, radius, subdivisions, 1);
    glPopMatrix();
}

void renderCylinderP(float x1, float y1, float z1,
    float x2, float y2, float z2, float radius, int subdivisions) {
    //the same quadric can be re-used for drawing many cylinders
    GLUquadricObj *quadric=gluNewQuadric();
    gluQuadricNormals(quadric, GLU_SMOOTH);
    renderCylinder(x1,y1,z1,x2,y2,z2,radius,subdivisions,quadric);
    gluDeleteQuadric(quadric);
}

```

Vectors in 3-d space will be a 3-tuple. The vector addition, scalar multiplication and properties associated with these operations generalised naturally to 3-d vector. In order to compare with 2-d vector, we look at the definition of vector addition and scalar multiplication in 3-d space.

Definition 20.7. Let $\mathbf{u} = (a, b, c)$ and $\mathbf{v} = (p, q, r)$ be two vectors in 3-d space and let k be any real number. The *sum* (resp. *scalar multiplication*) of the two vectors is defined as

$$\mathbf{u} + \mathbf{v} = (a + p, b + q, c + r) \quad (\text{resp. } k\mathbf{u} = (ka, kb, kc)).$$

Definition 20.8. The *standard basis* in the 3-d space are $\mathbf{i} = (1, 0, 0)$, $\mathbf{j} = (0, 1, 0)$, $\mathbf{k} = (0, 0, 1)$.

Since the notion of “linear combination” is important in linear algebra, we will study the 3-d example below.

Example 20.15. Determine whether the vector $\mathbf{x} = (-1, -2, 2)$ is a linear combination of $\mathbf{u} = (0, 1, 4)$, $\mathbf{v} = (-1, 1, 2)$ and $\mathbf{w} = (3, 1, -2)$.

Solution. Let a, b, c be real numbers. Then assume

$$(-1, -2, 2) = a(0, 1, 4) + b(-1, 1, 2) + c(3, 1, -2).$$

Solving these simultaneous gives a consistent solution $a = 1$, $b = -2$ and $c = -1$ which tells

that $\mathbf{x}$ is a linear combination of $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$. □

20.3 Length and Angle

First, let us state a theorem that allows to derive the length of 2-d and 3-d vectors.

Theorem 20.16 (Pythagorean Theorem). *In any right-angled triangle, the area of the square whose side is the hypotenuse (the side opposite the right angle) is equal to the sum of the areas of the squares whose sides are the two legs (the two sides that meet at a right angle).*

The length of a vector $\mathbf{u}$ is often called the *norm* of $\mathbf{u}$ and is denoted by $\|\mathbf{u}\|$.

It follows from the *Pythagorean Theorem* that the norm of

- a 2-d vector $\mathbf{u} = (a, b)$ is $\|\mathbf{u}\| = \sqrt{a^2 + b^2}$.
- a 3-d vector $\mathbf{u} = (a, b, c)$ is $\|\mathbf{u}\| = \sqrt{a^2 + b^2 + c^2}$.

Example 20.17. If $\mathbf{u} = (3, -4)$, then $\|\mathbf{u}\| = \sqrt{3^2 + (-4)^2} = 5$

Example 20.18. If $\mathbf{u} = (-1, 2, 1)$ and $\mathbf{v} = (1, 0, -1)$, then
 $\|\mathbf{u}\| = \sqrt{1 + 4 + 1} = \sqrt{6}$ $\|\mathbf{v}\| = \sqrt{1 + 0 + 1} = \sqrt{2}$

Definition 20.9. A *unit vector* is a vector of norm 1.

Example 20.19. In a 2-d plane, $\mathbf{i}$, $\mathbf{j}$ and $\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ are unit vectors.

Example 20.20. In a 3-d space, $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$, $\left(\frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}}\right)$ are unit vectors.

Mathematicians have obtained the characterisation of the 2-D unit vectors by using sine and cosine functions as follow.

thm:1.2.9

Theorem 20.21. *Any 2-dimensional unit vector is of the form $(\cos \theta, \sin \theta)$ for some angle $0 \leq \theta < 2\pi$.*

Proof: Let (a, b) be any fixed 2-d unit vector. Since

$$|a| \leq \sqrt{a^2 + b^2} = 1.$$

Based on a theory from calculus, the following equation

$$\cos \theta = a$$

has a solution θ , if we restrict ourselves to $0 \leq \theta < 2\pi$, there should be a solution but no more than 2 solutions. Since $\sin \theta = \sqrt{1 - \cos^2 \theta} = \sqrt{1 - a^2} = b$, we shown that there exists θ such that $(\cos \theta, \sin \theta) = (a, b)$.

thm:1.2.4

Theorem 20.22. *For a 2-d or 3-d vector $\mathbf{u}$, $\|c\mathbf{u}\| = |c|\|\mathbf{u}\|$.*

Proof: We will just prove the case for a 3-d vector $\mathbf{u} = (u_1, u_2, u_3)$ as follows. Let c be any

real number.

$$\begin{aligned}\|\mathbf{c}\mathbf{u}\| &= \|(cu_1, cu_2, cu_3)\| = \sqrt{(cu_1)^2 + (cu_2)^2 + (cu_3)^2} \\ &= \sqrt{c^2(u_1^2 + u_2^2 + u_3^2)} = \sqrt{c^2} \sqrt{u_1^2 + u_2^2 + u_3^2} = |c| \|\mathbf{u}\|\end{aligned}$$

A unit vector of a given vector is normally treated as the direction of the given vector. This is stated in the following theorem.

Theorem 20.23. For any vector $\mathbf{v} \neq \mathbf{0}$, the vector $\frac{1}{\|\mathbf{v}\|} \mathbf{v}$ is a unit vector in the direction of $\mathbf{v}$.

Proof. Let $c = \frac{1}{\|\mathbf{v}\|}$ and using Theorem [thm:1.2.4](#) 20.22, we have

$$\left\| \frac{1}{\|\mathbf{v}\|} \mathbf{v} \right\| = \|\mathbf{c}\mathbf{v}\| = |c| \|\mathbf{v}\| = \frac{1}{\|\mathbf{v}\|} \|\mathbf{v}\| = 1. \quad \square$$

Remark 20.24. $\frac{1}{\|\mathbf{v}\|} \mathbf{v}$ is often written as $\frac{\mathbf{v}}{\|\mathbf{v}\|}$.

Example 20.25. A unit vector in the opposite direction as $\mathbf{v}$ is $-\frac{\mathbf{v}}{\|\mathbf{v}\|}$.

Now, we define the notion of dot product between 2-D and 3-D vectors.

Definition 20.10. The *dot product* between two vectors is denoted as $\mathbf{u} \cdot \mathbf{v}$. The dot product (or *inner product*) of 2-dimensional vectors $\mathbf{u} = (a_1, b_1)$ and $\mathbf{v} = (a_2, b_2)$ is $a_1a_2 + b_1b_2$ and the dot product for 3-dimensional vectors $\mathbf{u} = (a_1, b_1, c_1)$ and $\mathbf{v} = (a_2, b_2, c_2)$ is $\mathbf{u} \cdot \mathbf{v} = a_1a_2 + b_1b_2 + c_1c_2$.

Example 20.26. If $\mathbf{u} = (-1, 2)$ and $\mathbf{v} = (4, -3)$, $\mathbf{u} \cdot \mathbf{v} = -4 - 6 = -10$

Example 20.27. If $\mathbf{u} = (-1, 0, 2)$ and $\mathbf{v} = (1, 2, 3)$, $\mathbf{u} \cdot \mathbf{v} = -1 + 0 + 6 = 5$

Theorem 20.28. For a 2-d or 3-d vector, $\|\mathbf{u}\|^2 = \mathbf{u} \cdot \mathbf{u}$, or equivalently, $\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}}$.

Proof: We will only consider the 3-d vector case. Let $\mathbf{u} = (u_1, u_2, u_3)$. Then

$$\mathbf{u} \cdot \mathbf{u} = u_1^2 + u_2^2 + u_3^2 = (\sqrt{u_1^2 + u_2^2 + u_3^2})^2 = \|\mathbf{u}\|^2.$$

Let us now investigate the notion of *angle between two vectors*. We fix the angle between vectors, θ , to be in the range $0 \leq \theta \leq \pi$. We will learn from the following theorem the reason for introducing the concept of “dot product”: It is related to the angle between two vectors.

[thm:cos-theta](#)

Theorem 20.29. For any 2-d or 3-d vectors $\mathbf{u}$, $\mathbf{v}$, let θ be the angle between $\mathbf{u}$ and $\mathbf{v}$, we have the following identity

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \|\mathbf{v}\| \cos \theta.$$

Proof: We will only show the proof for 2-d vectors. The proof for 3-d vectors is similar but slightly longer. Assume $\mathbf{u} = (x_1, y_1)$, $\mathbf{v} = (x_2, y_2)$.

Using the Cosine rule (a triangle in Euclidean geometry satisfies the formula: $a^2 = b^2 + c^2 - 2bc \cos A$. This formula is not true in non-Euclidean geometry), we have

$$\|\mathbf{u} - \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - 2\|\mathbf{u}\| \|\mathbf{v}\| \cos \theta$$

which implies

$$\begin{aligned}\cos \theta &= \frac{\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - \|\mathbf{u} - \mathbf{v}\|^2}{2\|\mathbf{u}\|\|\mathbf{v}\|} \\ &= \frac{x_1^2 + y_1^2 + x_2^2 + y_2^2 - [(x_1 - x_2)^2 + (y_1 - y_2)^2]}{2\|\mathbf{u}\|\|\mathbf{v}\|} \\ &= \frac{2x_1x_2 + 2y_1y_2}{2\|\mathbf{u}\|\|\mathbf{v}\|} \\ &= \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}\end{aligned}$$

Example 20.30. Find the angle between $\mathbf{u}$ and $\mathbf{v}$ if

1. $\mathbf{u} = (\sqrt{3}, 1)$ and $\mathbf{v} = (1, \sqrt{3})$. $\cos^{-1} \frac{\sqrt{3}+\sqrt{3}}{2 \cdot 2} = \cos^{-1} \frac{\sqrt{3}}{2} = \frac{\pi}{6}$
2. $\mathbf{u} = (1, 2, 3)$ and $\mathbf{v} = (-2, 1, -1)$. $\cos^{-1} \frac{-2+2-3}{\sqrt{14} \cdot \sqrt{6}} = \cos^{-1}(-0.327327) = 1.90427$

Definition 20.11. The vectors $\mathbf{u}$ and $\mathbf{v}$ are *orthogonal* (or *perpendicular*) iff $\mathbf{u} \cdot \mathbf{v} = 0$.

Example 20.31. The vector which is orthogonal to all other vectors is the zero vector.

By using the properties of vector operations and definition of dot product, we can obtain a theorem characterising the properties of dot product as follows.

thm:1.2.13

Theorem 20.32 (Properties of the Dot Product). *Let $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ be vectors. Let c be a scalar.*

1. (Commutative Law) $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$
2. (Distributive Law) $\mathbf{u} \cdot (\mathbf{v} + \mathbf{w}) = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w}$ and $(\mathbf{u} + \mathbf{v}) \cdot \mathbf{w} = \mathbf{u} \cdot \mathbf{w} + \mathbf{v} \cdot \mathbf{w}$
3. (Associative Law) $(c\mathbf{u}) \cdot \mathbf{v} = c\mathbf{u} \cdot \mathbf{v} = \mathbf{u} \cdot (c\mathbf{v})$.
4. $\mathbf{v} \cdot \mathbf{v} = \|\mathbf{v}\|^2$
5. $\mathbf{v} \cdot \mathbf{v} > 0$ for any non-zero vector $\mathbf{v}$ and $\mathbf{v} \cdot \mathbf{v} = 0$ iff $\mathbf{v} = \mathbf{0}$.

20.4 Lines, Cross Product and Planes

In this section, we study the geometrical objects (lines and planes) in a 3-d space. We will begin with the simplest geometrical object, the line. Before proceeding to the study of planes, we need to introduce the concept of cross product. With the help of cross product, we can perform calculations with planes.

Let ℓ be a line in 3-d space through the point $P_0(x_0, y_0, z_0)$ and parallel to the nonzero vector $\mathbf{v} = (a, b, c)$. The line ℓ consists of those points $P(x, y, z)$ for which the vector $\overrightarrow{P_0P}$ is parallel to $\mathbf{v}$. So there exists a scalar t such that $\overrightarrow{P_0P} = t\mathbf{v}$, that is,

$$(x - x_0, y - y_0, z - z_0) = (ta, tb, tc),$$

which implies

$$(x, y, z) = (x_0, y_0, z_0) + t(a, b, c) \quad (\text{vector equation})$$

or

$$\begin{aligned}x &= x_0 + ta, \\ y &= y_0 + tb, \\ z &= z_0 + tc.\end{aligned} \quad (\text{parametric equation})$$

eg:1.3.6

Example 20.33. Find the equation of the line through $P_0(-1, 2, 5)$ in the direction of $\mathbf{v} = (2, -3, 4)$. Does the line pass through the point $A(-7, 11, -7)$? Does it pass through $B(0, -1, 2)$?

Solution. The equation is $(x, y, z) = (-1, 2, 5) + t(2, -3, 4)$, where $t \in \mathbb{R}$.

A line passes through a point (a, b, c) iff there is a consistent t_0 such that $(a, b, c) = (-1, 2, 5) + t_0(2, -3, 4)$.

$(-7, 11, -7) = (-1, 2, 5) + t_0(2, -3, 4) \Rightarrow t_0 = -3$ which is consistent. Hence the line passes through A .

$(0, -1, 2) = (-1, 2, 5) + t_0(2, -3, 4) \Rightarrow t_0 = 1/2$ and $t_0 = 1$ which are inconsistent. Hence the line does not pass through B . $\square$

Example 20.34. Find the equation of the line through $P(2, 0, 1)$ and $Q(4, -1, 1)$.

Solution. The direction of the line is the same as $\overrightarrow{PQ} = (2, -1, 0)$. Therefore the equation of the line is

$$(x, y, z) = (2, 0, 1) + t(2, -1, 0).$$

$\square$

Example 20.35. Does the lines with equations

$$\ell_1 : (x, y, z) = (1, 2, 1) + t(-3, 5, 1)$$

and

$$\ell_2 : (x, y, z) = (-1, 3, 1) + t(1, -4, -1)$$

intersect? If so, find the intersection. [**Warning:** Beware that the two t 's in the equation should be treated as different.]

Solution. The direction vectors of the lines are $\mathbf{v}_1 = (-3, 5, 1)$ and $\mathbf{v}_2 = (1, -4, -1)$ which are not parallel since $\mathbf{v}_1 \neq k\mathbf{v}_2$ for any k .

The parametric equations of the lines are

$$\ell_1 : x = 1 - 3t_1, y = 2 + 5t_1, z = 1 + t_1$$

and

$$\ell_2 : x = -1 + t_2, y = 3 - 4t_2, z = 1 - t_2.$$

These lines intersect if there exists t_1 and t_2 such that

$$1 - 3t_1 = -1 + t_2$$

$$2 + 5t_1 = 3 - 4t_2$$

$$1 + t_1 = 1 - t_2.$$

Rearranging, we have

$$3t_1 + t_2 = 2 \tag{20.1}$$

$$5t_1 + 4t_2 = 1 \tag{20.2}$$

$$t_1 + t_2 = 0. \tag{20.3}$$

Solving equations (36.62) and (36.63) we have

$$12t_1 - 5t_1 = 8 - 1 \Rightarrow t_1 = 1, t_2 = -1$$

The *crucial step* in the process is to check equation (37.38).

If we substitute $t_1 = 1$ and $t_2 = -1$ into the left hand side of (37.38) we obtain $1 + (-1)$, which equals 0. Thus equation (37.38) is satisfied and the lines intersect.

Thus the lines meet where $t_1 = 1$ and $t_2 = -1$, i.e. at the point $(-2, 7, 2)$. $\square$

Let us now define the notion of “cross product”. It is unique to 3-d space since it cannot be generalised to n -dimensional space.

Definition 20.12. The *cross product* of $\mathbf{u} = (u_1, u_2, u_3)$ and $\mathbf{v} = (v_1, v_2, v_3)$ is the vector

$$\mathbf{u} \times \mathbf{v} = (u_2v_3 - v_2u_3)\mathbf{i} - (u_1v_3 - v_1u_3)\mathbf{j} + (u_1v_2 - v_1u_2)\mathbf{k} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix}.$$

Here, we use the notation of determinant which will be discussed in Chapter [22](#). [ch:eigen](#)

Example 20.36. Let $\mathbf{u} = (1, 3, 2)$ and $\mathbf{v} = (0, 2, -1)$. Find $\mathbf{u} \times \mathbf{v}$.

Solution. $\mathbf{u} \times \mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & 2 & 3 \\ 0 & 2 & -1 \end{vmatrix} = (-3 - 4)\mathbf{i} - (-1 - 0)\mathbf{j} + (2 - 0)\mathbf{k} = -7\mathbf{i} + \mathbf{j} + 2\mathbf{k}.$ □

The relationship involving cross product and dot product as summarised in the following theorem.

Theorem 20.37. If $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ are 3-dimensional vectors.

- a. $\mathbf{u} \cdot (\mathbf{u} \times \mathbf{v}) = 0$.
- b. $\mathbf{v} \cdot (\mathbf{u} \times \mathbf{v}) = 0$.
- c. $\|\mathbf{u} \times \mathbf{v}\|^2 = \|\mathbf{u}\|^2\|\mathbf{v}\|^2 - (\mathbf{u} \cdot \mathbf{v})^2$ (Lagrange's Identity).

Example 20.38. Find a unit vector orthogonal to both $\mathbf{u} = (1, -4, 1)$ and $\mathbf{v} = (2, 3, 0)$.

Solution. From the [Theorem 20.37](#), we know that $\mathbf{u} \times \mathbf{v}$ is orthogonal to $\mathbf{u}$ and $\mathbf{v}$. [thm:cross:properties](#)

$$\mathbf{u} \times \mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & -4 & 1 \\ 2 & 3 & 0 \end{vmatrix} = -3\mathbf{i} + 2\mathbf{j} + 11\mathbf{k}$$

The orthogonal unit vector is

$$\frac{\mathbf{u} \times \mathbf{v}}{\|\mathbf{u} \times \mathbf{v}\|} = \frac{1}{\sqrt{9 + 4 + 121}}(-3\mathbf{i} + 2\mathbf{j} + 11\mathbf{k}) = \frac{-3}{\sqrt{134}}\mathbf{i} + \frac{2}{\sqrt{134}}\mathbf{j} + \frac{11}{\sqrt{134}}\mathbf{k}$$

□

Theorem 20.39. If $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ are 3-dimensional vectors and c is a scalar.

1. $\mathbf{u} \times \mathbf{v} = -\mathbf{v} \times \mathbf{u}$.
2. $\mathbf{u} \times (\mathbf{v} + \mathbf{w}) = (\mathbf{u} \times \mathbf{v}) + (\mathbf{u} \times \mathbf{w})$.
3. $(c\mathbf{u}) \times \mathbf{v} = c(\mathbf{u} \times \mathbf{v}) = \mathbf{u} \times (c\mathbf{v})$.
4. $\mathbf{u} \times \mathbf{0} = \mathbf{0} = \mathbf{0} \times \mathbf{u}$.
5. $\mathbf{u} \times \mathbf{u} = \mathbf{0}$.
6. $\|\mathbf{u} \times \mathbf{v}\| = \|\mathbf{u}\|\|\mathbf{v}\|\sin\theta$, where $\theta \in [0, \pi)$ is the angle between $\mathbf{u}$ and $\mathbf{v}$.
7. $\|\mathbf{u} \times \mathbf{v}\|$ is the area of the parallelogram having $\mathbf{u}$ and $\mathbf{v}$ as adjacent sides.

Proof: 1~3 can be obtained conveniently using properties of determinants (Chapter 22). Otherwise, we can make use of Definition 20.12, which leads to long and tedious calculations.

1. Exchange the second and third row of the determinant introduces a negative sign.
2. Summation properties of determinant.
3. Take out c from the second or third row of the determinant.
4. Let $\mathbf{v} = \mathbf{w} = \mathbf{0}$ and obtain from 2.
5. Let $\mathbf{v} = \mathbf{u}$ and obtain from 1.
6. From $\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \|\mathbf{v}\| \cos \theta$, we obtain

$$(\mathbf{u} \cdot \mathbf{v})^2 = \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 \cos^2 \theta \Rightarrow (\mathbf{u} \cdot \mathbf{v})^2 = \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 (1 - \sin^2 \theta)$$

From the Lagrange's identity, we have

$$\|\mathbf{u}\|^2 \|\mathbf{v}\|^2 - \|\mathbf{u} \times \mathbf{v}\|^2 = \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 - \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 \sin^2 \theta$$

Therefore

$$\|\mathbf{u} \times \mathbf{v}\|^2 = \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 \sin^2 \theta \Rightarrow \|\mathbf{u} \times \mathbf{v}\| = \|\mathbf{u}\| \|\mathbf{v}\| \sin \theta$$

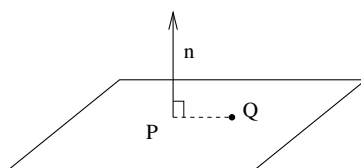
7. Area = $\|\mathbf{u}\| \underbrace{(\|\mathbf{v}\| \sin \theta)}_{\text{height}} = \|\mathbf{u} \times \mathbf{v}\|$ from 6.

Example 20.40. Find the area of the parallelogram with adjacent sides $\mathbf{u} = (-3, 2, 1)$ and $\mathbf{v} = (1, 0, -2)$.

Solution. Area = $\|\mathbf{u} \times \mathbf{v}\| = \left\| \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ -3 & 2 & 1 \\ 1 & 0 & -2 \end{vmatrix} \right\| = \|-4\mathbf{i} - 5\mathbf{j} - 2\mathbf{k}\| = \sqrt{16 + 25 + 4} = \sqrt{45} \quad \square$

Now, we want to find the equation of the plane passing through the point $P_0(x_0, y_0, z_0)$ and having the nonzero vector $\mathbf{n} = (a, b, c)$ as a normal. This plane consists of those points $P(x, y, z)$ for which the vector $\overrightarrow{P_0P}$ is orthogonal to $\mathbf{n}$, that is

$$\overrightarrow{P_0P} \cdot \mathbf{n} = 0.$$



So, the equation of a plane is of the form

$$ax + by + cz + d = 0$$

with the normal vector $\mathbf{n} = (a, b, c)$.

Example 20.41. Find the equation of the plane that

1. is perpendicular to $\mathbf{n} = (1, 2, 3)$ and containing the point $(4, 0, -1)$.
2. goes through $(4, -1, 5)$ and is parallel to the plane $2x - 3y = 6$.
3. contains the points $A(1, 3, -2)$, $B(1, 1, 5)$ and $C(2, -2, 3)$.

Solution.

$$1. (x - 4, y - 0, z - (-1)) \cdot (1, 2, 3) = 0 \Rightarrow x + 2y + 3z = 1.$$

$$2. (x - 4, y - (-1), z - 5) \cdot (2, -3, 0) = 0 \Rightarrow 2x - 3y = 11.$$

$$3. \mathbf{n} = \overrightarrow{AB} \times \overrightarrow{AC} = (0, -2, 7) \times (1, -5, 5) = (25, 7, 2).$$

$$(x - 1, y - 3, z - (-2)) \cdot (25, 7, 2) = 0 \Rightarrow 25x + 7y + 2z = 42.$$

□

20.5 Projections

Definition 20.13. The *projection* of a vector $\mathbf{u}$ on a nonzero vector $\mathbf{v}$ (or a line of non-zero direction $\mathbf{v}$) is the vector $\mathbf{p} = c\mathbf{v}$ with the property that $\mathbf{u} - \mathbf{p}$ is orthogonal to $\mathbf{v}$. It is denoted as $\text{proj}_{\mathbf{v}}\mathbf{u}$.

Using the knowledge of trigonometry, mathematicians have obtained the formula of projection using the concept of dot product.

Theorem 20.42. The projection of a vector $\mathbf{u}$ on a nonzero vector $\mathbf{v}$ is the vector

$$\text{proj}_{\mathbf{v}}\mathbf{u} = \frac{\mathbf{u} \cdot \mathbf{v}}{\mathbf{v} \cdot \mathbf{v}}\mathbf{v} = \mathbf{v}(\mathbf{v} \cdot \mathbf{v})^{-1}\mathbf{u} \cdot \mathbf{v}.$$

Proof. Since $\text{proj}_{\mathbf{v}}\mathbf{u} = \|\mathbf{u}\|(\cos \theta) \frac{\mathbf{v}}{\|\mathbf{v}\|}$ and $\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\|\|\mathbf{v}\| \cos \theta$, we have

$$\text{proj}_{\mathbf{v}}\mathbf{u} = \|\mathbf{u}\| \left(\frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} \right) \frac{\mathbf{v}}{\|\mathbf{v}\|} = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{v}\|^2} \mathbf{v} = \frac{\mathbf{u} \cdot \mathbf{v}}{\mathbf{v} \cdot \mathbf{v}} \mathbf{v}.$$

□

Example 20.43. Suppose $\mathbf{u} = (1, 1)$ and $\mathbf{v} = (3, 1)$. Find $\text{proj}_{\mathbf{v}}\mathbf{u}$.

$$\text{Solution: } \text{proj}_{\mathbf{v}}\mathbf{u} = \|\mathbf{u}\|(\cos \theta) \frac{\mathbf{v}}{\|\mathbf{v}\|} = \sqrt{1+1} \left(\frac{(1, 1) \cdot (3, 1)}{\sqrt{1+1} \cdot \sqrt{9+1}} \right) \frac{(3, 1)}{\sqrt{9+1}} = \frac{4}{10}(3, 1)$$

Example 20.44. Find the projection of $\mathbf{u} = (1, -3, 4)$ on $\mathbf{v} = (-3, 1, 2)$.

$$\text{Solution. } \text{proj}_{\mathbf{v}}\mathbf{u} = \frac{-3-3+8}{9+1+4}(-3, 1, 2) = \frac{2}{14}(-3, 1, 2) = \left(-\frac{3}{7}, \frac{1}{7}, \frac{2}{7}\right)$$

□

The distance from the point B to the line $(x, y, z) = (x_0, y_0, z_0) + t\mathbf{v}$ is

$$\|\mathbf{w} - \text{proj}_{\mathbf{v}}\mathbf{w}\|$$

where $\mathbf{w} = \overrightarrow{AB}$ and A is a point on the line.

Example 20.45. Find the distance from the point $P(1, 8, -5)$ to the line $(x, y, z) = (-1, 1, 6) + t(3, 0, -4)$.

Solution. Let $\mathbf{v} = (3, 0, -4)$ be the direction of the line. Take a point $P_0 = (-1, 1, 6)$ on the line. Then

$$\mathbf{w} = \overrightarrow{P_0P} = (1, 8, -5) - (-1, 1, 6) = (2, 7, -11).$$

Now, we calculate

$$\text{proj}_{\mathbf{v}}\mathbf{w} = \frac{6 + 44}{3^2 + 4^2}(3, 0, -4) = (6, 0, -8).$$

Finally, the distance is

$$\|\mathbf{w} - \text{proj}_{\mathbf{v}} \mathbf{w}\| = \|(2, 7, -11) - (6, 0, -8)\| = \sqrt{(-4)^2 + 7^2 + (-3)^2} = \sqrt{74}.$$

□

The distance from the point B to the plane with equation $ax + by + cz + d = 0$ is

$$\|\text{proj}_{\mathbf{n}} \mathbf{w}\|$$

where $\mathbf{n}$ is the normal vector of the plane, $\mathbf{w} = \overrightarrow{AB}$ and A is a point on the plane.

Example 20.46. Find the distance from the point $P(-1, 2, 1)$ to the plane with equation $5x + y + 3z = 7$.

Solution. The normal vector $\mathbf{n} = (5, 1, 3)$. Picking a point $(1, -1, 1)$ from the plane and the distance can be calculated:

$$\begin{aligned} \|\text{proj}_{\mathbf{n}} \mathbf{w}\| &= \|\text{proj}_{(5,1,3)}((-1, 2, 1) - (1, -1, 1))\| \\ &= \left\| \frac{(5, 1, 3) \cdot (-2, 3, 0)}{(5, 1, 3) \cdot (5, 1, 3)} (5, 1, 3) \right\| \\ &= \frac{|-10 + 3|}{\sqrt{5^2 + 1^2 + 3^2}} = \frac{7}{\sqrt{35}} \end{aligned}$$

□

Projections onto Planes through the Origin.

Definition 20.14. The set of all linear combination of vectors $\mathbf{u}$ and $\mathbf{v}$ is called the *plane spanned by $\mathbf{u}$ and $\mathbf{v}$* . We also say that $\mathbf{u}$ and $\mathbf{v}$ *span the plane*.

Example 20.47. The vectors $\mathbf{i} = (1, 0, 0)$ and $\mathbf{j} = (0, 1, 0)$ span the xy -plane in 3-space since the set of linear combinations of $\mathbf{i}$ and $\mathbf{j}$ is the set of vectors of the form

$$a(1, 0, 0) + b(0, 1, 0) = (a, b, 0).$$

rem:1.4.6

Remark 20.48. The plane spanned by vectors $\mathbf{u}$ and $\mathbf{v}$ always passes through the origin because when the scalars are chosen to be zero, we will obtain a zero vector.

eg:1.4.7

Example 20.49. Find an equation of the plane spanned by vectors $\mathbf{u} = (1, 3, 2)$ and $\mathbf{v} = (0, -2, 1)$.

Solution. The normal of the plane is $\mathbf{n} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & 3 & 2 \\ 0 & -2 & 1 \end{vmatrix} = 7\mathbf{i} - \mathbf{j} - 2\mathbf{k}.$

The equation of the plane is

$$7x - y - 2z = 0.$$

□

Definition 20.15. The *projection* of a vector $\mathbf{w}$ on a plane π is a vector $\mathbf{p}$ in with the property $\mathbf{w} - \mathbf{p}$ is orthogonal to every vector in π . We denote it as $\text{proj}_{\pi} \mathbf{w}$.

The projection of $\mathbf{w} = (a, b, c)$ on the xy -plane is $\mathbf{p} = (a, b, 0)$ since $\mathbf{p}$ is in the xy -plane and $\mathbf{w} - \mathbf{p} = (0, 0, c)$ is orthogonal to every vector on the plane.

thm:1.4.8

Theorem 20.50. Suppose a plane π is spanned by vectors $\mathbf{u}$ and $\mathbf{v}$. If a vector is orthogonal just to $\mathbf{u}$ and $\mathbf{v}$, then it is orthogonal to π , that is orthogonal to every vector in π .

Proof. If π is spanned by $\mathbf{u}$ and $\mathbf{v}$, then every vector in π can be represented as $a\mathbf{u} + b\mathbf{v}$ for some $a, b \in \mathbb{R}$.

If a vector $\mathbf{w}$ is orthogonal to $\mathbf{u}$ and $\mathbf{v}$, then

$$\mathbf{w} \cdot \mathbf{u} = 0, \quad \mathbf{w} \cdot \mathbf{v} = 0.$$

Since every vector in π has the form $a\mathbf{u} + b\mathbf{v}$ and

$$\mathbf{w} \cdot (a\mathbf{u} + b\mathbf{v}) = a\mathbf{w} \cdot \mathbf{u} + b\mathbf{w} \cdot \mathbf{v} = a \times 0 + b \times 0 = 0.$$

This shows that it is orthogonal to every vector in π . □

eg:vec+plane1

Example 20.51. Find the projection of $\mathbf{w} = (5, -3, -7)$ on the plane π with equation $3x - 2y + z = 0$.

Solution. $\text{proj}_{\pi} \mathbf{w} = \mathbf{w} - \text{proj}_{\mathbf{n}} \mathbf{w}$ □

Find angle between lines: Find directions $\mathbf{v}_1, \mathbf{v}_2$ of both lines and use dot product to determine the angle:

$$\theta = \cos^{-1} \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\| \|\mathbf{v}_2\|}.$$

Find angle between a line and a plane: Find direction $\mathbf{v}$ of the line and the normal vector $\mathbf{n}$ of the plane and use dot product to determine the angle:

$$\theta = \sin^{-1} \frac{\mathbf{v} \cdot \mathbf{n}}{\|\mathbf{v}\| \|\mathbf{n}\|}.$$

Find angle between a line and a plane: Find the normal directions $\mathbf{n}_1$ and $\mathbf{n}_2$ of both planes and use dot product to determine the angle:

$$\theta = \cos^{-1} \frac{\mathbf{n}_1 \cdot \mathbf{n}_2}{\|\mathbf{n}_1\| \|\mathbf{n}_2\|}.$$

20.6 Euclidean n -Space

a:euclidean-sp

The 2-d and 3-d vectors can be generalised to n -dimensional using the Cartesian product (Definition 4.2). def:set-op2

def:vector1

Definition 20.16. Let $n \geq 1$ be an integer. An n -dimensional vector is the element of $\mathbb{R} \times \cdots \times \mathbb{R} =: \mathbb{R}^n$ with the form $(u_1, u_2, \dots, u_n)$ where $u_i \in \mathbb{R}$ for $i \in \{1, 2, \dots, n\}$.

Using set theory, it is natural to see that two vectors $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_m)$ are equal iff $n = m$ and corresponding components are equal, i.e. $x_1 = y_1, x_2 = y_2, \dots, x_n = y_n$.

Remark 20.52. This is supported in the Scheme programming language as illustrated below.

```
(define v1 (list->vector '(1 2 3 4 5 6))) ; m=6
(define v2 (list->vector (cdr (iota 6)))) ; m=5, (1 2 3 4 5)
(define v3 (list->vector (map 1+ (iota 6)))) ; m=6, (1 2 3 4 5 6)
(equal? v1 v2) ; false
(equal? v1 v3) ; true
(equal? v1 (vector-append v2 (make-vector 1 6))) ; true
```

Definition 20.17. Let $n \geq 1$ be an integer. A (*Euclidean*) n -space is the set $\mathbb{R}^n$ together with the vector addition

$$\mathbf{x} + \mathbf{y} = (x_1 + y_1, \dots, x_n + y_n),$$

scalar multiplication

$$c\mathbf{x} = (cx_1, \dots, cx_n)$$

and **inner product** or **dot product**

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x} \cdot \mathbf{y} = x_1y_1 + x_2y_2 + \dots + x_ny_n.$$

Here, $\mathbf{x} = (x_1, x_2, \dots, x_n), \mathbf{y} = (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$ and $c \in \mathbb{R}$ are arbitrary.

The n -dimensional *zero vector* is $\mathbf{0} = (0, 0, \dots, 0)$.

Remark 20.53. All the vector addition and scalar multiplication properties in Theorem 20.7 and all the dot product properties in Theorem 20.32 hold in n -space. thm:1.1.6
thm:1.2.13

Definition 20.18. A *linear combination* of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n \in \mathbb{R}^n$ is a vector of the form

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \dots + c_n\mathbf{v}_n$$

where $c_1, c_2, \dots, c_n$ are scalars.

Example 20.54. Show that $(-11, 7, -4, 2)$ is a linear combination of $(-2, 0, 5, 1)$ and $(1, -1, 2, 0)$.

Solution.

$$(-11, 7, -4, 2) = 2(-2, 0, 5, 1) + (-7)(1, -1, 2, 0).$$

□

Example 20.55. Is $(1, 5, 2, 0, -1)$ a linear combination of $(1, 2, -1, 3, 4)$ and $(3, 5, 0, 1, 2)$?

Solution. Suppose we have a linear combination:

$$(1, 5, 2, 0, -1) = a(1, 2, -1, 3, 4) + b(3, 5, 0, 1, 2) = (a + 3b, 2a + 5b, -a, 3a + b, -a + 2b).$$

From the above equality, we know that $-a = 2$ and from $a + 3b = 1$, $b = 1$. However, $2a + 5b = 2(-2) + 5(1) = 1 = 5$ is a contradiction.

Therefore, $(1, 5, 2, 0, -1)$ is NOT a linear combination of $(1, 2, -1, 3, 4)$ and $(3, 5, 0, 1, 2)$.

□

Definition 20.19. The *standard basis vectors* in $\mathbb{R}^n$ are the n -dimensional vectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ where $\mathbf{e}_i$ has i th component 1 and all other components 0. Thus

$$\mathbf{e}_1 = (1, 0, 0, \dots, 0),$$

$$\mathbf{e}_2 = (0, 1, 0, \dots, 0),$$

...

$$\mathbf{e}_n = (0, 0, \dots, 0, 1).$$

Theorem 20.56. Every vector in $\mathbb{R}^n$ is a linear combination of the standard basis vectors.

Proof: This is because $(x_1, x_2, \dots, x_n) = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \dots + x_n\mathbf{e}_n$.

def:vec-norm

Definition 20.20. The *length* or *norm* of the n -dimensional vector $\mathbf{x} = (x_1, x_2, \dots, x_n)$ is the number

$$\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}} = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}.$$

The following is the generalisation of Theorem [thm:1.2.4](#) 20.22.

Theorem 20.57. For $\mathbf{u} \in \mathbb{R}^n$, $\|c\mathbf{u}\| = |c|\|\mathbf{u}\|$.

Proof.
$$\begin{aligned} \|c\mathbf{u}\| &= \|(cu_1, \dots, cu_n)\| = \sqrt{(cu_1)^2 + \dots + (cu_n)^2} \\ &= \sqrt{c^2(u_1^2 + \dots + u_n^2)} = \sqrt{c^2} \sqrt{u_1^2 + \dots + u_n^2} = |c|\|\mathbf{u}\| \end{aligned}$$

□

Theorem 20.58. $\|\mathbf{u}\|^2 = \mathbf{u} \cdot \mathbf{u}$, or equivalently, $\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}}$.

Proof.

$$\mathbf{u} \cdot \mathbf{u} = u_1^2 + \dots + u_n^2 = (\sqrt{u_1^2 + \dots + u_n^2})^2 = (\|\mathbf{u}\|)^2 = \|\mathbf{u}\|^2$$

□

If we generalised the Theorem [thm:cos-theta](#) 20.29 to $\mathbb{R}^n$, then we can define the concept of “angle” in n -dimensional space as follows.

Definition 20.21. If $\mathbf{u}$ and $\mathbf{v}$ are nonzero vectors in $\mathbb{R}^n$, the *angle* θ between $\mathbf{u}$ and $\mathbf{v}$ is given by

$$\cos \theta = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}, \quad 0 \leq \theta \leq \pi.$$

Example 20.59. If $\mathbf{u} = (-4, 0, 2, -2)$ and $\mathbf{v} = (2, 0, -1, 1)$, find the angle θ .

Solution.
$$\cos \theta = \frac{(-4, 0, 2, -2) \cdot (2, 0, -1, 1)}{\sqrt{16 + 0 + 4 + 4}\sqrt{4 + 0 + 1 + 1}} = \frac{-8 - 2 - 2}{\sqrt{24}\sqrt{6}} = \frac{-12}{12} = -1$$

Therefore $\theta = 180^\circ$.

□

Definition 20.22. Vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$ are said to be *orthogonal* if $\mathbf{u} \cdot \mathbf{v} = 0$.

Example 20.60. Are the vectors $\mathbf{u} = (3, 2, -1, 4)$ and $\mathbf{v} = (1, -1, 1, 0)$ orthogonal in $\mathbb{R}^4$?

Solution. Yes since $\mathbf{u} \cdot \mathbf{v} = 3 - 2 - 1 + 0 = 0$.

□

thm:1.4.9

Theorem 20.61. The projection of a vector $\mathbf{w}$ on the plane (in $\mathbb{R}^n$) spanned by orthogonal vectors $\mathbf{e}$ and $\mathbf{f}$ is

$$\text{proj}_\pi \mathbf{w} = \frac{\mathbf{w} \cdot \mathbf{e}}{\mathbf{e} \cdot \mathbf{e}} \mathbf{e} + \frac{\mathbf{w} \cdot \mathbf{f}}{\mathbf{f} \cdot \mathbf{f}} \mathbf{f} \quad (20.4)$$

e:1.4.9

Proof. Any vector on the plane π spanned by $\mathbf{e}$ and $\mathbf{f}$ has the form $a\mathbf{e} + b\mathbf{f}$, where $a, b \in \mathbb{R}$. Therefore

$$\text{proj}_\pi \mathbf{w} = a\mathbf{e} + b\mathbf{f} \quad (20.5)$$

e:1.4.9a

Let the normal vector to the plane π be $\mathbf{n} = \mathbf{e} \times \mathbf{f}$. Since π and $\mathbf{n}$ span $\mathbb{R}^3$, we have

$$\text{proj}_\pi \mathbf{w} = \mathbf{w} - \text{proj}_{\mathbf{n}} \mathbf{w} = \mathbf{w} - \frac{\mathbf{n} \cdot \mathbf{w}}{\mathbf{n} \cdot \mathbf{n}} \mathbf{n} = \mathbf{w} - \frac{\mathbf{e} \times \mathbf{f} \cdot \mathbf{w}}{(\mathbf{e} \times \mathbf{f}) \cdot (\mathbf{e} \times \mathbf{f})} \mathbf{e} \times \mathbf{f} \quad (20.6)$$

e:1.4.9b

From equations [\(20.5\)](#) and [\(20.6\)](#), we have

$$a\mathbf{e} + b\mathbf{f} = \mathbf{w} - \frac{\mathbf{e} \times \mathbf{f} \cdot \mathbf{w}}{(\mathbf{e} \times \mathbf{f}) \cdot (\mathbf{e} \times \mathbf{f})} \mathbf{e} \times \mathbf{f}. \quad (20.7)$$

[e:1.4.9c](#)

Taking dot product with $\mathbf{e}$ to the equation [\(20.7\)](#) and note that $\mathbf{e} \cdot \mathbf{f} = 0$, we have

$$a\mathbf{e} \cdot \mathbf{e} = \mathbf{w} \cdot \mathbf{e} - \frac{\mathbf{e} \times \mathbf{f} \cdot \mathbf{w}}{(\mathbf{e} \times \mathbf{f}) \cdot (\mathbf{e} \times \mathbf{f})} \mathbf{e} \times \mathbf{f} \cdot \mathbf{e} = \mathbf{w} \cdot \mathbf{e} \Rightarrow a = \frac{\mathbf{w} \cdot \mathbf{e}}{\mathbf{e} \cdot \mathbf{e}}.$$

Taking dot product with $\mathbf{f}$ to the equation [\(20.7\)](#), we have

$$b\mathbf{f} \cdot \mathbf{f} = \mathbf{w} \cdot \mathbf{f} \Rightarrow b = \frac{\mathbf{w} \cdot \mathbf{f}}{\mathbf{f} \cdot \mathbf{f}}.$$

Therefore equation [\(20.5\)](#) becomes equation [\(20.4\)](#). □

Example 20.62. Find the projection of a four-dimensional vector $\mathbf{w} = (5, -3, -7, 8)$ on the plane π spanned by $(1, 0, -3, 0)$ and $(3, 5, 1, 1)$.

Solution. Let $\mathbf{e} = (1, 0, -3, 0)$ and $\mathbf{f} = (3, 5, 1, 1)$. Since $\mathbf{e} \cdot \mathbf{f} = 3 - 3 = 0$, we can use the formula [\(20.4\)](#) to calculate

$$\begin{aligned} \text{proj}_{\pi} \mathbf{w} &= \frac{\mathbf{w} \cdot \mathbf{e}}{\mathbf{e} \cdot \mathbf{e}} \mathbf{e} + \frac{\mathbf{w} \cdot \mathbf{f}}{\mathbf{f} \cdot \mathbf{f}} \mathbf{f} \\ &= \frac{5 + 0 + 21}{1 + 0 + 9} (1, 0, -3) + \frac{3 - 3 - 7/5}{9/25 + 1 + 1/25} \left(\frac{3}{5}, 1, \frac{1}{5}\right) \\ &= (2, -1, -8, !!!) \end{aligned}$$

□

Chapter 21

Matrices and Linear Equations

matrix-lin-sys

Scheme does not support array or matrices. However, Common Lisp do support array.

In armadillo C++, matrices and the numerical methods for matrix decompositions, inverses and equation solvers are encoded as follows.

```
1 arma::cx_mat  = arma::Mat<cx_double>   A(m, n);  // 64-bit
2 arma::cx_fmat = arma::Mat<cx_float>    A(m, n);  // 32-bit
3
4 arma::lu       lower-upper decomposition
5 arma::chol     Cholesky decomposition
6 arma::solve    solve systems of linear equations
7 arma::inv      inverse of general square matrix
8 arma::inv_sympd inverse of symmetric positive definite matrix
9 arma::null     orthonormal basis of null space
10 arma::orth     orthonormal basis of range space
11 arma::pinv     pseudo-inverse / generalised inverse
```

Need to add complex matrices. References:

PaulRHalmos-FiniteDimensionalVectorSpaces-Springer1974;

LinearAlgebra2ndEd-KennethHoffman+RayKunze-1971 (inner product space over complex field)

May be we need to add Unitary matrices, a special kind of complex matrices, which are used in quantum mechanics and quantum computing?

This topic investigates a mathematical object called matrix. In the first two sections, we will discover that algebraic operations for vectors can be extended to matrix as well. We will also discover that matrix has some operations which are not applicable to vectors.

Next, we will investigate the “linear equation” for matrix, which has the abstract form $A\mathbf{x} = \mathbf{b}$. Mathematicians found that the LU factorisation and its variants are important in the theoretical study of the solution of the equation $A\mathbf{x} = \mathbf{b}$. Finally, we study how to apply the algebraic reduction technique to calculate the inverse of a matrix, etc.

21.1 Algebra of Matrices

Similar the case of vectors, we introduce “algebraic operations” to matrices. We will find that apart from addition and scalar multiplication, matrices can have other operations such as “product”, “transpose”, etc. which will be discussed in this section.

Definition 21.1. A *matrix* is a rectangular array of numbers, denoted by

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \quad \text{or} \quad A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

or (sometimes in Russian texts or older books, e.g. J. H. M. Wedderburn, “Lectures on Matrices”, AMS 1934)

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

The numbers in the array are called the *entries* in the matrix. The *entry* in row i and column j of the matrix A is denoted a_{ij} and we write $A = (a_{ij})$.

Definition 21.2. The i th row of A is $[a_{i1} \ a_{i2} \ \cdots \ a_{in}]$ for $i \in \{1, 2, \dots, m\}$.

Definition 21.3. The j th column of A is $\begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{bmatrix}$ for $j \in \{1, 2, \dots, n\}$.

Definition 21.4. The *size* of a matrix is described in terms of the number of rows and columns it contains. If a matrix A has m rows and n columns, then A is an m by n matrix (written $m \times n$).

The collection of all matrices of size $m \times n$ is denoted by $M_{m,n}(\mathbb{R})$ (or $\mathbb{M}_{m,n}$). When $m = n$, we denote it $M_n(\mathbb{R})$. The elements of $M_n(\mathbb{R})$ are matrices with n rows and n columns and are called a *square matrices of order n* . If $A \in M_n(\mathbb{R})$, the entries a_{ii} for $i \in \{1, 2, \dots, n\}$ are said to be on the *main diagonal* of the square matrix A .

Example 21.1. The matrix $A = \begin{bmatrix} 1 & 13 & 5 & 7 \\ -4 & 8 & 2 & 6 \\ 9 & 0 & -7 & 25 \end{bmatrix}$ is a 3×4 matrix with 3 rows and 4 columns with size 3×4 .

Let us introduce a matrix operation which change an $m \times n$ matrix to an $n \times m$ matrix.

Definition 21.5. The *transpose* of an $m \times n$ matrix A is the $n \times m$ matrix whose rows are the columns of A in the same order. The transpose of A is denoted A^T (or A'). If $A = (a_{ij})$, $B = (b_{ij})$ and $B = A^T$, then $b_{ij} = a_{ji}$ for $i = 1, \dots, m$ and $j = 1, \dots, n$.

eg:transpl

Example 21.2. Write down the transpose of $A = \begin{bmatrix} 6 & 3 & 5 \\ 4 & 7 & 2 \end{bmatrix}$ $A^T = \begin{bmatrix} 6 & 4 \\ 3 & 7 \\ 5 & 2 \end{bmatrix}$.

We will now define algebraic matrix operations involving two matrices.

Definition 21.6. Two matrices are said to be *equal* if they have the same size and their corresponding entries are equal.

Definition 21.7 (Addition of Matrices). Let A and B be matrices of the same size. The sum $A + B$ (difference $A - B$) is the matrix obtained by adding (subtracting) the entries of B to the corresponding entries of A .

Definition 21.8 (Scalar Multiplication of Matrices). If A is a matrix and c is any scalar, then the product cA is the matrix obtained by multiplying each entry of the matrix A by c . The matrix cA is called a *scalar multiple* of A .

eg:mat-op1

Example 21.3. Given $A = \begin{bmatrix} 6 & 3 & 5 \\ 4 & 7 & 2 \end{bmatrix}$, $B = \begin{bmatrix} 7 & 9 & 0 \\ 3 & 1 & 4 \end{bmatrix}$ and $C = \begin{bmatrix} 0 & 5 \\ 6 & 9 \end{bmatrix}$. Determine if $A + B$, $A - B$, $A + C$, $A - C$, $7A$ are meaningful and perform calculations if possible.

Solution: Since A and B are of the same size $A + B$ and $A - C$ are meaningful and since A and C has different size, $A + C$ and $A - C$ are not meaningful.

$$A + B = \begin{bmatrix} 13 & 12 & 5 \\ 7 & 8 & 6 \end{bmatrix}, \quad A - B = \begin{bmatrix} -1 & -6 & 5 \\ 1 & 6 & -2 \end{bmatrix}$$

$7A$ is defined to have the following meaning:

$$7A = \begin{bmatrix} 42 & 21 & 35 \\ 28 & 49 & 14 \end{bmatrix}.$$

Again, similar to the case of vectors, matrix addition and scalar multiplication are characterised by the following properties.

Theorem 21.4 (Properties of Matrix addition and Scalar Multiplication). Let A , B and C be matrices of the same size and let c and d be scalars.

1. (Closed under addition) $A + B$ is a matrix.
2. (Additive Commutative Law) $A + B = B + A$.
3. (Additive Associative Law) $(A + B) + C = A + (B + C)$.
4. (Zero) There is a zero matrix 0 with the property that $A + 0 = 0 + A = A$.
5. (Negatives) For every matrix A , there is a matrix $-A$ called the negative of A with the property that $A + (-A) = 0$ is the zero matrix.
6. (Closed under scalar multiplication) cA is a matrix.
7. (Scalar Associativity) $c(dA) = (cd)A$.
8. (Distributive Law) $c(A + B) = cA + cB$ and $(c + d)A = cA + dA$.
9. If $cA = 0$, then $c = 0$ or $A = 0$ (NB. These two zeros mean different things, one is a scalar, the other a matrix).

In the above theorem, we have a special matrix called the *zero matrix*, which is similar to zero vector.

Example 21.5. Let $A = \begin{bmatrix} 1 & 2 \\ 3 & 5 \\ -6 & -7 \end{bmatrix}$ and $B = \begin{bmatrix} -2 & 4 \\ 6 & 2 \\ -8 & 2 \end{bmatrix}$. Solve the equation $2A + 4X = 3B$.

Solution.

$$2A + 4X = 3B \Rightarrow 4X = 3B - 2A \Rightarrow X = \frac{1}{4} \left(\begin{bmatrix} -6 & 12 \\ 18 & 6 \\ -24 & 6 \end{bmatrix} - \begin{bmatrix} 2 & 4 \\ 6 & 10 \\ -12 & -14 \end{bmatrix} \right) = \begin{bmatrix} -2 & 2 \\ 3 & -1 \\ -3 & 5 \end{bmatrix}.$$

□

Matrix algebra has richer operations compare to vectors. Apart from the “transpose” operation mentioned previously, we can define the “multiplication” operation.

Definition 21.9 (Multiplication of Matrices). If $A = (a_{ij})$ is an $m \times k$ matrix and $B = (b_{ij})$ is an $k \times n$ matrix, then the *product* $AB = (c_{ij})$ is the $m \times n$ matrix whose entry in row i and column j is the i th row of A times the j th column of B and is determined by

$$AB = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & & & \\ a_{i1} & a_{i2} & \cdots & a_{ik} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mk} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1j} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2j} & \cdots & b_{2n} \\ \vdots & & & \vdots & & \\ b_{k1} & b_{k2} & \cdots & b_{kj} & \cdots & b_{kn} \end{bmatrix} = \begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & c_{ij} & \vdots \\ c_{m1} & \cdots & c_{mn} \end{bmatrix}.$$

where $c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{ik}b_{kj} = \sum_{k=1}^r a_{ik}b_{kj}$.

The following example will demonstrate the product of matrices concretely.

Example 21.6. Given $A = \begin{bmatrix} 6 & 3 & 5 \\ 4 & 7 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 7 & 3 \\ 9 & 1 \\ 0 & 4 \end{bmatrix}$. Find the product $C = AB$.

Solution: We form the element in the first row and first column of AB by performing a *dot product* between the first row of A and the first column of B ($(6, 3, 5) \cdot (7, 9, 0) = 42 + 27 + 0 = 69$):

$$AB = \begin{bmatrix} 6 & 3 & 5 \\ 4 & 7 & 2 \end{bmatrix} \begin{bmatrix} 7 & 3 \\ 9 & 1 \\ 0 & 4 \end{bmatrix} = \begin{bmatrix} \boxed{69} & \\ & \end{bmatrix},$$

To get the element in the first row and second column of AB we perform a dot product between the first row of A and the second column of B :

$$AB = \begin{bmatrix} 6 & 3 & 5 \\ 4 & 7 & 2 \end{bmatrix} \begin{bmatrix} 7 & 3 \\ 9 & 1 \\ 0 & 4 \end{bmatrix} = \begin{bmatrix} 69 & \boxed{41} \\ & \end{bmatrix},$$

Similarly, we get c_{21} :

$$AB = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 7 & 9 \\ 3 & 1 \\ 0 & 4 \end{bmatrix} = \begin{bmatrix} 69 & 41 \\ \boxed{91} & \end{bmatrix}, \text{ and finally } c_{22}:$$

$$AB = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 7 & 9 \\ 3 & 1 \\ 0 & 4 \end{bmatrix} = \begin{bmatrix} 69 & 41 \\ 91 & \boxed{27} \end{bmatrix} = C.$$

Remark 21.7. Matrix multiplication is closely related to the composition of linear transformation that we are going to explore in Chapter 24. It would be useful to give an interpretation of special square matrices.

- In Discrete Mathematics, we learn that a digraph with n vertices can be represented by a square matrix $A \in M_n(\mathbb{Z})$. If A is the adjacency matrix of a digraph, then the entry in row i and column j of the product A^2 gives the number of directed walks of length 2 from vertex i to vertex j . In general, the entry in row i and column j of the product A^n gives the number of directed walks of length n from vertex i to vertex j .
- In Markov Chain (mentioned in Discrete Mathematics), the *(right) stochastic matrix* (or *right transition matrix*) is a square matrix of nonnegative real numbers with each row summing to 1. A *doubly stochastic matrix* is a square matrix of nonnegative real numbers with each row and column summing to 1.

A common convention in English language mathematics literature is to use *row vectors of probability* (oppose to column vectors which is commonly used in linear algebra) and right stochastic matrix!

A stochastic matrix describes a Markov chain X_t over a *finite state space* S . If the probability of moving from i to j in one time step is $Pr[j|i] = p_{i,j}$, then the stochastic matrix

$$P = \begin{bmatrix} p_{1,1} & \cdots & p_{1,n} \\ \vdots & \ddots & \vdots \\ p_{n,1} & \cdots & p_{n,n} \end{bmatrix}$$

where $\sum_{j=1}^n p_{i,j} = 1$.

Now the entry of the multiplication of P with itself, P^2 , is the probability of transitioning from i to j in two steps! In general the probability of transitioning from any state to another state in finite Markov chain is given by P^k .

It is natural to ask: Is there a unique matrix, when multiply with A gives A ? The kind of matrix is defined below.

Definition 21.10 (Identity Matrix). A square matrix with 1's on the main diagonal and 0's off the main diagonal is called an *identity matrix* and is denoted by I . We use the notation I_n for $n \times n$ identity matrix.

Mathematicians have studied the properties of matrix multiplication together with matrix addition and scalar multiplication mentioned above. They are summarised below.

Theorem 21.8 (Properties of Matrix Multiplication). Let $A, \tilde{A} \in M_{m,k}$, $B, \tilde{B} \in M_{k,n}$ and $C \in M_{n,p}$.

1. Matrix multiplication is associative i.e. $(AB)C = A(BC)$.
2. Matrix multiplication is not commutative in general, i.e. $AB \neq BA$.
3. Matrix multiplication distributes over addition i.e. $(A + \tilde{A})B = AB + \tilde{A}B$ and $A(B + \tilde{B}) = AB + A\tilde{B}$.
4. $A(cB) = (cA)B = c(AB)$ for any scalar c .
5. There exists identity matrices I_k and I_m such that $AI_k = A$ and $I_mA = A$.

Example 21.9. Given $A = \begin{bmatrix} 1 & 3 & 4 \\ 6 & 2 & 5 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 1 \\ 3 & 8 \\ 9 & 2 \end{bmatrix}$. Calculate AB and BA , are they the same?

Solution:

$$AB = \begin{bmatrix} 1 & 3 & 4 \\ 6 & 2 & 5 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 3 & 8 \\ 9 & 2 \end{bmatrix} = \begin{bmatrix} 45 & 33 \\ 51 & 32 \end{bmatrix}.$$

$$BA = \begin{bmatrix} 0 & 1 \\ 3 & 8 \\ 9 & 2 \end{bmatrix} \begin{bmatrix} 1 & 3 & 4 \\ 6 & 2 & 5 \end{bmatrix} = \begin{bmatrix} 6 & 2 & 5 \\ 51 & 25 & 52 \\ 21 & 31 & 46 \end{bmatrix}.$$

Therefore, $AB \neq BA$ (and this is true for matrix multiplication in general).

What happens when we investigate the changes of matrix operations under matrix transpose? Mathematicians have obtained the following results.

Theorem 21.10. *Let A and B be matrices and let c be a scalar.*

1. *If A and B have the same size, then $(A + B)^T = A^T + B^T$;*
2. *$(cA)^T = cA^T$;*
3. *If AB is defined, $(AB)^T = B^T A^T$.*

Example 21.11. Suppose $A = \begin{bmatrix} 1 & 2 & 3 \\ -4 & 0 & -1 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 1 \\ 1 & -3 \\ 4 & 2 \end{bmatrix}$. Verify that $(AB)^T = B^T A^T$.

Solution.

$$(AB)^T = \begin{bmatrix} 14 & -4 \\ 1 & -6 \end{bmatrix}$$

$$B^T A^T = \begin{bmatrix} 0 & 1 & 4 \\ 1 & -3 & 2 \end{bmatrix} \begin{bmatrix} 1 & -4 \\ 2 & 0 \\ 3 & -1 \end{bmatrix} = \begin{bmatrix} 14 & -4 \\ 1 & -6 \end{bmatrix}$$

□

21.2 The Inverse of a Matrix

For a non-zero real number x , we can always find an multiplication inverse of x , which is $\frac{1}{x}$. This section tries to extend the concept of inverse for matrix.

Definition 21.11. A matrix A is *invertible* or *has an inverse* if there is a matrix B such that $AB = I$ and $BA = I$. The matrix B is called the *inverse of A* and we write $B = A^{-1}$.

Example 21.12. Given $A = \begin{bmatrix} 1 & -2 \\ 2 & -3 \end{bmatrix}$ and $B = \begin{bmatrix} -3 & 2 \\ -2 & 1 \end{bmatrix}$. Show that $AB = BA$ and show that A is invertible.

Solution:

$$AB = \begin{bmatrix} 1 & -2 \\ 2 & -3 \end{bmatrix} \begin{bmatrix} -3 & 2 \\ -2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad BA = \begin{bmatrix} -3 & 2 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ 2 & -3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Therefore $AB = BA$. By definition, when $AB = I$, B is the inverse of A . Hence, A is invertible.

Theorem 21.13. If A is invertible, so is A^{-1} , and $(A^{-1})^{-1} = A$.

Proof: Let $B = A^{-1}$. If A invertible, then $AB = BA = I$ by definition. Again by definition, A is the inverse of B . Therefore, $B^{-1} = A \Rightarrow (A^{-1})^{-1} = A$.

Theorem 21.14. If A is invertible, then its inverse is unique.

Proof: Assume A has two inverses B and C . So $AB = I = AC$. This implies $AB - AC = 0 \Rightarrow A(B - C) = 0$. Since A is invertible, $B - C = 0$ which implies $B = C$. The inverse is unique.

Example 21.15. Is $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$ invertible?

Solution. No. If A is invertible, there is a matrix $B = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ such that

$$AB = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a + 2c & b + 2d \\ 2a + 4c & 2b + 4d \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

From column 1, we have $a + 2c = 1$ and $2a + 4c = 2(a + 2c) = 0$. Therefore, $2 = 0$, a contradiction. $\square$

Theorem 21.16. The matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is invertible if $ad - bc \neq 0$, in which case the inverse is given by

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Proof: It is easy to verify that $AA^{-1} = I_2$ and $A^{-1}A = I_2$.

eg: inv1

Example 21.17. Let $A = \begin{bmatrix} 1 & 3 \\ -2 & 5 \end{bmatrix}$ and $B = \begin{bmatrix} 2 & -1 \\ 1 & -3 \end{bmatrix}$. Find A^{-1} , B^{-1} , $(AB)^{-1}$ and $B^{-1}A^{-1}$.

Solution. $A^{-1} = \frac{1}{5 - (-6)} \begin{bmatrix} 5 & -3 \\ 2 & 1 \end{bmatrix}$

$$B^{-1} = \frac{1}{-6 - (-1)} \begin{bmatrix} -3 & 1 \\ -1 & 2 \end{bmatrix}$$

$$(AB)^{-1} = \begin{bmatrix} 5 & -10 \\ 1 & -13 \end{bmatrix}^{-1} = \frac{1}{-65 - (-10)} \begin{bmatrix} -13 & 10 \\ -1 & 5 \end{bmatrix}$$

$$B^{-1}A^{-1} = \frac{1}{-55} \begin{bmatrix} -13 & 10 \\ -1 & 5 \end{bmatrix} \quad \square$$

Theorem 21.18. If matrices A and B are invertible and AB is defined, then $(AB)^{-1} = B^{-1}A^{-1}$.

Proof: Since A and B are invertible, A^{-1} and B^{-1} exist. Therefore

$$BB^{-1} = I \Rightarrow ABB^{-1} = AI = A \Rightarrow ABB^{-1}A^{-1} = AA^{-1} = I \Rightarrow AB(B^{-1}A^{-1}) = I.$$

By definition, $(AB)^{-1} = B^{-1}A^{-1}$.

Definition 21.12. If $A \in M_n(\mathbb{R})$, then we define the (*nonnegative integer*) powers of A to be

$$A^k = \begin{cases} I_n & k = 0, \\ A(A^{k-1}) & k > 0. \end{cases}$$

Theorem 21.19. If A is an invertible matrix, then

1. A^k is invertible, $(A^k)^{-1} = (A^{-1})^k$;
2. cA is invertible for any nonzero scalar c , and $(cA)^{-1} = \frac{1}{c}A^{-1}$.

Theorem 21.20. If A is invertible, then A^T is invertible and $(A^T)^{-1} = (A^{-1})^T$.

Proof: If A is invertible, then A^{-1} exists and $AA^{-1} = A^{-1}A = I$. This implies $(A^{-1})^T A^T = A^T(A^{-1})^T = I$. By definition, A^T has an inverse and $(A^T)^{-1} = (A^{-1})^T$.

21.3 Systems of Linear Equations and Linear Independence

With matrix algebra operations, we can define matrix equations. Some matrix equations come from linear approximation or statistical problems. We will begin with the simplest “linear equation” follow by a “system of linear equations” and study how to rewrite these equations in matrix form.

Definition 21.13. The equation $a_1x_1 + a_2x_2 + \cdots + a_nx_n = b$, which expresses b in terms of the variables $x_1, x_2, \dots, x_n$ and constants $a_1, a_2, \dots, a_n$, is called a *linear equation*.

Definition 21.14. A *solution* of a linear equation $a_1x_1 + a_2x_2 + \cdots + a_nx_n = b$ is a sequence of n numbers $s_1, s_2, \dots, s_n$ such that the equation is satisfied when we substitute $x_1 = s_1, x_2 = s_2, \dots, x_n = s_n$.

The set of all solutions of the equations is called its *solution set* or *general solution* of the equation.

Example 21.21. Find the solution set of $4x - 2y = 3$.

Solution. The solution set is $\{(t, \frac{4t-3}{2}) : t \in \mathbb{R}\}$. □

In the above, we discuss just a single linear equation. It can be viewed as the generalisation of 2-d plane to higher-dimensional plane in $\mathbb{R}^n$. In Chapter [20](#), we have learned that the intersection of planes can be found by solving a system of linear equations. The following definition provides a generalisation.

The matrix A is called the *coefficient matrix* of the system. The *augmented matrix* for the system (21.1) is the matrix obtained by combining the coefficient matrix A and matrix $\mathbf{b}$ i.e.

$$\left[\begin{array}{ccc|c} a_{11} & \dots & a_{1n} & b_1 \\ \vdots & \ddots & \vdots & \vdots \\ a_{m1} & \dots & a_{mn} & b_m \end{array} \right].$$

Example 21.24. Write the linear system

$$\begin{aligned} x + 2y &= 1 \\ 4x + 8y &= 3. \end{aligned}$$

in the form of (a) the single matrix equation; (b) the augmented matrix.

Solution. (a) In the single matrix equation form: $\begin{bmatrix} 1 & 2 \\ 4 & 8 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$

(b) In the augmented matrix form: $\left[\begin{array}{cc|c} 1 & 2 & 1 \\ 4 & 8 & 3 \end{array} \right]$ □

What is benefit of writing a linear equation (21.1) in matrix form (21.2)? For a computer point of view, a matrix form allows the data to be stored systematically. From a mathematical point of view, it is because the language of vector space (especially the concepts of linear combination and linearly independence) allow us to describe the solution space of the linear system.

Let us rewrite the matrix A as

$$\left[\begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} \quad \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} \quad \dots \quad \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix} \right].$$

If we denote each column vector as $\mathbf{a}_i$, then A can be viewed as a $1 \times n$ *block matrix* (or *partitioned matrix*) $[\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n]$. Therefore $A\mathbf{x}$ is a linear combination of column vectors $\mathbf{a}_i$:

$$A\mathbf{x} = x_1\mathbf{a}_1 + \dots + x_n\mathbf{a}_n.$$

There is a linear system which is special, called a homogeneous system defined below.

Definition 21.18. A linear system (21.1) (resp. a linear matrix equation (21.2)) is called *homogeneous* when $b_i = 0$ (resp. $\mathbf{b} = \mathbf{0}$).

Example 21.25. The linear system $\begin{cases} x_1 - x_2 - x_3 = 0 \\ 2x_1 - x_2 - 3x_3 = 0 \\ x_1 - 2x_3 = 0 \end{cases}$ is homogeneous.

Definition 21.19. The homogeneous system *corresponding* to the linear matrix equation (21.2) (resp. linear system (21.1)) is the linear matrix equation $A\mathbf{x} = \mathbf{0}$ (resp. $b_i = 0$), with the same matrix A (resp. coefficients).

The solution of the a homogeneous system naturally relates to the following concepts.

Definition 21.20 (Linear Dependence and Independence). Consider vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$.

1. $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ are *linearly independent* iff the equation

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \dots + c_n\mathbf{v}_n = \mathbf{0}$$

has only trivial solution i.e. $c_1 = 0, c_2 = 0, \dots, c_n = 0$;

2. $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ are *linearly dependent* iff the equation

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \dots + c_n\mathbf{v}_n = \mathbf{0}$$

has nontrivial solutions, i.e. there exist real numbers $c_1, c_2, \dots, c_n$, not all c_i 's are zero, such that $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \dots + c_n\mathbf{v}_n = \mathbf{0}$.

When the column vectors of A is “linearly independent”, we have a unique solution for the homogeneous system $A\mathbf{x} = \mathbf{0}$. But when column vectors of A is “linearly dependent”, we have many solutions for $A\mathbf{x} = \mathbf{0}$.

Example 21.26. Are the vectors $\mathbf{v}_1 = (1, 2)$ and $\mathbf{v}_2 = (2, 4)$ linearly independent?

Solution. No. Since $\mathbf{v}_2 = 2\mathbf{v}_1$, we have $2\mathbf{v}_1 + (-1)\mathbf{v}_2 = \mathbf{0}$ and by definition, they are linearly dependent. $\square$

This example has inspired mathematicians to obtain the following theorem.

Theorem 21.27. Two vectors are linearly dependent iff they are parallel, i.e. one is a scalar multiple of the other.

Example 21.28. Determine whether the (row) vectors $(1, -2, 3)$, $(5, 6, -1)$, $(3, 2, 1)$ are linearly independent.

Solution. Consider

$$a(1, -2, 3) + b(5, 6, -1) + c(3, 2, 1) = (0, 0, 0).$$

We can solve the above equation to obtain

$$a = -k, \quad b = -k, \quad c = 2k, \quad k \in \mathbb{R}.$$

Hence, the vectors are linearly dependent since there non-zero coefficients such that (e.g., take $k = -1$)

$$(1, -2, 3) + (5, 6, -1) + (-2)(3, 2, 1) = (0, 0, 0).$$

$\square$

We can generalise this example to obtain the following theorem.

Theorem 21.29. Three vectors in $\mathbb{R}^3$ are linearly dependent iff they all lie in the same plane, i.e. one is a linear combination of the other two.

Theorem 21.30. The standard basis vectors of $\mathbb{R}^n$ are linearly independent.

Proof: Let $\mathbf{e}_1, \dots, \mathbf{e}_n$ be the standard basis vectors. Then the equation

$$c_1 \mathbf{v}_1 + \dots + c_n \mathbf{v}_n = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \Rightarrow c_1 = \dots = c_n = 0.$$

By definition, $\mathbf{e}_1, \dots, \mathbf{e}_n$ are linearly independent.

Example 21.31. Decide whether the vectors $\mathbf{u}_1 = (2, 3, 1, 4)$, $\mathbf{u}_2 = (-1, 1, 2, 3)$ and $\mathbf{u}_3 = (4, 0, -1, 1)$ are linearly independent or linearly dependent.

Solution. Consider $a\mathbf{u}_1 + b\mathbf{u}_2 + c\mathbf{u}_3 = \mathbf{0}$. From this, we obtain a linear system

$$2a - b + 4c = 0$$

$$3a + b = 0$$

$$a + 2b - c = 0$$

$$4a + 3b + c = 0$$

Solving it gives $a = b = c = 0$. Therefore, $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ are linearly independent. $\square$

The following theorem says that the solution of a system (21.2) (and (21.1)) can be decomposed to two parts — a particular point and a vector subspace (which will be discussed in Chapter 23).

Theorem 21.32. Suppose $\mathbf{x}_p$ is a particular solution to the system $A\mathbf{x} = \mathbf{b}$. Then any solution to $A\mathbf{x} = \mathbf{b}$ is the sum $\mathbf{x}_p + \mathbf{x}_h$ where $\mathbf{x}_h$ is a solution to the homogeneous system $A\mathbf{x} = \mathbf{0}$.

Proof: From the assumption, we know: $A\mathbf{x}_p = \mathbf{b}$ and $A\mathbf{x}_h = \mathbf{0}$. Adding the two gives us: $A\mathbf{x}_p + A\mathbf{x}_h = \mathbf{b} + \mathbf{0}$, this implies $A(\mathbf{x}_p + \mathbf{x}_h) = \mathbf{b}$, which indicates $\mathbf{x}_p + \mathbf{x}_h$ is also a solution of $A\mathbf{x} = \mathbf{b}$.

21.4 Special Matrices and Row Echelon Form

We find that certain types of equations are easier to solve. Let us look at the following examples.

Example 21.33. Solve $\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 2 & 2 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \\ 2 \end{bmatrix}$.

Solution. It is easy to obtain the solution $\begin{cases} y_1 = 2 \\ y_2 = 4 - 2y_1 = 0 \\ y_3 = 2 - 2y_2 - 2y_1 = -2 \end{cases}$ $\square$

Example 21.34. Solve $\begin{bmatrix} 2 & 2 & 2 \\ 0 & 2 & 2 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ -2 \end{bmatrix}$.

Solution. It is easy to obtain the solution $\begin{cases} x_3 = -1 \\ x_2 = -2x_3/2 = 1 \\ x_1 = (2 - 2x_2 - 2x_3)/2 = 1 \end{cases}$ $\square$

Example 21.35. If $\begin{bmatrix} 2 & 2 & 2 \\ 4 & 6 & 6 \\ 4 & 8 & 10 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 2 & 2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 2 & 2 \\ 0 & 2 & 2 \\ 0 & 0 & 2 \end{bmatrix}$, solve $\begin{bmatrix} 2 & 2 & 2 \\ 4 & 6 & 6 \\ 4 & 8 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \\ 2 \end{bmatrix}$.

Solution. Let $A = LU$ and $\mathbf{y} = U\mathbf{x}$, then making use of the above two examples, we can find that the solution is $x_1 = 1, x_2 = 1, x_3 = -1$. $\square$

We find that Examples 21.33 and 21.34 are easy to solve because the coefficient matrix has special forms. The following definition defines a few types of commonly encounter special matrices.

Definition 21.21. Let $A \in M_n(\mathbb{R})$.

1. A is called *upper triangular* if $a_{ij} = 0$ whenever $j < i$ and *lower triangular* if $a_{ij} = 0$ whenever $j > i$.
2. A matrix P is called a *permutation matrix* if exactly one entry in each row and column is equal to 1, and all other entries are zero. Pre-multiplication by such a matrix interchanges the rows and post-multiplication, the columns.
3. A is called a *diagonal matrix* if $a_{ij} = 0$, whenever $i \neq j$.
4. A is called *symmetric* if $A^T = A$.
5. A is called *nil-potent* if there is a positive integer k such that $A^k = 0$.
6. A is called *idempotent* if $A^2 = A$.
7. A is called *orthogonal* if $A^T A = AA^T = I$.

Example 21.36. The coefficient matrix in Example 21.33 is lower triangular while the coefficient matrix of Example 21.34 is upper triangular.

When solving a linear system, what are the operations that do not change the solutions? There are three elementary row operations stated below which are found not to change the solutions of a linear system:

1. Multiply a row through by a nonzero constant $a \neq 0$: $aR_i \rightarrow R_i$.
2. Interchange two rows: $R_i \leftrightarrow R_j$ ($i \neq j$).
3. Add a multiple of one row to another row. $R_i + aR_j \rightarrow R_i$ ($i \neq j$).

Let us investigate how the elementary row operations work by studying the following example

Example 21.37. Write down the augmented matrix for the system of linear equations

$$\begin{aligned}x + y + z &= 0, \\ -2x + y + z &= -9. \\ x + 3y &= -3.\end{aligned}$$

Apply the elementary row operations on the augmented matrix to simplify the above equation and solve it.

Solution:

$$\begin{aligned}\left[\begin{array}{ccc|c} 1 & 1 & 1 & 0 \\ -2 & 1 & 1 & -9 \\ 1 & 3 & 0 & -3 \end{array} \right] & \xrightarrow[\substack{R_2+2R_1 \rightarrow R_2 \\ R_3-R_1 \rightarrow R_3}]{\substack{R_2+2R_1 \rightarrow R_2 \\ R_3-R_1 \rightarrow R_3}} \left[\begin{array}{ccc|c} 1 & 1 & 1 & 0 \\ 0 & 3 & 3 & -9 \\ 0 & 2 & -1 & -3 \end{array} \right] \\ & \xrightarrow{3R_3-2R_2 \rightarrow R_3} \left[\begin{array}{ccc|c} 1 & 1 & 1 & 0 \\ 0 & 3 & 3 & -9 \\ 0 & 0 & -9 & 9 \end{array} \right]\end{aligned}$$

The equation corresponding to the simplify augmented matrix is

$$\begin{aligned}x + y + z &= 0, \\3y + 3z &= -9. \\-9z &= 9.\end{aligned}$$

Solving it gives $z = -1$, $y = -2$, $x = 3$.

The following definition provides a generalisation of the upper triangular matrix mentioned above.

Definition 21.22. An $m \times n$ matrix A is said to be in *row echelon form* (r.e.f.) if it satisfies the following properties:

- (i) All rows that consist of *all* zeros, if any, are at the *bottom* of the matrix.
- (ii) The first nonzero entry in each row which is called the *leading entry* and leading entries step from left to right as we read down the matrix.
- (iii) All the entries in a column below a leading entry are 0.

Definition 21.23. A matrix in r.e.f. is said to be in *reduced row echelon form* (r.r.e.f.) if each of its leading entry is a 1 and the leading entry is the only nonzero entry in its column.

Example 21.38. Determine whether the following matrices are r.e.f, r.r.e.f or neither:

$$\begin{aligned} & \text{(a) } \begin{bmatrix} 1 & 0 & 0 & 3 \\ 0 & 1 & 0 & 7 \\ 0 & 0 & 1 & -1 \end{bmatrix}; \text{ (b) } \begin{bmatrix} 0 & 1 & -1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}; \text{ (c) } \begin{bmatrix} 0 & 1 & 2 & 6 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \\ & \text{(d) } \begin{bmatrix} 1 & 3 & 0 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 \end{bmatrix}; \text{ (e) } \begin{bmatrix} 1 & 0 & 2 & 3 \\ 0 & 4 & -4 & 2 \\ 0 & 0 & 2 & 9 \end{bmatrix}; \text{ (f) } \begin{bmatrix} 1 & 0 & 3 & 4 \\ 0 & 1 & -2 & 5 \\ 0 & 1 & 2 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{aligned}$$

Solution. (a) r.r.e.f (& r.e.f)
 (b) r.r.e.f (& r.e.f)
 (c) r.e.f
 (d) neither
 (e) r.e.f
 (f) neither

□

In what follows, we will be introducing two elimination techniques — the *Gaussian Elimination* is used to reduce a matrix to a r.e.f. The *Gauss-Jordan Elimination* is used to reduce a matrix to the r.r.e.f.

Example 21.39. Solve the system $\begin{cases} 2x - 3y + 6z = 14 \\ x + y - 2z = -3 \end{cases}$ by using Gauss-Jordan elimination.

eg:lin-sys2

Solution.

$$\begin{aligned} \left[\begin{array}{ccc|c} 2 & -3 & 6 & 14 \\ 1 & 1 & -2 & -3 \end{array} \right] &\xrightarrow{2R_2 - R_1 \rightarrow R_2} \left[\begin{array}{ccc|c} 2 & -3 & 6 & 14 \\ 0 & 5 & -10 & -20 \end{array} \right] \\ &\xrightarrow[R_1/2 \rightarrow R_1]{R_2/5 \rightarrow R_2} \left[\begin{array}{ccc|c} 1 & -1.5 & 3 & 7 \\ 0 & 1 & -2 & -4 \end{array} \right] \\ &\xrightarrow{R_1 + 1.5R_2 \rightarrow R_1} \left[\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & -2 & -4 \end{array} \right] \quad \text{in r.r.e.f.} \end{aligned}$$

Therefore, the solution is

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ -4 \\ 0 \end{bmatrix} + t \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}$$

□

Remark 21.40. According to Theorem [21.32](#), the solutions of the linear system Example [21.39](#) can be written as the sum of a particular solution and a homogeneous solution in the form

$$\underbrace{\begin{bmatrix} x \\ y \\ z \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} 1 \\ -4 \\ 0 \end{bmatrix}}_{\mathbf{x}_p} + t \underbrace{\begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}}_{\mathbf{x}_h}$$

Example 21.41. Solve the system $\begin{cases} x + 2y = 1 \\ 4x + 8y = 3 \end{cases}$ by using Gaussian elimination.

Solution.

$$\left[\begin{array}{cc|c} 1 & 2 & 1 \\ 4 & 8 & 3 \end{array} \right] \rightarrow \left[\begin{array}{cc|c} 1 & 2 & 1 \\ 0 & 0 & -1 \end{array} \right]$$

There is no solution to the system above.

□

Theoretically, when solving a linear system (21.1), we are faced with three types of situations:

1. the system has no solution; Example [21.41](#)
2. the system has a unique solution; and
3. the system has infinitely many solutions. Example [21.39](#)

We will be investigating the second and third situations with the following examples.

Example 21.42. Solve the system $\begin{cases} 2x - y + z = -7 \\ x + y + z = -2 \\ 3x + y + z = 0 \end{cases}$

Solution.

$$\left[\begin{array}{ccc|c} 2 & -1 & 1 & -7 \\ 1 & 1 & 1 & -2 \\ 3 & 1 & 1 & 0 \end{array} \right] \xrightarrow[2R_3 - 3R_1 \rightarrow R_3]{2R_2 - R_1 \rightarrow R_2} \left[\begin{array}{ccc|c} 2 & -1 & 1 & -7 \\ 0 & 3 & 1 & 3 \\ 0 & 5 & -1 & 21 \end{array} \right] \rightarrow \left[\begin{array}{ccc|c} 2 & -1 & 1 & -7 \\ 0 & 3 & 1 & 3 \\ 0 & 0 & -8 & 48 \end{array} \right]$$

The (unique) solution is $z = -6, y = 3, x = 1$.

□

Example 21.43. Solve the system $\begin{cases} x_1 - x_2 - x_3 = 2 \\ 2x_1 - x_2 - 3x_3 = 6 \\ x_1 - 2x_3 = 4 \end{cases}$

Solution.

$$\left[\begin{array}{ccc|c} 1 & -1 & -1 & 2 \\ 2 & -1 & -3 & 6 \\ 1 & 0 & -2 & 4 \end{array} \right] \rightarrow \left[\begin{array}{ccc|c} 1 & -1 & -1 & 2 \\ 0 & 1 & -1 & 2 \\ 0 & 1 & -1 & 2 \end{array} \right] \rightarrow \left[\begin{array}{ccc|c} 1 & -1 & -1 & 2 \\ 0 & 1 & -1 & 2 \\ 0 & 0 & 0 & 0 \end{array} \right]$$

There are multiple solutions: $z = t$, $y = 2 + t$, $x = 2 + (2 + t) + t = 4 + 2t$, where t is any real number. $\square$

21.5 LU Factorisation and Its Variants

In the previous section, we learn that there are three elementary row operations on a matrix which preserve the solution of a linear system:

1. the interchange of two rows;
2. multiplication of a row by a nonzero scalar; and
3. addition of a multiple of one row to another row.

These three operations are associated with the notion defined below.

Definition 21.24. An *elementary matrix* is a square matrix obtained from the identity matrix by a *single* elementary row operation.

1. Elementary matrix for interchanging two rows:

$$E_{(i),(j)} = \begin{bmatrix} 1 & & & \vdots & & \vdots & & \\ & \ddots & & \vdots & & \vdots & & \\ & & 1 & & & & & \\ \cdots & \cdots & 0 & \cdots & 1 & \cdots & \cdots & (i) \\ & & & 1 & & & & \\ & & & \vdots & \ddots & \vdots & & \\ \cdots & \cdots & 1 & \cdots & 0 & \cdots & \cdots & (j) \\ & & & & & 1 & & \\ & & & \vdots & & \vdots & \ddots & \\ & & (i) & & (j) & & & 1 \end{bmatrix}.$$

2. Elementary matrix for multiplication of a row by a nonzero scalar α :

$$E_{\alpha(i)} = \begin{bmatrix} 1 & & & \vdots & & \\ & \ddots & & \vdots & & \\ & & 1 & & & \\ \cdots & \cdots & \alpha & \cdots & \cdots & \\ & & & 1 & & \\ & & & \vdots & \ddots & \\ & & & \vdots & & 1 \end{bmatrix}.$$

3. Elementary matrix for addition of a multiple of one row to another row:

$$E_{(i)+\alpha(j)} = \begin{bmatrix} 1 & & & & \vdots & & \\ & \ddots & & & \vdots & & \\ \dots & \dots & 1 & \dots & \alpha & \dots & (i) \\ & & & \ddots & \vdots & & \\ & & & & 1 & & \\ & & & & \vdots & \ddots & \\ & & & & (j) & & 1 \end{bmatrix}.$$

Theorem 21.44. *If A is a matrix and E is an elementary matrix with EA defined, then EA is a matrix obtained by applying to A the elementary operation that defined E .*

Proof: Just multiply $E_{\alpha(i)}$, $E_{(i),(j)}$ and $E_{(i)+\alpha(j)}$ to A and observe the effect as follows:

The interchange of two rows R_i and R_j can be carried out by multiplying $E_{(i),(j)}$ to A :

$$A \xrightarrow{R_i \leftrightarrow R_j} E_{(i),(j)} A.$$

Multiplication of a row R_i by a nonzero scalar can be carried out by multiplying $E_{\alpha(i)}$ to A :

$$A \xrightarrow{\alpha R_i \rightarrow R_i} E_{\alpha(i)} A.$$

Addition of a multiple α to one row R_j to another row R_i can be carried out by:

$$A \xrightarrow{R_i + \alpha R_j \rightarrow R_i} E_{(i)+\alpha(j)} A.$$

thm:elem:inv

Theorem 21.45. *If E is an elementary matrix, then E has an inverse that is also an elementary matrix that reverses what E does.*

$$\begin{aligned} E_{(i),(j)}^{-1} &= E_{(i),(j)} \\ E_{\alpha(i)}^{-1} &= E_{\frac{1}{\alpha}(i)}, i \neq 0 \\ E_{(i)+\alpha(j)}^{-1} &= E_{(i)-\alpha(j)}, i \neq j. \end{aligned}$$

Proof: They can be derived from

$$\begin{aligned} E_{(i),(j)} E_{(i),(j)} &= I \\ E_{\alpha(i)} E_{\frac{1}{\alpha}(i)} &= I, \quad i \neq 0 \\ E_{(i)+\alpha(j)} E_{(i)-\alpha(j)} &= I, \quad i \neq j. \end{aligned}$$

The elementary row matrices allow us to factorise a large number of matrices into a product to lower triangular matrix and an r.e.f. (a generalisation of upper triangular matrix). The notion of this factorisation is defined below.

Definition 21.25. An LU factorisation of a matrix $A \in M_{m,n}(\mathbb{R})$ is the representation of A as the product $A = LU$ of a (square) lower triangular matrix $L \in M_m(\mathbb{R})$ and an r.e.f. $U \in M_{m,n}(\mathbb{R})$.

The following example shows that there are some matrices which do not have LU factorisation.

Example 21.46. Show that $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ does not have an LU factorisation.

Solution. Suppose A has an LU factorisation. Then

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} a & 0 \\ b & c \end{bmatrix} \begin{bmatrix} d & e \\ 0 & f \end{bmatrix} = \begin{bmatrix} ad & ae \\ bd & be + cf \end{bmatrix} \Rightarrow \begin{cases} ad = 0 \Rightarrow a = 0 \vee d = 0 \\ ae = 1 \Rightarrow a \neq 0 \wedge e \neq 0 \\ bd = 1 \Rightarrow b \neq 0 \wedge d \neq 0 \end{cases}$$

This is contradicting since $a \neq 0$ and $d \neq 0$ but $ad = 0$. □

The following theorem states the condition where LU factorisation exists for a given matrix.

Theorem 21.47. If a matrix A can be reduced to row echelon form without row interchanges, then A has an LU factorisation.

Suppose that A has an LU factorisation. Then there exists a sequence of elementary matrices $E_1, \dots, E_k$ such that

$$E_k \cdots E_1 A = U.$$

Mathematicians have discovered that the elementary matrices can be further grouped as

$$L_{n-1} \cdots L_1 A = U.$$

Therefore $A = LU$ where

$$L = L_1^{-1} \cdots L_{n-1}^{-1} = \begin{bmatrix} 1 & & & & \\ l_{21} & 1 & & & \\ l_{31} & l_{32} & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \cdots & l_{n,n-1} & 1 \end{bmatrix}, \quad L_k = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -l_{k+1,k} & 1 & \\ & & \vdots & & \ddots \\ & & -l_{nk} & & & 1 \end{bmatrix}.$$

Let us investigate the steps in LU factorisation with examples.

Example 21.48. Find an LU factorisation of $\begin{bmatrix} 1 & 2 \\ 4 & 6 \end{bmatrix}$.

Solution:

$$A = \begin{bmatrix} 1 & 2 \\ 4 & 6 \end{bmatrix} \xrightarrow{R_2 - 4R_1 \rightarrow R_2} \underbrace{\begin{bmatrix} 1 & 0 \\ -4 & 1 \end{bmatrix}}_{L_1=E_1} \underbrace{\begin{bmatrix} 1 & 2 \\ 4 & 6 \end{bmatrix}}_A = \underbrace{\begin{bmatrix} 1 & 2 \\ 0 & -2 \end{bmatrix}}_U \Rightarrow A = \begin{bmatrix} 1 & 0 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & -2 \end{bmatrix}.$$

Example 21.49. Write down the LU factorisation of $\begin{bmatrix} 1 & 2 & 3 \\ 4 & -6 & -6 \\ -7 & -7 & 9 \end{bmatrix}$.

Solution.

$$\begin{aligned} \begin{bmatrix} 1 & 2 & 3 \\ 4 & -6 & -6 \\ -7 & -7 & 9 \end{bmatrix} &= \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ -7 & 0 & 1 \end{bmatrix}}_{L_1^{-1}} \begin{bmatrix} 1 & 2 & 3 \\ 0 & -14 & -18 \\ 0 & 7 & 30 \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ -7 & -\frac{1}{2} & 1 \end{bmatrix}}_{L=L_2^{-1}L_1^{-1}} \begin{bmatrix} 1 & 2 & 3 \\ 0 & -14 & -18 \\ 0 & 0 & 21 \end{bmatrix}. \end{aligned}$$

□

Example 21.50. Solve the following linear system by using LU factorisation.

$$\begin{aligned} x_1 + 2x_2 + 3x_3 &= 0 \\ 4x_1 - 6x_2 - 6x_3 &= 1 \\ -7x_1 - 7x_2 + 9x_3 &= 0 \end{aligned}$$

Solution. Express the above equation as $A\mathbf{x} = \mathbf{b}$, where $\mathbf{b} = (0, 1, 0)^T$ and $\mathbf{x} = (x_1, x_2, x_3)^T$. First, the LU factorisation of A (this has been done in Example 21.49):

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & -6 & -6 \\ -7 & -7 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ -7 & -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 0 & -14 & -18 \\ 0 & 0 & 21 \end{bmatrix} = LU.$$

Let $\mathbf{y} = U\mathbf{x}$, we then solve $L\mathbf{y} = \mathbf{b}$ as follows:

$$\begin{aligned} y_1 &= 0, \\ y_2 &= 1 - 4y_1 = 1, \\ y_3 &= 7y_1 + \frac{1}{2}y_2 = \frac{1}{2}. \end{aligned}$$

We now solve $U\mathbf{x} = \mathbf{y}$ as follows:

$$\begin{aligned} 21x_3 &= \frac{1}{2} \Rightarrow x_3 = \frac{1}{42} \\ -14x_2 - 18x_3 &= 1 \Rightarrow x_2 = -\frac{5}{49} \\ x_1 + 2x_2 + 3x_3 &= 0 \Rightarrow x_1 = \frac{13}{98} \end{aligned}$$

□

Definition 21.26. An LDU factorisation of a matrix A is a presentation of $A = LDU$ as the product of a lower triangular matrix L with 1s on the main diagonal, a diagonal matrix D , and an upper triangular matrix U (the same size as A) with 1s on the diagonal.

Example 21.51. Find an LDU factorisation of $A = \begin{bmatrix} -1 & 2 & 0 & -2 \\ 2 & 4 & 6 & 8 \\ -3 & 2 & 4 & 1 \end{bmatrix}$.

Solution.

$$\begin{aligned}
 A &= \begin{bmatrix} 1 & & \\ -2 & 1 & \\ 3 & & 1 \end{bmatrix} \begin{bmatrix} -1 & 2 & 0 & -2 \\ 0 & 8 & 6 & 4 \\ 0 & -4 & 4 & 7 \end{bmatrix} \\
 &= \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 3 & -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} -1 & 2 & 0 & -2 \\ 0 & 8 & 6 & 4 \\ 0 & 0 & 7 & 9 \end{bmatrix} \\
 &= \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 3 & -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} -1 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 7 \end{bmatrix} \begin{bmatrix} 1 & -2 & 0 & 2 \\ 0 & 1 & \frac{3}{4} & \frac{1}{2} \\ 0 & 0 & 1 & \frac{9}{7} \end{bmatrix}
 \end{aligned}$$

□

In general, neither LU nor LDU factorisation are unique. This is demonstrated in the following examples.

Example 21.52. Show that $A = \begin{bmatrix} -4 & 2 \\ 3 & 1 \end{bmatrix}$ has two different LU factorisations.

Solution:

$$A = \begin{bmatrix} -1 & 0 \\ -\frac{3}{4} & 1 \end{bmatrix} \begin{bmatrix} -4 & 2 \\ 0 & \frac{5}{2} \end{bmatrix} = \begin{bmatrix} -4 & 0 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{2} \\ 0 & \frac{5}{2} \end{bmatrix}.$$

Example 21.53. Show that $A = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ has infinitely many different LDU factorisations.

Solution: For any real numbers a and b ,

$$A = \begin{bmatrix} 1 & 0 \\ a & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix}.$$

Theorem 21.54. If an invertible matrix A can be factored $A = LDU$ and also $A = L_1 D_1 U_1$ with L and L_1 lower triangular matrices with 1s on the main diagonal, U and U_1 upper triangular matrices with 1s on the main diagonal, and D and D_1 diagonal matrices, then $L = L_1$, $U = U_1$ and $D = D_1$.

Example 21.55. Find an LDU factorisation of the symmetric matrix $A = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 3 & 0 \\ -1 & 0 & 5 \end{bmatrix}$.

Solution.

$$\begin{aligned}
 A &= \begin{bmatrix} 1 & & \\ 2 & 1 & \\ -1 & & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & -1 \\ 0 & -1 & 2 \\ 0 & 2 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & -1 \\ 0 & -1 & 2 \\ 0 & 0 & 8 \end{bmatrix} \\
 &= \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 8 \end{bmatrix} \begin{bmatrix} 1 & 2 & -1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{bmatrix}
 \end{aligned}$$

□

PLU and $PLDU$ factorisations are the generalisation of LU and LDU factorisations. The P stands for permutation matrix which has been defined in Definition [21.21](#). After stating the

properties of permutation matrix, we state the theorem which says that all matrix has a PLU factorisation.

Theorem 21.56. *The product of permutation matrices is a permutation matrix.*

Theorem 21.57. *If P is a permutation matrix, then $P^{-1} = P^T$.*

Example 21.58. Given $A = \begin{bmatrix} 2 & 4 & 2 \\ 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix}$, find a permutation matrix P such that the (1,1)-element of PA is 1.

Solution.

$$PA = \underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_P \begin{bmatrix} 2 & 4 & 2 \\ 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 2 \\ 3 & 2 & 1 \end{bmatrix}.$$

□

Theorem 21.59. *If A is a nonsingular matrix, then there exists a permutation matrix P so that PA has an LU-factorisation.*

Example 21.60. Find the PLU factorisation of $A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & -1 & -1 & 4 \\ -2 & 3 & 2 & 1 \\ -6 & 4 & 2 & -8 \end{bmatrix}$

Solution.

$$\begin{aligned} \underbrace{\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{P_1} A &= \begin{bmatrix} -2 & 3 & 2 & 1 \\ 0 & -1 & -1 & 4 \\ 0 & 1 & 1 & 0 \\ -6 & 4 & 2 & -8 \end{bmatrix} = \begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ 0 & & 1 & \\ 3 & & & 1 \end{bmatrix} \begin{bmatrix} -2 & 3 & 2 & 1 \\ 0 & -1 & -1 & 4 \\ 0 & 1 & 1 & 0 \\ 0 & -5 & -4 & -11 \end{bmatrix} \\ &= \begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ 0 & -1 & 1 & \\ 3 & 5 & & 1 \end{bmatrix} \begin{bmatrix} -2 & 3 & 2 & 1 \\ 0 & -1 & -1 & 4 \\ 0 & 0 & 0 & 4 \\ 0 & 0 & 1 & -31 \end{bmatrix} \\ \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{P_2} P_1 A &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ 0 & -1 & 1 & \\ 3 & 5 & & 1 \end{bmatrix} \begin{bmatrix} -2 & 3 & 2 & 1 \\ 0 & -1 & -1 & 4 \\ 0 & 0 & 0 & 4 \\ 0 & 0 & 1 & -31 \end{bmatrix} \\ &= \begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ 3 & 5 & & 1 \\ 0 & -1 & 1 & \end{bmatrix} P_2^{-1} P_2 \begin{bmatrix} -2 & 3 & 2 & 1 \\ 0 & -1 & -1 & 4 \\ 0 & 0 & 0 & 4 \\ 0 & 0 & 1 & -31 \end{bmatrix} \end{aligned}$$

$$\underbrace{\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}}_{P_2 P_1} A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 3 & 5 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} -2 & 3 & 2 & 1 \\ 0 & -1 & -1 & 4 \\ 0 & 0 & 1 & -31 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

□

Example 21.61. Find a *PLDU* factorisation of the symmetric matrix $A = \begin{bmatrix} 2 & 4 & 2 \\ 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix}$.

Solution.

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} A = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \\ 2 & 4 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 0 & -4 & -8 \\ 0 & 0 & -4 \end{bmatrix}$$

□

Using *PLDU* factorisation, we can decompose a matrix into a product of elementary matrices. Decomposing a matrix as a product of elementary matrices is useful in the theoretical study of matrix inverse as we will see in the next section.

Example 21.62. Express $A = \begin{bmatrix} 2 & 3 \\ 1 & 1 \end{bmatrix}$ as a product of elementary matrices.

Solution.

$$\begin{aligned} A &= \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & -5 \end{bmatrix} \begin{bmatrix} 1 & 1.5 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -5 \end{bmatrix} \begin{bmatrix} 1 & 1.5 \\ 0 & 1 \end{bmatrix} \end{aligned}$$

□

21.6 Applications of Gaussian Elimination

In this section, we try to introduce some applications of Gaussian(-Jordan) Elimination. One application is in the calculation of matrix inverse, the other application is in the study of econometrics.

There are two techniques for finding the inverse of matrix when the size is small — one is by using formula, the other is by reducing the augmented matrix to r.r.e.f.

When $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, we learn earlier that the inverse is $A^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$. When

$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$, the formula of the inverse is

$$A^{-1} = \frac{1}{a(ei-fh) - b(di-fg) + c(dh-eg)} \begin{bmatrix} ei-fh & -(di-fg) & dh-eg \\ -(bi-ch) & ai-cg & -(ah-bg) \\ bf-ce & -(af-cd) & ae-bd \end{bmatrix}^T.$$

Due to the complexity of the above matrix, reducing the augmented matrix to r.r.e.f. as follows may be a better tactic:

$$[A|I] \rightarrow [I|A^{-1}].$$

Example 21.63. Find the inverse of $A = \begin{bmatrix} 1 & 4 \\ 2 & 7 \end{bmatrix}$ by reducing the matrix $\begin{bmatrix} 1 & 4 & | & 1 & 0 \\ 2 & 7 & | & 0 & 1 \end{bmatrix}$ to r.r.e.f.

Solution.

$$\begin{aligned} \begin{bmatrix} 1 & 4 & | & 1 & 0 \\ 2 & 7 & | & 0 & 1 \end{bmatrix} &\rightarrow \begin{bmatrix} 1 & 4 & | & 1 & 0 \\ 0 & -1 & | & -2 & 1 \end{bmatrix} \\ &\rightarrow \begin{bmatrix} 1 & 4 & | & 1 & 0 \\ 0 & -1 & | & -2 & 1 \end{bmatrix} \\ &\rightarrow \begin{bmatrix} 1 & 0 & | & -7 & 4 \\ 0 & 1 & | & 2 & -1 \end{bmatrix} \end{aligned}$$

Therefore $A^{-1} = \begin{bmatrix} -7 & 4 \\ 2 & -1 \end{bmatrix}$. □

eg:2.6.2

Example 21.64. Write the linear system

$$\begin{aligned} 2x_1 + 7x_2 &= -2 \\ x_1 + 4x_2 - x_3 &= 4 \\ x_1 + 3x_2 &= 5 \end{aligned}$$

in the form of $A\mathbf{x} = \mathbf{b}$. Find A^{-1} and hence solve the linear system.

Solution.

$$\begin{bmatrix} 2 & 7 & 0 \\ 1 & 4 & -1 \\ 1 & 3 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -2 \\ 4 \\ 5 \end{bmatrix}$$

Either using the formula $A^{-1} = \text{Adj}(A)/\det(A)$ or

$$\begin{bmatrix} 2 & 7 & 0 & | & 1 & 0 & 0 \\ 1 & 4 & -1 & | & 0 & 1 & 0 \\ 1 & 3 & 0 & | & 0 & 0 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 0 & | & -3 & 0 & 7 \\ 0 & 1 & 0 & | & 1 & 0 & -2 \\ 0 & 0 & 1 & | & 1 & -1 & -1 \end{bmatrix}.$$

Therefore

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -3 & 0 & 7 \\ 1 & 0 & -2 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} -2 \\ 4 \\ 5 \end{bmatrix} = \begin{bmatrix} 41 \\ -12 \\ -11 \end{bmatrix}.$$

□

Theorem 21.65. If A has an inverse, then $A\mathbf{x} = \mathbf{b}$ has the unique solution $\mathbf{x} = A^{-1}\mathbf{b}$.

Proof: Assuming A has an inverse, then its inverse A^{-1} exists and defined $\mathbf{x}_1 = A^{-1}\mathbf{b}$. Since $A\mathbf{x}_1 = AA^{-1}\mathbf{b} = \mathbf{b}$, $\mathbf{x}_1$ is a solution.

If there is another solution $\mathbf{x}_2$, then $A\mathbf{x}_1 = A\mathbf{x}_2 = \mathbf{b}$. So $A(\mathbf{x}_1 - \mathbf{x}_2) = \mathbf{0}$. Multiplying A^{-1} on both sides gives use $\mathbf{x}_1 - \mathbf{x}_2 = \mathbf{0}$ which shows them to be the same vector.

Remark 21.66. The formula $\mathbf{x} = A^{-1}\mathbf{b}$ is only of theoretical interest. Never use this in real life numerical calculation! For instance, in the statistical software *R*, the right way to solve Example 21.64 is

```
> A = matrix(c(2,7,0,1,4,-1,1,3,0),nrow=3,byrow=TRUE)
> b = c(-2,4,5)
> solve(A,b)
```

Theorem 21.67. *A matrix is invertible iff it can be written as the product of elementary matrices.*

Proof: Consider a square matrix A .

$\Rightarrow$: If A is invertible, then it is row-equivalent to the identity matrix I through the Gauss-Jordan elimination: there is a sequence of elementary operations that transform A into I :

$$E_k^{-1} \cdots E_2^{-1} E_1^{-1} A = I$$

So, we have

$$A = E_1 E_2 \cdots E_k$$

where E_i are elementary matrices.

$\Leftarrow$: If A can be written as the product of elementary matrices

$$A = E_1 E_2 \cdots E_k,$$

since elementary matrices are invertible (Theorem [21.45](#)), ^{thm:elem:inv} therefore

$$E_k^{-1} \cdots E_2^{-1} E_1^{-1} A = E_k^{-1} \cdots E_2^{-1} E_1^{-1} E_1 E_2 \cdots E_k = I.$$

By definition, A is invertible.

Theorem 21.68. *A square matrix is invertible iff its columns are linearly independent.*

Proof: Let the square matrix be A .

$\Rightarrow$: If A is invertible, A^{-1} exists. Consider

$$A\mathbf{x} = \text{linear combination of the columns} = \mathbf{0}.$$

Multiplying A^{-1} gives $\mathbf{x} = A^{-1}\mathbf{0} = \mathbf{0}$. By definition, the columns of A are linearly independent.

$\Leftarrow$: Suppose the columns of A are linearly independent, the column vectors of A form a base for $\mathbb{R}^n$ so we can write every vector $\mathbf{b} \in \mathbb{R}^n$ as a unique linear combination of the columns, therefore we can always find $\mathbf{x} \in \mathbb{R}^n$ such that $A\mathbf{x} = \mathbf{b}$. Consider the special case where $\mathbf{b} = \mathbf{e}_i$, it has a corresponding $\mathbf{x}_i$ such that $A\mathbf{x}_i = \mathbf{e}_i$, putting them in matrix form and we have $A[\mathbf{x}_1 \cdots \mathbf{x}_n] = I_n$, by definition, A is invertible.

In order to understand and be able to manipulate the economy of a country or a region, one needs to come up with a certain model based on the various sectors of this economy. The *Leontief model* is an attempt in this direction. Based on the assumption that each industry in the economy has two types of demands: *external demand* (from outside the system) and *internal demand* (demand placed on one industry by another in the same system), the Leontief model represents the economy as a system of linear equations. The Leontief model was invented in the 1930's by Professor Wassily Leontief who developed an economic model of the United States economy by dividing it into 500 economic sectors. On 18 October, 1973, Professor Leontief was awarded the Nobel Prize in economy for his effort.

Consider an economy consisting of n interdependent industries (or sectors) $S_1, \dots, S_n$. That means that each industry consumes some of the goods produced by the other industries, including itself (for example, a power-generating plant uses some of its own power for production). We say that such an economy is *closed* if it satisfies its own needs; that is, no goods leave or enter the system. Let a_{ij} be the number of units produced by industry S_i and necessary to produce one unit of industry S_j . If x_k is the production level of industry S_k , then $a_{ij}x_j$ represents the

number of units produced by industry S_i and consumed by industry S_j . Then the total number of units produced by industry S_i is given by:

$$a_{i1}x_1 + \cdots + a_{in}x_n$$

In order to have a balanced economy, the total production of each industry must be equal to its total consumption. This gives the linear system:

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= x_1 \\ &\vdots \\ a_{n1}x_1 + \cdots + a_{nn}x_n &= x_n \end{aligned}$$

The above system can be written in matrix form:

$$A\mathbf{x} = \mathbf{x}$$

where the matrix A is called the *input-output matrix*.

Example 21.69. Suppose that the economy of a certain region depends on three industries: service, electricity and oil production. Monitoring the operations of these three industries over a period of one year, we were able to come up with the following observations:

1. To produce 1 unit worth of service, the service industry must consume 0.3 units of its own production, 0.3 units of electricity and 0.3 units of oil to run its operations.
2. To produce 1 unit of electricity, the power-generating plant must buy 0.4 units of service, 0.1 units of its own production, and 0.5 units of oil.
3. Finally, the oil production company requires 0.3 units of service, 0.6 units of electricity and 0.2 units of its own production to produce 1 unit of oil.

Find the production level of each of these industries in order to satisfy the external and the internal demands assuming that the above model is closed, that is, no goods leave or enter the system.

Solution: Let x_1 , x_2 and x_3 be the production level for the service industry, the power-generating plant (electricity) and the oil production company, respectively. Since the model is closed, the total consumption of each industry must equal its total production. This gives the following linear system (in matrix form):

$$\underbrace{\begin{bmatrix} 0.3 & 0.3 & 0.3 \\ 0.4 & 0.1 & 0.5 \\ 0.3 & 0.6 & 0.2 \end{bmatrix}}_{\text{input-output matrix}} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Rewriting the above equation as a homogeneous equation $(A - I)\mathbf{x} = \mathbf{0}$:

$$\begin{bmatrix} -0.7 & 0.3 & 0.3 \\ 0.4 & -0.9 & 0.5 \\ 0.3 & 0.6 & -0.8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Reducing the coefficient matrix above to r.r.e.f:

$$\begin{bmatrix} 1 & -\frac{3}{7} & -\frac{3}{7} \\ 0 & -\frac{51}{70} & \frac{47}{70} \\ 0 & \frac{51}{70} & -\frac{47}{70} \end{bmatrix} \rightarrow \begin{bmatrix} 1 & -\frac{3}{7} & -\frac{3}{7} \\ 0 & 1 & -\frac{47}{51} \\ 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & -\frac{14}{17} \\ 0 & 1 & -\frac{47}{51} \\ 0 & 0 & 0 \end{bmatrix}$$

Let $x_3 = t$, then the solution is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 14/17 \\ 47/51 \\ 1 \end{bmatrix} t$$

The values of the variables x_i in this system must be nonnegative in order for the model to make sense; in other words, $t \geq 0$.

21.7 Working with Maxima

and are encoded in Maxima as follows.

```
/*Examples eg:tralegmat-op1
21.2, 21.3 */
A : matrix([6,3,5], [4,7,2]);
B : matrix([7,9,0], [3,1,4]);

transpose(A);
A + B;
A - B;

/*Example eg:mat-prod1
21.6 */
A . transpose(B);

ident(3); /* 3 x 3 identity matrix */

/*Example eg:inv1
21.17 */
invert(matrix([1, 2], [2, 4])); /* Error when no inverse */

invert(matrix([1, 3], [-2, 5]));

determinant(matrix([1, 3], [-2, 5]));

invert(matrix([1, 3], [-2, 5])), detout; /* Show determinant */

/*Examples eg:lin-egs1eg:lin-sys2
21.23, 21.39 */
solve([x + y = 3, 3*x + 3*y = 6], [x, y]);
solve([x - 3*y = -3, 2*x + y = 8], [x, y]);
solve([2*x-3*y+6*z=14, x+y-2*z=-3], [x,y,z]);

/*Examples eg:lu1eg:2.5.8
21.48, 21.49 */
lu_factor(matrix([1, 2], [4, 6]));
lu_factor(matrix([1, 2, 3], [4, -6, -6], [-7, -7, 9]));
```

Chapter 22

Determinants, Eigenvalues, Eigenvectors

ch:eigen

A determinant calculates the “directional volume” of “vectors”. It is used in the hand calculation of eigenvalues. For an eigenvalue, the row echelon form (Section [21.4](#)) is used to reduce the system of linear equations for the computation of corresponding eigenvector(s). Finally, we apply the eigenvectors to simplify some matrix algebra calculations.

22.1 The Determinant of a Matrix

Historically, determinants were considered without reference to matrices: originally, a determinant was defined as a property of a system of linear equations. The determinant “determines” whether the system has a unique solution (which occurs precisely if the determinant is non-zero).

Definition 22.1. The *determinant* of a square matrix $A \in M_n(\mathbb{R})$ is the alternated sum

$$\sum_{\sigma} (-1)^{\sigma} a_{1\sigma(1)} a_{2\sigma(2)} \cdots a_{n\sigma(n)}$$

where the summation is over all permutations σ (Recall the concept of permutation from Discrete Mathematics). We denote the determinant of matrix A as $\det(A)$, $\det A$ or $|A|$.

If $\det(A) \neq 0$, then A is called *non-singular*. If $\det(A) = 0$, then A is called *singular*.

Theorem 22.1. If A is a square matrix, then $\det(A^T) = \det(A)$.

From the definition, mathematicians obtained a “recursive” formula for an $n \times n$ matrix A .

Theorem 22.2. Let m_{ij} be the determinant of the matrix obtained from A by crossing out the i th row and the j th column of A . Then $\det(A)$ has an expansion with respect to the i th row:

$$\det(A) = \sum_{j=1}^n (-1)^{i+j} a_{ij} m_{ij}.$$

Corollary 22.3. The determinant of any (square) matrix that has a row of zeros is 0.

Hence, any property of determinant proved for rows (for instance, the expansion with respect to the i th row) is also valid for columns (for instance, the expansion with respect to the j th column).

Corollary 22.4. The determinant of a square triangular matrix is a product of the entries on its main diagonal. In particular, the determinant of a diagonal matrix is the product of its main diagonal entries.

Classical linear algebra has developed a “general formulation” for finding the inverse of a square matrix and the solution of a system of linear equations. In order to state the inverse and the solution, we will define a few concepts below.

Definition 22.2. If A is an $n \times n$ matrix, the (i, j) -minor m_{ij} is the determinant of the $(n-1) \times (n-1)$ matrix obtained from A by deleting row i and column j . The (i, j) -cofactor c_{ij} of A is $(-1)^{i+j}m_{ij}$. The cofactors of A forms the *cofactor matrix* $C = (c_{ij})$.

Example 22.5. Find all minors and cofactors of $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$.

Solution:

All the minors of A are

$$m_{11} = \det(d) = d$$

$$m_{12} = \det(c) = c$$

$$m_{21} = \det(b) = b$$

$$m_{22} = \det(a) = a$$

All the cofactors of A are

$$c_{11} = (-1)^{1+1} \det(d) = d$$

$$c_{12} = (-1)^{1+2} \det(c) = -c$$

$$c_{21} = (-1)^{2+1} \det(b) = -b$$

$$c_{22} = (-1)^{2+2} \det(a) = +a$$

Example 22.6. Find all minors and cofactors of $B = \begin{bmatrix} 1 & 2 & 3 \\ 0 & -2 & 2 \\ 3 & 7 & -4 \end{bmatrix}$.

Solution. All the minors of B are

$$m_{11} = \begin{vmatrix} -2 & 2 \\ 7 & -4 \end{vmatrix} = -6$$

$$m_{12} = -6$$

$$m_{13} = 6$$

$$m_{21} = \begin{vmatrix} 2 & 3 \\ 7 & -4 \end{vmatrix} = -29$$

$$m_{22} = -13$$

$$m_{23} = 1$$

$$m_{31} = 10$$

$$m_{32} = 2$$

$$m_{33} = -2$$

All the cofactors of B are

$$c_{11} = (-1)^{1+1}(-6) = -6$$

$$c_{12} = (-1)^{1+2}(-6) = 6$$

$$c_{13} = (-1)^{1+3}(6) = 6$$

$$c_{21} = (-1)^{2+1}(29) = 29$$

$$c_{22} = (-1)^{2+2}(-13) = -13$$

$$c_{23} = (-1)^{2+3}(1) = -1$$

$$c_{31} = (-1)^{3+1}(10) = 10$$

$$c_{32} = (-1)^{3+2}(2) = -2$$

$$c_{33} = (-1)^{3+3}(-2) = -2$$

□

Definition 22.3. The transpose of the cofactor matrix of A is called the *adjoint of matrix* A and is denoted as $\text{adj}A$.

eg:3.1.2

Example 22.7. Let $A = \begin{bmatrix} -1 & 0 & 2 \\ 3 & -7 & 8 \\ -4 & 4 & -5 \end{bmatrix}$. Find $\text{adj}A$ and calculate $A(\text{adj}A)$ and $(\text{adj}A)A$.

Solution. The matrix of cofactors of A is

$$C = \begin{bmatrix} (-1)^{1+1} \begin{vmatrix} -7 & 8 \\ 4 & -5 \end{vmatrix} & (-1)^{1+2} \begin{vmatrix} 3 & 8 \\ -4 & -5 \end{vmatrix} & (-1)^{1+3} \begin{vmatrix} 3 & -7 \\ -4 & 4 \end{vmatrix} \\ (-1)^{2+1} \begin{vmatrix} 0 & 2 \\ 4 & -5 \end{vmatrix} & (-1)^{2+2} \begin{vmatrix} -1 & 2 \\ -4 & -5 \end{vmatrix} & (-1)^{2+3} \begin{vmatrix} -1 & 0 \\ -4 & 4 \end{vmatrix} \\ (-1)^{3+1} \begin{vmatrix} 0 & 2 \\ -7 & -8 \end{vmatrix} & (-1)^{3+2} \begin{vmatrix} -1 & 2 \\ 3 & 8 \end{vmatrix} & (-1)^{3+3} \begin{vmatrix} -1 & 0 \\ 3 & -7 \end{vmatrix} \end{bmatrix} = \begin{bmatrix} 3 & -17 & -16 \\ 8 & 13 & 4 \\ 14 & 14 & 7 \end{bmatrix}$$

Therefore

$$\text{adj}A = C^T = \begin{bmatrix} 3 & 8 & 14 \\ -17 & 13 & 14 \\ -16 & 4 & 7 \end{bmatrix}$$

After calculation, we will find that

$$A(\text{adj}A) = (\text{adj}A)A = \begin{bmatrix} -35 & 0 & 0 \\ 0 & -35 & 0 \\ 0 & 0 & -35 \end{bmatrix}$$

□

This example is a special case of the following theorem.

thm:det:adj

Theorem 22.8. Let $A \in M_n(\mathbb{R})$ and I be an $n \times n$ identity matrix. Then

$$A(\text{adj}A) = (\text{adj}A)A = \det(A)I.$$

Proof: Let c_{ij} be the (i, j) -cofactor of A . According to the expansion of determinant of A at the k th row, we have

$$r_{ij} = \sum_{k=1}^n a_{ik}c_{jk} = \delta_{ij} \det A$$

where δ_{ij} equals zero when $i \neq j$, and equals one when $i = j$.

$$\begin{aligned} \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nn} \end{bmatrix} &= \begin{bmatrix} \sum_{k=1}^n a_{1k}c_{1k} & \cdots & \sum_{k=1}^n a_{1k}c_{nk} \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^n a_{nk}c_{1k} & \cdots & \sum_{k=1}^n a_{nk}c_{nk} \end{bmatrix} \\ &= \begin{bmatrix} \delta_{11} \det(A) & \cdots & \delta_{1n} \det(A) \\ \vdots & \ddots & \vdots \\ \delta_{n1} \det(A) & \cdots & \delta_{nn} \det(A) \end{bmatrix} = \det(A) \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} \end{aligned}$$

Example 22.9. Let $A = \begin{bmatrix} -1 & 2 & 2 \\ 3 & 0 & -1 \\ 1 & 4 & 3 \end{bmatrix}$. Find $A \text{adj}A$.

Solution. From Theorem [22.8](#), we can obtain

$$A \operatorname{adj} A = \det(A) I_3 = 0 I_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

□

Corollary 22.10. If $A \in M_n(\mathbb{R})$, A is invertible iff A is non-singular. Equivalently, A is non-invertible iff A is singular.

Corollary 22.11. If A is a square matrix and $\det(A) \neq 0$, then A is invertible and

$$A^{-1} = \frac{1}{\det(A)} \operatorname{adj} A.$$

Example 22.12. Let $A = \begin{bmatrix} -1 & 0 & 2 \\ 3 & -7 & 8 \\ -4 & 4 & -5 \end{bmatrix}$, find A^{-1} .

Solution. From Example [22.7](#) and Corollary [22.11](#), we have

$$A^{-1} = \frac{1}{\det(A)} (\operatorname{adj} A) = \frac{1}{-35} \begin{bmatrix} 3 & 8 & 14 \\ -17 & 13 & 14 \\ -16 & 4 & 7 \end{bmatrix}.$$

□

With the concept of determinant and adjoint, mathematicians developed the Cramer's Rule.

Theorem 22.13 (Cramer's Rule). Consider a system $A\mathbf{x} = \mathbf{b}$ in vector form:

$$x_1 \mathbf{a}_1 + \cdots + x_n \mathbf{a}_n = \mathbf{b}$$

where $\mathbf{a}_1 = \begin{bmatrix} a_{11} \\ \vdots \\ a_{n1} \end{bmatrix}$, $\cdots$, $\mathbf{a}_n = \begin{bmatrix} a_{1n} \\ \vdots \\ a_{nn} \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} b_{11} \\ \vdots \\ b_{n1} \end{bmatrix}$. If the determinant of A of the system $A\mathbf{x} = \mathbf{b}$ is non-zero, then

$$x_i = \frac{\det[\mathbf{a}_1, \cdots, \mathbf{a}_{i-1}, \mathbf{b}, \mathbf{a}_{i+1}, \cdots, \mathbf{a}_n]}{\det A}. \quad (*)$$

Proof: $(\operatorname{adj} A)A\mathbf{x} = (\operatorname{adj} A)\mathbf{b} \Rightarrow \det(A)I\mathbf{x} = (\operatorname{adj} A)\mathbf{b} \Rightarrow \mathbf{x} = \frac{1}{\det(A)}(\operatorname{adj} A)\mathbf{b} \Rightarrow (*)$

Example 22.14. Solve the following linear system of equations using Cramer's Rule.

$$\begin{cases} 2x - 4y + z = 1 \\ x - 5y + 3z = 2 \\ x - y + z = -1 \end{cases}$$

Solution.

$$x = \frac{\begin{vmatrix} 1 & -4 & 1 \\ 2 & -5 & 3 \\ -1 & -1 & 1 \end{vmatrix}}{\begin{vmatrix} 2 & -4 & 1 \\ 1 & -5 & 3 \\ 1 & -1 & 1 \end{vmatrix}} = \frac{11}{-8}, \quad y = \frac{\begin{vmatrix} 2 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & -1 & 1 \end{vmatrix}}{\begin{vmatrix} 2 & -4 & 1 \\ 1 & -5 & 3 \\ 1 & -1 & 1 \end{vmatrix}} = \frac{9}{-8}, \quad z = \frac{\begin{vmatrix} 2 & -4 & 1 \\ 1 & -5 & 2 \\ 1 & -1 & -1 \end{vmatrix}}{\begin{vmatrix} 2 & -4 & 1 \\ 1 & -5 & 3 \\ 1 & -1 & 1 \end{vmatrix}} = \frac{6}{-8}.$$

□

22.2 Properties of Determinants

This section studies the algebraic and geometric properties of determinants which are useful in the calculation of determinants.

Theorem 22.15. *Let A , B and C be $n \times n$ matrices that differ only in a single row, say the r th row, and assume that the r th row of C can be obtained by adding corresponding entries in the r th rows of A and B . Then $\det(C) = \det(A) + \det(B)$.*

Example 22.16. Verify that $\begin{vmatrix} 1 & 7 & 5 \\ 2 & 0 & 3 \\ 1 & 5 & 6 \end{vmatrix} = \begin{vmatrix} 1 & 7 & 5 \\ 2 & 0 & 3 \\ 1 & 4 & 7 \end{vmatrix} + \begin{vmatrix} 1 & 7 & 5 \\ 2 & 0 & 3 \\ 0 & 1 & -1 \end{vmatrix}.$

Solution. $LHS = \begin{vmatrix} 1 & 7 & 5 \\ 2 & 0 & 3 \\ 1 & 5 & 6 \end{vmatrix} = -28$

$$RHS = \begin{vmatrix} 1 & 7 & 5 \\ 2 & 0 & 3 \\ 1 & 4 & 7 \end{vmatrix} + \begin{vmatrix} 1 & 7 & 5 \\ 2 & 0 & 3 \\ 0 & 1 & -1 \end{vmatrix} = -49 + 21 = -28$$

Therefore $LHS = RHS$.

□

The following theorem characterises the effect of elementary row (column) operations on the determinant of a matrix.

Theorem 22.17. *Let A be an $n \times n$ matrix. Suppose B is the matrix obtained from A by*

1. *multiplying a single row (column) by a scalar k , then $\det(B) = k \det(A)$.*
2. *interchanging two rows (columns) of A , then $\det(B) = -\det(A)$.*
3. *adding a multiple of a row (column) of A to another, then $\det(B) = \det(A)$.*

Corollary 22.18. *The determinant of any permutation matrix is ± 1 .*

Corollary 22.19. *Let A be an $n \times n$ matrix.*

1. *If A has 2 identical rows or columns, then $\det(A) = 0$.*
2. *If A has 2 proportional rows or columns, then $\det(A) = 0$.*
3. $\det(kA) = k^n \det(A)$.

Example 22.20. What is the relationship between $\det(A)$ and $\det(-A)$?

Solution. $\det(-A) = \det((-1)A) = (-1)^n \det(A)$. This is a special case of 3. of Corollary 22.19. □

Example 22.21. Find (a) $\begin{vmatrix} 3 & -1 & 4 & 5 \\ 6 & -2 & 5 & 2 \\ 5 & 8 & 1 & 4 \\ -9 & 3 & -12 & -15 \end{vmatrix}$ and (b) $\begin{vmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{vmatrix}$.

Solution. (a) Since row 4 is -3 times of row 1, so the determinant is zero.

(b) Swapping row 1 and row 4 follow by swapping row 2 and row 3, we obtain $\det(I_4)$, which is equal to 1. □

Example 22.22. Evaluate the following by elementary row operations.

1. $\begin{vmatrix} 1 & -1 & 4 \\ 2 & 1 & -1 \\ -2 & 0 & 3 \end{vmatrix} = \begin{vmatrix} 1 & -1 & 4 \\ 0 & 3 & -9 \\ 0 & -2 & 11 \end{vmatrix} = \begin{vmatrix} 1 & -1 & 4 \\ 0 & 3 & -9 \\ 0 & 0 & 5 \end{vmatrix} = 15$

2. $\begin{vmatrix} 1 & 0 & 3 & 1 \\ 2 & 2 & 6 & 0 \\ 4 & 0 & -3 & 1 \\ 4 & 1 & 12 & 1 \end{vmatrix} = \begin{vmatrix} 1 & 0 & 3 & 1 \\ 0 & 2 & 0 & -2 \\ 0 & 0 & -15 & -3 \\ 0 & 1 & 0 & -3 \end{vmatrix} = - \begin{vmatrix} 1 & 0 & 3 & 1 \\ 0 & 1 & 0 & -3 \\ 0 & 0 & -15 & -3 \\ 0 & 2 & 0 & -2 \end{vmatrix} = - \begin{vmatrix} 1 & 0 & 3 & 1 \\ 0 & 1 & 0 & -3 \\ 0 & 0 & -15 & -3 \\ 0 & 0 & 0 & 4 \end{vmatrix} = 60$

Example 22.23. Find the determinant of $A = \begin{bmatrix} -1 & 2 & 3 & 4 \\ 0 & -2 & 3 & 0 \\ 1 & 2 & 3 & 5 \\ 8 & 9 & -1 & 2 \end{bmatrix}$.

[Remark: A wise tactic may be to apply “elementary row operation” and then calculate the determinant using column expansion]

Solution.

$$\det A = \begin{vmatrix} -1 & 2 & 3 & 4 \\ 0 & -2 & 3 & 0 \\ 0 & 4 & 6 & 9 \\ 0 & 25 & 23 & 34 \end{vmatrix} = (-1)[-2(-3) - 3(-89)] = -273$$

□

Example 22.24. Suppose $A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$ is a 3×3 matrix with $\det(A) = 2$. Find the determinant of $B = \begin{bmatrix} -g & -h & -i \\ 3d + 5a & 3e + 5b & 3f + 5c \\ 7d & 7e & 7f \end{bmatrix}$.

Solution.

$$\begin{aligned}
\det(B) &= \begin{vmatrix} -g & -h & -i \\ 3d & 3e & 3f \\ 7d & 7e & 7f \end{vmatrix} + \begin{vmatrix} -g & -h & -i \\ 5a & 5b & 5c \\ 7d & 7e & 7f \end{vmatrix} \\
&= (-1)(3)(7) \begin{vmatrix} g & h & i \\ d & e & f \\ d & e & f \end{vmatrix} + (-1)(5)(7) \begin{vmatrix} g & h & i \\ a & b & c \\ d & e & f \end{vmatrix} \quad \square \\
&= (-1)(3)(7)(0) + (-1)(5)(7)(-1) \begin{vmatrix} a & b & c \\ g & h & i \\ d & e & f \end{vmatrix} = (-1)(5)(7)(-1)(-1) \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} \\
&= -35 \times 2 = -70
\end{aligned}$$

The following theorem is useful in the proof of some mathematical theorems and is stated here for reference.

Theorem 22.25. *Let A and B be $n \times n$ matrices. Then $\det(AB) = \det(A) \det(B)$.*

Corollary 22.26. *If $A \in M_n(\mathbb{R})$ is invertible, then $\det(A^{-1}) = \frac{1}{\det(A)}$.*

22.3 Geometric Applications of Determinants

In this section, we will study some geometric applications of determinants. In Chapter [20](#), we have defined the cross product of vectors $\mathbf{u} = (u_1, u_2, u_3)$ and $\mathbf{v} = (v_1, v_2, v_3)$ to be

$$\mathbf{u} \times \mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix}$$

We can define the triple scalar product using the dot product and the cross product as follows.

Let $\mathbf{u} = (u_1, u_2, u_3)$, $\mathbf{v} = (v_1, v_2, v_3)$ and $\mathbf{w} = (w_1, w_2, w_3)$ be vectors from $\mathbb{R}^3$. The *triple scalar product* is defined as

$$\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}) = \begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix}.$$

It is equal to the *sign volume of the parallelepiped characterised by $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$* . The positive and negative signs of the triple scalar product are related to the left- and right-handed system. We will take absolute value when we want to find the volume of a 3-D object.

eg:daddel

Example 22.27. Given vectors $\mathbf{u} = (3, 2, 1)$, $\mathbf{v} = (-1, 3, 0)$ and $\mathbf{w} = (2, 2, 5)$.

1. Find the volume of the parallelepiped constructed by $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$.
2. Find the volume of the tetrahedron constructed by the $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$.

Solution. 1. The volume is

$$\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}) = \begin{vmatrix} 3 & 2 & 1 \\ -1 & 3 & 0 \\ 2 & 2 & 5 \end{vmatrix} = 3(15) - 2(-5) + 1(-8) = 47 \text{ (cubic units)}$$

2. Since each parallelepiped consists of four tetrahedron of equal volume, the volume of

the tetrahedron is $\frac{1}{4}(\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w})) = \frac{47}{4}$.

□

Example 22.28. Show that the three vectors $(3, 4, -3)$, $(5, -1, 2)$, $(-11, 16, -17)$ lie on a plane.

Solution: The three vectors $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ lie in the same plane iff the volume of the parallelepiped is zero, i.e. $\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}) = 0$.

$$\text{Since } \mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}) = \begin{vmatrix} 3 & 4 & -3 \\ 5 & -1 & 2 \\ -11 & 16 & -17 \end{vmatrix} = 0, \text{ the three vectors lie on a plane.}$$

As a special case of 3-D geometric problem, the calculation of the *area of parallelogram* constructed by two 3-d vectors $\mathbf{u} = (u_1, u_2, u_3)$ and $\mathbf{v} = (v_1, v_2, v_3)$ is

$$\underbrace{\frac{\mathbf{u} \times \mathbf{v}}{\|\mathbf{u} \times \mathbf{v}\|}}_{\text{height} = 1} \cdot (\mathbf{u} \times \mathbf{v}) = \frac{\|\mathbf{u} \times \mathbf{v}\|^2}{\|\mathbf{u} \times \mathbf{v}\|} = \|\mathbf{u} \times \mathbf{v}\|.$$

Example 22.29. Given vectors $\mathbf{u} = (3, 2, 1)$ and $\mathbf{v} = (-1, 3, 0)$.

1. Find the area of the parallelogram constructed by $\mathbf{u}$ and $\mathbf{v}$.
2. Find the area of the triangle constructed by the $\mathbf{u}$ and $\mathbf{v}$.

Solution:

$$1. \left\| \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 3 & 2 & 1 \\ -1 & 3 & 0 \end{vmatrix} \right\| = \sqrt{(-3)^2 + (1)^2 + 11^2} = \sqrt{131}$$

$$2. \text{ Area of triangle} = \frac{1}{2} \left\| \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 3 & 2 & 1 \\ -1 & 3 & 0 \end{vmatrix} \right\| = \frac{\sqrt{131}}{2}$$

Many problems in linear algebra (and many other branches of science) boil down to solving a system of linear equations in a number of variables. This in turn means finding common solutions to some “polynomial” equations of degree 1 (hyperplanes). However, in many cases, we are faced with “nonlinear” system of polynomial equations in more than one variable. Geometrically, this means finding common points to some “surfaces”. Like the Gaussian elimination for linear systems, the elimination theory in general is about eliminating a number of unknowns from a system of polynomial equations in one or more variables to get an easier equivalent system.

One way of find common solutions to polynomial equations is to solve each equation separately and then compare all the solutions. This is not an efficient way especially if the goal is only to show the existence of a solution.

To learn the elimination theory, let us start by considering the following “simple” example.

Example 22.30. Let $f(x) = a_1x^2 + b_1x + c_1$, $g(x) = a_2x^2 + b_2x + c_2$. Consider a system of two quadratic equations in one variable x :

$$f(x) = 0, \quad g(x) = 0.$$

Find a necessary and sufficient condition for the existence of a common solution for the system.

Solution: If $f(x)$ and $g(x)$ have a common solution x_0 , then they must have a common linear factor, say, $L = x - x_0$. Dividing $f(x)$ and $g(x)$ with L gives:

$$\frac{f(x)}{L} = A_1x + B_1, \quad \frac{g(x)}{L} = -(A_2x + B_2).$$

The negative sign is chosen above to simplify some calculations below. Therefore,

$$\begin{aligned} L &= g(x)(A_1x + B_1) = -f(x)(A_2x + B_2) \\ \Rightarrow (a_1x^2 + b_1x + c_1)(A_2x + B_2) &+ (a_2x^2 + b_2x + c_2)(A_1x + B_1) \\ &= (a_1A_2 + a_2A_1)x^3 + (b_2A_1 + b_1A_2 + a_1B_2 + a_2B_1)x^2 \\ &+ (c_1A_2 + c_2A_1 + b_2B_1 + b_1B_2)x + (c_1B_2 + c_2B_1) = 0. \end{aligned}$$

This equation is only possible if the coefficients of x , x^2 , x^3 and the constant term are all equal to 0. This gives the following homogeneous system with variables A_1 , A_2 , B_1 , B_2 :

$$\begin{aligned} a_1A_2 + a_2A_1 &= 0 \\ b_2A_1 + b_1A_2 + a_1B_2 + a_2B_1 &= 0 \\ c_1A_2 + c_2A_1 + b_2B_1 + b_1B_2 &= 0 \\ c_1B_2 + c_2B_1 &= 0 \end{aligned}$$

In order for this system to have a nontrivial solution, its coefficient matrix must be singular:

$$\begin{vmatrix} a_1 & 0 & a_2 & 0 \\ b_1 & a_1 & b_2 & a_2 \\ c_1 & b_1 & c_2 & b_2 \\ 0 & c_1 & 0 & c_2 \end{vmatrix} = 0 \Rightarrow \begin{vmatrix} a_1 & b_1 & c_1 & 0 \\ 0 & a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 & 0 \\ 0 & a_2 & b_2 & c_2 \end{vmatrix} = 0 \quad (\because \det(A) = \det(A^T))$$

This is the condition for $f(x)$ and $g(x)$ to have a common solution.

Mathematicians have extracted the condition above to obtain the following definition and theorem.

Definition 22.4. Let $f(x) = a_mx^m + \cdots + a_1x + a_0$ and $g(x) = b_nx^n + \cdots + b_1x + b_0$ be two polynomials of degree m and n respectively such that $a_m \neq 0$ or $b_n \neq 0$. If $m \leq n$, we define the *resultant* of $f(x)$ and $g(x)$ to be the following determinant of a square matrix of size $m + n$:

$$\text{Res}(f(x), g(x)) = \begin{vmatrix} a_m & a_{m-1} & \cdots & a_1 & a_0 & 0 & \cdots & 0 \\ 0 & a_m & a_{m-1} & \cdots & a_1 & a_0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & a_m & a_{m-1} & \cdots & a_1 & a_0 \\ b_n & b_{n-1} & \cdots & b_1 & b_0 & 0 & \cdots & 0 \\ 0 & b_n & b_{n-1} & \cdots & b_1 & b_0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & b_n & b_{n-1} & \cdots & b_1 & b_0 \end{vmatrix}.$$

Theorem 22.31. Let $f(x) = a_mx^m + \cdots + a_1x + a_0$ and $g(x) = b_nx^n + \cdots + b_1x + b_0$ be two polynomials of degree m and n respectively such that $a_m \neq 0$ or $b_n \neq 0$. Then the system

$$f(x) = 0, \quad g(x) = 0$$

has a solution iff $\text{Res}(f(x), g(x)) = 0$.

One can use the above theorem to determine if a polynomial system in more than one variable has a solution. The trick is to look at the polynomials in the system as polynomials in one variable with coefficients polynomials in the other variables. The next example illustrates this idea.

Example 22.32. State the condition for

$$\begin{cases} 0 = f(x, y) = 9x^2 + 4y^2 - 18x + 16y - 11 \\ 0 = g(x, y) = x^2 + y^2 - 9 \end{cases}$$

to have a solution.

Solution: From

$$\text{Res}_y(f(x, y), g(x, y)) = \begin{vmatrix} 4 & 16 & 9x^2 - 18x - 11 & 0 \\ 0 & 4 & 16 & 9x^2 - 18x - 11 \\ 1 & 0 & x^2 - 9 & 0 \\ 0 & 1 & 0 & x^2 - 9 \end{vmatrix} = 0,$$

we obtain the condition for existence: $25x^4 - 180x^3 + 574x^2 - 900x + 625 = 0$.

Thus, the problem of solving two-variable equations is reduced to solving a polynomial equation in one variable x . Although the solution of this equation does not look easy, one can use a numerical approach to estimate the solutions. These solutions correspond to the x -coordinates of the intersection points.

With a little bit of algebra techniques, we can obtain the “discriminant” for a quadratic equation. The discriminant is used to determine if a polynomial has a unique solution.

Example 22.33. Find the discriminant of the quadratic equation $q(x) = ax^2 + bx + c = 0$, $a \neq 0$.

Solution: For $q(x)$ to have a unique solution, we learn from calculus that $q(x) = 0$ and $q'(x) = 2ax + b = 0$. Therefore,

$$\text{Res}(q(x), q'(x)) = \begin{vmatrix} a & b & c \\ 2a & b & 0 \\ 0 & 2a & b \end{vmatrix} = 0 \Rightarrow a(b^2) - b(2ab) + c(4a^2) = -a(b^2 - 4ac) = 0.$$

Since $a \neq 0$, therefore, when the discriminant $b^2 - 4ac$ is zero, $q(x) = 0$ has a unique solution.

22.4 Eigenvalues and Eigenvectors

sec:eigval

Eigenvalues have their greatest importance in *dynamic problems*, where the solution is changing with time — growing or decaying or oscillating. Eigenvalues and eigenvectors are very useful in the solution of ordinary differential equations in some engineering problems. In statistics, they are useful in the analysis of long term behaviour of a (discrete-time) Markov chain.

Definition 22.5. If A is an $n \times n$ matrix, a real number λ is called a *eigenvalue* of A if $A\mathbf{x} = \lambda\mathbf{x}$ for some *nonzero* vector $\mathbf{x}$. The nonzero vector $\mathbf{x}$ is called the *eigenvector* of A corresponding to λ . The set of all solutions to $A\mathbf{x} = \lambda\mathbf{x}$ is called the *eigenspace* of A corresponding to λ .

Example 22.34. Let $A = \begin{bmatrix} 1 & 4 \\ 2 & 3 \end{bmatrix}$. Find the eigenvector and eigenspace of A corresponding to the eigenvalue $\lambda = 5$.

Solution.

$$\begin{bmatrix} 1 & 4 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 5 \begin{bmatrix} x \\ y \end{bmatrix} \Rightarrow \begin{bmatrix} 1-5 & 4 \\ 2 & 3-5 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} c \\ c \end{bmatrix}$$

where $c \in \mathbb{R}$ is a parameter.

Taking $c = 1$, we obtain an eigenvector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

The eigenspace of A corresponding to $\lambda = 5$ is

$$\{\mathbf{x} \in \mathbb{R}^2 : A\mathbf{x} = 5\mathbf{x}\} = \left\{ \begin{bmatrix} c \\ c \end{bmatrix} : c \in \mathbb{R} \right\}$$

□

Theorem 22.35. *Eigenvectors correspond to lines which are left invariant under multiplication by A .*

Proof: Suppose $\mathbf{x}$ is an eigenvector of A corresponding to the eigenvalue λ . Then $A\mathbf{x} = \lambda\mathbf{x}$. Since for any value $c \in \mathbb{R}$, $A(c\mathbf{x}) = c(A\mathbf{x}) = c(\lambda\mathbf{x}) = \lambda(c\mathbf{x})$. Therefore, $c\mathbf{x}$ is an eigenvector as well. The set $\{c\mathbf{x} : A\mathbf{x} = \lambda\mathbf{x}\}$ is a line which is left invariant under multiplication by A .

Example 22.36. Show that the matrix $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ does not have real eigenvalue.

Solution. Suppose A has a real eigenvalue λ and $\begin{bmatrix} x \\ y \end{bmatrix}$ is the corresponding eigenvector (which must be non-zero by definition). Then

$$A\mathbf{x} = \lambda\mathbf{x} \Rightarrow A^2\mathbf{x} = A(\lambda\mathbf{x}) = \lambda(A\mathbf{x}) = \lambda^2\mathbf{x} \Rightarrow \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -x \\ -y \end{bmatrix} = \lambda^2 \begin{bmatrix} x \\ y \end{bmatrix}$$

This implies $(\lambda^2 + 1) \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

If λ is real, then $\lambda^2 + 1 > 0$ and $\begin{bmatrix} x \\ y \end{bmatrix}$ is a zero vector, which is a contradiction to the requirement that eigenvector is non-zero. □

Theorem 22.37. *If λ is an eigenvalue of A , then $\det(A - \lambda I) = 0$.*

Proof: If λ is an eigenvalue of A , then

$$A\mathbf{x} = \lambda\mathbf{x} \Rightarrow (A - \lambda I_n)\mathbf{x} = \mathbf{0}.$$

If $\mathbf{x}$ is non-zero, then $A - \lambda I_n$ must be singular and therefore $\det(A - \lambda I) = 0$.

Definition 22.6. The expression $\det(A - \lambda I)$ is a polynomial in the variable λ called the *characteristic polynomial* of A .

Theorem 22.38. The roots of a characteristic polynomial are the eigenvalues of A .

Remark 22.39. Not all matrices have real eigenvalues. For example, $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ has the characteristic polynomial

$$\det(A - \lambda I) = \begin{vmatrix} -\lambda & 1 \\ -1 & -\lambda \end{vmatrix} = \lambda^2 + 1$$

which does not have a real root.

eg:eig1

Example 22.40. Let $A = \begin{bmatrix} 1 & 4 \\ 2 & 3 \end{bmatrix}$. Find the characteristic polynomial of A . Hence, find the eigenvalues and the corresponding eigenspaces.

Solution. To find the eigenvalues, we solve $\begin{vmatrix} 1-\lambda & 4 \\ 2 & 3-\lambda \end{vmatrix} = \lambda^2 - 4\lambda + 3 - 8 = \lambda^2 - 4\lambda - 5 = (\lambda - 5)(\lambda + 1) = 0$.

The characteristic polynomial is $\lambda^2 - 4\lambda - 5$ and the eigenvalues are $\lambda = -1, 5$.

When $\lambda = 5$,

$$A\mathbf{x} = \lambda\mathbf{x} \Rightarrow \begin{bmatrix} -4 & 4 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} t$$

where $t \in \mathbb{R}$. The corresponding eigenspace is $\left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix} t : t \in \mathbb{R} \right\}$.

When $\lambda = -1$,

$$A\mathbf{x} = \lambda\mathbf{x} \Rightarrow \begin{bmatrix} 2 & 4 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \end{bmatrix} t$$

where $t \in \mathbb{R}$. The corresponding eigenspace is $\left\{ \begin{bmatrix} 2 \\ -1 \end{bmatrix} t : t \in \mathbb{R} \right\}$. □

eg:3.4.8

Example 22.41. Find all eigenvalues and the corresponding eigenspaces of $A = \begin{bmatrix} 5 & 8 & 16 \\ 4 & 1 & 8 \\ -4 & -4 & -11 \end{bmatrix}$.

Solution. To find the eigenvalues, we solve

$$\begin{vmatrix} 5-\lambda & 8 & 16 \\ 4 & 1-\lambda & 8 \\ -4 & -4 & -11-\lambda \end{vmatrix} = -\lambda^3 - 5\lambda^2 - 3\lambda + 9 = -(\lambda - 1)(\lambda + 3)^2 = 0.$$

Therefore, the eigenvalues are $\lambda = 1, -3$.

When $\lambda = 1$,

$$\begin{bmatrix} 5-1 & 8 & 16 \\ 4 & 1-1 & 8 \\ -4 & -4 & -11-1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Reducing the equation to r.e.f, we obtain

$$\begin{bmatrix} 4 & 8 & 16 \\ 0 & -8 & -8 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Solving it gives the corresponding eigenspace is $\left\{ \begin{bmatrix} -2t_1 \\ -t_1 \\ t_1 \end{bmatrix} : t_1 \in \mathbb{R} \right\}$

When $\lambda = -3$,

$$\begin{bmatrix} 5+3 & 8 & 16 \\ 4 & 1+3 & 8 \\ -4 & -4 & -11+3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Reducing the equation to r.e.f, we obtain

$$\begin{bmatrix} 1 & 1 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -s-2t_2 \\ s \\ t_2 \end{bmatrix}$$

where $s, t_2 \in \mathbb{R}$ are parameters. The corresponding eigenspace is $\left\{ \begin{bmatrix} -s-2t_2 \\ s \\ t_2 \end{bmatrix} : s, t_2 \in \mathbb{R} \right\}$. $\square$

The following two theorems characterise the relation of eigenvalues to determinant and trace.

thm:tr:prod

Theorem 22.42. *If $A \in M_n(\mathbb{C})$, the product of n eigenvalues equals the determinant of A .*

Example 22.43. Show that the equality stated in the above theorem is valid for $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$

Proof: It is easy to find $\det(A) = 0 - (-1) = 1$.

The eigenvalues can be found from

$$\det(A - \lambda I) = \lambda^2 + 1 = 0 \Rightarrow \lambda_1 = \sqrt{-1}, \lambda_2 = -\sqrt{-1}.$$

Therefore, $\lambda_1 \cdot \lambda_2 = -(\sqrt{-1})^2 = -(-1) = 1 = \det(A)$.

Remark 22.44. The above example demonstrates why complex numbers are important in advanced linear algebra.

Definition 22.7. If $A = (a_{ij}) \in M_n(\mathbb{C})$, the *trace* of A , denoted as $\text{tr}A$ (or $\text{tr}(A)$) is defined as the sum of main diagonal entries of A , i.e. $\text{tr}A = a_{11} + \cdots + a_{nn}$.

thm:tr:sum

Theorem 22.45. *If $A \in M_n(\mathbb{C})$, the sum of n eigenvalues equals the trace of A .*

Remark 22.46. We can check it for Example ^{eg:3.4.8} 22.41, $\lambda_1 + \lambda_2 + \lambda_3 = 1 + (-3) + (-3) = -5 = 5 + 1 + (-11)$.

S. Karlin and M. Taylor. A First Course in Stochastic Processes. Academic Press, 2nd edition, 1975.

A.M. Langville and C.D. Meyer. Google's PageRank and Beyond: The Science of Search Engine Rankings. Princeton University Press, 2006.

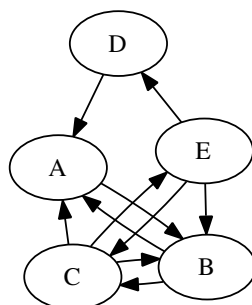
L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998.

S.M. Ross. Stochastic Processes. Wiley & Sons, 2nd edition, 1996. (A more advanced book than that of Karlin and Taylor.)

One of the most famous application of eigenvalue and eigenvector is Google's *pagerank algorithm*. It is described in the article "The \$25,000,000,000 Eigenvector The Linear Algebra Behind Google".

So, how does the PageRank algorithm work? To explain the algorithm, we model the web as an oriented graph. The *vertices* are *the pages*, and the *oriented edges* are the *links between pages*. Each page corresponds to a different url. Hence, a website may contain many pages. The model makes no difference between the individual pages of a website and its front page. But, most likely, the algorithm will rank better the front page of an important website.

Let us look at the simple web with five pages named A, B, C, D, and E.



This web has few links. If we are on page A, then there is only one link to page B, while, if we are on page C, we find three links, and we can choose to move to either page A, or B, or E. Notice that there is at least one link from each page.

If we walk randomly on the oriented graph, starting from a page, at each step we choose at random a link from the page where we are, and we follow it. For instance, if we start on page B, then we can go to A or to C with probability $1/2$ for each case while, if we start on D, then we necessarily go to A with probability 1. Where will we be after n steps? This can be characterised by the left transition matrix (this is the transpose of the right transition matrix mentioned in Chapter 21 and Probabilistic State Machine)

$$P = \begin{matrix} & \begin{matrix} A & B & C & D & E \end{matrix} \\ \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{3} & 1 & 0 \\ 1 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} & 0 & 0 \end{pmatrix} & \begin{matrix} A \\ B \\ C \\ D \\ E \end{matrix} \end{matrix} = (p_{ij}). \quad (22.1) \quad \text{e:pagerank}$$

If we consider a random variable X_n with values in the set of pages $\{A, B, C, D, E\}$, which contains $N = 5$ pages. X_n represents the page where we are after n steps of the random walk. Hence, p_{ij} , the entry of the matrix P located on the i -th row and j -th column, is the conditional probability that we are on the i -th page at step $n + 1$, given that we were on the j -th page at step n :

$$p_{ij} = \text{Prob}(X_{n+1} = i \mid X_n = j).$$

Note that the sum of entries of any column of P is equal to 1 and that all entries are greater or equal to zero. Matrices having these two properties are very special: each such matrix is the matrix of a *Markov chain process*, also called *Markov transition matrix*. It always has 1 as an

eigenvalue and there exists an eigenvector with eigenvalue 1, all the components of which are less than or equal to 1 and greater than or equal to 0, with sum equal to 1.

As mentioned in Discrete Mathematics, the entry $(p_{ij}^{(m)})$ of the matrix P^m describes the probability $\text{Prob}(X_{n+m} = i \mid X_n = j)$. When strange conditions do not happen, after n (which is sufficiently large) steps, the probability of being on a page is independent from where we started! In fact the eigenvector π corresponding to the eigenvalue $\lambda = 1$ is the *stationary vector* of P :

$$P\pi = \pi.$$

The following deep theorem of Frobenius guarantees that for the matrix associated to a web graph, we will always find a stationary solution.

Theorem 22.47 (Frobenius Theorem). *Consider an $N \times N$ Markov transition matrix $P = (p_{ij})$ (i.e. $p_{ij} \in [0, 1]$ for all i, j , and the sum of entries on each column is equal to 1, namely $\sum_{i=1}^N p_{ij} = 1$). Then*

1. $\lambda = 1$ is one eigenvalue of P .
2. Any eigenvalue λ of P satisfies $|\lambda| \leq 1$. There exists an eigenvector π of the eigenvalue 1, all the coordinates of which are greater or equal than zero. Without loss of generality we can suppose that the sum of its coordinates be 1.

It is now time to admire the power of this theorem. For this purpose, we will make the simplifying hypothesis that the matrix P has a basis of eigenvectors $\mathcal{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ and we suppose that $\mathbf{v}_1$ is the vector π of Frobenius theorem. For each $\mathbf{v}_i$, there exists λ_i such that $P\mathbf{v}_i = \lambda_i\mathbf{v}_i$. Let us take any nonzero vector X with $X^T = (x_1, \dots, x_N)$, where $x_i \in [0, 1]$ and $\sum_{i=1}^N x_i = 1$. We decompose X in the basis $\mathcal{V}$:

$$X = \sum_{i=1}^N a_i \mathbf{v}_i.$$

Mathematicians prove that $a_1 = 1$. Now, let us calculate PX :

$$PX = \sum_{i=1}^N a_i P\mathbf{v}_i = \sum_{i=1}^n a_i \lambda_i \mathbf{v}_i,$$

since $\mathbf{v}_i$ is an eigenvector of the eigenvalue λ_i . And if we iterate we get

$$P^n X = \sum_{i=1}^n a_i \lambda_i^n \mathbf{v}_i.$$

If all λ_i for $i > 1$ satisfy $|\lambda| < 1$, then

$$\lim_{n \rightarrow \infty} P^n X = a_1 \mathbf{v}_1 = \pi.$$

Notice however that the theorem does not guarantee that any matrix P satisfying the hypotheses of the theorem will have this property. Let us describe the possible pathologies and the remedy.

Possible pathologies:

1. The eigenvalue 1 may be a multiple root of the characteristic polynomial $\det(\lambda I - A) = 0$.
2. The matrix P may have other eigenvalues λ than 1, with modulus equal to 1. Here the eigenvalues can be complex numbers.

What do we do in this case? First, we define what a “good” matrix with no pathology.

Definition 22.8. A Markov transition matrix is *well-behaved* if

1. The eigenvalue 1 is a simple root of the characteristic polynomial $\det(\lambda I - A) = 0$.
2. All other eigenvalues λ of P than 1 have modulus less than 1.

Mathematicians discover that most matrices P are well-behaved. So, if the Markov transition matrix of a web is not well-behaved, then the strategy is to deform it slightly to become well-behaved. The **remedy**: consider the $N \times N$ matrix $Q = (q_{ij})$, with $q_{ij} = \frac{1}{N}$ for all i, j and replace the matrix P of the web by the matrix

$$P_\beta = (1 - \beta)P + \beta Q \quad (22.2) \quad \text{e:1}$$

for some small $\beta \in [0, 1]$. The value $\beta = 0.15$ has been used by Google.

Notice that the matrix P_β still has nonnegative entries and that the sum of the entries of each column is still equal to 1, so it is still a Markov transition matrix. The following theorem guarantees that there exists some small β for which we have fixed the pathologies.

Theorem 22.48. *Given any Markov transition matrix P , there always exists a positive β , as small as we wish, such that the matrix P_β is well-behaved. Let π be the eigenvector of 1 for the matrix P_β , normalized so that the sum of its coordinates be 1. For the matrix P_β , given any nonzero vector X , where $X^T = (p_1, \dots, p_N)$ with $p_i \in [0, 1]$ and $\sum_{i=1}^N p_i = 1$, then*

$$\lim_{n \rightarrow \infty} (P_\beta)^n X = \pi. \quad (22.3) \quad \text{e:2}$$

Example 22.49. Find the stationary vector π for the Markov transition matrix (41.38). le:pagerank

Solution:

$$P\pi = \pi \Rightarrow (P - I)\pi = \begin{bmatrix} -1 & \frac{1}{2} & \frac{1}{3} & 1 & 0 \\ 1 & -1 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & -1 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & -1 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} & 0 & -1 \end{bmatrix} \pi = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Solving it gives

$$\pi = \begin{bmatrix} 4 \\ 16/3 \\ 3 \\ 1/3 \\ 1 \end{bmatrix} t.$$

Pick $t = 3/41$, we have the stationary distribution (a unit vector): $\pi = \begin{bmatrix} 12/41 \\ 16/41 \\ 9/41 \\ 1/41 \\ 3/41 \end{bmatrix}$.

Remark 22.50. Finding the stationary distribution π , i.e. an eigenvector of the eigenvalue 1 for the matrix P_β in Equation (41.39), is not at all a trivial task when the matrix has billions of lines and columns: both the time of the computation and the memory space to achieve it represent real challenges. The usual method of Gauss elimination (the centre of attention of Chapter 21) is useless in this case, both because of the size of the computation, and because it requires to divide by small coefficients. A more efficient algorithm makes use of the property

^{e:2}(41.40) (see the paper by A. N. Langville and C. D. Meyer, “A Survey of Eigenvector Methods for Web Information Retrieval”, SIAM Review, Volume 47, Issue 1, (2005), pp. 135–161).

Another important application of eigenvalues and eigenvectors is in the study of long term behaviour of Markov chains. If X_{n+1} and X_n are two consecutive state vectors of a Markov chain with transition matrix P , then $X_{n+1} = PX_n$. What makes X_n approach a limiting vector as $n \rightarrow \infty$? Mathematicians discovered that a type of Markov chain define below has a limiting vector.

Definition 22.9. A transition matrix P is called *regular* if, for some integer r , all entries of P^r are strictly positive. Beware that 0 is not strictly positive.

A Markov chain process is called *regular* if its transition matrix is regular.

Example 22.51. Show that $A = \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0 & 1 \\ 0 & 0.5 & 0 \end{bmatrix}$ is regular.

Proof: Since $A^4 = \begin{bmatrix} 0.4375 & 0.3125 & 0.5 \\ 0.3125 & 0.625 & 0.125 \\ 0.25 & 0.0625 & 0.375 \end{bmatrix}$ has all positive entries, A is regular.

Theorem 22.52 (Main Theorem in Markov Chain Theory). 1. If P is a regular transition matrix, then as n approaches infinity, $P^n \rightarrow S$ where S is a matrix of the form $[\mathbf{v}, \mathbf{v}, \dots, \mathbf{v}]$ with $\mathbf{v}$ being a constant vector.

2. If P is a regular transition matrix of a Markov chain process, and if X is any state vector, then as n approaches infinity, $P^n X \rightarrow \mathbf{p}$, where $\mathbf{p}$ is a fixed probability vector (the sum of its entries is 1), all of whose entries are positive. This $\mathbf{p}$ is called the steady-state vector of the system.

Example 22.53. People who lived in Ottawa for a while are concerned with the weather. An unofficial study of the weather in the city in early spring yields the following observations:

- It is almost impossible to have two nice days in a row.
- If we have a nice day, we just as likely to have snow or rain the next day.
- If we have snow or rain, then we have an even chance to have the same the next day.
- If there is a change from snow or rain, only half of the time is this a change to a nice day.

1. Write the (left) transition matrix to model this system.
2. If it is nice today, what is the probability of being nice after one week?
3. If there is no sudden climate change (due to global warming or pollution), find the long time behaviour of the weather.

Solution:

1. Since the weather tomorrow depends only on today, this is a Markov chain process. The (left) transition matrix of this system is

$$\begin{matrix} & N & R & S \\ \begin{matrix} N \\ R \\ S \end{matrix} & \begin{pmatrix} 0 & 0.25 & 0.25 \\ 0.5 & 0.5 & 0.25 \\ 0.5 & 0.25 & 0.5 \end{pmatrix} \end{matrix}$$

where the letters N , R , S represent Nice, Rain, Snow respectively.

Note that this is NOT the way we learned in Discrete Mathematics (adjacency matrix in digraph) and it is NOT the way statistician represent transition matrix. Since we are writing vectors in column, therefore the matrix is a transpose of the “adjacency matrix”.

2. If it is nice today, then the initial state-vector is

$$X_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

After seven days (one week), the state-vector would be

$$X_7 = \begin{bmatrix} 0 & 0.25 & 0.25 \\ 0.5 & 0.5 & 0.25 \\ 0.5 & 0.25 & 0.5 \end{bmatrix}^7 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.19995117 \\ 0.40002441 \\ 0.40002441 \end{bmatrix}$$

So, there is about 20% chance of being nice in one week.

3. Notice first that we are dealing with a regular Markov chain since the transition matrix is regular, so we are sure that the steady-state vector exists. To find it we solve the homogeneous system

$$(P - I)X = 0 \Rightarrow \begin{bmatrix} -1 & 0.25 & 0.25 \\ 0.5 & -0.5 & 0.25 \\ 0.5 & 0.25 & -0.5 \end{bmatrix} X = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Reducing the coefficient matrix to r.r.e.f. gives

$$\begin{bmatrix} 1 & 0 & -0.5 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix}$$

The general solution of this system is $X = \begin{bmatrix} 0.5t \\ t \\ t \end{bmatrix}$.

So what solution do we choose? Remember that a steady-state vector is in particular a probability vector; that is the sum of its components is 1: $0.5t + t + t = 1$ gives $t = 0.4$. Thus, the steady-state vector is

$$X = \begin{bmatrix} 0.2 \\ 0.4 \\ 0.4 \end{bmatrix}.$$

In the long term, there is 20% chance of getting a nice day, 40% chance of having a rainy day and 40% chance of having a snowy day.

22.5 Similarity and Diagonalisation

The easiest type of matrix to calculate is the diagonal matrix. When we use the eigenvectors properly, a matrix can be transformed into a diagonal matrix. This “transformation” is called the similarity relation (to be defined below) and finding the diagonal matrix corresponding to a

given matrix is called “diagonalisation”.

Definition 22.10. Matrices A and B are *similar* if there exists an invertible matrix P such that $P^{-1}AP = B$.

Theorem 22.54. *Similar relation is an equivalence relation (reflexive, symmetric and transitive).*

Example 22.55. Let $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $B = \begin{bmatrix} -55 & -97 \\ 34 & 60 \end{bmatrix}$ and $P = \begin{bmatrix} 3 & 5 \\ 1 & 2 \end{bmatrix}$. Show that A is similar to B by verifying $P^{-1}AP = B$.

Solution. $P^{-1}AP = \begin{bmatrix} 2 & -5 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 3 & 5 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} -55 & -97 \\ 34 & 60 \end{bmatrix}$ □

In order to prove some properties of related to similarity, we introduce the following theorem.

Theorem 22.56. *If $A \in M_{m,n}(\mathbb{R})$ and $B \in M_{n,m}(\mathbb{R})$, then $\text{tr}(AB) = \text{tr}(BA)$.*

Theorem 22.57 (Properties of Similar Matrices). *Let A and B be $n \times n$ matrices such that A is similar to B .*

1. $\det(A) = \det(B)$.
2. A is invertible iff B is invertible.
3. A and B have the same characteristic polynomial.
4. A and B have the same eigenvalues.
5. $\text{tr}(A) = \text{tr}(B)$.

Proof:

1. Since $B = P^{-1}AP$,

$$\det(B) = \det(P^{-1}AP) = \det(P^{-1})\det(A)\det(P) = \frac{1}{\det(P)}\det(A)\det(P) = \det(A).$$

2. If A is invertible, $\det(A) \neq 0$ and $\det(B) = \det(A) \neq 0$, therefore B is invertible. The case is similar when B is invertible.
3. $\det(A - \lambda I) = \frac{1}{\det(P)}\det(A - \lambda I)\det(P) = \det(P^{-1})\det(A - \lambda I)\det(P) = \det(P^{-1}(A - \lambda I)P) = \det(P^{-1}AP - \lambda P^{-1}IP) = \det(B - \lambda I)$
4. Since the characteristic polynomials are the same for A and B , the zeroes of the characteristic polynomials, which are the eigenvalues are the same.
5. $\text{tr}(B) = \text{tr}(P^{-1}AP) = \text{tr}(P^{-1}PA) = \text{tr}(A)$.

Definition 22.11. A matrix A is *diagonalisable* or *semi-simple* if A is similar to a diagonal matrix.

Why is “diagonalisable” useful? Let us look at an example involving diagonal matrix.

Example 22.58. Evaluate $A = \begin{pmatrix} 2 & 0 \\ 0 & -5 \end{pmatrix} \begin{pmatrix} -17 & 0 \\ 0 & 8 \end{pmatrix}^{100}$.

Solution. $A = \begin{bmatrix} -34 & 0 \\ 0 & -40 \end{bmatrix}^{100} = \begin{bmatrix} 34^{100} & 0 \\ 0 & 40^{100} \end{bmatrix}$ □

From the above example, we learn that algebraic calculations involving diagonal matrix are relatively easy. We are interested to know that if there are ways to transform a general matrix into diagonal matrix? Mathematicians has discovered the following theorems.

Theorem 22.59. *A square $n \times n$ matrix is diagonalisable iff it has n linearly independent eigenvectors.*

thm:3.5.10

Theorem 22.60. *An $n \times n$ matrix with n different eigenvalues is diagonalisable.*

Proof: Assume the $n \times n$ matrix A has n eigenvalues $\lambda_1, \dots, \lambda_n$. If $\mathbf{v}_i$ and $\mathbf{v}_j$ are the eigenvectors corresponding to λ_i and λ_j respectively, then the eigenvectors $\mathbf{v}_i$ and $\mathbf{v}_j$ are linearly independent. Otherwise, $\mathbf{v}_i = k\mathbf{v}_j$ for some real number $k \neq 0$ and

$$A(\mathbf{v}_i - \mathbf{v}_j) = A\mathbf{v}_i - A\mathbf{v}_j = \lambda_i\mathbf{v}_i - \lambda_j\mathbf{v}_j \Rightarrow (k-1)A(\mathbf{v}_j) = (k-1)\lambda_j\mathbf{v}_j = (\lambda_i k - \lambda_j)\mathbf{v}_j$$

Hence, $(k-1)\lambda_j - \lambda_i k + \lambda_j = 0 \Rightarrow k\lambda_i - k\lambda_j = 0 \Rightarrow \lambda_i = \lambda_j$.

Theorem 22.61. *Let D be a diagonal matrix. $P^{-1}AP = D$ iff the columns of P are eigenvectors of A and the main diagonal entries of D are eigenvalues that correspond in order to the columns of P .*

Let us study the following example to learn how to find P and D and learn that *the diagonalising matrix P is not unique*.

Example 22.62. Let $A = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}$. Find the matrices P and D such that $P^{-1}AP = D$.

Solution:

$$\begin{vmatrix} 1-\lambda & 2 \\ 2 & 3-\lambda \end{vmatrix} = \lambda^2 - 4\lambda + 3 - 4 = 0 \Rightarrow \lambda = \frac{4 \pm \sqrt{16 - 4(-1)}}{2} = 2 \pm \sqrt{5}$$

For $\lambda_1 = 2 + \sqrt{5}$, the eigenvector $\mathbf{v}_1 = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$ can be found by solving

$$\begin{bmatrix} -1 - \sqrt{5} & 2 \\ 2 & 1 - \sqrt{5} \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Solving it gives

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{5}-1}{2}t \\ t \end{bmatrix}$$

where $t \neq 0$ is a parameter.

For $\lambda_2 = 2 - \sqrt{5}$, the eigenvector $\mathbf{v}_2 = \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}$ can be found by solving

$$\begin{bmatrix} -1 + \sqrt{5} & 2 \\ 2 & 1 + \sqrt{5} \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Solving it gives

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{5}+1}{2}s \\ s \end{bmatrix}$$

where $s \neq 0$ is a parameter.

Take $s = t = 2$, we obtain one possible solution

$$P = \begin{bmatrix} \sqrt{5}-1 & \sqrt{5}+1 \\ 2 & 2 \end{bmatrix}, \quad D = \begin{bmatrix} 2+\sqrt{5} & 0 \\ 0 & 2-\sqrt{5} \end{bmatrix}$$

Another possible solution is

$$P = \begin{bmatrix} \sqrt{5}+1 & \sqrt{5}-1 \\ 2 & 2 \end{bmatrix}, \quad D = \begin{bmatrix} 2-\sqrt{5} & 0 \\ 0 & 2+\sqrt{5} \end{bmatrix}$$

Example 22.63. Show that $A = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix}$ is diagonalisable.

Solution.

$$\begin{vmatrix} 1-\lambda & 2 \\ 1 & 1-\lambda \end{vmatrix} = \lambda^2 - 2\lambda + 1 - 2 = 0 \Rightarrow \lambda = \frac{2 \pm \sqrt{4 - 4(-1)}}{2} = 1 \pm \sqrt{2}$$

Since there are two different eigenvalues, by Theorem [22.60](#), the matrix is diagonalisable. $\square$

Example 22.64. Let $A = \begin{bmatrix} 5 & 8 & 16 \\ 4 & 1 & 8 \\ -4 & -4 & -11 \end{bmatrix}$. Find a diagonal matrix D which is similar to A .

Solution. From Example [22.41](#), we know that A has 3 linearly independent eigenvectors (take $t_1 = -1$, $(s, t_2) = (1, 0), (0, 1)$)

$$\begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix}, \quad \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}$$

Therefore, we can have a few different P and the corresponding D :

$$P = \begin{bmatrix} 2 & -1 & -2 \\ 1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & -3 \end{bmatrix}$$

or

$$P = \begin{bmatrix} -1 & -2 & 2 \\ 1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix}, \quad D = \begin{bmatrix} -3 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

or

$$P = \begin{bmatrix} -1 & 2 & -2 \\ 1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} -3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -3 \end{bmatrix}$$

□

When A is similar to a diagonal matrix Λ , then calculation of the power of A , i.e. A^k , can be simplified as calculated exactly as follows:

$$A^k = (P\Lambda P^{-1})(P\Lambda P^{-1}) \cdots (P\Lambda P^{-1}) = P \underbrace{\Lambda \cdots \Lambda}_k P^{-1} = P\Lambda^k P^{-1}$$

eg:wp:wp:p244

Example 22.65. Consider $A = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$, find a general formula for the entries of A^k where the integer $k \geq 0$.

Solution: It is easy to see that the eigenvalues of A is 1 and 2. We need to determine the eigenvectors:

$$\text{When } \lambda = 1, \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \mathbf{v}_1 = \mathbf{0} \Rightarrow \mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\text{When } \lambda = 2, \begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{v}_2 = \mathbf{0} \Rightarrow \mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\text{Therefore } A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} = P\Lambda P^{-1} \text{ and}$$

$$A^k = P\Lambda^k P^{-1} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2^k \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2^k - 1 \\ 0 & 2^k \end{bmatrix}.$$

eg:wikibook:la

Example 22.66. Given the matrix $A = \begin{bmatrix} 4 & -2 \\ 1 & 1 \end{bmatrix}$. Calculate $A^{1023} + 2^{489} A^{534}$ exactly.

$$\text{Solution. } \begin{vmatrix} 4 - \lambda & -2 \\ 1 & 1 - \lambda \end{vmatrix} = \lambda^2 - 5\lambda + 4 + 2 = (\lambda - 2)(\lambda - 3) = 0$$

$$\text{When } \lambda = 2: \begin{bmatrix} 2 & -2 \\ 1 & -1 \end{bmatrix} \mathbf{x} = \mathbf{0} \Rightarrow \mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\text{When } \lambda = 3: \begin{bmatrix} 1 & -2 \\ 1 & -2 \end{bmatrix} \mathbf{x} = \mathbf{0} \Rightarrow \mathbf{x} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Therefore

$$\begin{bmatrix} 4 & -2 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix}^{-1} = P\Lambda P^{-1}$$

and

$$\begin{aligned} A^{1023} + 2^{489} A^{534} &= \underbrace{P\Lambda P^{-1} \cdots P\Lambda P^{-1}}_{1023} + 2^{489} \underbrace{P\Lambda P^{-1} \cdots P\Lambda P^{-1}}_{534} \\ &= P\Lambda^{1023} P^{-1} + 2^{489} P\Lambda^{534} P^{-1} = P(\Lambda^{1023} + 2^{489} \Lambda^{534}) P^{-1} \\ &= \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 2^{1023} + 2^{489}(2^{534}) & 0 \\ 0 & 3^{1023} + 2^{489}(3^{534}) \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix}^{-1} \\ &= \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 2^{1024} & 0 \\ 0 & 3^{534}(3^{489} + 2^{489}) \end{bmatrix} \begin{bmatrix} -1 & 2 \\ 1 & -1 \end{bmatrix} \\ &= \begin{bmatrix} -2^{1024} + 2(3^{534})(3^{489} + 2^{489}) & 2^{1025} - 2(3^{534})(3^{489} + 2^{489}) \\ -2^{1024} + 3^{534}(3^{489} + 2^{489}) & 2^{1025} - 3^{534}(3^{489} + 2^{489}) \end{bmatrix} \end{aligned}$$

Remark 22.67. Not all matrices are diagonalisable in $\mathbb{R}$. For example, $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ cannot be diagonalised in $\mathbb{R}$ since the characteristic polynomial

$$\det(A - \lambda I) = \begin{vmatrix} -\lambda & -1 \\ 1 & \lambda \end{vmatrix} = \lambda^2 + 1$$

does not have real number solution.

In fact, things are far more complicated. The following example demonstrates without sufficient “eigenvectors”, we cannot diagonalise a matrix.

Example 22.68. Let $A = \begin{bmatrix} 2 & 1 & -12 \\ 0 & 1 & 11 \\ 1 & 1 & 4 \end{bmatrix}$. Find the eigenvalues and the corresponding eigenvectors. Determine whether A is diagonalisable.

Solution.

$$\begin{aligned} \begin{vmatrix} 2-\lambda & 1 & -12 \\ 0 & 1-\lambda & 11 \\ 1 & 1 & 4-\lambda \end{vmatrix} &= (1-\lambda) \begin{vmatrix} 2-\lambda & -12 \\ 1 & 4-\lambda \end{vmatrix} - 11 \begin{vmatrix} 2-\lambda & 1 \\ 1 & 1 \end{vmatrix} \\ &= (1-\lambda)(\lambda^2 - 6\lambda + 8 + 12) - 11(2 - \lambda - 1) \\ &= (1-\lambda)[\lambda^2 - 6\lambda + 20 - 11] = (1-\lambda)[\lambda - 3]^2 = 0 \end{aligned}$$

For the eigenvalue $\lambda = 1$, the eigenvector can be found by solving

$$\begin{bmatrix} 1 & 1 & -12 \\ 0 & 0 & 11 \\ 1 & 1 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} t, \quad t \in \mathbb{R}.$$

For the eigenvalue $\lambda = 3$, the eigenvector can be found by solving

$$\begin{bmatrix} -1 & 1 & -12 \\ 0 & -2 & 11 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -13/2 \\ 11/2 \\ 1 \end{bmatrix} s, \quad s \in \mathbb{R}.$$

The two eigenvalues just provide two eigenvectors, which insufficient to diagonalise A .

A cannot be diagonalised because A does not have sufficient eigenvectors! □

22.6 Jordan Canonical Form

TO WORK ON. References

<https://www.math.uci.edu/~ndonalds/math121b/3canonical.pdf>

<https://www.math.utah.edu/~gustafso/s2018/2280/chapter11-systems/Ch11-6-Jordan-Form-and-Eigenanalysis.pdf>

<https://iuuk.mff.cuni.cz/~rakdver/linalg/lesson15-8.pdf>

<https://math.stackexchange.com/questions/1304165/name-of-jordan-canonical-form-in-infinite-dimensions>

<https://math.stackexchange.com/questions/907376/is-there-an-infinite-dimensional-jordan-decomposition>

<https://mathoverflow.net/questions/207495/spectral-theorem-from-jordan-decomposition-in-infinite-dimensions>

ToeplitzAndCirculantMatricesReview
Applications of Jordan Canonical Form:
Techniques of diagonalisation to solve difference equations and system of differential equations
Application of diagonalization on Markov matrices

22.7 Working with Maxima

```
/*Example eg:3.1.2 22.7 */  
A : matrix([-1, 0, 2], [3, -7, 8], [-4, 4, -5]);  
adjoint(A);  
A . adjoint(A) = determinant(A)*ident(3);  
  
/*Example eg:inv2 22.12 */  
invert(matrix([-1, 0, 2], [3, -7, 8], [-4, 4, -5]));  
  
/*Example eg:eig1 22.40 */  
A : matrix([1, 4], [2, 3]);  
charpoly(A, lambda), expand;  
[evals, multiplicities] : eigenvalues(A);  
[evals, evecs] : eigenvectors(matrix([1, 4], [2, 3]));  
evecs[1]; /* or part(evecs, 1) for first eigenvector */  
evecs[2]; /* or part(evecs, 2) for second eigenvector */  
  
/*Example eg:3.4.8 22.41 */  
eigenvectors(matrix([5, 8, 16], [4, 1, 8], [-4, -4, -11]));
```

Chapter 23

Vector Spaces

ch:vec-sp

What is the benefit of introducing the abstract concept of “vector space”? For mathematicians, this abstraction allows them to study the algebraic properties of “vector space”. For physicists, “vector space” is a language for them to describe certain physical phenomena independent of coordinate system. For statisticians, it is a convenient language to understand how the ranks of matrices change.

23.1 The Theory of Linear Equations

Mathematicians discovered that abstract “vector space” is a good language to study of the theory of linear equations:

$$A\mathbf{x} = \mathbf{b}, \quad A \in M_{m,n}(\mathbb{R}). \quad (23.1) \quad \text{e:lineq}$$

The columns of $A\mathbf{x}$ can be viewed as linear combinations of n “column” vectors $\mathbf{a}_1, \dots, \mathbf{a}_n$ of A :

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \dots + x_n\mathbf{a}_n.$$

Instead of looking at individual columns, we look at a set of column vectors defined below.

def:colsp

Definition 23.1. The *column space* of a matrix A is the set of all linear combinations of the columns of A , i.e. $\{A\mathbf{x} : \mathbf{x} \in \mathbb{R}^n\}$. It is denoted as $\text{colsp}(A)$.

Now, Equation (23.1) ^{e:lineq} has a solution if $\mathbf{b}$ is in the set $\text{colsp}(A)$.

When $\mathbf{b} = \mathbf{0}$ in the linear equation (23.1) ^{e:lineq}, mathematicians discovered that the solution set is also a vector space, they called it the null space of A , it is formally defined below.

Definition 23.2. The *null space* of a matrix A is the set $\{\mathbf{x} : A\mathbf{x} = \mathbf{0}\}$. We denote it as $\text{nullsp}(A)$.

Remark 23.1. Since A is $m \times n$, the columns of A are vectors in $\mathbb{R}^m$, so column space of A is a collection of vectors in $\mathbb{R}^m$, where as the null space of A consists of vectors in $\mathbb{R}^n$.

Example 23.2. Find the column space and null space of $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

Solution: By definition,

$$\text{colsp}(A) = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} : x, y \in \mathbb{R} \right\} = \mathbb{R}^2.$$

and

$$\text{nullsp}(A) = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} : A \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\} = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}$$

Example 23.3. Suppose $A = \begin{bmatrix} 1 & 5 \\ 0 & 1 \end{bmatrix}$. Determine $\text{colsp}(A)$.

Solution: The columns of this matrix are nonparallel non-zero vectors in $\mathbb{R}^2$, so they span the xy -plane. The column space of the matrix is also $\mathbb{R}^2$.

Example 23.4. Identify geometrically the null space and the column space of $A = \begin{bmatrix} 1 & 3 \\ 2 & 6 \end{bmatrix}$.

Solution. Notice that the second row is twice the first row, they are linearly dependent.

$$\text{nullsp}(A) = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} : \begin{bmatrix} 1 & 3 \\ 2 & 6 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\} = \left\{ \begin{bmatrix} -3 \\ 1 \end{bmatrix} t : t \in \mathbb{R} \right\}.$$

It is a line perpendicular to the “vector” $\begin{bmatrix} 1 \\ 3 \end{bmatrix}$ (the first row).

Think of it as a dot product: $\begin{bmatrix} 1 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} -3 \\ 1 \end{bmatrix} t = \begin{bmatrix} 1 \\ 3 \end{bmatrix}^T \begin{bmatrix} -3 \\ 1 \end{bmatrix} t = 0$.

Notice that the second column is twice the first column, the column space is a line:

$$\text{colsp}(A) = \left\{ \begin{bmatrix} 1 \\ 2 \end{bmatrix} t : t \in \mathbb{R} \right\}$$

□

The above examples are square matrices. What happens when A is a general $m \times n$ matrix? Mathematicians have obtained the following theorems that characterise $\text{colsp}(A)$ and $\text{nullsp}(A)$.

Theorem 23.5. The column space of an $m \times n$ matrix A is the set of all $A\mathbf{x}$ as $\mathbf{x}$ ranges over all vectors in $\mathbb{R}^n$.

Theorem 23.6. If A is an $m \times n$ matrix and $m < n$, then A has a nontrivial null space; there exists a vector $\mathbf{x} \neq \mathbf{0}$ such that $A\mathbf{x} = \mathbf{0}$.

Example 23.7. Find the column space and null space of $A = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 3 \end{bmatrix}$.

Solution. By definition,

$$\text{colsp}(A) = \left\{ \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = x \begin{bmatrix} 1 \\ 0 \end{bmatrix} + y \begin{bmatrix} 0 \\ 1 \end{bmatrix} + z \begin{bmatrix} 2 \\ 3 \end{bmatrix} : x, y, z \in \mathbb{R} \right\} = \left\{ s \begin{bmatrix} 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} 0 \\ 1 \end{bmatrix} : s, t \in \mathbb{R} \right\}.$$

and

$$\text{nullsp}(A) = \left\{ \begin{bmatrix} x \\ y \\ z \end{bmatrix} : A \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\} = \left\{ \begin{bmatrix} -2t \\ -3t \\ t \end{bmatrix} : t \in \mathbb{R} \right\}$$

□

eg:4.1.10

Example 23.8. Let $A = \begin{bmatrix} 1 & -3 & -2 & 4 \\ 2 & 0 & 2 & 2 \\ 0 & 4 & 4 & -4 \end{bmatrix}$. Find the column space and null space of A .

Write down the basis and the dimension of the null space and column space of A . [Remark: The definition of basis and dimension will be discussed in Section 23.4]

Solution. The column space is

$$\text{colsp}(A) = \left\{ a \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} + b \begin{bmatrix} -3 \\ 0 \\ 4 \end{bmatrix} + c \begin{bmatrix} -2 \\ 2 \\ 4 \end{bmatrix} + d \begin{bmatrix} 4 \\ 2 \\ -4 \end{bmatrix} : a, b, c, d \in \mathbb{R} \right\}$$

To find the basis and dimension, we need the r.e.f. of A :

$$\begin{bmatrix} 1 & -3 & -2 & 4 \\ 0 & 6 & 6 & -6 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Therefore the basis of the column space is

$$\begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}, \begin{bmatrix} 0 \\ 3 \\ 1 \end{bmatrix}$$

and the dimension is 2.

The null space can be found by solving

$$\begin{bmatrix} 1 & -3 & -2 & 4 \\ 0 & 6 & 6 & -6 \\ 0 & 0 & 0 & 0 \end{bmatrix} \mathbf{x} = \mathbf{0} \Rightarrow \mathbf{x} = \begin{bmatrix} 3(-s+t) + 2s - 4t \\ -s+t \\ s \\ t \end{bmatrix} = s \begin{bmatrix} -1 \\ -1 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} -1 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

Therefore

$$\text{nullsp}(A) = \{ \mathbf{x} : A\mathbf{x} = \mathbf{0} \} = \left\{ s \begin{bmatrix} -1 \\ -1 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} -1 \\ 1 \\ 0 \\ 1 \end{bmatrix} : s, t \in \mathbb{R} \right\}$$

The basis is $\begin{bmatrix} -1 \\ -1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \\ 0 \\ 1 \end{bmatrix}$ and the dimension is 2. □

The column space provides a “geometrical language” to describe when a linear equation has a solution and when a solution does not exist. This is described by the following theorem.

Theorem 23.9. A vector $\mathbf{b}$ is in the column space of A iff the linear system $A\mathbf{x} = \mathbf{b}$ has a solution.

def:rank

Definition 23.3. The (*column*) rank (resp. *row rank*) of a matrix A (denoted as $\text{rank}(A)$) is the maximum number of linearly independent columns (resp. rows).

Mathematicians have discovered the following theorem characterising the column rank and row rank.

Theorem 23.10. For a matrix $A \in M_{m,n}(\mathbb{R})$,

column rank of A = row rank of A = number of leading entries in its r.e.f.

This is the theorem which we will constantly use to solve problems involving the computation of rank and finding spanning set of a column space.

Example 23.11. Find the rank of $A = \begin{bmatrix} 4 & 0 & 2 \\ 0 & 1 & 3 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{bmatrix}$.

Solution: Since A is in r.e.f., we only need to identify that there are two leading entries, therefore $\text{rank}(A) = 2$.

Transforming B into r.e.f.: $B \rightarrow \begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 0 \end{bmatrix}$
there is only one leading entry, therefore $\text{rank}(B) = 1$.

Example 23.12. Is $\mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}$ in the column space of $A = \begin{bmatrix} 1 & 0 & -2 & -1 \\ -1 & 3 & 5 & 4 \\ 2 & 1 & -3 & -1 \end{bmatrix}$? What is the column space of A ?

Solution. Consider the augmented matrix and its r.e.f.

$$\left[\begin{array}{cccc|c} 1 & 0 & -2 & -1 & 1 \\ -1 & 3 & 5 & 4 & 2 \\ 2 & 1 & -3 & -1 & 2 \end{array} \right] \rightarrow \left[\begin{array}{cccc|c} 1 & 0 & -2 & -1 & 1 \\ 0 & 3 & 3 & 3 & 3 \\ 0 & 0 & 0 & 0 & -1 \end{array} \right]$$

The last row indicates that $0 = -1$, therefore, $A\mathbf{x} = \mathbf{b}$ does not have a solution, by the above theorem, $\mathbf{b} \notin \text{colsp}(A)$.

The column space of A is

$$\text{colsp}(A) = \left\{ a \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix} + b \begin{bmatrix} 0 \\ 3 \\ 1 \end{bmatrix} : a, b \in \mathbb{R} \right\} = \text{span} \left\{ \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}, \begin{bmatrix} 0 \\ 3 \\ 1 \end{bmatrix} \right\}$$

□

eg: single-soln

Example 23.13. Let $A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 0 & 12 & 11 & 10 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \in \mathbb{R}^3$. Solve $A\mathbf{x} = \mathbf{b}$. Does a solution exist for any $\mathbf{b} \in \mathbb{R}^3$?

Solution. After reducing the augmented matrix to r.e.f.

$$[A|\mathbf{b}] \rightarrow \left[\begin{array}{cccc|c} 1 & 2 & 3 & 4 & b_1 \\ 0 & -4 & -8 & -12 & b_2 - 5b_1 \\ 0 & 12 & 11 & 10 & b_3 \end{array} \right] \rightarrow \left[\begin{array}{cccc|c} 1 & 2 & 3 & 4 & b_1 \\ 0 & -4 & -8 & -12 & b_2 - 5b_1 \\ 0 & 0 & -13 & -26 & b_3 + 3b_2 - 15b_1 \end{array} \right]$$

Since there are only 3 leading entries, there is a free parameter $x_4 = t$, $t \in \mathbb{R}$ and the solution is

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -(2b_3 - 7b_2 + 9b_1)/26 \\ (-8b_3 - 11b_2 + 55b_1 - 52t)/52 \\ (-b_3 - 3b_2 + 15b_1 - 26t)/13 \\ t \end{bmatrix}.$$

Yes, from the augmented matrix, we know that a solution exists for any $\mathbf{b} \in \mathbb{R}^3$. $\square$

eg:no-soln

Example 23.14. Let $C = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$. Solve $C\mathbf{x} = \mathbf{b}$ and determine if a solution exists for any $\mathbf{b} \in \mathbb{R}^3$?

Solution. After reducing the augmented matrix to r.e.f.

$$[C|\mathbf{b}] \rightarrow \left[\begin{array}{cc|c} 1 & 2 & b_1 \\ 0 & -2 & b_2 - 3b_1 \\ 0 & -4 & b_3 - 5b_1 \end{array} \right] \rightarrow \left[\begin{array}{cc|c} 1 & 2 & b_1 \\ 0 & -2 & b_2 - 3b_1 \\ 0 & 0 & b_3 - 2b_2 + b_1 \end{array} \right]$$

From the augmented matrix, we learn that:

- No. There is no solution when $\mathbf{b}$ is chosen such that $b_3 - 2b_2 + b_1 \neq 0$.
- A solution only exists when $b_3 - 2b_2 + b_1 = 0$, solving it leads to

$$\mathbf{x} = \begin{bmatrix} b_2 - 2b_1 \\ (3b_1 - b_2)/2 \end{bmatrix}$$

$\square$

From the Example [eg:single-soln](#) 23.13 and Example [eg:no-soln](#) 23.14, we learn that

- A is a 3×4 matrix, $\text{rank}(A) = 3 = m$, for any $\mathbf{b} \in \mathbb{R}^m$, $A\mathbf{x} = \mathbf{b}$ has a solution.
- C is a 3×2 matrix, $\text{rank}(C) = 2 = n$, for any $\mathbf{b} \in \mathbb{R}^n$, $C\mathbf{x} = \mathbf{b}$ has at most one solution.

Generalising the above result, we obtain the following theorem.

Theorem 23.15. Let A be an $m \times n$ matrix.

1. Let $\text{rank}(A) = m$. Then, for any vector $\mathbf{b} \in \mathbb{R}^m$, $A\mathbf{x} = \mathbf{b}$ has at least one solution.
2. Let $\text{rank}(A) = n$. Then, for any vector $\mathbf{b} \in \mathbb{R}^m$, $A\mathbf{x} = \mathbf{b}$ has at most one solution.
3. If $m = n$, the following statements are equivalent:
 - (a) $\text{rank}(A) = n$.
 - (b) For any vector $\mathbf{b} \in \mathbb{R}^m$, $A\mathbf{x} = \mathbf{b}$ has a unique solution.
 - (c) A is invertible.
 - (d) $\det(A) \neq 0$.

Corollary 23.16. If A is an $n \times n$ matrix, then the following statements are equivalent:

1. For any vector $\mathbf{b} \in \mathbb{R}^n$, $A\mathbf{x} = \mathbf{b}$ has a unique solution.
2. The homogeneous system $A\mathbf{x} = \mathbf{0}$ has just the trivial solution $\mathbf{x} = \mathbf{0}$.
3. $\text{rank}(A) = n$.

23.2 Vector Spaces and Subspaces

In the previous section, some abstract concepts such as vector space and basis are used. Here, we go further by abstracting algebraic objects that behave very similar to the vector space $\mathbb{R}^n$.

One famous example is the set of polynomials with coefficients from $\mathbb{R}$ can be viewed as a vector space over $\mathbb{R}$. If one is mathematically-inclined, one may want to get a taste of the application of vector space theory in algebraic number theory by reading the Karkheinz Spindler's "Abstract Algebra with Applications, Volume II: Rings and Fields", page 222, where an example is given to explain how one can find the basis of the "vector space" $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ over $\mathbb{Q}$.

The following is a definition of vector space in the language of first-order logic (Chapter 2). What mathematician did is just write down the abstract rules that make an algebraic object behaving just like $\mathbb{R}^n$.

ch:pred-logic

axiom:vec-sp

Definition 23.4 (Vector Space Axioms). A $\mathbb{R}$ -vector space V is a tuple with $(V, \mathbb{R}, +_V, \cdot, +_{\mathbb{R}}, \cdot_{\mathbb{R}})$ such that

$$+_V : V \times V \rightarrow V, \quad \cdot : \mathbb{R} \times V \rightarrow V$$

and the operations $+_V$ satisfies the following axioms:

1. $\forall \mathbf{x} \forall \mathbf{y} \forall \mathbf{z} (\mathbf{x} +_V (\mathbf{y} +_V \mathbf{z}) = (\mathbf{x} +_V \mathbf{y}) +_V \mathbf{z})$
2. $\forall \mathbf{x} \forall \mathbf{y} (\mathbf{x} +_V \mathbf{y} = \mathbf{y} +_V \mathbf{x})$
3. $\forall \mathbf{x} (\mathbf{x} +_V \mathbf{0} = \mathbf{x})$
4. $\forall \mathbf{x} (\mathbf{x} +_V (-\mathbf{x}) = \mathbf{0})$
5. $\forall \lambda \forall \mu \forall \mathbf{x} (\lambda \cdot (\mu \cdot \mathbf{x}) = (\lambda \cdot_{\mathbb{R}} \mu) \cdot \mathbf{x})$
6. $\forall \lambda \forall \mu \forall \mathbf{x} (\lambda \cdot \mathbf{x} +_V \mu \cdot \mathbf{x} = (\lambda +_{\mathbb{R}} \mu) \cdot \mathbf{x})$
7. $\forall \lambda \forall \mathbf{x} (\lambda \cdot (\mathbf{x} +_V \mathbf{y}) = \lambda \cdot \mathbf{x} +_V \lambda \cdot \mathbf{y})$
8. $\forall \mathbf{x} (1 \cdot \mathbf{x} = \mathbf{x})$

The elements of V are called *vectors*, the operation $+_V$ is called *vector addition* and is usually denoted as $+$. The operation $\cdot$ is called *scalar multiplication*. $\mathbf{0} \in V$ is called the *zero vector*. The vector $-\mathbf{x}$ is called the *negative* of the vector $\mathbf{x}$.

Example 23.17. Consider the set of polynomials of at most degree n , $P_n = \{a_0 + a_1x + \cdots + a_nx^n : a_0, a_1, \dots, a_n \in \mathbb{R}\}$. Show that P_n together with polynomial addition and scalar multiplication is a vector space over $\mathbb{R}$.

Proof: Let $V = P_n$. Since for any $a_0 + a_1x + \cdots + a_nx^n$, $b_0 + b_1x + \cdots + b_nx^n$ and $c_0 + c_1x + \cdots + c_nx^n \in V$ and any $\lambda, \mu \in \mathbb{R}$, we have

1. $a_0 + a_1x + \cdots + a_nx^n + (b_0 + b_1x + \cdots + b_nx^n + c_0 + c_1x + \cdots + c_nx^n) = a_0 + a_1x + \cdots + a_nx^n + (b_0 + c_0) + (b_1 + c_1)x + \cdots + (b_n + c_n)x^n = (a_0 + b_0 + c_0) + (a_1 + b_1 + c_1)x + \cdots + (a_n + b_n + c_n)x^n = ((a_0 + b_0) + (a_1 + b_1)x + \cdots + (a_n + b_n)x^n) + c_0 + c_1x + \cdots + c_nx^n = (a_0 + a_1x + \cdots + a_nx^n + b_0 + b_1x + \cdots + b_nx^n) + c_0 + c_1x + \cdots + c_nx^n$
2. $a_0 + a_1x + \cdots + a_nx^n + b_0 + b_1x + \cdots + b_nx^n = (a_0 + b_0) + (a_1 + b_1)x + \cdots + (a_n + b_n)x^n = (b_0 + a_0) + (b_1 + a_1)x + \cdots + (b_n + a_n)x^n = b_0 + b_1x + \cdots + b_nx^n + a_0 + a_1x + \cdots + a_nx^n$
3. $a_0 + a_1x + \cdots + a_nx^n + 0 = a_0 + a_1x + \cdots + a_nx^n$
4. $a_0 + a_1x + \cdots + a_nx^n + (-a_0) + (-a_1)x + \cdots + (-a_n)x^n = (a_0 - a_0) + (a_1 - a_1)x + \cdots + (a_n - a_n)x^n = 0 + 0x + \cdots + 0x^n = 0$
5. $\lambda(\mu(a_0 + a_1x + \cdots + a_nx^n)) = \lambda((\mu a_0) + (\mu a_1)x + \cdots + (\mu a_n)x^n) = (\lambda \cdot_{\mathbb{R}} \mu \cdot_{\mathbb{R}} a_0) + (\lambda \cdot_{\mathbb{R}} \mu a_1)x + \cdots + (\lambda \cdot_{\mathbb{R}} \mu \cdot_{\mathbb{R}} a_n)x^n = (\lambda \cdot_{\mathbb{R}} \mu)(a_0 + a_1x + \cdots + a_nx^n)$

6. $\lambda(a_0 + a_1x + \cdots + a_nx^n) + \mu(a_0 + a_1x + \cdots + a_nx^n) = (\lambda a_0 + \lambda a_1x + \cdots + \lambda a_nx^n) + (\mu a_0 + \mu a_1x + \cdots + \mu a_nx^n) = (\lambda a_0 + \mu a_0) + (\lambda a_1 + \mu a_1)x + \cdots + (\lambda a_n + \mu a_n)x^n = ((\lambda + \mu)a_0) + ((\lambda + \mu)a_1)x + \cdots + ((\lambda + \mu)a_n)x^n = (\lambda + \mu)(a_0 + a_1x + \cdots + a_nx^n)$
7. $\lambda(a_0 + a_1x + \cdots + a_nx^n + b_0 + b_1x + \cdots + b_nx^n) = \lambda((a_0 + b_0) + (a_1 + b_1)x + \cdots + (a_n + b_n)x^n) = (\lambda(a_0 + b_0)) + (\lambda(a_1 + b_1))x + \cdots + (\lambda(a_n + b_n))x^n = (\lambda a_0 + \lambda b_0) + (\lambda a_1 + \lambda b_1)x + \cdots + (\lambda a_n + \lambda b_n)x^n = (\lambda a_0 + \lambda a_1x + \cdots + \lambda a_nx^n) + (\lambda b_0 + \lambda b_1x + \cdots + \lambda b_nx^n) = \lambda(a_0 + a_1x + \cdots + a_nx^n) + \lambda(b_0 + b_1x + \cdots + b_nx^n)$
8. $1(a_0 + a_1x + \cdots + a_nx^n) = (1 \cdot_{\mathbb{R}} a_0) + (1 \cdot_{\mathbb{R}} a_1)x + \cdots + (1 \cdot_{\mathbb{R}} a_n)x^n = a_0 + a_1x + \cdots + a_nx^n$

As we can observe from the above proof that showing a given set with algebraic operations can be very long and tedious. We will hence ignored the proof for the following examples, which have been somehow explore in Chapter 20 and Chapter 21. [ch:linalg:vec] [ch:matrix-lin-sys]

Example 23.18. $(\mathbb{R}^n, +, \cdot, +_{\mathbb{R}}, \cdot_{\mathbb{R}})$ is a $\mathbb{R}$ -vector space under the standard operations of addition + and scalar multiplication $\cdot$ defined in Chapter 20. [ch:linalg:vec]

Example 23.19. The set of all $m \times n$ matrices $M_{m,n}(\mathbb{R})$ is a vector space under the addition and scalar multiplication defined in Chapter 21. [ch:matrix-lin-sys]

Note that “addition” and “scalar multiplication” are just algebraic operations. They can be defined in a different manner in which we are not familiar. In this case, we have to check and see if the set together with the new algebraic operations is still a vector space.

Example 23.20. Let $V = \mathbb{R}^2$ and define addition and scalar multiplication operations as follows: If $\mathbf{u} = (u_1, u_2)$ and $\mathbf{v} = (v_1, v_2)$, then define

$$\mathbf{u} + \mathbf{v} = (u_1 + v_1, u_2 + v_2)$$

and if $k \in \mathbb{R}$, then define

$$k\mathbf{u} = (ku_1, 0).$$

Is V a vector space under the stated operations?

Solution. No, since it violates axiom 8. in Definition 23.4: Let $\mathbf{u} = (1, 1)$, then [axiom:vec-sp]

$$LHS = 1(1, 1) = (1, 0), \quad RHS = (1, 1)$$

Since $(1, 0) \neq (1, 1)$, $LHS \neq RHS$ and therefore V not a vector space under the stated operations. □

Most of the time, we will be dealing with the subset of a vector space having vector space structure. Therefore, we define the concept of subspace as follows.

Definition 23.5. A subset W of a vector space V is called a *subspace* of V if W is itself a vector space under the addition and scalar multiplication defined on V .

If would be too tedious to verify all the 8 axioms in Definition 23.4 every time just to show that something is indeed a subspace. Mathematicians have discovered the following theorem, which reduces the verification process to just checking two conditions. [axiom:vec-sp]

Theorem 23.21 (Subspace Theorem). *If W is a subset of a vector space V , then W is a subspace of V iff the following conditions hold.*

(a) *If $\mathbf{u}$ and $\mathbf{v}$ are vectors in W , then $\mathbf{u} + \mathbf{v} \in W$.*

(b) *If k is any scalar and $\mathbf{u}$ is any vector in W , then $k\mathbf{u}$ is in W .*

Proof: Let $\mathbf{x}, \mathbf{y}, \mathbf{z}$ be any vectors in W and we check if W satisfies all the rules in the axioms:

1. By associativity of V ,

$$\mathbf{x} +_V (\mathbf{y} +_V \mathbf{z}) = (\mathbf{x} +_V \mathbf{y}) +_V \mathbf{z}.$$

By condition (a), $\mathbf{y} +_V \mathbf{z}, \mathbf{x} +_V \mathbf{y} \in W$. Again by condition (a), $\mathbf{x} +_V (\mathbf{y} +_V \mathbf{z}), (\mathbf{x} +_V \mathbf{y}) +_V \mathbf{z} \in W$.

2. By commutativity of V ,

$$\mathbf{x} +_V \mathbf{y} = \mathbf{y} +_V \mathbf{x}.$$

By condition (a), $\mathbf{x} +_V \mathbf{y}, \mathbf{y} +_V \mathbf{x} \in W$.

3. Take $\mathbf{x}$ as an element of W and $k = 0$, by condition (b), $0\mathbf{x} = \mathbf{0} \in W$ and $\forall \mathbf{x}(\mathbf{x} +_V \mathbf{0} = \mathbf{x})$.

4. Take $\mathbf{x}$ as an element of W and $k = -1$, condition (b) says that $(-1)\mathbf{x} \in W$. Therefore, by condition (a), $\mathbf{x} +_V (-1)\mathbf{x} = (1)\mathbf{x} +_V (-1)\mathbf{x} = (1 + (-1))\mathbf{x} = 0\mathbf{x} = \mathbf{0}$.

5. Take $\mathbf{x}$ as an element of W and $\lambda, \mu \in \mathbb{R}$. By condition (b), $\mu \cdot \mathbf{x} \in W$ and $(\lambda \cdot \mu) \cdot \mathbf{x} \in W$. By the scalar product associativity,

$$\lambda(\mu \cdot \mathbf{x}) = (\lambda \cdot \mu) \cdot \mathbf{x} \in W.$$

6. By condition (a) and scalar product scalar distributivity of V ,

$$W \ni \lambda \cdot \mathbf{x} +_V \mu \cdot \mathbf{x} = (\lambda + \mu) \cdot \mathbf{x}$$

7. By scalar product scalar distributivity of V ,

$$\lambda \cdot (\mathbf{x} +_V \mathbf{y}) = \lambda \cdot \mathbf{x} + \lambda \cdot \mathbf{y}$$

By condition (b), $\lambda \cdot \mathbf{x}, \lambda \cdot \mathbf{y} \in W$. Then, by condition (a), $\lambda \cdot \mathbf{x} + \lambda \cdot \mathbf{y} \in W$.

8. Identity property is inherited from V .

The subset W satisfies the vector space axiom and is hence a subspace of V .

We will practise how to apply Subspace Theorem to show that a subset is a subspace and we will also learn to apply the axioms in Definition 23.4 to show that a subset is not a subspace.

eg:4.2.3a

Example 23.22. Let $V = \mathbb{R}^3$ and $W = \{(x, y, z) | x + y - z = 0\}$. Show that W is a subspace of V .

Proof: Let $(x_1, y_1, z_1), (x_2, y_2, z_2) \in W$, by definition,

$$x_1 + y_1 - z_1 = 0, \quad x_2 + y_2 - z_2 = 0.$$

1. For $(x_1, y_1, z_1) + (x_2, y_2, z_2) = (x_1 + x_2, y_1 + y_2, z_1 + z_2)$,

$$(x_1 + x_2) + (y_1 + y_2) - (z_1 + z_2) = (x_1 + y_1 - z_1) + (x_2 + y_2 - z_2) = 0 + 0 = 0.$$

2. For $k(x_1, y_1, z_1) = (kx_1, ky_1, kz_1)$,

$$(kx_1) + (ky_1) - (kz_1) = k(x_1 + y_1 - z_1) = k0 = 0.$$

By Subspace Theorem, W is a subspace of $\mathbb{R}^3$.

eg:4.2.3b

Example 23.23. Let $V = M_2(\mathbb{R})$ and W be the set of all 2×2 diagonal matrices. Show that W is a subspace of V .

Solution. Let $\begin{bmatrix} a_1 & 0 \\ 0 & d_1 \end{bmatrix}, \begin{bmatrix} a_2 & 0 \\ 0 & d_2 \end{bmatrix} \in W$.

$$1. \begin{bmatrix} a_1 & 0 \\ 0 & d_1 \end{bmatrix} + \begin{bmatrix} a_2 & 0 \\ 0 & d_2 \end{bmatrix} = \begin{bmatrix} a_1 + a_2 & 0 \\ 0 & d_1 + d_2 \end{bmatrix}$$

is a diagonal matrix and therefore is an element of W .

2.

$$k \begin{bmatrix} a_1 & 0 \\ 0 & d_1 \end{bmatrix} = \begin{bmatrix} ka_1 & 0 \\ 0 & kd_1 \end{bmatrix}$$

is a diagonal matrix and therefore is an element of W .

By Subspace Theorem, W is a subspace of $M_2(\mathbb{R})$. □

exam:201201q1a

Example 23.24 (April 2012 Final Exam Q1(a)(i)). Show that the subset $W = \{A \in M_3(\mathbb{R}) : A^T = -A\}$ is a subspace of $M_3(\mathbb{R})$.

Solution. For any $A, B \in W$ and any $k \in \mathbb{R}$, $A^T = -A$ and $B^T = -B$ by definition.

- $(A + B)^T = A^T + B^T = -A - B = -(A + B)$. By definition, $A + B \in W$.
- $(kA)^T = k(A^T) = k(-A) = -(kA)$. By definition, $kA \in W$.

By Subspace Theorem, W is a subspace of $M_4(\mathbb{R})$. □

eg:4.2.3c

Example 23.25. Let $V = \mathbb{R}^2$ and $W = \{(x, y) | x = 1 \text{ and } y \in \mathbb{R}\}$. Is W a subspace of V ?

Solution. Since $0 \neq 1$, $(0, 0) \notin W$, W is not a vector space and is therefore not a subspace of $\mathbb{R}^2$. □

eg:4.2.3d

Example 23.26. Let $V = \mathbb{R}^3$ and $W = \{(x, y, z) | x < y < z\}$. Is W a subspace of V ?

Solution. Since $0 < 0 < 0$ is false, $(0, 0, 0) \notin W$ and W is not a vector space, therefore W is not a subspace of $\mathbb{R}^3$. □

def:span

Definition 23.6. Let V be a vector space. If $S = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ is a **subset** of V , then the subspace

$$W = \{c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + \dots + c_n\mathbf{u}_n | c_1, c_2, \dots, c_n \in \mathbb{R}\}$$

of V is called the *space spanned* by S (or by $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$) and we write

$$W = \text{span}(S) \text{ or } W = \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}.$$

We also say that S *spans* W or W is a *spanning set* (of S).

When S is an empty set (which is denoted as $\emptyset$ or $\{\}$), we define the space spanned by an empty set as $\text{span}(\emptyset) = \{\mathbf{0}\}$.

Example 23.27. The standard basis vectors $S = \{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ span $\mathbb{R}^n$, i.e. $\text{span}\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\} = \mathbb{R}^n$.

Example 23.28. Let $V = M_2(\mathbb{R})$ and $S = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\}$. Every element in $\text{span}(S)$ is of the form

$$a \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + b \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}.$$

where a, b are any real numbers. So,

$$\text{span}(S) = \left\{ \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \mid a, b \in \mathbb{R} \right\},$$

which is the set of all 2×2 diagonal matrices.

Example 23.29. The span of a single vector $\mathbf{v}$ is $\{a\mathbf{v} \mid a \in \mathbb{R}\}$ is just the set of all scalar multiples of $\mathbf{v}$, a line passes through the origin.

Example 23.30. The plane with equation $2x - y + 3z = 0$ is spanned by $\mathbf{u} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} 0 \\ 3 \\ 1 \end{bmatrix}$.

In fact any two nonparallel vectors in this plane span it.

Theorem 23.31. *The span W of any set of vectors $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ is a subspace.*

Proof: By definition, $W = \{a_1\mathbf{u}_1 + \dots + a_n\mathbf{u}_n : a_1, \dots, a_n \in \mathbb{R}\}$. For any $\mathbf{x} = a_1\mathbf{u}_1 + \dots + a_n\mathbf{u}_n$, $\mathbf{y} = b_1\mathbf{u}_1 + \dots + b_n\mathbf{u}_n$,

$$1. \mathbf{x} + \mathbf{y} = a_1\mathbf{u}_1 + \dots + a_n\mathbf{u}_n + b_1\mathbf{u}_1 + \dots + b_n\mathbf{u}_n = (a_1 + b_1)\mathbf{u}_1 + \dots + (a_n + b_n)\mathbf{u}_n$$

Since $a_i + b_i \in \mathbb{R}$, $(a_1 + b_1)\mathbf{u}_1 + \dots + (a_n + b_n)\mathbf{u}_n \in W$.

$$2. k\mathbf{x} = k(a_1\mathbf{u}_1 + \dots + a_n\mathbf{u}_n) = (ka_1)\mathbf{u}_1 + \dots + (ka_n)\mathbf{u}_n$$

Since $ka_i \in \mathbb{R}$, $(ka_1)\mathbf{u}_1 + \dots + (ka_n)\mathbf{u}_n \in W$.

With the help of Subspace Theorem, mathematicians have proven that the column space, row space and the null space of a matrix have vector space structures. The row space is defined similar to column space (Definition 23.1) as follows.

def:rowsp

Definition 23.7. The *row space* of a matrix A is the set of all linear combinations of the rows of A , i.e. $\{\mathbf{x}^T A : \mathbf{x} \in \mathbb{R}^m\}$. It is denoted as $\text{rowsp}(A)$.

thm:vs:space

Theorem 23.32. *Given $A \in M_{m,n}(\mathbb{R})$. Then*

1. $\text{colsp}(A)$ is subspace of $\mathbb{R}^m$.

2. $\text{rowsp}(A)$ is subspace of $\mathbb{R}^n$.

3. $\text{nullsp}(A)$ is subspace of $\mathbb{R}^n$.

In Chapter 21, we found that in order to find the solution of Equation (23.1), we will solve the following equation:

$$U\mathbf{x} = \mathbf{c}, \quad \text{ref}([A|\mathbf{b}]) = [U|\mathbf{c}]. \quad (23.2)$$

e:lineq-ref

And, in Example 23.8, in order to find the column space and the null space, we have also solved Equation 23.2. What is the theoretical foundation for this to work???

Mathematicians discovered that this has something to do with the row equivalence between A and $\text{ref}(A)$, here $\text{ref}(A)$ is the row echelon form of A .

Definition 23.8. Two matrices $A, B \in M_{m,n}(\mathbb{R})$ are said to be *row equivalent* if there is a sequence of elementary matrices $E_1, \dots, E_k$ such that

$$B = E_1 \cdots E_k A.$$

From Chapter 21, we know that $\text{ref}(A) = E_1 \cdots E_k A$, therefore, A and $\text{ref}(A)$ are row equivalent.

In general, it is discovered that two row equivalent matrices has the same row space.

Theorem 23.33. *If A and B are row equivalent matrices, then the row space of A is the row space of B . In other words, elementary row operations preserve the row space of a matrix.*

Example 23.34. Suppose $A = \begin{bmatrix} 1 & 3 & 1 & -1 & 1 \\ -1 & -1 & 3 & 1 & 3 \\ 1 & 4 & 3 & -1 & 3 \end{bmatrix}$. Reduce matrix A to r.e.f. and hence verify that the row space of $\text{ref}(A)$ is the row space of A . Can we use two vectors to span $\text{rowsp}(A)$?

Solution. $\text{rowsp}(A) = \text{span}\{(1, 3, 1, -1, 1), (-1, -1, 3, 1, 3), (1, 4, 3, -1, 3)\}$.

$$\text{ref}(A) = \begin{bmatrix} 1 & 3 & 1 & -1 & 1 \\ 0 & 2 & 4 & 0 & 4 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$\text{rowsp}(\text{ref}(A)) = \text{span}\{(1, 3, 1, -1, 1) + (0, 2, 4, 0, 4)\}$.

Since $(-1, -1, 3, 1, 3) = -(1, 3, 1, -1, -1) + (0, 2, 4, 0, 4)$ and $(1, 4, 3, -1, 3) = (1, 3, 1, -1, -1) + \frac{1}{2}(0, 2, 4, 0, 4)$,

$$\begin{aligned} \text{rowsp}(A) &= \{a(1, 3, 1, -1, 1) + b(-1, -1, 3, 1, 3) + c(1, 4, 3, -1, 3) : a, b, c \in \mathbb{R}\} \\ &= \{a(1, 3, 1, -1, 1) + b[-(1, 3, 1, -1, -1) + (0, 2, 4, 0, 4)] \\ &\quad + c[(1, 3, 1, -1, -1) + \frac{1}{2}(0, 2, 4, 0, 4)] : a, b, c \in \mathbb{R}\} \\ &= \{(a - b + c)(1, 3, 1, -1, 1) + (b + \frac{c}{2})(0, 2, 4, 0, 4) : a, b, c \in \mathbb{R}\} \\ &= \text{span}\{(1, 3, 1, -1, 1), (0, 2, 4, 0, 4)\} = \text{rowsp}(\text{ref}(A)). \end{aligned}$$

□

23.3 Linear Independence

Let us further formalise the definitions of linear independence and dependence defined in Chapter 21 as follows.

Definition 23.9. A (multi-)set of vectors $\{\mathbf{x}_i : i \in I\}$ are called *linearly independent* if

$$\forall i \left(i \in I \rightarrow \left(\sum_{i \in I} c_i \mathbf{x}_i = \mathbf{0} \rightarrow c_i = 0 \right) \right). \quad (*)$$

$\{\mathbf{x}_i : i \in I\}$ are called *linearly dependent* if

$$\exists i \left(i \in I \wedge \left(\sum_{i \in I} c_i \mathbf{x}_i = \mathbf{0} \wedge c_i \neq 0 \right) \right).$$

Remark 23.35. If you feel that the above definition is too abstract, refer back to Definition 21.20. The definition tells that **an empty set $\emptyset$ is linearly independent!!!** This is because $i \in \emptyset$ is always false and by predicate logic, the proposition $(*)$ is true.

Mathematicians introduced the above abstract definition just to allow the notion of linear independence to be available for infinite dimensional vector spaces. For practical calculations,

all verification of linear independence is the same as what we have done in Chapter 21, as demonstrated by the following examples. [ch:matrix-lin-sys](#)

Example 23.36. Determine whether $\mathbf{v}_1 = (1, 2, 5)$, $\mathbf{v}_2 = (1, 0, 0)$, $\mathbf{v}_3 = (-2, 1, 3)$ are linearly independent.

Solution: Consider $a, b, c \in \mathbb{R}$ and

$$a(1, 2, 5) + b(1, 0, 0) + c(-2, 1, 3) = (0, 0, 0) \Rightarrow \begin{cases} a + b - 2c = 0 \\ 2a + c = 0 \\ 5a + 3c = 0 \end{cases}$$

Simplifying the augmented matrix as r.e.f.:

$$\left[\begin{array}{ccc|c} 1 & 1 & -2 & 0 \\ 0 & -2 & 5 & 0 \\ 0 & -5 & 13 & 0 \end{array} \right] \rightarrow \left[\begin{array}{ccc|c} 1 & 1 & -2 & 0 \\ 0 & -2 & 5 & 0 \\ 0 & 0 & 1/2 & 0 \end{array} \right]$$

Since there are 3 leading entries, the only solution is $a = b = c = 0$. By definition, $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ are linearly independent.

Example 23.37. Show that $(0, 3, 4)$, $(1, 1, 2)$, $(3, 0, 2)$ are linearly dependent.

Proof: For

$$a(0, 3, 4) + b(1, 1, 2) + c(3, 0, 2) = (0, 0, 0)$$

we can find non-zero values $a = 1$, $b = -3$ and $c = 1$ such that

$$(0, 3, 4) - 3(1, 1, 2) + (3, 0, 2) = (0, 0, 0)$$

By definition, the three vectors are linearly dependent.

Example 23.38. Are $\mathbf{v}_1 = (1, 2, 5)$, $\mathbf{v}_2 = (1, 0, 0)$, $\mathbf{v}_3 = (-2, 1, 3)$ and $\mathbf{v}_4 = (1, 1, 1)$ linearly independent?

Solution. Consider $a, b, c, d \in \mathbb{R}$ and the linear equation

$$a(1, 2, 5) + b(1, 0, 0) + c(-2, 1, 3) + d(1, 1, 1) = (0, 0, 0) \Rightarrow \begin{cases} a + b - 2c + d = 0 \\ 2a + c + d = 0 \\ 5a + 3c + d = 0 \end{cases} \quad (*)$$

Simplifying the augmented matrix as r.e.f.:

$$\left[\begin{array}{cccc|c} 1 & 1 & -2 & 1 & 0 \\ 0 & -2 & 5 & -1 & 0 \\ 0 & -5 & 13 & -4 & 0 \end{array} \right] \rightarrow \left[\begin{array}{cccc|c} 1 & 1 & -2 & 1 & 0 \\ 0 & -2 & 5 & -1 & 0 \\ 0 & 0 & 1/2 & -3/2 & 0 \end{array} \right]$$

There are 3 leading entries but 4 variables, hence, one variable is a free parameter. This means that we are able to find a non-zero real number solution (e.g. take $t = 1$) for the above linear equation (*):

$$d = t = 1, \quad c = 3t = 3, \quad b = 7t = 7, \quad a = -2t = -2.$$

By definition, $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4$ are linearly dependent. □

The above example informs us that 4 vectors in $\mathbb{R}^3$ are linearly dependent. In fact, we can generalise this as follows.

Theorem 23.39. *If $n > m$, any n vectors in $\mathbb{R}^m$ are linearly dependent.*

Theorem 23.40. *Vectors are linearly dependent iff one of them is a linear combination of the others.*

Proof:

$\Rightarrow$: If the vectors $\mathbf{u}_1, \dots, \mathbf{u}_n$ is linear dependent, then there are $c_i \neq 0$ such that $c_1\mathbf{u}_1 + \dots + c_n\mathbf{u}_n = \mathbf{0}$. Therefore

$$\mathbf{u}_i = (c_1\mathbf{u}_1 + \dots + c_{i-1}\mathbf{u}_{i-1} + c_{i+1}\mathbf{u}_{i+1} + \dots + c_n\mathbf{u}_n)/c_i.$$

$\Leftarrow$: If one of them is a linear combination of the others, then

$$\mathbf{u}_i = a_1\mathbf{u}_1 + \dots + a_{i-1}\mathbf{u}_{i-1} + a_{i+1}\mathbf{u}_{i+1} + \dots + a_n\mathbf{u}_n.$$

This implies

$$\mathbf{0} = a_1\mathbf{u}_1 + \dots + a_{i-1}\mathbf{u}_{i-1} + \mathbf{u}_i + a_{i+1}\mathbf{u}_{i+1} + \dots + a_n\mathbf{u}_n.$$

Since the coefficient of $\mathbf{u}_i$ is 1, by definition, $\mathbf{u}_1, \dots, \mathbf{u}_n$ are linearly dependent.

When we want to determine whether a given set of (column) vectors $\mathbf{v}_1, \dots, \mathbf{v}_n$ are linearly independent or not, we always end up solving the homogeneous equation

$$c_1\mathbf{v}_1 + \dots + c_n\mathbf{v}_n = \mathbf{0}.$$

In solving this problem, we always reduce the coefficient matrix $A = [\mathbf{v}_1, \dots, \mathbf{v}_n]$ to the row echelon form and try to obtain a general solution for $c_1, \dots, c_n$. However, mathematicians discover that it is possible to avoid finding the general solution by just extracting information from the row echelon form as follows.

thm:4.3.8

Theorem 23.41. *Let U be a matrix in r.e.f.*

1. *The columns that contain leading entries of U are linearly independent.*
2. *The nonzero rows of U are linearly independent.*

Example 23.42. Given the matrix $U = \begin{bmatrix} \boxed{3} & -1 & 4 & 1 & 2 & 8 \\ 0 & \boxed{-2} & 0 & 5 & 7 & 5 \\ 0 & 0 & 0 & \boxed{8} & 3 & -2 \\ 0 & 0 & 0 & 0 & 0 & \boxed{9} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$, determine column vectors

which are linearly independent.

Solution: Since U is in r.e.f., by the above theorem, the leading entries, 3, -2, 8, 9 are in column 1, 2, 4, 6, therefore, the columns 1, 2, 4, 6 are linearly independent.

What is so nice and special about linear independence? Well, if a vector $\mathbf{x} = a_1\mathbf{v}_1 + \dots + a_n\mathbf{v}_n$ and $\mathbf{v}_1, \dots, \mathbf{v}_n$ are linearly independent, then the coefficients $a_1, \dots, a_n$ is unique as stated below.

Theorem 23.43. *If $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ are linearly independent vectors in vectors space V , then any vector in their span can be expressed in exactly one way as a linear combination of these vectors; the coefficient are unique.*

23.4 Basis and Dimension, Change of Basis

sec:basis

Let us investigate the property of spanning set (recall Definition [23.6](#)) and linear independence (Section [23.3](#)) for a given vector space $\mathbb{R}^m$. How many elements should there be if the spanning set spans $\mathbb{R}^m$ and how many elements should there be so that a set is linear independent in $\mathbb{R}^m$?

Theorem 23.44. Any linearly independent set of vectors in $\mathbb{R}^m$ contains **at most** m vectors. Any spanning set for $\mathbb{R}^m$ contains **at least** m vectors.

Example 23.45. Let $\mathbf{u}_1 = (1, 2, 3, 4)$, $\mathbf{u}_2 = (5, 6, 7, 8)$, $\mathbf{u}_3 = (9, 10, 11, 12)$, $\mathbf{u}_4 = (13, 14, 15, 16)$. Do $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4$ span $\mathbb{R}^4$?

Solution.

$$\begin{bmatrix} 1 & 5 & 9 & 13 \\ 2 & 6 & 10 & 14 \\ 3 & 7 & 11 & 15 \\ 4 & 8 & 12 & 16 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 5 & 9 & 13 \\ 0 & -4 & -8 & -12 \\ 0 & -8 & -16 & -24 \\ 0 & -12 & -24 & -36 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 5 & 9 & 13 \\ 0 & -4 & -8 & -12 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

According to the r.e.f, there are only two independent vectors. (Theorem [23.41](#))
No. They span only a two-dimensional subspace in $\mathbb{R}^4$. □

By combining the notions of spanning set and linear independence, we obtain the notion of basis.

Definition 23.10. Let $S = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ be a subset of a vector space V . Then S is called a *basis* for V if

1. S spans V and
2. S is linearly independent.

Before proceeding to general case, we will study the basis for a zero vector space.

Example 23.46. The basis of a zero vector space $\{\mathbf{0}\}$ is an empty set $\emptyset$.

Reason:

1. By Definition [23.6](#), $\text{span}(\emptyset) = \{\mathbf{0}\}$.
2. By Remark [23.35](#), $\emptyset$ is linearly independent.

Note that since a zero vector is always linearly dependent by definition, hence a basis cannot contain a zero vector. A basis for a vector space V contains the smallest possible number of vectors that can span V .

Example 23.47. Show that

- (1) the standard basis vectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n \in \mathbb{R}^n$ form a basis of $\mathbb{R}^n$;
- (2) the standard basis matrices $E_{11}, \dots, E_{mn} \in M_{m,n}(\mathbb{R})$ form a basis of $M_{m,n}(\mathbb{R})$;
- (3) $\{1, t, \dots, t^n\}$ is a basis for P_n .

Proof: For (1), We need to show that

1. $\mathbf{e}_1, \dots, \mathbf{e}_n$ spans $\mathbb{R}^n$. This can be shown from the equation below, where for every vector $(x_1, \dots, x_n) \in \mathbb{R}^n$,

$$(x_1, \dots, x_n) = x_1 \mathbf{e}_1 + \dots + x_n \mathbf{e}_n$$

This means that $(x_1, \dots, x_n)$ can be written as a linear combination of $\mathbf{e}_1, \dots, \mathbf{e}_n$.

2. $\mathbf{e}_1, \dots, \mathbf{e}_n$ are linearly independent because

$$a_1\mathbf{e}_1 + \dots + a_n\mathbf{e}_n = (a_1, \dots, a_n) = (0, \dots, 0) \Rightarrow a_1 = \dots = a_n = 0.$$

There are no non-zero values for a_i . By definition, $\mathbf{e}_1, \dots, \mathbf{e}_n$ are linearly independent.

In general, a vector space has many different bases. The following example investigates how to show that a given set of vectors are the basis of a vector space.

Example 23.48. Show that $\{(1, 1), (1, -1)\}$ is a basis for $\mathbb{R}^2$.

Proof. We need to show that

1. $(1, 1), (1, -1)$ spans $\mathbb{R}^2$. This because for every vector $(x_1, x_2) \in \mathbb{R}^2$,

$$(x_1, x_2) = \left(\frac{x_1 + x_2}{2}\right)(1, 1) + \left(\frac{x_1 - x_2}{2}\right)(1, -1).$$

This means that (x_1, x_2) can be written as a linear combination of $(1, 1), (1, -1)$.

2. $(1, 1), (1, -1)$ are linearly independent because

$$a_1(1, 1) + a_2(1, -1) = (0, 0) \Rightarrow a_1 = a_2 = 0.$$

There are no non-zero values for a_i . By definition, $(1, 1), (1, -1)$ are linearly independent.

□

How do we show that a given set of vectors is *not* a basis? Well, by definition, we just need to show that either the given set of vectors are *not a spanning set* or *not linearly independent*.

Example 23.49. Is $S = \{(1, 1, 1, 1), (0, 0, 1, 2), (-1, 0, 0, 1)\}$ a basis for $\mathbb{R}^4$?

Solution. There are only 3 vectors in S , but $\mathbb{R}^4$ is 4-dimensional, there are vectors in $\mathbb{R}^4$ which cannot be written as a linear combination of vectors in S : For example, consider $(1, 0, 0, 0)$, if it can be written as a linear combination of vectors in S , then there are $a, b, c \in \mathbb{R}$ such that

$$(1, 0, 0, 0) = a(1, 1, 1, 1) + b(0, 0, 1, 2) + c(-1, 0, 0, 1)$$

$$\text{This can be simplified to the equations } \begin{cases} 1 = a - c \Rightarrow 1 = 0 - 0 \text{ inconsistent} \\ 0 = a \\ 0 = a + b \Rightarrow b = 0 \\ 0 = a + 2b + c \Rightarrow c = 0 \end{cases}$$

Therefore, S is not a basis of $\mathbb{R}^4$.

□

Although the basis of a vector space is not unique, there is a concept unique to the given vector space, which is the dimension.

Definition 23.11. A nonzero vector space V is called *finite dimensional* if it has a basis with finite number of vectors. If V does not have a finite basis, then V is called *infinite dimensional*.

Definition 23.12. The *dimension* of a finite dimensional vector space V , denoted by $\dim(V)$, is defined to be the number of vectors in a basis for V . In addition, the dimension of the zero vector space is defined to be zero.

Let $\#S$ be the number of elements in the set S .

Example 23.50. Show that $\dim(\{\mathbf{0}\}) = 0$.

Proof: Since the basis is empty, there is no element in it, and therefore $\dim(\{\mathbf{0}\}) = \#\{\} = 0$.

Example 23.51. Show that $\dim(\mathbb{R}^n) = n$, $\dim(M_{m,n}(\mathbb{R})) = mn$, $\dim(P_n) = n + 1$.

Proof: Since $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ is the basis of $\mathbb{R}^n$, therefore, $\dim(\mathbb{R}^n) = \#\{\mathbf{e}_1, \dots, \mathbf{e}_n\} = n$.

Example 23.52. Find a basis for and determine the dimension of the subspace $W = \{(x, y, z) \mid x, y, z \in \mathbb{R} \text{ and } y = z\}$ of $\mathbb{R}^3$.

Solution. W has a basis $\{(1, 1, 1), (1, -1, -1)\}$ and $\dim W = 2$. □

Example 23.53. Find a basis of the plane $2x - y + z = 0$ and determine its dimension.

Solution. A basis of the plane is $(1, 2, 0), (0, 1, 1)$. Therefore, its dimension is 2. □

Example 23.54. Find a basis for and determine the dimension of the subspace

$$W = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \mid a, b, c, d \in \mathbb{R} \text{ and } b + c = 0 \right\},$$

of $M_{2 \times 2}$.

Solution. From the constraint $b + c = 0$, we get $c = -b$, therefore elements of W has the form

$$\begin{bmatrix} a & b \\ -b & d \end{bmatrix} = a \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + b \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} + d \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

The above relation says that element of W can be expressed by the “basis” below (it is not difficult to show that they are linearly independent):

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Therefore $\dim W = 3$. □

The dimensions of the three basic vector spaces associated with a given matrix A mentioned in Theorem 23.32 are known to relate to (column) rank (rank $A = \dim \text{colsp}(A)$) and row rank (row rank $A = \dim \text{rowsp}(A)$) defined in Definition 23.3 and the nullity defined below.

Definition 23.13. The *nullity* of a matrix A is the dimension of its null space.

Example 23.55. Find the row rank and the nullity

$$A = \begin{bmatrix} 1 & 2 & -1 & 1 & 3 \\ -3 & -6 & 4 & -13 & -17 \\ 4 & 8 & -6 & 25 & 29 \\ -1 & -2 & 0 & 10 & 6 \end{bmatrix}.$$

Solution. Note that:

- row rank = max. number of independent rows = dimension of the row space of A .
- nullity = dimension of null space of A .

Reducing the matrix A to r.e.f.:

$$A \rightarrow \begin{bmatrix} 1 & 2 & -1 & 1 & 3 \\ 0 & 0 & 1 & -10 & -8 \\ 0 & 0 & -2 & 21 & 17 \\ 0 & 0 & -1 & 11 & 9 \end{bmatrix} \rightarrow \begin{bmatrix} \boxed{1} & 2 & -1 & 1 & 3 \\ 0 & 0 & \boxed{1} & -10 & -8 \\ 0 & 0 & 0 & \boxed{1} & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Since there are only 3 leading entries, there are only 3 independent row vectors, therefore

$$\dim \text{rowsp}(A) = 3.$$

Solving $A\mathbf{x} = \mathbf{0}$, we obtain the solution set, which is the null space of A :

$$\text{nullsp}(A) = \left\{ \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} s + \begin{bmatrix} -4 \\ 0 \\ -2 \\ -1 \\ 1 \end{bmatrix} t : s, t \in \mathbb{R} \right\}.$$

Therefore, $\dim \text{nullsp}(A) = 2$. □

Mathematicians have found the relation between the rank and nullity as follow.

thm:fundamental

Theorem 23.56. For any $A \in M_{m,n}(\mathbb{R})$,

$$\begin{aligned} \text{rank}(A) + \text{nullity}(A) &= n; \\ \text{rank}(A^T) + \text{nullity}(A^T) &= m. \end{aligned}$$

Note that m and n are numbers of rows and columns of A respectively.

Remark 23.57. This is a very important theorem in the eye of mathematicians because it is the “geometric” version of the following principle that we have been applying in Chapter 21 to solve Equation (23.1): ch:matrix-lin-sys

$$\underbrace{n \text{ variables}}_n - \underbrace{\text{number of leading entries in ref}(A)}_{\text{rank}(A)} = \underbrace{\text{number of parameters}}_{\text{nullity}(A)}$$

Example 23.58. Show that the identity $\text{rank}(B) + \text{nullity}(B) = n$ is true for $B = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$.

Proof: B has 3 columns, therefore $n = 3$. Reducing B to r.e.f., we have

$$B \rightarrow \begin{bmatrix} \boxed{1} & 2 & 3 \\ 0 & \boxed{-3} & -6 \\ 0 & 0 & 0 \end{bmatrix}$$

There are two leading entries, therefore $\text{rank}(B) = 2$.

The null space of B is found to be

$$\text{nullsp}(B) = \left\{ \begin{bmatrix} t \\ -2t \\ t \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} t : t \in \mathbb{R} \right\}.$$

Therefore $\text{nullity}(B) = 1$.
The identity is true since $2 + 1 = 3$.

Example 23.59. Show that the identity $\text{rank}(A) + \text{nullity}(A) = n$ is valid for

$$A = \begin{bmatrix} -3 & 4 & 5 & 8 & 1 \\ 1 & 0 & -2 & 1 & -1 \\ -2 & 4 & 3 & 9 & 0 \\ 4 & -4 & -7 & -7 & 3 \end{bmatrix}.$$

Solution. A has 5 columns, therefore $n = 5$. Exchange the first and the second rows and then reduce A to r.e.f., we have

$$A \rightarrow \begin{bmatrix} 1 & 0 & -2 & 1 & -1 \\ 0 & 4 & -1 & 11 & -2 \\ 0 & 4 & -1 & 11 & -2 \\ 0 & -4 & 1 & -11 & 7 \end{bmatrix} \rightarrow \begin{bmatrix} \boxed{1} & 0 & -2 & 1 & -1 \\ 0 & \boxed{4} & -1 & 11 & -2 \\ 0 & 0 & 0 & 0 & \boxed{5} \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Therefore $\text{rank}(A) = 3$.

$$\text{The null space is } \text{nullsp}(A) = \left\{ \begin{bmatrix} 2s - t \\ (s - 11t)/4 \\ s \\ t \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 1/4 \\ 1 \\ 0 \\ 0 \end{bmatrix} s + \begin{bmatrix} -1 \\ -11/4 \\ 0 \\ 1 \\ 0 \end{bmatrix} t : s, t \in \mathbb{R} \right\}$$

Therefore $\text{nullity}(A) = 2$ and the identity is valid since $3 + 2 = 5$. $\square$

When we are given a basis, we can “represent” the a vector in terms of the basis. In order to make this precise, we introduce the following definition.

def:coord-basis

Definition 23.14. Let $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ be a basis for a vector space V and $\mathbf{v}$ be a vector in V . If

$$\mathbf{v} = c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + \dots + c_n\mathbf{u}_n,$$

then the coefficients $c_1, c_2, \dots, c_n$ are called the *coordinates* of $\mathbf{v}$ relative to the basis $\mathcal{U}$. The vector

$$(\mathbf{v})_{\mathcal{U}} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}$$

in $\mathbb{R}^n$ is called the *coordinate vector* of $\mathbf{v}$ relative to the basis $\mathcal{U}$.

Remark 23.60. Note that some books use the “square brackets” notation $[\mathbf{v}]_{\mathcal{U}}$. Sometimes, we write the vector in the row form just to save some space, writing vectors in column form is wasting space as can be observed from the above.

Example 23.61. Let $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2\}$ and $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2\}$, where $\mathbf{u}_1 = (-1, 2)$, $\mathbf{u}_2 = (2, -1)$, $\mathbf{v}_1 = (1, 0)$, $\mathbf{v}_2 = (1, 1)$. Find $(\mathbf{x})_{\mathcal{V}}$, given $(\mathbf{x})_{\mathcal{U}} = (1, 3)$.

Solution. Given $(\mathbf{x})_{\mathcal{U}} = (1, 3)$. By Definition ^{def:coord-basis} 23.14, this means

$$\mathbf{x} = 1\mathbf{u}_1 + 3\mathbf{u}_2 = 1(-1, 2) + 3(2, -1) = (5, -1).$$

Let $(\mathbf{x})_{\mathcal{V}} = (a, b)$. By Definition [23.14](#), ^{def:coord-basis}

$$\mathbf{x} = a(1, 0) + b(1, 1) = (a + b, b).$$

Since $\mathbf{x} = \mathbf{x}$, $(a + b, b) = (5, -1)$. Solving it gives $a = 6$, $b = -1$. $\therefore (\mathbf{x})_{\mathcal{V}} = (6, -1)$. $\square$

Example 23.62. Let $\mathcal{V} = \{(1, 1, 0), (0, 1, 1), (1, 0, 1)\}$. Find the coordinate vector of $\mathbf{v} = (1, 2, -1)$ w.r.t $\mathcal{V}$.

Solution. We are given $\mathbf{v} = (1, 2, -1)$ and are asked to find the unknown coordinate vector $(\mathbf{v})_{\mathcal{V}} = (a, b, c)$. By definition,

$$\mathbf{v} = a(1, 1, 0) + b(0, 1, 1) + c(1, 0, 1) = (1, 2, -1).$$

From here, we can obtain a simultaneous equation. Solving the equation, we obtain $a = 2$, $b = 0$ and $c = -1$. Therefore,

$$(\mathbf{v})_{\mathcal{V}} = (2, 0, -1).$$

$\square$

[Holtman-Hill:p253](#) **Example 23.63.** Consider $\mathbf{v} = 5t - 2 \in P_1$. If $\mathcal{U} = \{t + 1, t - 1\}$. Find $(\mathbf{v})_{\mathcal{U}}$.

Solution. Let $(\mathbf{v})_{\mathcal{U}} = \begin{bmatrix} a \\ b \end{bmatrix}$. Then

$$\mathbf{v} = a(t + 1) + b(t - 1) \Rightarrow 5t - 2 = (a + b)t + (a - b) \Rightarrow a = \frac{3}{2}, b = \frac{7}{2} \Rightarrow (\mathbf{v})_{\mathcal{U}} = \begin{bmatrix} 3/2 \\ 7/2 \end{bmatrix}$$

$\square$

Mathematicians have studied how vector algebra behaves under a given basis. This is stated as follows.

[thm:basis1](#)

Theorem 23.64. Let $\mathcal{U}$ be a basis of a vector space V , then for all $\mathbf{x}, \mathbf{y} \in V$ and for any $k \in \mathbb{R}$,

- $(\mathbf{x} + \mathbf{y})_{\mathcal{U}} = (\mathbf{x})_{\mathcal{U}} + (\mathbf{y})_{\mathcal{U}}$.
- $(k\mathbf{x})_{\mathcal{U}} = k(\mathbf{x})_{\mathcal{U}}$.

We will come across the calculation of coordinate vector in Chapter [24](#) and we will make use of the above theorem for the computation of linear transformation. ^{ch:lin-transf}

Chapter 24

Linear Transformations

ch:lin-transf

In the previous topic, we have just studied a single vector space V and various objects in it, such as subspaces and bases. Here, we are going to study a special relation between two vector spaces. A linear transformation is a special type of function / mapping which transforms vectors “linearly”. It has applications in computer graphics, ordinary differential equations (therefore it is important in engineering). However, linear transformation is also used in statistics to perform suitable scaling of data to extract useful statistical information.

24.1 Linear Transformations

sec:lin-trans

What is the most fundamental function between two vector spaces? Mathematicians discovered that the most fundamental function should be the one that preserves two most important properties of vector spaces — vector addition and scalar multiplication. With this in mind, the introduce the following concept.

linear-transform

Definition 24.1. A function T from a vector space V to another vector space W , i.e. $T : V \rightarrow W$, is *linear* or is called a *linear transformation* (or *homomorphism*) if T satisfies two conditions:

1. $T(\mathbf{u} + \mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v})$ and
2. $T(c\mathbf{u}) = cT(\mathbf{u})$

for all vectors $\mathbf{u}, \mathbf{v}$ in V and all scalar c . The set of linear transformations from V to W is denoted by $\text{Hom}(V, W)$.

Remark 24.1. For all $S, T \in \text{Hom}(V, W)$ and $\lambda \in \mathbb{R}$, if we define $S + T$ and λT as $(S + T)(\mathbf{v}) = S(\mathbf{v}) + T(\mathbf{v})$, $(\lambda T)(\mathbf{v}) = \lambda T(\mathbf{v})$, for all $\mathbf{v} \in V$, then $\text{Hom}(V, W)$ is a vector space.

Example 24.2. Show that the function $T : \mathbb{R} \rightarrow \mathbb{R}$ defined by $T(x) = kx$, where the constant $k \in \mathbb{R}$, is a linear transformation.

Proof: To show that T is a linear transformation, we just need to verify that it satisfies the two condition given in Definition 24.1: For any x and y in $V = \mathbb{R}$ and for any $c \in \mathbb{R}$, we have

1. $T(x + y) = k(x + y) = kx + ky = T(x) + T(y)$;
2. $T(cx) = k(cx) = c(kx) = cT(x)$.

Therefore T is a linear transformation.

Example 24.3. Let $\mathbf{x} \in \mathbb{R}^2$ be a fixed (constant) vector. Show that the function $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ defined by $T(\mathbf{v}) = \mathbf{v} \cdot \mathbf{x}$ is linear.

Proof: Consider $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$ and $c \in \mathbb{R}$, since

- $T(\mathbf{u} + \mathbf{v}) = (\mathbf{u} + \mathbf{v}) \cdot \mathbf{x} = \mathbf{u} \cdot \mathbf{x} + \mathbf{v} \cdot \mathbf{x} = T(\mathbf{u}) + T(\mathbf{v})$; and
- $T(c\mathbf{u}) = (c\mathbf{u}) \cdot \mathbf{x} = c(\mathbf{u} \cdot \mathbf{x}) = cT(\mathbf{u})$,

T is a linear transformation.

eg:5.1.4

Example 24.4. Show that the function $T : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ defined by $T\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right) = \begin{bmatrix} x_1 - x_3 \\ x_1 + 2x_2 \end{bmatrix}$ is a linear transformation.

Solution. Consider $\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}, \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} \in \mathbb{R}^3$ and $c \in \mathbb{R}$, since

$$\begin{aligned} \bullet \quad T\left(\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} + \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}\right) &= T\left(\begin{bmatrix} u_1 + v_1 \\ u_2 + v_2 \\ u_3 + v_3 \end{bmatrix}\right) = \begin{bmatrix} (u_1 + v_1) - (u_3 + v_3) \\ (u_1 + v_1) + 2(u_2 + v_2) \end{bmatrix} \\ &= \begin{bmatrix} u_1 - u_3 \\ u_1 + 2u_2 \end{bmatrix} + \begin{bmatrix} v_1 - v_3 \\ v_1 + 2v_2 \end{bmatrix} = T\left(\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}\right) + T\left(\begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}\right) \\ \bullet \quad T\left(c \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}\right) &= T\left(\begin{bmatrix} cu_1 \\ cu_2 \\ cu_3 \end{bmatrix}\right) = \begin{bmatrix} cu_1 - cu_3 \\ cu_1 + 2cu_2 \end{bmatrix} = c \begin{bmatrix} u_1 - u_3 \\ u_1 + 2u_2 \end{bmatrix} = cT\left(\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}\right). \end{aligned}$$

□

The above analysis is true for any matrix. The following is a generalisation of the above result.

Theorem 24.5. Any $m \times n$ matrix A defines a linear transformation $T_A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ by $T_A(\mathbf{x}) = A\mathbf{x}$.

Proof: This function is linear because

$$T_A(\mathbf{x} + \mathbf{y}) = A(\mathbf{x} + \mathbf{y}) = A\mathbf{x} + A\mathbf{y}$$

and

$$T_A(c\mathbf{x}) = A(c\mathbf{x}) = c(A\mathbf{x})$$

for all vectors $\mathbf{x}, \mathbf{y}$ in $\mathbb{R}^n$ and all scalars c .

The following theorem is a characterisation of linear mappings.

characterisation

Theorem 24.6. A linear transformation $T : V \rightarrow W$ preserves

1. the zero vector: $T(\mathbf{0}_V) = \mathbf{0}_W$.
2. negatives: $T(-\mathbf{x}) = -T(\mathbf{x})$.
3. linear combinations: $T(c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_k\mathbf{v}_k) = c_1T(\mathbf{v}_1) + c_2T(\mathbf{v}_2) + \cdots + c_kT(\mathbf{v}_k)$ for any vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ and scalars $c_1, c_2, \dots, c_k$.

Proof:

1. $T(\mathbf{0}_V) = T(0 \cdot \mathbf{0}_V) = 0 \cdot T(\mathbf{0}_V) = \mathbf{0}_W$.
2. $T(-\mathbf{x}) = T((-1) \cdot \mathbf{x}) = (-1) \cdot T(\mathbf{x}) = -T(\mathbf{x})$
3. By definition and associative property, $T(c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k) = T(c_1\mathbf{v}_1) + \cdots + T(c_k\mathbf{v}_k)$.
By definition, $T(c_i\mathbf{v}_i) = c_iT(\mathbf{v}_i)$ for $i = 1, \dots, k$. Therefore,

$$T(c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k) = c_1T(\mathbf{v}_1) + \cdots + c_kT(\mathbf{v}_k).$$

The theorem above can be used to show that for a given mapping, when one of the characterisation is not met, the mapping is **not** linear.

Example 24.7. Determine if the following transformation is linear:

$$T : \mathbb{R}^2 \rightarrow \mathbb{R}^2, \quad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mapsto \begin{bmatrix} x_1 + 1 \\ x_2 + 2 \end{bmatrix}.$$

Proof: Since

$$T\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

therefore T is not a linear transformation.

The theorem above can also be used to express $T(\mathbf{v})$ for any $\mathbf{v} \in V$ if a basis of V is given.

eg:5.1.9

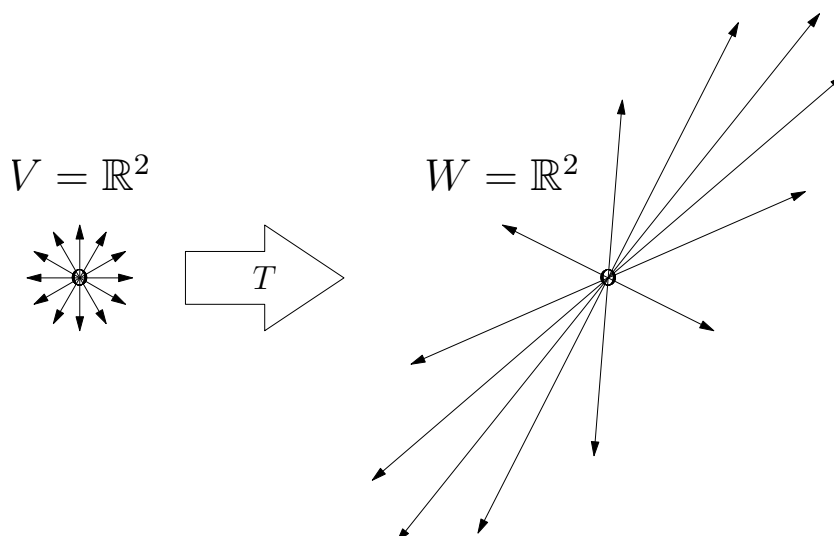
Example 24.8. Suppose $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is a linear transformation such that $T((1, 0)) = (2, -1)$ and $T((0, 1)) = (4, 5)$. Find $T(\mathbf{v})$ for any $\mathbf{v} = (a, b) \in \mathbb{R}^2$.

Solution. Note that $\{(1, 0), (0, 1)\}$ is a basis of $\mathbb{R}^2$. Therefore,

$$\begin{aligned} T(\mathbf{v}) &= T((a, b)) = T(a(1, 0) + b(0, 1)) \\ &= aT((1, 0)) + bT((0, 1)) \\ &= a(2, -1) + b(4, 5) = (2a + 4b, 5b - a). \end{aligned}$$

□

Remark 24.9. Informally, we can sketch the linear transformations for two- and three-dimensional spaces as a mapping of vectors. In the case of Example 24.8, the sketch is roughly



Notice that the centre “zero vector” is always mapped to a “zero vector” in the other vector space as stated in Theorem [24.6](#). [chm:characterisation](#)

We learn from set theory that if we are given two functions $f : U \rightarrow V$, $g : V \rightarrow W$, then we can define a “composition” function $g \circ f : U \rightarrow W$.

def:composition

Definition 24.2. Let $S : U \rightarrow V$ and $T : V \rightarrow W$ be linear transformations. The *composition* of S and T , denoted by $T \circ S$ (or TS), is defined as

$$T \circ S : U \rightarrow W, \quad \mathbf{u} \mapsto T(S(\mathbf{u})).$$

Pictorially, this is represented as

$$\begin{array}{ccc} U & \xrightarrow{S} & V \xrightarrow{T} W \\ & \searrow & \nearrow \\ & T \circ S & \end{array}$$

Example 24.10. Given the linear transformation $S : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ and $T : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ such that

$$S\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} x - 2y \\ 3x + 4y \end{bmatrix}, \quad T\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} 3x + 2y \\ 4x - 5y \\ 2x - y \end{bmatrix}$$

Find $(T \circ S)\left(\begin{bmatrix} x \\ y \end{bmatrix}\right)$.

Solution. By definition,

$$(T \circ S)\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = T\left(S\left(\begin{bmatrix} x \\ y \end{bmatrix}\right)\right) = T\left(\begin{bmatrix} x - 2y \\ 3x + 4y \end{bmatrix}\right) = \begin{bmatrix} 3(x - 2y) + 2(3x + 4y) \\ 4(x - 2y) - 5(3x + 4y) \\ 2(x - 2y) - (3x + 4y) \end{bmatrix} = \begin{bmatrix} 9x + 2y \\ -11x - 28y \\ -x - 8y \end{bmatrix}$$

□

24.2 Kernel and Range

This section discuss the concept which is closely related the column space and null space of a matrix — the range and kernel of a linear transformation. Mathematically, kernel and range determines how does a linear transformation different from a one-to-one linear transformation. However, this section just introduce some calculation examples and defer the discussion of the structures of a linear transformation using kernel and range to more advanced subjects.

Definition 24.3. Let T be a linear transformation from V to W . The *kernel* of T is the set of vectors in V that T sends to the zero vector of W :

$$\ker(T) := \{\mathbf{v} \in V | T(\mathbf{v}) = \mathbf{0}\}.$$

The *range* (or *image*) of T is the set of vectors in W that are T of some vector in V :

$$T[V] = \{\mathbf{w} \in W | \mathbf{w} = T(\mathbf{v}) \text{ for some } \mathbf{v} \in V\}.$$

Note that some books used a slightly different notation for kernel and range: $\text{Ker } T$ and $\text{Im } T$.

Theorem 24.11. If $T : V \rightarrow W$ is a linear transformation, the kernel of T is a subspace of V and the range of T is a subspace of W .

Proof:

1. Consider $\mathbf{u}, \mathbf{v} \in \ker(T)$ and $c \in \mathbb{R}$. Then $T(\mathbf{u}) = \mathbf{0}$ and $T(\mathbf{v}) = \mathbf{0}$. Since $T(\mathbf{u} + \mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v}) = \mathbf{0} + \mathbf{0} = \mathbf{0}$, $\mathbf{u} + \mathbf{v} \in \ker(T)$.

Since $T(c\mathbf{u}) = cT(\mathbf{u}) = c\mathbf{0} = \mathbf{0}$, $c\mathbf{u} \in \ker(T)$. Therefore, $\ker(T)$ is a subspace of V .

2. Consider $\mathbf{u}, \mathbf{v} \in T[V]$ and $c \in \mathbb{R}$. Then there exists $\mathbf{x}$ such that $T(\mathbf{x}) = \mathbf{u}$ and there exists $\mathbf{y}$ such that $T(\mathbf{y}) = \mathbf{v}$.

Since $\mathbf{u} + \mathbf{v} = T(\mathbf{x}) + T(\mathbf{y}) = T(\mathbf{x} + \mathbf{y})$, $\mathbf{u} + \mathbf{v} \in T[V]$ by definition.

Since $c\mathbf{u} = cT(\mathbf{x}) = T(c\mathbf{x})$, $c\mathbf{u} \in T[V]$. Therefore, $T[V]$ is a subspace of W .

Example 24.12. Let $T : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ be a linear transformation defined by $T\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} x - y \\ 2x + y \\ -x + 4y \end{bmatrix}$.

Find $\ker(T)$ and $T[\mathbb{R}^2]$.

Solution.

$$\ker(T) = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} : x - y = 0, 2x + y, -x + 4y = 0 \right\} = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}$$

$$T[V] = \left\{ x \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix} + y \begin{bmatrix} -1 \\ 1 \\ 4 \end{bmatrix} : x, y \in \mathbb{R} \right\}.$$

□

Remark 24.13. If $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is left multiplication by the matrix A , i.e. for $\mathbf{x} \in \mathbb{R}^n$, if $T(\mathbf{x}) = A\mathbf{x}$, then $\ker(T) = \text{nullsp}(A)$ and $T[\mathbb{R}^n] = \text{colsp}(A)$.

An application of the concept of kernel and range is to determine if a linear transformation is one-to-one or onto.

Definition 24.4. A linear transformation $T : V \rightarrow W$ is *one-to-one* if $T(\mathbf{v}_1) = T(\mathbf{v}_2)$ implies $\mathbf{v}_1 = \mathbf{v}_2$.

eg:5.2.6

Example 24.14. Show that the linear transformation $T : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ defined by $T\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) =$

$$\begin{bmatrix} x + y \\ x - y \\ 2x \end{bmatrix} \text{ is one-to-one.}$$

Proof [by definition]: Consider two vectors $\mathbf{v}_1 = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$ and $\mathbf{v}_2 = \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}$. Let us investigate

$$T(\mathbf{v}_1) = T(\mathbf{v}_2) \Rightarrow T\left(\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}\right) = T\left(\begin{bmatrix} x_2 \\ y_2 \end{bmatrix}\right) \Rightarrow \begin{bmatrix} x_1 + y_1 \\ x_1 - y_1 \\ 2x_1 \end{bmatrix} = \begin{bmatrix} x_2 + y_2 \\ x_2 - y_2 \\ 2x_2 \end{bmatrix}$$

Therefore

$$\begin{cases} x_1 + y_1 = x_2 + y_2 \\ x_1 - y_1 = x_2 - y_2 \Rightarrow x_1 = x_2, y_1 = y_2 \\ 2x_1 = 2x_2 \end{cases} \Rightarrow \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}$$

Note that two vectors are the same when all the components are the same.

Therefore, T is a one-to-one linear transformation.

Theorem 24.15. A linear transformation $T : V \rightarrow W$ is one-to-one iff $\ker(T) = \{\mathbf{0}\}$.

Example 24.16. Show that the linear transformation in Example 24.14 using Theorem 24.15.

Solution.

$$\ker(T) = \{\mathbf{v} : T(\mathbf{v}) = \mathbf{0}\} = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} : \begin{bmatrix} x+y \\ x-y \\ 2x \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \right\} = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} : x=0, y=0 \right\} = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}.$$

By Theorem 24.15, T is one-to-one. □

Definition 24.5. A linear transformation $T : V \rightarrow W$ is *onto* if $T[V] = W$, i.e. every $\mathbf{w}$ in W is of the form $\mathbf{w} = T(\mathbf{v})$ for some $\mathbf{v} \in V$.

Example 24.17. Determine whether the following linear transformations are onto.

1. $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ defined by $T\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} x+y \\ x-y \end{bmatrix}$.

Solution: If T is onto then for every $\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \in \mathbb{R}^2$, we can find $\mathbf{v} = \begin{bmatrix} x \\ y \end{bmatrix} \in \mathbb{R}^2$ such that

$$T(\mathbf{v}) = \mathbf{w} \Rightarrow T\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \Rightarrow \begin{bmatrix} x+y \\ x-y \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

Solving the last equation gives

$$x = \frac{w_1 + w_2}{2}, \quad y = \frac{w_1 - w_2}{2}.$$

Therefore, for every $\mathbf{w} \in \mathbb{R}^2$, there is a vector $\mathbf{v} = \begin{bmatrix} \frac{w_1 + w_2}{2} \\ \frac{w_1 - w_2}{2} \end{bmatrix}$ such that $T(\mathbf{v}) = \mathbf{w}$, which means that T is onto.

2. $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ defined by $T\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} x \\ 2x \end{bmatrix}$.

Solution. If T is onto then for every $\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \in \mathbb{R}^2$, we can find $\mathbf{v} = \begin{bmatrix} x \\ y \end{bmatrix} \in \mathbb{R}^2$ such that

$$T(\mathbf{v}) = \mathbf{w} \Rightarrow T\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \Rightarrow \begin{bmatrix} x \\ 2x \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

Since w_1 and w_2 are arbitrary, we can choose $w_1 = 1$, $w_2 = 3$. This means that $(x = 1 \wedge x = 1.5)$.

Therefore, for the vector $\mathbf{w}_0 = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$, there is NO vector $\mathbf{v} \in \mathbb{R}^2$ such that $T(\mathbf{v}) = \mathbf{w}_0$.

By definition, T is NOT onto. □

To work on:

Rank and Nullity of a linear transformation
Rank-nullity theorem and its applications

24.3 Matrix Representation of Linear Transformation

The abstract definition [24.1](#) for linear transformation is difficult to calculate. Mathematicians have discovered that the linear transformation for finite dimensional spaces can be “represented” by a matrix as follows.

def:mat-rep

Definition 24.6. Let $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ and $\mathcal{W}$ be ordered bases for vector spaces V and W , of dimensions n and m , respectively. Moreover, let $T : V \rightarrow W$ be a linear transform. Let $(\)_{\mathcal{V}} : V \rightarrow \mathbb{R}^n$ and $(\)_{\mathcal{W}} : W \rightarrow \mathbb{R}^m$ be the coordinate maps of V and W respectively. Then the *matrix representation of T relative to $\mathcal{V}$ and $\mathcal{W}$* (or the *matrix associated to T*), denoted by $M_{\mathcal{W} \leftarrow \mathcal{V}}(T)$ (some books denoted by $[T]_{\mathcal{W}}^{\mathcal{V}}$), is characterised by

$$(T(\mathbf{x}))_{\mathcal{W}} = M_{\mathcal{W} \leftarrow \mathcal{V}}(T)(\mathbf{x})_{\mathcal{V}}$$

for any vector $\mathbf{x}$ in V .

Remark 24.18. Beware that the basis $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ or $\mathbf{v}_1, \dots, \mathbf{v}_n$ (without the brackets) is an “ordered” sequence of vectors and is not a set. Therefore, the order is *essential*! As for the matrix representation of T given the bases $\mathcal{V}$ and $\mathcal{W}$ can be represented pictorially as follows:

$$\begin{array}{ccc} V & \xrightarrow{T} & W \\ (\)_{\mathcal{V}} \downarrow & & \downarrow (\)_{\mathcal{W}} \\ \mathbb{R}^n & \xrightarrow{M_{\mathcal{W} \leftarrow \mathcal{V}}(T)} & \mathbb{R}^m \end{array}$$

How do we calculate $M_{\mathcal{W} \leftarrow \mathcal{V}}$? The Matrix Representation Theorem below gives us a formula to perform the calculation.

thm:mat-rep-trans

Theorem 24.19 (Matrix Representation Theorem). *Let $T : V \rightarrow W$ be a linear transform and let $\mathcal{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ and $\mathcal{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_m\}$ be bases of V and W , respectively. If $M_{\mathcal{W} \leftarrow \mathcal{V}}(T)$ is the matrix representation of T with respect to these bases, then the columns of $M_{\mathcal{W} \leftarrow \mathcal{V}}(T)$ are the images of the basis vectors $\mathbf{v}_j$ under T , expressed in terms of the basis $\mathcal{W}$ of W :*

$$M_{\mathcal{W} \leftarrow \mathcal{V}}(T) = [(T(\mathbf{v}_1))_{\mathcal{W}}, \dots, (T(\mathbf{v}_n))_{\mathcal{W}}].$$

Proof: For every vector $\mathbf{x} \in V$, let $(\mathbf{x})_{\mathcal{V}} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$, then by definition, $\mathbf{x}$ can be written as a linear combination of $\mathcal{V}$:

$$\mathbf{x} = x_1\mathbf{v}_1 + \dots + x_n\mathbf{v}_n.$$

Therefore,

$$T(\mathbf{x}) = T(x_1\mathbf{v}_1 + \dots + x_n\mathbf{v}_n) = x_1T(\mathbf{v}_1) + \dots + x_nT(\mathbf{v}_n). \quad (*)$$

Let $\mathbf{y} \in W$ and $(\mathbf{y})_{\mathcal{W}} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$. Representing $(*)$ relative to the basis $\mathcal{W}$ and applying

Theorem ^{thm:basis1}23.64, we have

$$\begin{aligned}(T(\mathbf{x}))_{\mathcal{W}} &= (x_1 T(\mathbf{v}_1) + \cdots + x_n T(\mathbf{v}_n))_{\mathcal{W}} \\ &= x_1 (T(\mathbf{v}_1))_{\mathcal{W}} + \cdots + x_n (T(\mathbf{v}_n))_{\mathcal{W}} \\ &= [T(\mathbf{v}_1))_{\mathcal{W}}, \cdots, (T(\mathbf{v}_n))_{\mathcal{W}}] \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \underbrace{[T(\mathbf{v}_1))_{\mathcal{W}}, \cdots, (T(\mathbf{v}_n))_{\mathcal{W}}]}_{M_{\mathcal{W} \leftarrow \mathcal{V}}(T)} (\mathbf{x})_{\mathcal{V}}\end{aligned}$$

$\mathbb{R}^n$, $M_{m,n}(\mathbb{R})$ and P_n (the set of polynomials of degree no more than n are examples of vector space (Chapter ^{ch:vec-sp}23). The following example shall demonstrate that the linear transformation “differentiation” from P_3 to P_2 has a matrix representation and can be obtained using the Matrix Representation Theorem.

Example 24.20. Let P_n be the vector space of all polynomial functions of degree less than or equal n . Consider the linear transformation

$$D : P_3 \rightarrow P_2, \quad f(x) \mapsto \frac{d}{dx} f(x).$$

Find the matrix representation for D if P_3 has a basis $\mathcal{V} = \{1, x, x^2, x^3\}$ and P_2 has a basis $\mathcal{W} = \{1, x, x^2\}$.

Solution: By the Matrix Representation Theorem, we have

$$M_{\mathcal{W} \leftarrow \mathcal{V}}(D) = [(D(1))_{\mathcal{W}}, (D(x))_{\mathcal{W}}, (D(x^2))_{\mathcal{W}}, (D(x^3))_{\mathcal{W}}]$$

By definition, D is a derivative, therefore,

$$D(1) = 0, \quad D(x) = 1, \quad D(x^2) = 2x, \quad D(x^3) = 3x^2.$$

Again, by definition,

$$\begin{aligned}(D(1))_{\mathcal{W}} &= (0(1) + 0(x) + 0(x^2))_{\mathcal{W}} = [0, 0, 0]^T, \\ (D(x))_{\mathcal{W}} &= (1(1) + 0(x) + 0(x^2))_{\mathcal{W}} = [1, 0, 0]^T, \\ (D(x^2))_{\mathcal{W}} &= (0(1) + 2(x) + 0(x^2))_{\mathcal{W}} = [0, 2, 0]^T, \\ (D(x^3))_{\mathcal{W}} &= (0(1) + 0(x) + 3(x^2))_{\mathcal{W}} = [0, 0, 3]^T.\end{aligned}$$

Therefore, the matrix representation under the given bases is

$$M_{\mathcal{W} \leftarrow \mathcal{V}}(D) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}.$$

The above example demonstrates that a mathematical example can be rather complicated even vector space is considered to be a rather simple mathematical object. In statistics, most of the time, we will just be dealing with $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$ and therefore, finding matrix representations are not as complicated as the example given above.

Let us look at three examples where standard basis are implicitly assumed. Therefore, what we are calculating is the matrix $M_{\mathcal{E} \leftarrow \mathcal{E}}(T)$, in which we denote it A .

Example 24.21. Suppose $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is a linear transformation such that $T(\mathbf{i}) = (2, -1)$ and $T(\mathbf{j}) = (4, 7)$. Find the matrix A that represents T .

Solution: Note that in this question, standard basis $\mathcal{V} = \mathcal{W} = \{\mathbf{i}, \mathbf{j}\}$ are assumed.

According to the Matrix Representation Theorem, all vectors are written as column

vectors, therefore, we rewrite the vectors in this example as column vectors:

$$T(\mathbf{i}) = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \quad T(\mathbf{j}) = \begin{bmatrix} 4 \\ 7 \end{bmatrix}.$$

By the Matrix Representation Theorem,

$$A = [T(\mathbf{i}) \quad T(\mathbf{j})] = \begin{bmatrix} 2 & 4 \\ -1 & 7 \end{bmatrix}$$

Example 24.22. Let $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ be the linear transformation defined $T\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} 3x - y \\ 4x + 5y \end{bmatrix}$. Find a matrix representation A for the linear transformation T .

Solution. Let us calculate

$$T(\mathbf{i}) = T\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \quad T(\mathbf{j}) = T\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} -1 \\ 5 \end{bmatrix}.$$

Therefore

$$A = [T(\mathbf{i}) \quad T(\mathbf{j})] = \begin{bmatrix} 3 & -1 \\ 4 & 5 \end{bmatrix}$$

□

Sometimes standard basis problems can also be tricky. The following example make use of the “coordinate vector” (Definition 23.14). To solve the following example, we have to be careful about the meaning of $(\mathbf{x})_{\mathcal{V}}$.

Example 24.23. Let $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2\}$ where $\mathbf{v}_1 = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$, $\mathbf{v}_2 = \begin{bmatrix} 2 \\ -3 \end{bmatrix}$ is a basis for $\mathbb{R}^2$. Find the linear transformation $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ defined by $T(\mathbf{x}) = (\mathbf{x})_{\mathcal{V}}$ and hence find the matrix representation of T .

Solution. Let $T(\mathbf{x}) = \begin{bmatrix} a \\ b \end{bmatrix}$. By definition,

$$(\mathbf{x})_{\mathcal{V}} = T(\mathbf{x}) = \begin{bmatrix} a \\ b \end{bmatrix} \Rightarrow \mathbf{x} = a \begin{bmatrix} -1 \\ 2 \end{bmatrix} + b \begin{bmatrix} 2 \\ -3 \end{bmatrix} = \begin{bmatrix} -1 & 2 \\ 2 & -3 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

Therefore

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{3-4} \begin{bmatrix} -3 & -2 \\ -2 & -1 \end{bmatrix} \mathbf{x} \Rightarrow T(\mathbf{x}) = \begin{bmatrix} 3 & 2 \\ 2 & 1 \end{bmatrix} \mathbf{x}$$

The matrix representation of T is $\begin{bmatrix} 3 & 2 \\ 2 & 1 \end{bmatrix}$.

□

In Section 24.1, we have define the composition of two linear mappings (see Definition 24.2). In the following example, we will learn that the *composition of linear mappings is closely related to the multiplication of matrices!*

Example 24.24. Let $S : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ and $T : \mathbb{R}^3 \rightarrow \mathbb{R}^4$ be the linear transformations defined by

$$S\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} 2x_1 - 3x_2 \\ x_2 \\ x_1 + x_2 \end{bmatrix}, \quad T\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right) = \begin{bmatrix} x_1 - x_2 - x_3 \\ 2x_2 + x_3 \\ x_1 - x_3 \\ 3x_1 + x_2 - 2x_3 \end{bmatrix}.$$

Find the matrix representation C of the composition $T \circ S$. If the matrix representation of S is A and the matrix representation of T is B , then show that $C = BA$.

Solution. First, we must notice that the composition $T \circ S : \mathbb{R}^2 \rightarrow \mathbb{R}^4$ and then we calculate

$$\begin{aligned}(T \circ S)(\mathbf{x}) &= T(S(\mathbf{x})) = T\left(\begin{bmatrix} 2x_1 - 3x_2 \\ x_2 \\ x_1 + x_2 \end{bmatrix}\right) = \begin{bmatrix} (2x_1 - 3x_2) - x_2 - (x_1 + x_2) \\ 2x_2 + (x_1 + x_2) \\ (2x_1 - 3x_2) - (x_1 + x_2) \\ 3(2x_1 - 3x_2) + x_2 - 2(x_1 + x_2) \end{bmatrix} \\ &= \begin{bmatrix} x_1 - 5x_2 \\ x_1 + 3x_2 \\ x_1 - 4x_2 \\ 4x_1 - 10x_2 \end{bmatrix}.\end{aligned}$$

Therefore,

$$C = M_{\mathcal{E} \leftarrow \mathcal{E}}(T \circ S) = [(T \circ S)(\mathbf{i}), (T \circ S)(\mathbf{j})] = \begin{bmatrix} 1 & -5 \\ 1 & 3 \\ 1 & -4 \\ 4 & -10 \end{bmatrix}.$$

It is easy to get

$$A = \begin{bmatrix} 2 & -3 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & -1 & -1 \\ 0 & 2 & 1 \\ 1 & 0 & -1 \\ 3 & 1 & -2 \end{bmatrix}$$

By performing matrix multiplication, it is easy to show that $C = BA$. □

Mathematicians have discovered that when $W = V$, the matrix representation of the linear transformation T is a square matrix and under certain condition, the inverse exists.

Theorem 24.25. *Let V be a vector space and $T : V \rightarrow V$ a linear transformation with matrix representation A . If T is invertible, then A is invertible and the matrix representation of T^{-1} is A^{-1} .*

Example 24.26. Suppose $T : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ is defined by $T\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right) = \begin{bmatrix} x_1 + x_2 \\ x_1 - x_2 \\ x_1 + x_2 + x_3 \end{bmatrix}$. Find the matrix representation A of T and find A^{-1} and hence write down the expression for T^{-1} .

Solution.

$$A = [T(\mathbf{v}_1) \quad T(\mathbf{v}_2) \quad T(\mathbf{v}_3)] = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

The inverse of matrix A is

$$A^{-1} = \frac{1}{2} \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ -2 & 0 & 2 \end{bmatrix}.$$

The expression for T^{-1} is

$$T^{-1}(\mathbf{x}) = A^{-1}\mathbf{x} = \begin{bmatrix} (x_1 + x_2)/2 \\ (x_1 - x_2)/2 \\ x_3 - x_1 \end{bmatrix}.$$

□

Until now, we have discussed a lot of examples where both the bases for V and W are implicitly assumed to be standard. In general, the basis may not be the standard basis and calculations would be more complicated as the following examples demonstrated.

eg:5.3.8

Example 24.27. Let $V = \mathbb{R}^2$ and let $T : V \rightarrow V$ be the linear transformation defined by $T\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_2 \\ x_1 \end{bmatrix}$. The set $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2\}$ where $\mathbf{v}_1 = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$ and $\mathbf{v}_2 = \begin{bmatrix} 2 \\ -3 \end{bmatrix}$ is a basis for V . Let $\mathcal{E} = \{\mathbf{e}_1, \mathbf{e}_2\}$ be the standard basis for V . Find $M_{\mathcal{E} \leftarrow \mathcal{V}}(T)$, $M_{\mathcal{E} \leftarrow \mathcal{V}}(T)(\mathbf{x})_{\mathcal{V}}$.

Solution: By Matrix Representation Theorem, we have

$$M_{\mathcal{E} \leftarrow \mathcal{V}}(T) = [(T(\mathbf{v}_1))_{\mathcal{E}}, (T(\mathbf{v}_2))_{\mathcal{E}}] = \left[T\left(\begin{bmatrix} -1 \\ 2 \end{bmatrix}\right), T\left(\begin{bmatrix} 2 \\ -3 \end{bmatrix}\right) \right] = \begin{bmatrix} 2 & -3 \\ -1 & 2 \end{bmatrix}.$$

When $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and $(\mathbf{x})_{\mathcal{V}} = \begin{bmatrix} c \\ d \end{bmatrix}$, we have

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = c \begin{bmatrix} -1 \\ 2 \end{bmatrix} + d \begin{bmatrix} 2 \\ -3 \end{bmatrix} \Rightarrow \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} -1 & 2 \\ 2 & -3 \end{bmatrix}^{-1} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

and

$$M_{\mathcal{E} \leftarrow \mathcal{V}}(T)(\mathbf{x})_{\mathcal{V}} = \begin{bmatrix} 2 & -3 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} 2 & -3 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} = T(\mathbf{x}).$$

eg:5.3.9

Example 24.28. Let $V = \mathbb{R}^2$ and let $T : V \rightarrow V$ be the linear transformation defined by $T\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_2 \\ x_1 \end{bmatrix}$. The set $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2\}$ where $\mathbf{v}_1 = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$ and $\mathbf{v}_2 = \begin{bmatrix} 2 \\ -3 \end{bmatrix}$ is a basis for V . Let $\mathcal{E} = \{\mathbf{e}_1, \mathbf{e}_2\}$ be the standard basis for V . Find $M_{\mathcal{V} \leftarrow \mathcal{E}}(T)$.

Solution. However, by using Matrix Representation Theorem, the calculation is as follows:

$$M_{\mathcal{V} \leftarrow \mathcal{E}}(T) = [(T(\mathbf{e}_1))_{\mathcal{V}}, (T(\mathbf{e}_2))_{\mathcal{V}}]$$

Let $(T(\mathbf{e}_1))_{\mathcal{V}} = \begin{bmatrix} a \\ b \end{bmatrix}$. Then

$$T(\mathbf{e}_1) = a \begin{bmatrix} -1 \\ 2 \end{bmatrix} + b \begin{bmatrix} 2 \\ -3 \end{bmatrix} \Rightarrow \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -a + 2b \\ 2a - 3b \end{bmatrix} \Rightarrow a = 2, b = 1.$$

Let $(T(\mathbf{e}_2))_{\mathcal{V}} = \begin{bmatrix} c \\ d \end{bmatrix}$. Then

$$T(\mathbf{e}_2) = c \begin{bmatrix} -1 \\ 2 \end{bmatrix} + d \begin{bmatrix} 2 \\ -3 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -c + 2d \\ 2c - 3d \end{bmatrix} \Rightarrow c = 3, d = 2.$$

Therefore,

$$M_{\mathcal{V} \leftarrow \mathcal{E}}(T) = \begin{bmatrix} 2 & 3 \\ 1 & 2 \end{bmatrix}.$$

□

We have explored the composition of linear transformations in Example 24.24. The following theorem characterises the matrix representation of the composition of linear transformations as the product of matrix representations of the linear transformations. leg:composition:std

Theorem 24.29. Let $T : V \rightarrow W$ and $S : W \rightarrow U$ be linear transformations. Let $\mathcal{V}$, $\mathcal{W}$, and $\mathcal{U}$ be bases for V , W , and U , respectively and $\dim V = n$, $\dim W = m$, $\dim U = \ell$. Then

$$\underbrace{M_{\mathcal{U} \leftarrow \mathcal{V}}(ST)}_{\ell \times n \text{ matrix}} = \underbrace{M_{\mathcal{U} \leftarrow \mathcal{W}}(S)}_{\ell \times m \text{ matrix}} \underbrace{M_{\mathcal{W} \leftarrow \mathcal{V}}(T)}_{m \times n \text{ matrix}}.$$

Remark 24.30. Mathematicians visualise the above equation using the following diagram

$$\begin{array}{ccccc} V & \xrightarrow{T} & W & \xrightarrow{S} & U \\ \downarrow (\cdot)_{\mathcal{V}} & & \downarrow (\cdot)_{\mathcal{W}} & & \downarrow (\cdot)_{\mathcal{U}} \\ \mathbb{R}^n & \xrightarrow{M_{\mathcal{W} \leftarrow \mathcal{V}}(T)} & \mathbb{R}^m & \xrightarrow{M_{\mathcal{U} \leftarrow \mathcal{W}}(S)} & \mathbb{R}^{\ell} \end{array}$$

At the above of the diagram, we have abstract vector spaces and the associated linear transformations. At the bottom of the diagram, we have Euclidean spaces with matrix multiplications between them.

We will explore how to perform calculations in the case of non-standard bases in the following example.

Example 24.31. Let $V = \mathbb{R}^2$, $W = \mathbb{R}^3$ and $U = \mathbb{R}^4$. Let $T : V \rightarrow W$ and $S : W \rightarrow U$ be the linear transformations defined by

$$T\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} 2x_1 - 3x_2 \\ x_2 \\ x_1 + x_2 \end{bmatrix} \quad \text{and} \quad S\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right) = \begin{bmatrix} x_1 - x_2 - x_3 \\ 2x_2 + x_3 \\ x_1 - x_3 \\ 3x_1 + x_2 - 2x_3 \end{bmatrix}.$$

Let $\mathcal{V}$, $\mathcal{W}$, $\mathcal{U}$ have, respectively, the ordered bases

$$\mathcal{V} = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right\}, \quad \mathcal{W} = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\}, \quad \mathcal{U} = \{\mathbf{e}_4, \mathbf{e}_3, \mathbf{e}_2, \mathbf{e}_1\}.$$

Find $M_{\mathcal{W} \leftarrow \mathcal{V}}(T)$, $M_{\mathcal{U} \leftarrow \mathcal{W}}(S)$, $M_{\mathcal{U} \leftarrow \mathcal{V}}(ST)$.

Solution. Assume A and B are the matrix representation of T and S respectively. Using the Theorem thm:mat-rep-trans 24.19, we have

$$\begin{aligned} M_{\mathcal{W} \leftarrow \mathcal{V}}(T) &= [(T(\mathbf{v}_1))_{\mathcal{W}} \quad (T(\mathbf{v}_2))_{\mathcal{W}}] = (T([\mathbf{v}_1 \quad \mathbf{v}_2]))_{\mathcal{W}} = [\mathbf{w}_1 \quad \mathbf{w}_2 \quad \mathbf{w}_3]^{-1} A [\mathbf{v}_1 \quad \mathbf{v}_2] \\ &= \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 2 & -3 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} -2 & -6 \\ -1 & 1 \\ 2 & 0 \end{bmatrix} \end{aligned}$$

$$\begin{aligned}
M_{\mathcal{U} \leftarrow \mathcal{W}}(S) &= [(S(\mathbf{w}_1))_{\mathcal{U}} \quad (S(\mathbf{w}_2))_{\mathcal{U}} \quad (S(\mathbf{w}_3))_{\mathcal{U}}] \\
&= (S([\mathbf{w}_1 \quad \mathbf{w}_2 \quad \mathbf{w}_3]))_{\mathcal{U}} = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \mathbf{u}_3 \quad \mathbf{u}_3]^{-1} B [\mathbf{w}_1 \quad \mathbf{w}_2 \quad \mathbf{w}_3] \\
&= \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}^{-1} \begin{bmatrix} 1 & -1 & -1 \\ 0 & 2 & 1 \\ 1 & 0 & -1 \\ 3 & 1 & -2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 4 & 2 \\ 1 & 1 & 0 \\ 0 & 2 & 3 \\ 1 & 0 & -1 \end{bmatrix}
\end{aligned}$$

By Theorem [24.29](#), [\[thm:mat-rep-comp\]](#)

$$M_{\mathcal{U} \leftarrow \mathcal{V}}(ST) = M_{\mathcal{U} \leftarrow \mathcal{W}}(S) M_{\mathcal{W} \leftarrow \mathcal{V}}(T) = \begin{bmatrix} 3 & 4 & 2 \\ 1 & 1 & 0 \\ 0 & 2 & 3 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} -2 & -6 \\ -1 & 1 \\ 2 & 0 \end{bmatrix} = \begin{bmatrix} -6 & -14 \\ -3 & -5 \\ 4 & 2 \\ -4 & -6 \end{bmatrix}.$$

□

To work on:

Application of linear transformation to geometry

24.4 Changing Coordinates

This section discuss a special case of the previous section, when the linear transformation is an identity. In this case, we will have two bases and it has a special name defined below.

Definition 24.7. Let $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ and $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ be bases for a vector space V . Then $n \times n$ matrix whose columns are the coordinate vectors $(\mathbf{u}_1)_{\mathcal{V}}, (\mathbf{u}_2)_{\mathcal{V}}, \dots, (\mathbf{u}_n)_{\mathcal{V}}$ is denoted by $P_{\mathcal{V} \leftarrow \mathcal{U}}$ and is called the *change-of-basis matrix* from $\mathcal{U}$ to $\mathcal{V}$. Formally, we have

$$P_{\mathcal{V} \leftarrow \mathcal{U}} = [(\mathbf{u}_1)_{\mathcal{V}}, (\mathbf{u}_2)_{\mathcal{V}}, \dots, (\mathbf{u}_n)_{\mathcal{V}}] = M_{\mathcal{V} \leftarrow \mathcal{U}}(id)$$

where $id : V \rightarrow V$ is the identity transformation on V such that $id(\mathbf{x}) = \mathbf{x}$.

Theorem 24.32. Let $\mathcal{U} = \{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ and $\mathcal{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ be bases for a vector space V and let $P_{\mathcal{V} \leftarrow \mathcal{U}}$ be the change-of-basis matrix from $\mathcal{U}$ to $\mathcal{V}$. Then

- (a) $P_{\mathcal{V} \leftarrow \mathcal{U}}(\mathbf{x})_{\mathcal{U}} = (\mathbf{x})_{\mathcal{V}}$ for all $\mathbf{x}$ in V .
- (b) $P_{\mathcal{U} \leftarrow \mathcal{V}}$ is the unique matrix P with the property that $P(\mathbf{x})_{\mathcal{U}} = (\mathbf{x})_{\mathcal{V}}$ for all $\mathbf{x}$ in V .
- (c) $P_{\mathcal{V} \leftarrow \mathcal{U}}$ is invertible and $(P_{\mathcal{V} \leftarrow \mathcal{U}})^{-1} = P_{\mathcal{U} \leftarrow \mathcal{V}}$.

Example 24.33. Given the (ordered) basis $\mathcal{U} = \left\{ \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} -2 \\ 3 \end{bmatrix} \right\}$ and $\mathcal{V} = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right\}$ for $\mathbb{R}^2$. Find the change-of-basis matrix from $\mathcal{U}$ to $\mathcal{V}$, $P_{\mathcal{V} \leftarrow \mathcal{U}}$.

Solution. $P_{\mathcal{V} \leftarrow \mathcal{U}} = M_{\mathcal{V} \leftarrow \mathcal{U}}(id) = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & -2 \\ 2 & 3 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 3 & 1 \\ -1 & -5 \end{bmatrix}.$

□

Example 24.34. Let $\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$, $\mathbf{v}_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$, $\mathbf{v}_3 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}$, $\mathbf{v}_4 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}$. These four vectors form a basis $\mathcal{V}$ for $\mathbb{R}^4$. Find the coordinates of a vector $\mathbf{x} \in \mathbb{R}^4$ relative to $\mathcal{V}$ i.e. $(\mathbf{x})_{\mathcal{V}}$.

Solution. Note that $\mathbf{x}$ is written in standard basis. Therefore,

$$(\mathbf{x})_{\mathcal{V}} = P_{\mathcal{V} \leftarrow \mathcal{E}} \mathbf{x} = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & -1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} x_1 - x_4 \\ -x_2 + x_3 + x_4 \\ x_3 \\ x_2 - x_3 \end{bmatrix}.$$

□

Chapter 25

Inner Product Spaces and Orthogonality

In statistics, especially in linear regression problems, we often encounter a linear system $A\mathbf{x} = \mathbf{b}$ with $A \in M_{m,n}(\mathbb{R})$ and $m > n$, which *does not have a solution* in general. However, by introducing the concept of “total squared error” $(\mathbf{b} - A\mathbf{x})^T(\mathbf{b} - A\mathbf{x})$, we can try to find an approximation called the “least square solution” to the linear system. The least square solution is obtained using the “generalised” projection (Remark [25.4](#)).

The concept of orthogonality used to be useful in the numerical solution of least square solution. However, in real world application, it is more important in the solution of eigenvalue problem using iterative QR factorisation. The numerical eigenvalue problem is too complex, therefore only the basic calculations related to orthogonality is introduced, which are found in Sections [25.3](#) to Section [25.5](#).

25.1 Inner Product and Norm

Recall the inner product and norm of $\mathbb{R}^n$ from Definition [20.17](#) and Definition [20.20](#)

Definition 25.1. For any $\mathbf{x} = (x_1, x_2, \dots, x_n)$, $\mathbf{y} = (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$, the **inner product** or **dot product** is defined as

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x} \cdot \mathbf{y} = x_1 y_1 + x_2 y_2 + \dots + x_n y_n.$$

Definition 25.2. The **norm** of the n -dimensional vector $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ is the number

$$\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}} = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}.$$

25.2 Least Square Approximations and Projection

square-problem

Consider the following system:

$$\begin{cases} x_1 + x_2 = 2 \\ x_1 - x_2 = 0 \\ 2x_1 + x_2 = -4. \end{cases} \quad (25.1) \quad \text{e:6.1}$$

Using the linear algebra techniques (r.e.f. of augmented matrix) from Chapter [21](#), it easy to know that the above system is inconsistent. Therefore, for all $\mathbf{x} \in \mathbb{R}^2$, $\mathbf{b} - A\mathbf{x} \neq \mathbf{0}$. However, we can try to find a $\hat{\mathbf{x}}$ such that the distance (error) between $\mathbf{b}$ and $A\hat{\mathbf{x}}$ — $\|\mathbf{b} - A\hat{\mathbf{x}}\|$ — is as small as possible.

Definition 25.3. A vector $\hat{\mathbf{x}} \in \mathbb{R}^n$ is called a *least square approximation* or *least square solution* to the linear system $A\mathbf{x} = \mathbf{b}$ if the *total squared error* defined below is minimum:

$$(\mathbf{b} - A\hat{\mathbf{x}})^T(\mathbf{b} - A\hat{\mathbf{x}}) = \min_{\mathbf{x} \in \mathbb{R}^n} (\mathbf{b} - A\mathbf{x})^T(\mathbf{b} - A\mathbf{x}) \quad (*)$$

The vector $\hat{\mathbf{x}}$ is also denoted $\hat{\mathbf{x}} = \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{b} - A\mathbf{x}\|$.

Remark 25.1. For any given $\mathbf{x} \in \mathbb{R}^n$, let $\mathbf{r} = \mathbf{b} - A\mathbf{x}$. Then $\mathbf{r}$ is called the *residue error* of the linear system, then the total squared error is related to the square of the norm of $\mathbf{r}$ by

$$\mathbf{r}^T \mathbf{r} = \mathbf{r} \cdot \mathbf{r} = \|\mathbf{r}\|^2 = \|\mathbf{b} - A\mathbf{x}\|^2.$$

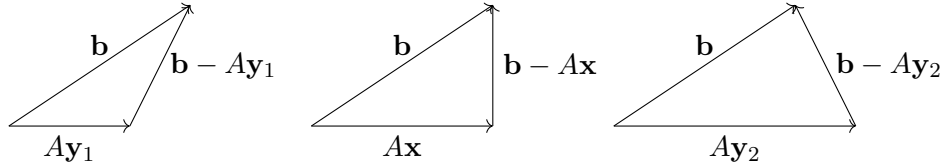
In solving of least square problem, mathematicians have discovered that the least square solution can be calculated from a normal equation, defined below.

Definition 25.4. The linear system $A^T A\mathbf{x} = A^T \mathbf{b}$ is called the *normal equation* to the linear system $A\mathbf{x} = \mathbf{b}$.

thm:lsqsoln

Theorem 25.2. Consider $A \in M_{m,n}(\mathbb{R})$ (and assume that A is full rank). The solution of the normal equation $A^T A\mathbf{x} = A^T \mathbf{b}$ is the least square solution of the linear system $A\mathbf{x} = \mathbf{b}$.

Idea of the Proof: Since $A\mathbf{x} - \mathbf{b}$ is a error vector. For any $\mathbf{y} \in \mathbb{R}^n$, $A\mathbf{y}$ is a vector in the column space $\operatorname{colsp}(A)$. To make $\|A\mathbf{x} - \mathbf{b}\|$ minimum, we need to make sure that $A\mathbf{x} - \mathbf{b}$ is orthogonal to all vectors in $\operatorname{colsp}(A)$, i.e. $(A\mathbf{y})^T(A\mathbf{x} - \mathbf{b}) = 0$ for every $\mathbf{y} \in \mathbb{R}^n$. This implies $A^T(A\mathbf{x} - \mathbf{b}) = \mathbf{0}$.



Example 25.3. Find the least square solution of the system $A\mathbf{x} = \mathbf{b}$ given by

$$\begin{cases} x_1 - x_2 = 4, \\ 3x_1 + 2x_2 = 1, \\ -2x_1 + 4x_2 = 3. \end{cases}$$

Solution: According to the Theorem [25.2](#), transform the problem of finding a least square approximation to

$$\begin{bmatrix} 1 & 3 & -2 \\ -1 & 2 & 4 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 3 & 2 \\ -2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 3 & -2 \\ -1 & 2 & 4 \end{bmatrix} \begin{bmatrix} 4 \\ 1 \\ 3 \end{bmatrix}$$

After simplification, we obtain the normal equation

$$\begin{bmatrix} 14 & -3 \\ -3 & 21 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 10 \end{bmatrix}$$

Solving it by back substitution gives the least square solution:

$$x_2 = \frac{51}{285}, \quad x_1 = \frac{143}{285}.$$

rem:proj

Remark 25.4. When A is full-rank (which is assumed implicitly), the solution of the normal equation $A^T A \mathbf{x} = A^T \mathbf{b}$ can be written as

$$\hat{\mathbf{x}} = (A^T A)^{-1} A^T \mathbf{b}.$$

The linear combination of the column vectors of A that is closest to $\mathbf{b}$ is given by

$$A\hat{\mathbf{x}} = \underbrace{A(A^T A)^{-1} A^T}_{\text{projection matrix}} \mathbf{b}.$$

The matrix $A(A^T A)^{-1} A^T$ can be viewed as projection of vectors onto the vector space $\text{colsp}(A)$. When $A = \mathbf{v}$ is an $n \times 1$ matrix (a column vector), then the projection of another vector $\mathbf{u}$ onto $\text{colsp}(A) = \text{span}\{\mathbf{v}\}$ is going to be

$$\text{proj}_{\mathbf{v}} \mathbf{u} = \mathbf{v}(\mathbf{v}^T \mathbf{v})^{-1} \mathbf{v}^T \mathbf{u}.$$

How does mathematicians discovered this???

Answer: They combine the knowledge of subspaces, bases and projections to generalise the projection of a vector $\mathbf{u}$ onto $\mathbf{v}$ — $\text{proj}_{\mathbf{v}} \mathbf{u}$ to the projection of a vector $\mathbf{u}$ onto a subspace $W = \text{span}\{\mathbf{a}_1, \dots, \mathbf{a}_k\}$ — $\text{proj}_W \mathbf{u}$. Since $\text{proj}_W \mathbf{u} \in W$, $\mathbf{u} - \text{proj}_W \mathbf{u}$ should be orthogonal to W , therefore, for every $\mathbf{w} \in W$,

$$(\mathbf{u} - \text{proj}_W \mathbf{u}) \cdot \mathbf{w} = \mathbf{w}^T (\mathbf{u} - \text{proj}_W \mathbf{u}) = 0.$$

Since $\mathbf{a}_i \in W$, we have $\mathbf{a}_i^T (\mathbf{u} - \text{proj}_W \mathbf{u}) = 0 \Rightarrow \mathbf{a}_i^T \text{proj}_W \mathbf{u} = \mathbf{a}_i^T \mathbf{u}$.

Since $\text{proj}_W \mathbf{u} \in W$, it can be written as a linear combination of $\mathbf{a}_1, \dots, \mathbf{a}_k$:

$$\text{proj}_W \mathbf{u} = c_1 \mathbf{a}_1 + \dots + c_k \mathbf{a}_k,$$

or in coordinate system representation (let $\mathcal{U}$ be ordered basis $\mathbf{a}_1, \dots, \mathbf{a}_k$),

$$(\text{proj}_W \mathbf{u})_{\mathcal{U}} = (c_1, \dots, c_k)^T$$

and so

$$\mathbf{a}_i^T (c_1 \mathbf{a}_1 + \dots + c_k \mathbf{a}_k) = \mathbf{a}_i^T \mathbf{u} \Rightarrow c_1 \mathbf{a}_i^T \mathbf{a}_1 + \dots + c_k \mathbf{a}_i^T \mathbf{a}_k = \mathbf{a}_i^T \mathbf{u}.$$

Writing in matrix form and note that $A = [\mathbf{a}_1, \dots, \mathbf{a}_k]$, we have

$$\begin{bmatrix} c_1 \mathbf{a}_1^T \mathbf{a}_1 + \dots + c_k \mathbf{a}_1^T \mathbf{a}_k \\ \vdots \\ c_1 \mathbf{a}_k^T \mathbf{a}_1 + \dots + c_k \mathbf{a}_k^T \mathbf{a}_k \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1^T \mathbf{u} \\ \vdots \\ \mathbf{a}_k^T \mathbf{u} \end{bmatrix} \Rightarrow \begin{bmatrix} \mathbf{a}_1^T \mathbf{a}_1 & \dots & \mathbf{a}_1^T \mathbf{a}_k \\ \vdots & \ddots & \vdots \\ \mathbf{a}_k^T \mathbf{a}_1 & \dots & \mathbf{a}_k^T \mathbf{a}_k \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_k \end{bmatrix} = [\mathbf{a}_1, \dots, \mathbf{a}_k]^T \mathbf{u}$$

Therefore,

$$(A^T A)(\text{proj}_W \mathbf{u})_{\mathcal{U}} = A^T \mathbf{u} \Rightarrow (\text{proj}_W \mathbf{u})_{\mathcal{U}} = (A^T A)^{-1} A^T \mathbf{u} \Rightarrow \text{proj}_W \mathbf{u} = A(A^T A)^{-1} A^T \mathbf{u}$$

Let us apply this idea to investigate Example [20.51](#). eg:vec+plane1

Example 25.5 (Example [20.51](#)). Find the projection of $\mathbf{w} = (5, -3, -7)$ on the plane π with equation $3x - 2y + z = 0$. eg:vec+plane1

Solution (using projection matrix): Take any two vectors that span the plane — $\mathbf{a}_1 = (1, 0, -3)$ and $\mathbf{a}_2 = (0, 1, 2)$. Then, the projection is

$$\text{proj}_{\pi} \mathbf{w} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -3 & 2 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 & -3 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -3 & 2 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 0 & -3 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 5 \\ -3 \\ -7 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \\ -8 \end{bmatrix}$$

Example 25.6. Find the least square solution of the system $A\mathbf{x} = \mathbf{b}$ given by

$$\begin{cases} 5x_1 - 3x_2 = 2, \\ 2x_1 + x_2 = 1, \\ 3x_1 - x_2 = -3, \\ 2x_1 + 4x_2 = 5. \end{cases}$$

Hence, find the projection of $(2, 1, -3, 5)^T$ to the column space of the coefficient matrix.

Solution. Writing the above linear equation in matrix form, we have

$$A\mathbf{x} = \begin{bmatrix} 5 & -3 \\ 2 & 1 \\ 3 & -1 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ -3 \\ 5 \end{bmatrix} = \mathbf{b}.$$

According to Remark [25.4](#), the least square solution is

$$\hat{\mathbf{x}} = (A^T A)^{-1} A^T \mathbf{b} = \begin{bmatrix} 42 & -8 \\ -8 & 27 \end{bmatrix}^{-1} \begin{bmatrix} 5 & 2 & 3 & 2 \\ -3 & 1 & -1 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ -3 \\ 5 \end{bmatrix} = \begin{bmatrix} 99/214 \\ 86/107 \end{bmatrix}$$

Let $W = \text{colsp}(A)$, the projection is

$$\text{proj}_W \mathbf{b} = A(A^T A)^{-1} A^T \mathbf{b} = \begin{bmatrix} 5 & -3 \\ 2 & 1 \\ 3 & -1 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} 99/214 \\ 86/107 \end{bmatrix} = \begin{bmatrix} -21/214 \\ 185/107 \\ 125/214 \\ 443/107 \end{bmatrix}$$

□

exam:201205q2c

Example 25.7 (September 2012 Final Exam Q2(c)). Use the least square method to find the straight line that best fits these three points: $(1, 0)$, $(-2, -3)$ and $(1, 1)$.

Solution. Assume the formula of the line to be $y = mx + c$. Then the least square problem is formalised as

$$\left. \begin{array}{l} 0 = m(1) + c \\ -3 = m(-2) + c \\ 1 = m(1) + c \end{array} \right\} \Rightarrow \begin{bmatrix} 1 & 1 \\ -2 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} m \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ -3 \\ 1 \end{bmatrix}$$

From the normal equation

$$\begin{bmatrix} 6 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} m \\ c \end{bmatrix} = \begin{bmatrix} 7 \\ -2 \end{bmatrix} \Rightarrow \begin{bmatrix} m \\ c \end{bmatrix} = \begin{bmatrix} 7/6 \\ -2/3 \end{bmatrix}.$$

□

25.3 Orthogonality in $\mathbb{R}^n$

orthogonality

ch:linear:vec

In Chapter 20, we defined the dot product of vectors in n -space for $n \geq 1$, i.e. if $\mathbf{u} = (u_1, u_2, \dots, u_n)$ and $\mathbf{v} = (v_1, v_2, \dots, v_n)$ are in $\mathbb{R}^n$, then $\mathbf{u} \cdot \mathbf{v} = u_1v_1 + u_2v_2 + \dots + u_nv_n$ and if $\mathbf{u} \cdot \mathbf{v} = 0$, $\mathbf{u}$ and $\mathbf{v}$ are said to be *orthogonal*. In this section, we generalise the notion of orthogonality to subspace as follows.

Definition 25.5. Let $V = \mathbb{R}^n$ be a (column) vector space and let U and W be subspace of V .

1. A vector $\mathbf{u} \in V$ is said to be *orthogonal* to W if it is orthogonal to every vector in W .
2. The set of all vectors in V that are orthogonal to W is called the *orthogonal complement* of W , denoted by $W^\perp$.
3. The vector space U and the vector space W are called *orthogonal* if every vector $\mathbf{u} \in U$ and every vector $\mathbf{w} \in W$ are orthogonal:

$$\mathbf{u} \cdot \mathbf{w} = \mathbf{u}^T \mathbf{w} = 0.$$

Theorem 25.8 (Properties of Orthogonal Complements). *If W is a subspace of a finite-dimensional vector space V , then*

1. $W^\perp$ is a subspace of V .
2. The only vector common to W and $W^\perp$ is $\mathbf{0}$, i.e. $W \cap W^\perp = \{\mathbf{0}\}$
3. The orthogonal complement of $W^\perp$ is W , i.e. $(W^\perp)^\perp = W$.

Proof:

1. $W^\perp = \{x \in V : x \cdot w = 0, \forall w \in W\}$, $\forall x, y \in W^\perp$, $(ax + by) \cdot w = a(x \cdot w) + b(y \cdot w) = a \cdot 0 + b \cdot 0 = 0$.
2. When $x \in W \cap W^\perp$, $x \cdot x = 0$. By the property of dot product, x has to be a zero vector.
3. For any $x \in W$, $x \cdot y = 0$ for every $y \in W^\perp$. Therefore, for $y \in V$, $y \cdot x = 0$, $x \in V^\perp = (W^\perp)^\perp$.

Making use of the fact that every vector space can be decomposed into the sum of two orthogonal subspace, we know that for any $x \in V$, $x = x_1 + x_2$ where $x_1 \in W$ and $x_2 \in W^\perp$ and $x_1 \cdot x_2 = 0$.

For any $x \in (W^\perp)^\perp$, $x \cdot x_2 = 0$ by definition. This implies $(x_1 + x_2) \cdot x_2 = 0$. Therefore, $x_1 \cdot x_2 + x_2 \cdot x_2 = x_2 \cdot x_2 = 0$, which implies $x_2 = 0$. Hence, $x = x_1 + 0 = x_1 \in W$.

Since every row of a matrix A is perpendicular to every solution of the linear system $A\mathbf{x} = \mathbf{0}$ and since every column of a matrix A is perpendicular to every solution of the linear system $A^T\mathbf{y} = \mathbf{0}$. Mathematicians have concluded this as a fundamental theorem stated below.

Theorem 25.9 (Fundamental Theorem of Linear Algebra). *For any $A \in M_{m,n}(\mathbb{R})$,*

$$\begin{aligned} \text{nullsp}(A) &= (\text{Range}(A^T))^\perp, & \text{nullsp}(A^T) &= (\text{Range}(A))^\perp, \\ \text{Range}(A) &= (\text{nullsp}(A^T))^\perp, & \text{Range}(A^T) &= (\text{nullsp}(A))^\perp. \end{aligned}$$

Note that $\text{Range}(A)$ is the “column space” of A , so $\text{Range}(A^T)$ is the “row space” of A . As for $\text{nullsp}(A^T)$, it is similar to $\text{nullsp}(A)$ and is called the “left null space” of A .

Example 25.10. Let W be a subspace of $\mathbb{R}^5$ spanned by the vectors

$$\mathbf{w}_1 = (2, 2, -1, 0, 1), \quad \mathbf{w}_2 = (-1, -1, 2, -3, 1), \quad \mathbf{w}_3 = (1, 1, -2, 0, -1), \quad \mathbf{w}_4 = (0, 0, 1, 1, 1).$$

Find a basis for the orthogonal complement of W .

Solution. This is equivalent to finding all the vectors $(x_1, x_2, x_3, x_4, x_5)$ orthogonal to $\mathbf{w}_1$ to $\mathbf{w}_4$:

$$\begin{cases} 2x_1 + 2x_2 - x_3 + x_5 = 0 \\ -x_1 - x_2 + 2x_3 - 3x_4 + x_5 = 0 \\ x_1 + x_2 - 2x_3 - x_5 = 0 \\ x_3 + x_4 + x_5 = 0 \end{cases}$$

After reducing to the r.e.f. $\begin{bmatrix} 2 & 2 & -1 & 0 & 1 \\ 0 & 0 & 1.5 & -3 & 1.5 \\ 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$ and solving it, we obtain

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} s + \begin{bmatrix} -1 \\ 0 \\ -1 \\ 0 \\ 1 \end{bmatrix} t.$$

Hence, the basis is $(-1, 1, 0, 0, 0)$, $(-1, 0, -1, 0, 1)$. □

Definition 25.6. A set $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k\}$ of vectors is said to be *orthogonal* iff the vectors are orthogonal pairwise, i.e. iff $\mathbf{u}_i \cdot \mathbf{u}_j = 0$ when $i \neq j$. If each $\mathbf{u}_i$ has norm 1 (i.e. $\|\mathbf{u}_i\| = 1$), then the set is called *orthonormal*.

Example 25.11. Any subset of the set of standard basis vectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ is an orthonormal set in $\mathbb{R}^n$. For example, $\{(0, 1, 0), (0, 0, 1)\} \in \mathbb{R}^3$.

Example 25.12. Show that $\{(1, 1, 1, -1), (1, 0, 1, 2), (-1, 0, 1, 0), (-1, 3, -1, 1)\}$ are orthogonal in $\mathbb{R}^4$.

Solution. Proof:

- $(1, 1, 1, -1) \cdot (1, 0, 1, 2) = 1 + 1 - 2 = 0$
 - $(1, 1, 1, -1) \cdot (-1, 0, 1, 0) = 0$
 - $(1, 1, 1, -1) \cdot (-1, 3, -1, 1) = 0$
 - $(1, 0, 1, 2) \cdot (-1, 0, 1, 0) = 0$
 - $(1, 0, 1, 2) \cdot (-1, 3, -1, 1) = 0$
 - $(-1, 0, 1, 0) \cdot (-1, 3, -1, 1) = 0$
-

Theorem 25.13. An orthogonal set of nonzero vectors in $\mathbb{R}^n$ is a linearly independent set.

Proof: Let the elements of orthogonal set be $\mathbf{v}_1, \dots, \mathbf{v}_n$. By definition, $\mathbf{v}_i \cdot \mathbf{v}_i \neq 0$ and $\mathbf{v}_i \cdot \mathbf{v}_j = 0$ for $i \neq j$.

Let $c_1, \dots, c_n$ be any numbers such that satisfy the linear equation

$$c_1 \mathbf{v}_1 + \dots + c_n \mathbf{v}_n = \mathbf{0}.$$

Taking dot product with the above equation, and we have

$$c_1 \mathbf{v}_1 \cdot \mathbf{v}_i + \cdots + c_i \mathbf{v}_i \cdot \mathbf{v}_i + \cdots + c_n \mathbf{v}_n \cdot \mathbf{v}_i = \mathbf{0} \cdot \mathbf{v}_i \Rightarrow c_i \mathbf{v}_i \cdot \mathbf{v}_i = 0 \Rightarrow c_i = 0$$

since $\mathbf{v}_i \cdot \mathbf{v}_i \neq 0$. Since we only have the zero solution, by definition, $\mathbf{v}_1, \dots, \mathbf{v}_n$ are linearly independent.

Definition 25.7. An *orthogonal matrix* is a square matrix with *orthonormal* columns.

Example 25.14. Show that the following are orthogonal matrices:

$$R = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix}, \quad P = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Proof: Since the column vectors in R are orthonormal, R is an orthogonal matrix.
Since the column vectors in P are orthonormal, P is an orthogonal matrix.

The following theorems characterise the basic properties of orthogonal matrices.

Theorem 25.15. Q is an orthogonal matrix iff $Q^{-1} = Q^T$.

Theorem 25.16. Multiplication by an orthogonal matrix preserves the dot product and the norm of vectors, i.e. for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$,

$$\mathbf{u} \cdot \mathbf{v} = (Q\mathbf{u}) \cdot (Q\mathbf{v}), \quad \|\mathbf{u}\| = \|Q\mathbf{u}\|.$$

25.4 The Gram-Schmidt Process and the QR Factorisation

sec:qr

From Chapter [20](#), [ch:linalg:vec](#) two linearly independent vectors $\mathbf{u}$ and $\mathbf{v}$ can be turned into orthogonal vectors $\mathbf{e}$ and $\mathbf{f}$ by applying projection:

$$\mathbf{e} = \mathbf{u} \quad \text{and} \quad \mathbf{f} = \mathbf{v} - \text{proj}_{\mathbf{u}} \mathbf{v}.$$

We can also turn $\mathbf{u}, \mathbf{v}$ and $\mathbf{w}$ to three orthogonal vectors by applying projection:

$$\mathbf{e} = \mathbf{u}, \mathbf{f} = \mathbf{v} - \text{proj}_{\mathbf{u}} \mathbf{v} \quad \text{and} \quad \mathbf{g} = \mathbf{w} - \text{proj}_{\mathbf{e}} \mathbf{w}.$$

We will now generalise this to the general n -dimensional space. The technique is called **Gram-Schmidt Algorithm**. The Gram-Schmidt Algorithm defines a procedure for turning a set of k linearly independent vectors $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k$ into a set of k orthogonal vectors $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_k$. The orthogonal vectors are constructed as follows.

Suppose $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k\}$ is a set of linearly independent vectors in $\mathbb{R}^n$. Let

$$\begin{aligned} U_1 &= \text{span}\{\mathbf{u}_1\} \\ U_2 &= \text{span}\{\mathbf{u}_1, \mathbf{u}_2\} \\ U_3 &= \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\} \\ &\vdots \\ U_k &= \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k\} \end{aligned}$$

Define vectors $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_k$ as follows:

$$\begin{aligned}\mathbf{f}_1 &= \mathbf{u}_1 \\ \mathbf{f}_2 &= \mathbf{u}_2 - \mathbf{f}_1(\mathbf{f}_1 \cdot \mathbf{f}_1)^{-1} \mathbf{f}_1 \cdot \mathbf{u}_2 \\ \mathbf{f}_3 &= \mathbf{u}_3 - \mathbf{f}_1(\mathbf{f}_1 \cdot \mathbf{f}_1)^{-1} \mathbf{f}_1 \cdot \mathbf{u}_3 - \mathbf{f}_2(\mathbf{f}_2 \cdot \mathbf{f}_2)^{-1} \mathbf{f}_2 \cdot \mathbf{u}_3 \\ &\vdots \\ \mathbf{f}_k &= \mathbf{u}_k - \mathbf{f}_1(\mathbf{f}_1 \cdot \mathbf{f}_1)^{-1} \mathbf{f}_1 \cdot \mathbf{u}_k - \mathbf{f}_2(\mathbf{f}_2 \cdot \mathbf{f}_2)^{-1} \mathbf{f}_2 \cdot \mathbf{u}_k - \dots - \mathbf{f}_{k-1}(\mathbf{f}_{k-1} \cdot \mathbf{f}_{k-1})^{-1} \mathbf{f}_{k-1} \cdot \mathbf{u}_k\end{aligned}$$

Then $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_k\}$ is an orthogonal set of vectors with the property that

$$\begin{aligned}U_1 &= \text{span}\{\mathbf{f}_1\} \\ U_2 &= \text{span}\{\mathbf{f}_1, \mathbf{f}_2\} \\ U_3 &= \text{span}\{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\} \\ &\vdots \\ U_k &= \text{span}\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_k\}\end{aligned}$$

Example 25.17. Suppose $\mathbf{u}_1 = (1, 0, -1, 1)$, $\mathbf{u}_2 = (1, -2, 1, -1)$ and $\mathbf{u}_3 = (2, 0, -1, 0)$. Use the Gram-Schmidt algorithm to find the orthogonal vectors $\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3$.

Solution.

$$\begin{aligned}\mathbf{f}_1 &= \mathbf{v}_1 = (1, 0, -1, 1) \\ \mathbf{f}_2 &= \mathbf{v}_2 - \mathbf{f}_1(\mathbf{f}_1 \cdot \mathbf{f}_1)^{-1} \mathbf{f}_1 \cdot \mathbf{v}_2 = (1, -2, 1, -1) - \frac{1-1-1}{3}(1, 0, -1, 1) \\ &= \left(\frac{4}{3}, -2, \frac{2}{3}, -\frac{2}{3}\right) \\ \mathbf{f}_3 &= \mathbf{v}_3 - \mathbf{f}_1(\mathbf{f}_1 \cdot \mathbf{f}_1)^{-1} \mathbf{f}_1 \cdot \mathbf{v}_3 - \mathbf{f}_2(\mathbf{f}_2 \cdot \mathbf{f}_2)^{-1} \mathbf{f}_2 \cdot \mathbf{v}_3 \\ &= (2, 0, -1, 0) - \frac{3}{3}(1, 0, -1, 1) - \frac{2}{20/3}\left(\frac{4}{3}, -2, \frac{2}{3}, -\frac{2}{3}\right) \\ &= \left(\frac{3}{5}, \frac{3}{5}, -\frac{1}{5}, -\frac{4}{5}\right)\end{aligned}$$

□

With the Gram-Schmidt algorithm, we can factorise any matrix A as a multiplication of orthogonal matrix and upper triangular matrix. This is called a *QR factorisation* and it is useful when in finding the least square solution of linear algebra problem $A\mathbf{x} = \mathbf{b}$ numerically.

Theorem 25.18. Let A be an $m \times n$ matrix with linearly independent columns. Then A can be factored $A = QR$ where Q is an $m \times n$ matrix with orthonormal columns and R is a square $n \times n$ upper triangular matrix.

Theorem 25.19. The least square solution of the linear equation $A\mathbf{x} = \mathbf{b}$, is a solution of

$$\underbrace{R}_{n \times n} \underbrace{\mathbf{x}}_{n \times 1} = \underbrace{Q^T}_{n \times m} \underbrace{\mathbf{b}}_{m \times 1}$$

where $0 = b'_i$ are ignored from the equation.

Example 25.20. Rewrite Equation (41.14) (refer to Section 25.2 on page 500 in matrix form and use *QR factorisation* to find the least square solution.

Solution: Equation (e:6.1) in matrix form:

$$A\mathbf{x} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ -4 \end{bmatrix} = \mathbf{b} \quad (25.2)$$

Let $A = [\mathbf{v}_1, \mathbf{v}_2]$. Then

$$\mathbf{f}_1 = \mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$$

$$\mathbf{f}_2 = \mathbf{v}_2 - \mathbf{f}_1(\mathbf{f}_1 \cdot \mathbf{f}_1)^{-1} \mathbf{f}_1 \cdot \mathbf{v}_2 = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} - \frac{2}{6} \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} \frac{2}{3} \\ -\frac{4}{3} \\ \frac{1}{3} \end{bmatrix}$$

Let $\mathbf{q}_1 = \mathbf{f}_1/\|\mathbf{f}_1\|$ and $\mathbf{q}_2 = \mathbf{f}_2/\|\mathbf{f}_2\|$, where $\|\mathbf{f}_1\| = \sqrt{6}$, $\|\mathbf{f}_2\| = \sqrt{7/3}$. Then

$$\mathbf{v}_1 = \mathbf{f}_1 = \sqrt{6}\mathbf{q}_1$$

$$\mathbf{v}_2 = \mathbf{f}_2 + \frac{1}{3}\mathbf{f}_1 = \frac{\sqrt{6}}{3}\mathbf{q}_1 + \sqrt{\frac{7}{3}}\mathbf{q}_2$$

Therefore $A = [\mathbf{v}_1, \mathbf{v}_2] = [\mathbf{q}_1, \mathbf{q}_2] \begin{bmatrix} \sqrt{6} & \sqrt{6}/3 \\ 0 & \sqrt{3} \end{bmatrix} = \begin{bmatrix} 1/\sqrt{6} & 2/\sqrt{21} \\ 1/\sqrt{6} & -4/\sqrt{21} \\ 2/\sqrt{6} & 1/\sqrt{21} \end{bmatrix} \begin{bmatrix} \sqrt{6} & \sqrt{6}/3 \\ 0 & \sqrt{7/3} \end{bmatrix}$.

Since $Q^T Q = I$, Equation (e:6.2) becomes

$$\begin{bmatrix} \sqrt{6} & \sqrt{6}/3 \\ 0 & \sqrt{7/3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{6} & 1/\sqrt{6} & 2/\sqrt{6} \\ 2/\sqrt{21} & -4/\sqrt{21} & 1/\sqrt{21} \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ -4 \end{bmatrix}$$

Simplifying it by multiplying first row with $\sqrt{6}$ and second row with $\sqrt{21}$, we have

$$\begin{bmatrix} 6 & 2 \\ 0 & 7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 2 \\ 2 & -4 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ -4 \end{bmatrix}$$

The least square solution is

$$x_2 = \frac{0}{7}, \quad x_1 = \frac{-6}{6} = -1.$$

From the above example, we learn that QR factorisation is very complicated from calculation point of view. However, from computer software point of view, it is a rather good algorithm.

In order to be more familiar with QR factorisation. Let us practise with two more examples.

Example 25.21. Find the QR factorisation of $A = \begin{bmatrix} 2 & 1 \\ 1 & 1 \\ -2 & 0 \end{bmatrix}$.

Solution. Let $A = [\mathbf{a}_1, \mathbf{a}_2]$ and apply the Gram-Schmidt algorithm to these vectors. Then

$$\mathbf{f}_1 = \mathbf{a}_1,$$

$$\mathbf{f}_2 = \mathbf{a}_2 - \mathbf{f}_1(\mathbf{f}_1 \cdot \mathbf{f}_1)^{-1} \mathbf{f}_1 \cdot \mathbf{a}_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} - \frac{3}{9} \begin{bmatrix} 2 \\ 1 \\ -2 \end{bmatrix} = \begin{bmatrix} 1/3 \\ 2/3 \\ 2/3 \end{bmatrix}$$

We now calculate the orthonormal vectors:

$$\mathbf{q}_1 = \frac{\mathbf{f}_1}{\|\mathbf{f}_1\|} = \frac{1}{3} \mathbf{f}_1 = \frac{1}{3} \begin{bmatrix} 2 \\ 1 \\ -2 \end{bmatrix}$$

$$\mathbf{q}_2 = \frac{\mathbf{f}_2}{\|\mathbf{f}_2\|} = \frac{1}{3} \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}.$$

Now we need to express $\mathbf{a}_1$ and $\mathbf{a}_2$ as the linear combination of $\mathbf{q}_1$ and $\mathbf{q}_2$:

$$\mathbf{a}_1 = 3\mathbf{q}_1$$

$$\mathbf{a}_2 = \mathbf{f}_2 + \frac{1}{3}\mathbf{f}_1 = \mathbf{q}_2 + \mathbf{q}_1 = \mathbf{q}_1 + \mathbf{q}_2$$

From here, we obtain the QR factorisation of matrix A :

$$[\mathbf{a}_1, \mathbf{a}_2] = [\mathbf{q}_1, \mathbf{q}_2] \begin{bmatrix} 3 & 1 \\ 0 & 1 \end{bmatrix} \Rightarrow A = \begin{bmatrix} 2 & 1 \\ 1 & 1 \\ -2 & 0 \end{bmatrix} = \underbrace{\begin{bmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \\ -2/3 & 2/3 \end{bmatrix}}_Q \underbrace{\begin{bmatrix} 3 & 1 \\ 0 & 1 \end{bmatrix}}_R$$

□

Example 25.22. Let $A = \begin{bmatrix} 1 & 0 & 3 \\ 0 & 2 & 1 \\ -1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$. Find a QR factorisation of A .

Solution. Let $A = [\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]$. Then

$$\mathbf{f}_1 = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix}, \quad \mathbf{f}_2 = \begin{bmatrix} 0 \\ 2 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{f}_3 = \begin{bmatrix} 2 \\ 0 \\ 2 \\ 0 \end{bmatrix}.$$

From here,

$$\mathbf{q}_1 = \mathbf{f}_1 / \|\mathbf{f}_1\| = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix}, \quad \mathbf{q}_2 = \mathbf{f}_2 / \|\mathbf{f}_2\| = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{q}_3 = \mathbf{f}_3 / \|\mathbf{f}_3\| = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}.$$

Expressing $\mathbf{a}_i$ in terms of $\mathbf{q}_i$, we have

$$\begin{aligned}\mathbf{a}_1 &= \mathbf{f}_1 = \sqrt{2}\mathbf{q}_1 \\ \mathbf{a}_2 &= \mathbf{f}_2 = 2\mathbf{q}_2 \\ \mathbf{a}_3 &= \mathbf{f}_3 + \frac{1}{2}\mathbf{f}_2 + \mathbf{f}_1 = 2\sqrt{2}\mathbf{q}_3 + \mathbf{q}_2 + \sqrt{2}\mathbf{q}_1\end{aligned}$$

Therefore, the QR factorisation is

$$A = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 & 1 \\ 0 & \sqrt{2} & 0 \\ -1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 & \sqrt{2} \\ 0 & 2 & 1 \\ 0 & 0 & 2\sqrt{2} \end{bmatrix}.$$

□

25.5 Orthogonal Diagonalisation of Symmetric Matrices

Symmetric matrix is the most important type of matrix in application. From Section 25.2, it appears in statistics (linear regression) as normal equations. In engineering problems, the discretisation of differential equations lead to symmetric matrix. Therefore, it is important to study the properties of symmetric matrices.

To study the properties of symmetric matrices, we ask: What is special about $A\mathbf{x} = \lambda\mathbf{x}$ when A is symmetric? From Chapter 22, we know that this is closely related to the concept similarity matrix:

$$A = P\Lambda P^{-1} \Rightarrow A^T = (P\Lambda P^{-1})^T = (P^{-1})^T \Lambda P^T = A.$$

Notice that $\Lambda^T = \Lambda$ since it is diagonal. We would suspect that maybe $P^T = P^{-1}$ and $(P^{-1})^T = P$. Mathematicians have proved that this is indeed the case. By definition, $P^T P = P^{-1} P = I$, therefore P is an orthogonal matrix. Hence, we introduce the following definition.

Definition 25.8. An $n \times n$ matrix A is *orthogonal diagonalisable* if there exists an orthogonal matrix Q such that $Q^T A Q = D$ where D is a diagonal matrix.

Mathematicians have discovered the following theorem which says that they are a lot of orthogonal diagonalisable matrices.

Theorem 25.23 (Spectral Theorem / Principal Axis Theorem). *A symmetric matrix is orthogonal diagonalisable. Furthermore,*

1. *symmetric matrix has real eigenvalues;*
2. *the eigenvectors of a symmetric matrix is orthogonal.*

The proof of the above algorithm is closely related to eigenvalues and eigenvectors of the symmetric matrix. We will investigate this in the following two examples.

Example 25.24. Let $A = \begin{bmatrix} 1 & -2 \\ -2 & -2 \end{bmatrix}$. Find an orthogonal matrix Q such that $Q^T A Q$ is a diagonal matrix.

Solution: The eigenvalues of A are $-3, 2$.

The eigenspace of A corresponding to $\lambda = -3$ is $\left\{ \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix} t \mid t \in \mathbb{R} \right\}$.

The eigenspace of A corresponding to $\lambda = 2$ is $\left\{ \begin{bmatrix} -2 \\ 1 \end{bmatrix} t \mid t \in \mathbb{R} \right\}$.

Let $\mathbf{f}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\mathbf{f}_2 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$. Note that $\mathbf{f}_1$ and $\mathbf{f}_2$ are orthogonal vectors. The orthonormal vectors are $\mathbf{q}_1 = \frac{\mathbf{f}_1}{\|\mathbf{f}_1\|}$, $\mathbf{q}_2 = \frac{\mathbf{f}_2}{\|\mathbf{f}_2\|}$ and the orthogonal matrix

$$Q = [\mathbf{q}_1, \mathbf{q}_2] = \begin{bmatrix} \frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{bmatrix}$$

whose columns are eigenvectors corresponding to -3 and 2 , respectively. Thus

$$Q^T A Q = \begin{bmatrix} -3 & 0 \\ 0 & 2 \end{bmatrix}.$$

Example 25.25. Let $A = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. Find an orthogonal matrix Q such that $Q^T A Q$ is a diagonal matrix.

Solution. After some calculations, we can find that the eigenvalues are

$$\lambda = 1, 1, 3.$$

The eigenspace corresponding to 1 is

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -s \\ s \\ t \end{bmatrix} = s \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

The eigenspace corresponding to 3 is

$$\begin{bmatrix} -1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & -2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} t \\ t \\ 0 \end{bmatrix}$$

Therefore, the orthogonal matrix is

$$Q = \begin{bmatrix} 0 & -1/\sqrt{2} & 1/\sqrt{2} \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad Q^T A Q = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

□

We remark that QR factorisation and Cholesky method (it is similar to LU factorisation with $U = L^T$) are two classical methods for solving the normal equation associated with the least-square problem in Section [25.2](#). [sec:least-square-problem](#)

Finally, we conclude this subject by listing some classical linear algebra problems:

$$A\mathbf{x} = \mathbf{b}, \quad A \in M_{m,n}(\mathbb{R}), \quad \mathbf{b} \in \mathbb{R}^m, \quad r = \text{rank}(A).$$

1. Compute the *leverage scores* of A , i.e. the L_2 -norm of the standard basis vectors when projected onto the subspace spanned by the columns of A .

2. *Least squares regression*¹: Compute $\hat{\mathbf{x}} = \underset{\mathbf{x} \in \mathbb{R}^n}{\operatorname{argmin}} \|A\mathbf{x} - \mathbf{b}\|_2$.
3. *Low-rank approximation*: Given $1 \leq k \leq n$, compute $A_k = \underset{\operatorname{rank}(B) \leq k}{\operatorname{argmin}} \|A - B\|_F$.
4. *Preconditioning*: Compute $R \in M_n(\mathbb{R})$ so that $\|AR\mathbf{x}\|_2 \approx \|\mathbf{x}\|_2$ for every $\mathbf{x} \in \mathbb{R}^n$.

A modern method to solve to the above problems is based on the singular value decomposition (SVD):

$$A = U\Sigma V^T, \quad r = \operatorname{rank}(A), \quad U \in M_{m,r}(\mathbb{R}), \quad \Sigma \in M_r(\mathbb{R}), \quad V \in M_{n,r}(\mathbb{R}).$$

as follows:

1. *Leverage scores*: Output row norms of U .
2. *Least squares regression*: $\hat{\mathbf{x}} = V\Sigma^{-1}U^T\mathbf{b}$.
3. *Low-rank approximation*: $A_k = U\Sigma_k V^T$.
4. *Preconditioning*: $R = V\Sigma^{-1}$.

The concept of *leverage scores* was introduced in 1978 by Hoaglin and Welsh in “The Hat matrix in regression and ANOVA”, Amer. Statist., 32 (1978), pp. 17–22. The purpose is to detect outliers when computing regression diagnostics, the *hat matrix* is the projection matrix mentioned in Remark 25.4. The diagonal elements of the hat matrix are called *leverage scores* of A , and is denoted by $\ell_j(A)$, $1 \leq j \leq m$. The value $\ell_j(A)$ reflects the leverage of the j th point b_j on the corresponding fit $\hat{b}_j$.

As mentioned earlier, symmetric matrix is important because it has many applications in statistics (e.g. covariant matrix), engineering (e.g. finite difference equation of partial differential equation), etc. According to Strang and Borre [1997, p. 249], when A is symmetric, all its eigenvalues are real and the eigenvectors are orthogonal. This implies that A is *orthogonal diagonalisable*, i.e. there is an orthogonal matrix Q such that $Q^T A Q = D$ where D is a diagonal matrix. This leads to an important theorem below.

Theorem 25.26 (Spectral Theorem / Principal Axis Theorem). *A symmetric matrix is orthogonal diagonalisable with real eigenvalues and orthogonal eigenvectors.*

The spectral theorem is the foundation of the *QR algorithm* (developed in the late 1950s) for finding the eigenvalues and eigenvectors of a symmetric matrix [Francis, 1961], [Francis, 1962], [Kublanovskaya, 1962]:

1. Set $A_0 = A$, $U_0 = I$
2. for $k = 1, 2, \dots$ do
 - (a) $A_{k-1} = Q_k R_k$ (Perform *QR* factorisation)
 - (b) $A_k = R_k Q_k$
 - (c) $U_k = U_{k-1} Q_k$ (Update the transform matrix)
3. Set $T := A_\infty$ and $U := U_\infty$.
4. Return the Schur decomposition of $A = UTU^*$.

¹We can have a general L_p regression: Compute $\hat{\mathbf{x}} = \underset{\mathbf{x} \in \mathbb{R}^n}{\operatorname{argmin}} \|A\mathbf{x} - \mathbf{b}\|_p$. However, it is much more difficult to obtain a solution.

Part IV

Multivariable Calculus

Chapter 26

Vector Calculus

Vector calculus works Euclidean spaces $\mathbb{R}^n$ (Section [20.6](#)) from nonlinear (rather than linear as in Chapter [24](#)) perspective.

I have not taught vector calculus before, so I need to work on the notes when I have the time. I will follow the following references:

fitzpatrick-AnalyticalClassicalDynamics.pdf
GeraldJaySussman-JackWisdom-FunctionalDifferentialGeometry-MIT2013
StevenGKrantz-HaroldRParks-APrimerOfRealAnalyticFunctions2e-2002
RobertSStichartz-TheWayOfAnalysis-Bartlett2000
MarianMuresan-AConcreteApproachToClassicalAnalysis-Springer2008
JohnDePree-CharlesSwartz-IntroductionToRealAnalysis-Wiley1988
LangS-UndergraduateAnalysis-2005
SergeLang-LynnH.Loomis-CalculusOfSeveralVariables
A Course of Higher Mathematics vol 2 - V. Smirnov
TomApostol-Calculus-v1, TomApostol-Calculus-v2
ZorichVA-MathematicalAnalysisI, II 2004
JerroldMarsden-ElementaryClassicalAnalysis-Freeman1974
ACompanionToAnalysisASecondFirstAndFirstSecondCourseInAnalysis-Draft-AMS2004
RudinW-PrinciplesOfMathematicalAnalysis3e
IvarEkeland-ExteriorDifferentialCalculusAndApplicationsToEconomicTheory-2007
EfeAOk-RealAnalysisWithEconomicApplications-PUP2007

26.1 Euclidean Space

Real numbers and complex numbers are regarded as scalars and vectors is a generalisation of Definition [20.16](#) such that it has the property of *direction* and *magnitude* (or *norm*).

Euclidean space (Section [20.6](#)) is a natural place to hold vectors. A vector is usually denoted by $\vec{v}$ (letter with an over-arrow), or $\mathbf{v}$ (bold face type letter); whereby its magnitude is denoted by $\|\vec{v}\|$ or $\|\mathbf{v}\|$.

A summary of the properties of real Euclidean spaces (Section [20.6](#)) is given below.

For any $m, n \in \mathbb{R}$ and any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$,

1. $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$
2. $\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z}$
3. $m\mathbf{x} = \mathbf{x}m$
4. $m(n\mathbf{x}) = (mn)\mathbf{x}$
5. $(m + n)\mathbf{x} = m\mathbf{x} + n\mathbf{x}$
6. $m(\mathbf{x} + \mathbf{y}) = m\mathbf{x} + m\mathbf{y}$

7. $\mathbf{x} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{x}$
8. $\mathbf{x} \cdot (\mathbf{y} + \mathbf{z}) = \mathbf{x} \cdot \mathbf{y} + \mathbf{x} \cdot \mathbf{z}$
9. $m(\mathbf{x} \cdot \mathbf{y}) = (m\mathbf{x}) \cdot \mathbf{y} = \mathbf{x} \cdot (m\mathbf{y}) = (\mathbf{x} \cdot \mathbf{y})m$ for $m \in \mathbb{R}$

Definition 26.1. A **unit vector** $\mathbf{v}$ is a vector whose magnitude, $\|\mathbf{v}\|$, is 1.

Definition 26.2. The vectors $\mathbf{x}$ and $\mathbf{y}$ are *perpendicular*, denoted by $\mathbf{x} \perp \mathbf{y}$ if $\mathbf{x} \cdot \mathbf{y} = 0$.

26.1.1 Special Case $n = 3$: $\mathbb{R}^3$

Definition 26.3. The **cross** or **vector product** of $\mathbf{A}$ and $\mathbf{B}$ is a **vector** $\mathbf{z} = \mathbf{A} \times \mathbf{B}$ (read $\mathbf{A}$ cross $\mathbf{B}$). The magnitude of the cross product $\|\mathbf{z}\| = |\mathbf{A}| |\mathbf{B}| \sin \theta$. The direction of $\mathbf{z}$ is perpendicular to the plane containing $\mathbf{A}$ and $\mathbf{B}$ such that $\mathbf{A}, \mathbf{B}, \mathbf{z}$ form a right-handed system. In symbols $\mathbf{A} \times \mathbf{B} = AB \sin \theta \hat{\mathbf{n}}$, $0 \leq \theta \leq \pi$, where $\hat{\mathbf{n}}$ is a unit vector indicating the direction of $\mathbf{A} \times \mathbf{B}$.

Remark 26.1. Let $k \in \mathbb{R}^*$, if $\mathbf{A} = k\mathbf{B}$ (i.e. $\mathbf{A}$ is parallel to $\mathbf{B}$), then $\theta = 0$ or π , $\sin \theta = 0$, and we have $\mathbf{A} \times \mathbf{B} = 0$.

Let $\mathbf{i}, \mathbf{j}$ and $\mathbf{k}$ be the standard basis of $\mathbb{R}^3$ (Definition [20.8](#) def:basis-std-3). Some laws on cross product:

1. $\mathbf{A} \times \mathbf{B} = -(\mathbf{B} \times \mathbf{A}) = (-\mathbf{B}) \times \mathbf{A} = \mathbf{B} \times (-\mathbf{A})$
2. $\mathbf{A} \times (\mathbf{B} + \mathbf{C}) = \mathbf{A} \times \mathbf{B} + \mathbf{A} \times \mathbf{C}$
3. $m(\mathbf{A} \times \mathbf{B}) = (m\mathbf{A}) \times \mathbf{B} = \mathbf{A} \times (m\mathbf{B}) = (\mathbf{A} \times \mathbf{B})m$ for $m \in \mathbb{R}$
4. $\mathbf{i} \times \mathbf{i} = \mathbf{j} \times \mathbf{j} = \mathbf{k} \times \mathbf{k} = 0$
5. $\mathbf{i} \times \mathbf{j} = \mathbf{k}, \mathbf{j} \times \mathbf{k} = \mathbf{i}, \mathbf{k} \times \mathbf{i} = \mathbf{j}$ (cyclic rule)
6. If $\mathbf{A} = (a_1, a_2, a_3)$, $\mathbf{B} = (b_1, b_2, b_3)$, then $\mathbf{A} \times \mathbf{B} = \mathbf{i} \begin{vmatrix} a_2 & a_3 \\ b_2 & b_3 \end{vmatrix} - \mathbf{j} \begin{vmatrix} a_1 & a_3 \\ b_1 & b_3 \end{vmatrix} + \mathbf{k} \begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix} =:$
 $\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix}$ (Definition [20.12](#) def:cross-product)

7. The magnitude of $\mathbf{A} \times \mathbf{B}$ is the same as the area of a parallelogram with sides $\mathbf{A}$ and $\mathbf{B}$.

Definition 26.4. The combinations of dot and cross multiplication of three vectors $\mathbf{A}, \mathbf{B}, \mathbf{C}$ gives the **scalar triple products** $\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C})$ and $(\mathbf{A} \times \mathbf{B}) \cdot \mathbf{C}$ and the **vector triple products** $\mathbf{A} \times (\mathbf{B} \times \mathbf{C})$ and $(\mathbf{A} \times \mathbf{B}) \times \mathbf{C}$.

Sometimes, the parentheses (or brackets) of $\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C})$ are omitted and it is written as $\mathbf{A} \cdot \mathbf{B} \times \mathbf{C}$ for simplicity.

Some important laws:

1. cyclic rule: $\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) = \mathbf{B} \cdot (\mathbf{C} \times \mathbf{A}) = \mathbf{C} \cdot (\mathbf{A} \times \mathbf{B})$
2. The (directional) volume of a parallelepiped having $\mathbf{A}, \mathbf{B}, \mathbf{C}$ as edges is $|\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C})|$.
3. For nonzero vectors $\mathbf{A}, \mathbf{B}, \mathbf{C}$, if $\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) = 0$, $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are coplanar, i.e. they lie on a same plane.

4. If $\mathbf{A} = (a_1, a_2, a_3)$, $\mathbf{B} = (b_1, b_2, b_3)$ and $\mathbf{C} = (c_1, c_2, c_3)$, then

$$\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) = \begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix}$$

5. $\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) \neq (\mathbf{A} \times \mathbf{B}) \times \mathbf{C}$ in general

6. $\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = (\mathbf{A} \cdot \mathbf{C}) \mathbf{B} - (\mathbf{A} \cdot \mathbf{B}) \mathbf{C}$

7. $(\mathbf{A} \times \mathbf{B}) \times \mathbf{C} = (\mathbf{A} \cdot \mathbf{C}) \mathbf{B} - (\mathbf{B} \cdot \mathbf{C}) \mathbf{A}$

Definition 26.5. Consider the vector $\overrightarrow{OA} = \mathbf{a}$ with an initial point O , and a terminal point $A(a_1, a_2, a_3)$. The direction of $\mathbf{a}$ can be determined by the **directional angles**:

$$\alpha = \angle AOX, \quad \beta = \angle AOY, \quad \gamma = \angle AOZ$$

specified by OA and the three axes in a standard basis $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$.

The directional angles can be calculated using the **direction cosines** of $\mathbf{a}$:

$$\cos \alpha = \frac{a_1}{\|\mathbf{a}\|}, \quad \cos \beta = \frac{a_2}{\|\mathbf{a}\|}, \quad \cos \gamma = \frac{a_3}{\|\mathbf{a}\|}.$$

Note: $\cos^2 \alpha + \cos^2 \beta + \cos^2 \gamma = 1$ and the ratio $a_1 : a_2 : a_3$ is called the **direction ratio** for vector $\mathbf{a}$.

26.2 Parametric Curves

- Introduction to parametric and polar curves
- Calculus with parametric curves
- Areas and lengths of parametric and polar curves
- Arc lengths and areas of smooth curves
- Conic sections in polar coordinates

Definition 26.6. A *parametric curve* in $\mathbb{R}^n$ is a mapping

$$C : \mathbb{R} \rightarrow \mathbb{R}^n, \quad t \mapsto (p_1(t), \dots, p_n(t))$$

where $p_i : \mathbb{R} \rightarrow \mathbb{R}$ are some functions.

Example 26.2. A *straight line* is a parametric curve with the form below:

$$L : \mathbb{R} \rightarrow \mathbb{R}^n, \quad t \mapsto (a_1 + b_1 t, \dots, a_n + b_n t).$$

Example 26.3. Find the equation of the line that passes through points $A(-1, 5, 2)$ and $B(1, 2, 3)$.

***Figure Eg 7

$$\overrightarrow{AB} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} - \begin{pmatrix} -1 \\ 5 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ -3 \\ 1 \end{pmatrix} \equiv \text{direction of the line}$$

$$\mathbf{r} = \mathbf{a} + \lambda \mathbf{b} \quad \Rightarrow \quad \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -1 \\ 5 \\ 2 \end{pmatrix} + \lambda \begin{pmatrix} 2 \\ -3 \\ 1 \end{pmatrix}$$

$$\left. \begin{array}{l} x = -1 + 2\lambda \\ y = 5 - 3\lambda \\ z = 2 + \lambda \end{array} \right\} \quad \lambda = \frac{x+1}{2} = \frac{y-5}{-3} = \frac{z-2}{1}$$

$$\text{or } \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + \lambda \begin{pmatrix} 2 \\ -3 \\ 1 \end{pmatrix} \text{ vector form}$$

$$\left. \begin{array}{l} x = 1 + 2\lambda \\ y = 2 - 3\lambda \\ z = 3 + \lambda \end{array} \right\} \quad \lambda = \frac{x-1}{2} = \frac{y-2}{-3} = \frac{z-3}{1}$$

parametric form

Cartesian form

Example 26.4. Find the line that passes point $(2, 3, 0)$ and perpendicular to both $\mathbf{a} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$

and $\mathbf{b} = \begin{pmatrix} 3 \\ 4 \\ 5 \end{pmatrix}$.

***Figure Eg 8

$$\mathbf{n} = \mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & 2 & 3 \\ 3 & 4 & 5 \end{vmatrix} = \begin{pmatrix} -2 \\ 4 \\ -2 \end{pmatrix} = 2 \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}$$

$$\text{line: } \mathbf{r} = \mathbf{a} + \lambda \mathbf{b} \quad \equiv \quad \mathbf{r} = \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} + \lambda \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix} \quad \text{or} \quad \mathbf{r} = \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} + \beta \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$$

26.3 Multivariable Functions

TO WORK ON:

- Vector Functions
- Derivatives and Integrals of Vector Functions
- The TNB Frame and Curvature
- Motion in Space
- Cartesian, Cylindrical and Spherical Coordinates
- Change of coordinate systems: polar, spherical, cylindrical and Jacobian
- Review of Triple Integrals in Cartesian Coordinates
- Triple Integrals involving Cylindrical and Spherical Coordinates
- Vector Fields
- Line Integrals
- The Fundamental Theorem for Line Integrals
- Green's Theorem
- Divergence and Curl: Vector Forms of Green's Theorem
- Parametric Surfaces and Surface Integrals
- Stokes' Theorem
- The Divergence Theorem

- Orthogonal Curvilinear Coordinates
- Applications of Vector Calculus

The generalisation of function to a subset $D \subset \mathbb{R}^n$ is the **multivariable function**:

$$f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}, \quad (x_1, \dots, x_n) \mapsto f(x_1, \dots, x_n) \quad (26.1)$$

e:multi-fn

A function f can be regarded as a **scalar field** on D . In the language of physics, if we regard $(x_1, \dots, x_n)$ as a point in D , then $f(x_1, \dots, x_n)$ is the “potential” (or “density” or “temperature” or “pressure”) at the point $(x_1, \dots, x_n)$.

In the language of statisticians, the function (26.1) will be expressed as

$$y = f(x_1, \dots, x_n)$$

and the variables $x_1, \dots, x_n$ are called **independent variables** or **factors**, while y is called the **dependent variable** or **response**.

Example 26.5. The following are some functions on $\mathbb{R}^2$:

- $(x_1, x_2) \mapsto y = 3x_1^2 + 4x_2^3$
- $(x_1, x_2) \mapsto y = \tan(x_1 x_2) + e^{x_1 + 2x_2}$

They can be expressed in the Scheme programming language as follows:

```
1 (define (f1 x1 x2)
2   (+ (* 3 (* x1 x1) (* 4 (* x2 x2 x2)))))
3 (define (f2 x1 x2)
4   (+ (tan (* x1 x2)) (exp (+ x1 (* 2 x2)))))
```

Example 26.6. The following is a function on $\mathbb{R}^3$:

$$(x_1, x_2, x_3) \mapsto (x_2^2 - x_3) \sin(x_1)$$

Example 26.7. The following is a function on $\mathbb{R}^4$:

$$(x_1, x_2, x_3, x_4) \mapsto (x_1^2 + x_2^2 + x_3^2) e^{-x_4}$$

Example 26.8. Some multivariable functions from physics:

$$\begin{aligned} T &= \frac{1}{2} m v^2 \text{ (kinetic energy),} \\ V &= x y z \text{ (volume of a cuboid)} \\ R &= \frac{\sum_{i=1}^N m_i r_i}{\sum_{i=1}^N m_i} \text{ (center of mass),} \\ Z &= \sum_i e^{-E_i/kT} \text{ (partition function).} \end{aligned}$$

26.3.1 Graphing two-variable functions in $\mathbb{R}^3$

This section is contributed by Dr Lem Kok Hong.

The graph of a two-variable function $f(x, y)$ is the set of all points (x, y, z) such that $z = f(x, y)$. Generally, the graph of a two-variable function is a surface in 3-space. The domain of a two-variable function is a set of points (or duplets) in the xy plane such as $(0, 0)$, $(1, 2)$; points on the line $y = 2x$; points inside the unit circle $x^2 + y^2 = 1$. For instance, the domain of function $z = \sqrt{1 - x^2 - y^2}$ is the set of (x, y) points satisfying $x^2 + y^2 < 1$, i.e. points inside the unit circle.

The three mutually perpendicular coordinate axes meet at the origin: a vertical z -axis, and two horizontal x - and y - axes form a 3-dimensional rectangular or Cartesian coordinate system. The order of the x -, y - and z - axes follows the right-hand rule.

Each pair of coordinate axes form a plane, called a coordinate plane. The three coordinate planes are the xy -plane (with equation $z = 0$); the xz -plane (with equation $y = 0$) and the yz -plane. The coordinate planes divide 3-space into eight octants. The first octant is the one for which the three coordinates are positive.

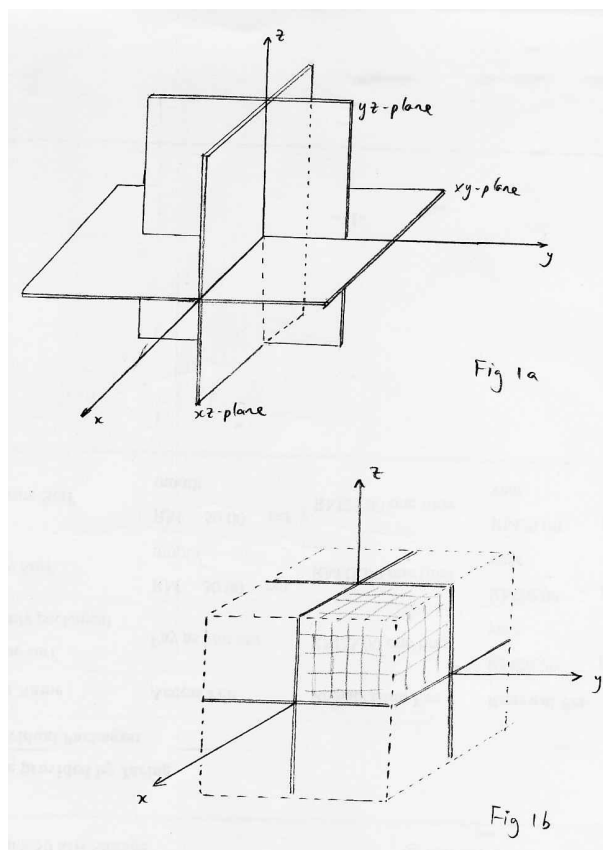


Fig1: The xy ; xz and yz planes
The first octant is darken

We specify a point P by giving its coordinates in form of the ordered triple (x, y, z) . The coordinates instruct us how to reach to the point P : start at the origin, go x units along the x -axis, then y units in the direction parallel to the y -axis and finally z units in the direction parallel to the z -axis. The following figure shows the coordinates $(2, 3, 5)$ and $(-1, 2, -5)$.

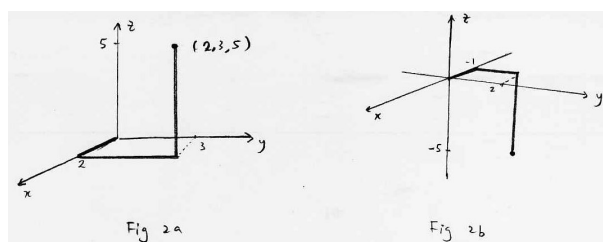


Fig2
a) $(2, 3, 5)$
b) $(-1, 2, -5)$

26.3.2 Linear function

Linear function has the general form $z = px + qy + r$ where some of the constants p , q , and r can be zero. For graphing purpose, a more useful form is obtained after some algebraic rearrangement:

$$ax + by + cz = 1. \quad (26.2) \quad \text{linearfn}$$

The graphs of linear functions yield planes in space. The graph of equation (26.2) is a plane which passes through $1/a$ at the x -axis, $1/b$ at the y -axis, $1/c$ at the z -axis as shown in the following figure.

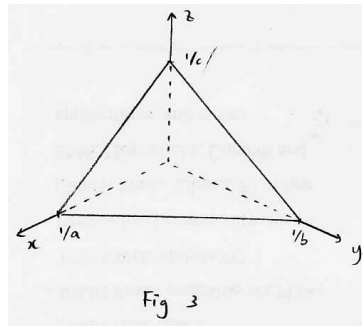


Fig3

For example, the function $z = 6x - 8y + 2$, after dividing both sides by two, may be rearranged as $-3x + 4y + \frac{1}{2}z = 1$. Its graph can be sketched easily as below.

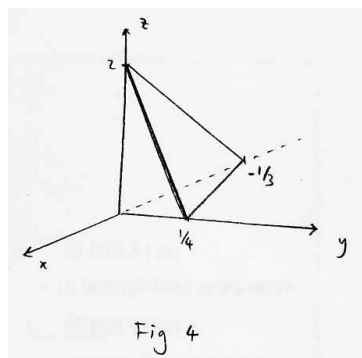


Fig4

Situation where some variables are missing

In the case of missing one variable (there is a zero among a, b, c), the plane will not pass through the corresponding (missing) axis but expand parallel with it. For example, $\frac{1}{3}x + y = 1$ passes through 3 at the x -axis and 1 at the y -axis and expand parallel with the z -axis as shown below.

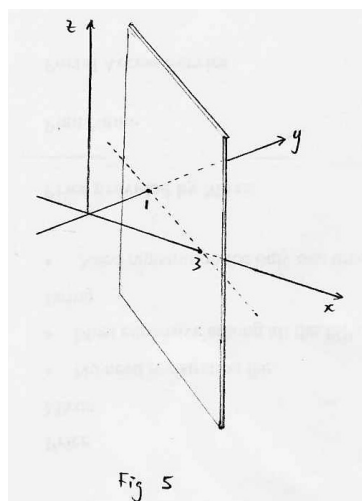


Fig 5

When any of the two variables in $ax + by + cz = 1$ is missing, the equation takes one of these forms: $x = k$, $y = k$, $z = k$. The graph of these three equations are planes which parallel with the yz -plane, xz -plane and xy -plane respectively. A few examples are shown in the following figures.

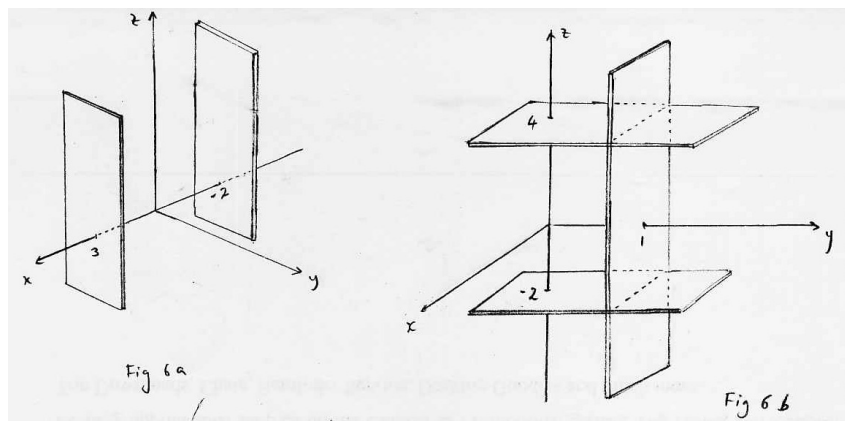


Fig 6a: $x=3, x=-2$
 Fig 6b: $y=1; z=-2, z=4$

26.3.3 Visualizing and sketching graphs of two-variable (nonlinear) functions

The ability to visualize and sketch graphs of some basic two-variable functions is important in vector calculus(our next chapter). The graphs of two-variable functions are generally surfaces in 3-space. For functions of three or more variables, their graphs are hypersurfaces in four or higher dimensional space. Hypersurfaces cannot be sketched and are mainly studied theoretically.

Cross-section, level curve, level surface, contour curve, contour surface

The following concepts are essential prior to sketching of two-variable functions graphs . Consider a two-variable function $f(x, y)$. The **cross-section** of f with x fixed is the function obtained when x is held fixed while y is varying. The cross-section of f with y fixed is defined similarly. If we hold x at k , then the graph of the cross-section is the intersecting curve between the graph of f and the plane $x = k$. The **level curve**(also called **contour curve**) of function f is the locus or curve satisfying $f = \text{constant}$.

Although the graphs of three-variable function are four dimensional hypersurface that we cannot sketch, we may lower one dimension to sketch their **level surfaces**(also named **contour surfaces**: $f(x, y, z) = \text{constants}$. Level surfaces help us visualize the functions better. One good illustration of level surfaces are the surfaces of constant gravitational potential(approximately spheres) about the earth. Surfaces of constant temperature or pressure also illustrate this.

To visualize or sketch graphs of a two-variable function, one usually take turns substituting constants to the variables x, y and z . This give us the cross-sections and the level curves of the function. These are then combined using some imagination to give us the whole picture of how the graph looks like. This process is illustrated by the following examples.

Egb1

Example 26.9. Consider the function $z(x, y) = x^2 + y^2$. Letting x, y equal zero yields two parabolas $z = y^2$ and $z = x^2$ respectively. These two cross-sections are shown in **figure 7a**. Letting z equal 1 and 4 yield the two level curves $x^2 + y^2 = 1$ and $x^2 + y^2 = 4$. They are just circles of radius one and two as shown in **figure 7b**. Combining these information, it is not difficult to imagine that the graph is a bowl-like surface(called paraboloid) as indicated in **figure 7c**.

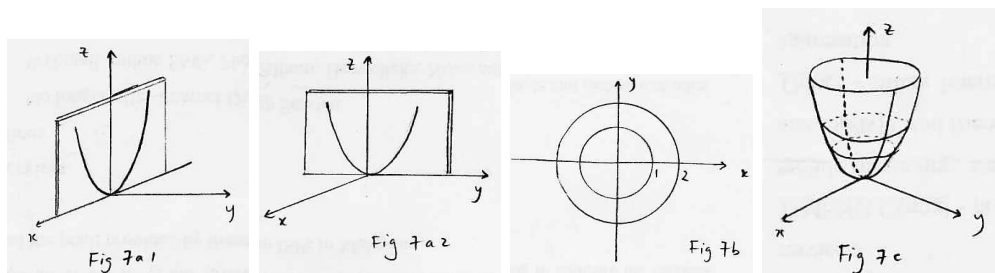


Fig 7a: corss-sections

Fig 7b: lever curves
Fig 7c: combined picture, paraboloid

Plotting of the graphs, cross-sections, level-curves and even the contour diagram of two-variable functions can now be done easily using many mathematical softwares such as Matlab, Maple and Mathematica, available in the market. **Contour diagram** is the projection of a collection of level curves onto the xy -plane. One can also imagine that a contour(or level curve) is obtained by slicing the the surface $z = f(x, y)$ with a horizontal plane with equation $z = k$. A few more relevant plots of $z = x^2 + y^2$ is presented below for better visualization.

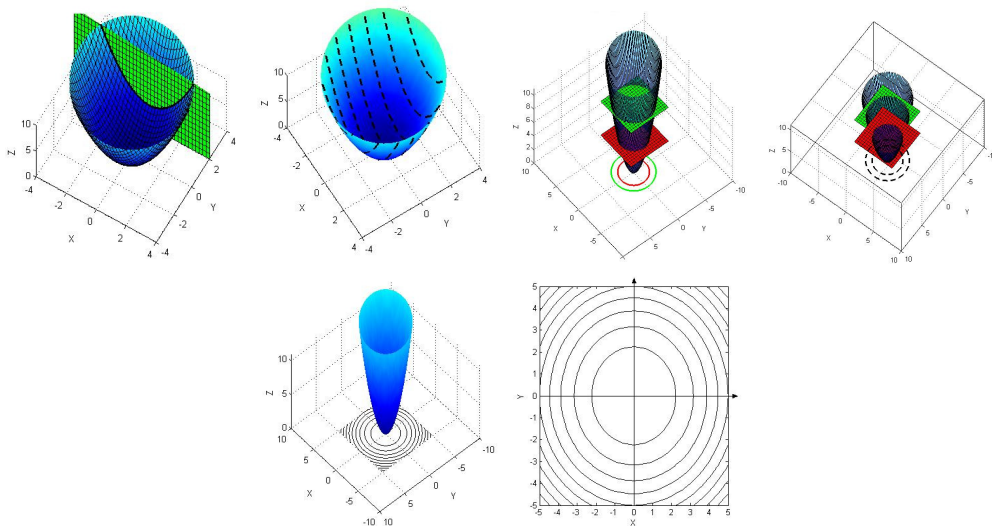


Fig 8a: corss-sections
Fig 8b: lever curves
Fig 8c: contour diagrams

□

Egb2

Example 26.10. Consider next the function $z(x, y) = y^2 - x^2$. Letting $y = 0$ yields $z = -x^2$ which is concave downward while letting $x = 0$ yields $z = y^2$ which is concave upward as indicated in **Fig 9a**. We further let $y = \pm 1, y = \pm 2, \dots, y = \pm 6$ to obtain $z = 1 - x^2$, $z = 4 - x^2$ and so on, see **Fig 9b**. To visualize better, we add a few more cross-sections by letting $x = \pm 1, x = \pm 2, \dots, x = \pm 4$. A sketch of all the cross-sections are shown in **Fig 9c**. The full picture of the graph of z is shown in **Fig 9d**. Note that here the turning point $(0, 0, 0)$ is called the **saddle point** which bends up in one direction and bends down in a different direction. The situation is reminiscent of a saddle, a seat for the rider of a horse, therefore such a name is given.

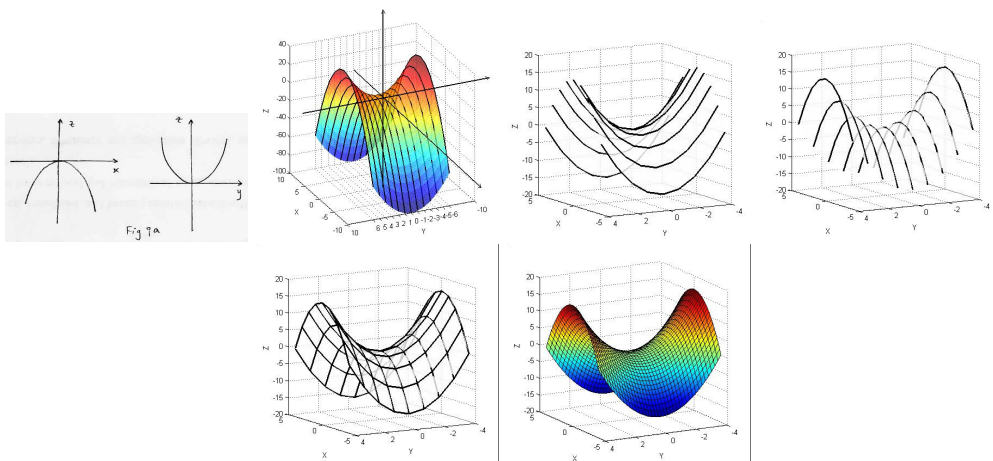


Fig 9a: corss-sections
Fig 9b: more cross-sections

Fig 9c: cross-sections of two directions

Fig 9d: the full picture

□

26.3.4 Graphs of some common non-linear functions

In studying multivariable calculus, surfaces in 3-space will be encountered quite often. Surfaces defined by equation of the following are common.

$$Ax^2 + By^2 + Cz^2 + Dx + Ey + Fz = G \quad (26.3)$$

quadrics

where some of the constants $A, B, C \dots, G$ can be zero. Such surfaces are generally known as **quadric surfaces**. They are listed briefly here for later references.

I. Sphere

$$x^2 + y^2 + z^2 = r^2$$

The sphere is centered at the origin and having radius r .

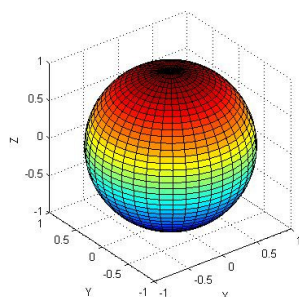


Fig 10: Sphere

II. Ellipsoid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$$

It is centered at the origin and symmetric about the x , y and z coordinate planes. The numbers a , b and c are the semi-axes of the ellipsoid. In the special case where $a = b = c = r$, then it is a sphere centered at the origin with radius r .

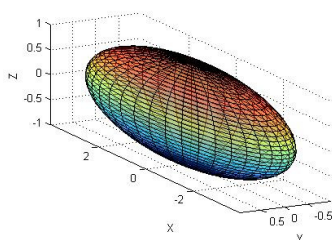


Fig 11: Ellipsoid

III. Quadric cone

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = z^2$$

The surface passes $(0,0,0)$ and is symmetric about the three coordinate planes. The level curves(setting $z = r$) are the circles $x^2 + y^2 = r^2$. The cross-section with the yz -plane(setting $x = 0$) are the lines $z = \pm \frac{1}{b}y$.

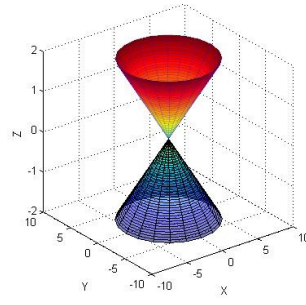


Fig 12: Quadric cone

IV. Hyperboloid of one sheet

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1$$

The equation can be rearranged in form of $z = \pm c\sqrt{\frac{x^2}{a^2} + \frac{y^2}{b^2} - 1}$ which implies x, y should be larger than certain values to avoid square root of negative. The level curve, by setting $z = 0$, is the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$. Note that as both x and y approach infinity, the effect of the ‘-1’ inside the square root is suppressed and the equation approaches the quadric cone form: $z^2 = \frac{x^2}{\alpha^2} + \frac{y^2}{\beta^2}$.

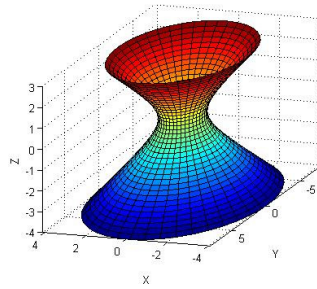


Fig 13: Hyperboloid of one sheet

V. Hyperboloid of two sheets

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = -1$$

The equation can be rearranged in form of $z = \pm c\sqrt{\frac{x^2}{a^2} + \frac{y^2}{b^2} + 1}$ which implies at least $|z| > c$ for any x, y values. Hence the surface appears in two sheets. Again as $x, y \rightarrow \infty$, the quadric cone form $z^2 = \frac{x^2}{\alpha^2} + \frac{y^2}{\beta^2}$ will dominate.

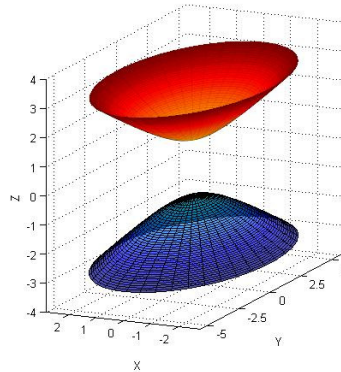


Fig 14: Hyperboloid of two sheet

VI. Elliptic paraboloid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = z$$

Since $z \geq 0 \forall x, y$, the surface is unbounded above the xy -plane but will never extend below it. The level curves are ellipses while the cross-sections with other coordinate planes are parabolas curving up.

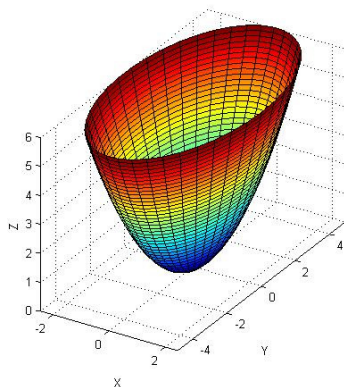


Fig 15: Elliptic paraboloid

VII. Hyperboloid paraboloid

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = z$$

The cross-sections with planes parallel to the xz -planes are curving up parabolas whereas the cross-sections with planes parallel to the yz -planes are parabolas that curve down. The surface takes the shape of a saddle – seat for a horse rider. The origin is a minimum point viewed in y direction but is a maximum point viewed in x direction. It is termed ‘saddle point’ of the surface.

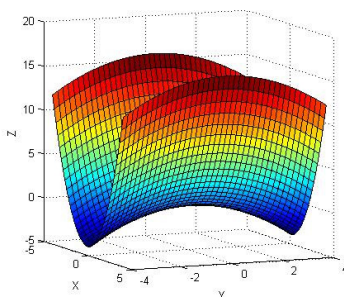


Fig 16: Hyperboloid paraboloid

Missing one variable situations

In particular when the variable z is missing, equation (26.3) ^{quadrics} takes the form

$$Ax^2 + By^2 + Cx + Dy = E, \quad (26.4)$$

cylind

where some of the constants $A, B, \dots, E$ can be zero. Quadric surfaces defined by this equation are generally known as **cylinders** ^{cylind} (quadric cylinders). Equation (26.4) involving only two variables form the **base curves** of a cylinder. They lie in the xy -plane or in planes parallel to the xy -plane. The missing variable z is left unrestricted, it varies freely up and down to sweep out the ‘wall’ of a cylinder. Similar concept applies to situation where other variable (x or y) is missing instead of variable z . We list three most common types of quadric cylinders for later references.

I. Parabolic cylinder

$$y = ax^2$$

It looks like a parabolically bent lamina which stands vertically. All the level curves are described by the equation $y = ax^2$.

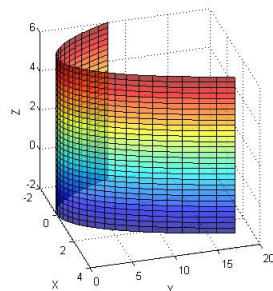


Fig 17: Parabolic cylinder

II. Elliptic cylinder

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

The elliptic cylinder has a parabolic level curve (or cross-section in general). If $a = b = r$, it forms an ordinary right circular cylinder of radius r having circular level curves.

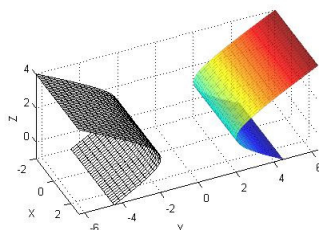


Fig 18: Elliptic cylinder (LiewHH: The above diagram is wrong)

III. Hyperbolic cylinder

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

This equation can be rearranged as $y = \pm \frac{b}{a} \sqrt{x^2 - a^2}$. It requires $|x| > a$ to avoid complex. So the surface has two parts, on the positive and negative sides of the x -axis respectively.

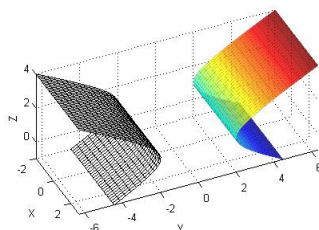


Fig 19: Hyperbolic cylinder

26.4 Differentiation of multivariable functions

26.4.1 Partial derivative

Let $f(x, y)$ be defined in a domain D of the xy -plane and let (x_0, y_0) be a point within D . By holding y fixed at y_0 and letting x vary in an interval about x_0 , the function $f(x, y_0)$ is then

defined. It is a function depends on x alone. Its derivative with respect to x at $x = x_0$, if exists, is called the **partial derivative of $f(x, y)$ with respect to x at (x_0, y_0)** and is denoted by

$$\frac{\partial f}{\partial x}(x_0, y_0) \text{ or by } \frac{\partial f}{\partial x} \Big|_{(x_0, y_0)} \text{ or by } f_x(x_0, y_0).$$

By the definition of derivative:

$$\frac{\partial f}{\partial x}(x_0, y_0) = \lim_{\Delta x \rightarrow 0} \frac{f(x_0 + \Delta x, y_0) - f(x_0, y_0)}{\Delta x}. \quad (26.5) \quad \boxed{f_x}$$

Similarly, the partial derivative of $f(x, y)$ with respect to y at (x_0, y_0) is defined by

$$\frac{\partial f}{\partial y}(x_0, y_0) = \lim_{\Delta y \rightarrow 0} \frac{f(x_0, y_0 + \Delta y) - f(x_0, y_0)}{\Delta y}. \quad (26.6) \quad \boxed{f_y}$$

The above definitions extend at once to function of three or more variables. Let $g = g(u, v, w, x, y)$ be a five-variable function, then one partial derivative, say with respect to w , at (u, v, w, x, y) is

$$\frac{\partial g}{\partial w} = \lim_{\Delta w \rightarrow 0} \frac{g(u, v, w + \Delta w, x, y) - g(u, v, w, x, y)}{\Delta w}. \quad (26.7) \quad \boxed{g_w}$$

Evaluation of partial derivative

For explicit functions in the form of $z = f(x, y)$, evaluation of their partial derivatives is carried out as in ordinary calculus, with one variable being differentiated while *the other being treated as constant*. One direct example: if $f = x^2 - 2y^3$, then

$$\frac{\partial f}{\partial x} = f_x = 2x \quad \text{and} \quad \frac{\partial f}{\partial y} = f_y = -6y^2.$$

All the differentiation formulas from single variable calculus can be used in computing partial derivatives.

EgPartial1

Example 26.11. Find the partial derivatives of the function $f(x, y, z) = 3x^2y + (2xy - z^2)^4 + \sin(yz)$ with respect to each of the variables.

Solution

When evaluating $\partial f / \partial x$, all other variables (i.e. y and z), except x , are treated as constants. Hence

$$\begin{aligned} \frac{\partial f}{\partial x} &= 6xy + 4(2xy - z^2)^3 \cdot 2y + 0 \\ &= 6xy + 8y(2xy - z^2)^3. \end{aligned}$$

Similarly,

$$\begin{aligned} \frac{\partial f}{\partial y} &= 3x^2 + 4(2xy - z^2)^3 \cdot 2x + \cos(yz) \cdot z = 3x^2 + 8x(2xy - z^2)^3 + z \cos(yz) \\ \frac{\partial f}{\partial z} &= 0 + 4(2xy - z^2)^3 \cdot -2z + \cos(yz) \cdot y = -8z(2xy - z^2)^3 + y \cos(yz) \end{aligned}$$

□

EgPartial2

Example 26.12. Compute $\partial z / \partial x$ at the point $(1, 0)$ if $z = xe^{-2xy}$.

Solution

$$\begin{aligned} \frac{\partial z}{\partial x} &= xe^{-2xy} \cdot -2y + e^{-2xy} = e^{-2xy}(1 - 2xy) \quad [\text{Note: } d(uv) = u dv + v du] \\ \therefore \frac{\partial z}{\partial x} \Big|_{(1,0)} &= e^0(1 - 0) = 1 \end{aligned}$$

□

Geometric interpretation of partial derivative

In single variable calculus, derivative of a function means the slope of its graph. Here, the graph of $f(x, y)$ is a surface in 3-space whereas the equation $y = y_0$ represents a plane cutting the surface in a curve as shown in **Fig 20**. The partial derivative $\partial f / \partial x$ at (x_0, y_0) can then be interpreted as the slope of the tangent to that intersecting curve at (x_0, y_0) . The partial derivative $\partial f(x_0, y_0) / \partial y = f_y(x_0, y_0)$ is defined similarly. Its geometric interpretation is: the slope of the tangent of the curve (at (x_0, y_0)) in which the plane $x = x_0$ cuts the surface $z = f(x, y)$. If the point (x_0, y_0) varies, wherever the derivative exists, we obtain a new two-variable function: $f_y(x, y)$. Similarly, the derivative $\partial f / \partial x$ at variable point (x, y) is a function $f_x(x, y)$.

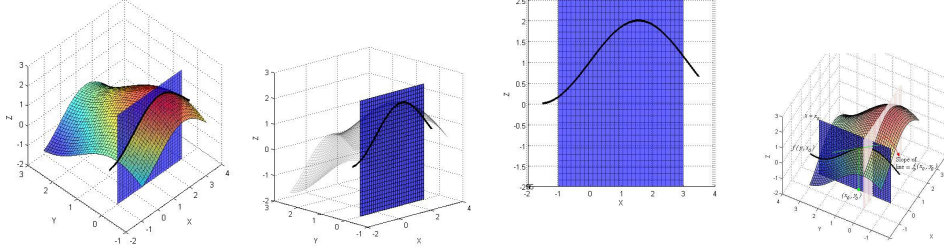


Fig 20: $f_x(x_0, y_0)$

Partial derivatives of higher order

When $f_x(x, y)$ or $f_y(x, y)$ is differentiated partially again, we have the following four second-order partial derivatives:

$$\frac{\partial^2 f}{\partial x^2} = f_{xx}, \quad \frac{\partial^2 f}{\partial y^2} = f_{yy}, \quad \frac{\partial^2 f}{\partial x \partial y} = f_{yx}, \quad \frac{\partial^2 f}{\partial y \partial x} = f_{xy}. \quad (26.8)$$

The mixed partial derivative $\partial^2 f / \partial x \partial y$ is the result of differentiating $\partial f / \partial y$ with respect to x , i.e. $\frac{\partial}{\partial x} \left(\frac{\partial f}{\partial y} \right)$. Care should be taken to its subscript format: f_{yx} means $(f_y)_x$. Within the domain considered, if all derivatives concerned are continuous, then the order of differentiation is unimportant:

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x} \quad \text{or} \quad f_{yx} = f_{xy}. \quad (26.9)$$

Third and higher order partial derivatives are defined in a similar way. Again, under appropriate continuity assumption, the order of differentiation does not matter. For example:

$$\frac{\partial^3 f}{\partial y \partial^2 x} = \frac{\partial^3 f}{\partial^2 x \partial y} = \frac{\partial^3 f}{\partial x \partial y \partial x} \quad \text{or equivalently} \quad f_{xxy} = f_{yxx} = f_{xyx}.$$

EgPartial3

Example 26.13. For the function $f(x, y) = x^2y + xy^3$, we have

$$\begin{aligned} f_x &= 2xy + y^3, & f_{xx} &= 2y, \\ f_y &= x^2 + 3y^2x, & f_{yy} &= 6xy, \\ f_{xy} &= 2x + 3y^2, & f_{yx} &= 2x + 3y^2. \end{aligned}$$

□

26.4.2 Total Differential

In forming the above partial derivatives, changes in x and in y were considered separately. The influence of changing of x and y together on the value of the function $f(x, y)$ is consider here. Suppose (x, y) is a fixed point and (x_1, y_1) is a nearby point. The the function $f(x, y)$ changes by an amount Δf moving from point (x, y) to point (x_1, y_1) :

$$\Delta f = f(x_1, y_1) - f(x, y).$$

The symbol ‘ Δ ’ means small change in certain variable. The notations

$$\Delta x = x_1 - x \quad \text{and} \quad \Delta y = y_1 - y$$

are then use to rewrite Δf as

$$\Delta f = f(x + \Delta x, y + \Delta y) - f(x, y). \quad (26.10)$$

This defines Δf as a function of Δx and Δy ($f(x, y)$ is regarded as a constant since (x, y) is fixed).



Change Δf in f can be understood as a change in x alone plus a subsequent change in y alone:

$$\Delta f = \{ f(x + \Delta x, y) - f(x, y) \} + \{ f(x + \Delta x, y + \Delta y) - f(x + \Delta x, y) \}.$$

The definition of partial derivative implies

$$f(x + \Delta x, y) - f(x, y) \approx f_x(x, y)\Delta x$$

and

$$f(x + \Delta x, y + \Delta y) - f(x + \Delta x, y) \approx f_y(x + \Delta x, y)\Delta y \approx f_y(x, y)\Delta y$$

for small $\Delta x, \Delta y$. Thus,

$$\Delta f(x, y) \approx f_x(x, y)\Delta x + f_y(x, y)\Delta y. \quad (26.11)$$



When Δx and Δy approach zero, Δf approaches df , a quantity known as **the total differential of f** (or just the **differential** of f). Under this limit, Δx and Δy can be replaced by dx and dy , thus one has

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy = f_x dx + f_y dy \quad (26.12)$$

Generalization to functions of three and more variables can be extended at once, for example, if $f = f(p, q, r, s)$ then

$$df = \frac{\partial f}{\partial p} dp + \frac{\partial f}{\partial q} dq + \frac{\partial f}{\partial r} dr + \frac{\partial f}{\partial s} ds.$$

Example 26.14. If $f(x, y, z) = z \sin(xy)$, then $df = yz \cos(xy)dx + xz \cos(xy)dy + \sin(xy)dz$. □

Total differential is often used to estimate small changes in physical quantities and errors in scientific measurements.

Example 26.15. The dimensions of a right circular cylinder are measured (with errors or uncertainties) and recorded as $r = (5 \pm 0.05)\text{cm}$ and $h = (10 \pm 0.05)\text{cm}$, then the uncertainty in its volume, $V = \pi r^2 h$ can be estimated using total differential:

$$\begin{aligned} \Delta V \approx dV &= \frac{\partial V}{\partial r} dr + \frac{\partial V}{\partial h} dh \\ &= 2\pi r h dh + \pi r h^2 dr \\ &= 2\pi \times 5 \times 10 \times 0.05 + \pi \times 25 \times 0.05 \\ &= 6.25\pi \\ &= 19.63 \text{ (2d.p.)} \end{aligned} \quad \left[\begin{array}{l} dr \approx \Delta r = 0.05 \\ dh \approx \Delta h = 0.05 \\ r = 5 \\ h = 10 \\ V = \pi r^2 h = 250\pi = 785.40 \text{ (2d.p.)} \end{array} \right.$$

Hence, the volume is recorded as $V = (250.00 \pm 6.25)\pi \text{ cm}^2$. □

26.4.3 Chain rule and tree diagram

The chain rule is basic for us to differentiate composite functions (or functions of functions). Let $f(x, y)$ be a two-variable function and now if both $x = x(t)$ and $y = y(t)$ are functions of t , then one has a composite function in which f is a function of t :

$$f = f(x, y) = f(x(t), y(t)).$$

On the other hand, if $x = x(u, v)$, $y = y(u, v)$, then one forms a different composite function where f is a function of u and v :

$$f = f(x, y) = f(x(u, v), y(u, v)).$$

Finding the dependence of a variable (a function) with respect to another in a chain of composite function is the subject of study in this subsection.

Suppose $f = f(x(t), y(t))$ is a function of t and we are interested in df/dt . To study how a small change in t is propagated in f , we use the differential of f to approximate small change in f :

$$\Delta f \approx \frac{\partial f}{\partial x} \Delta x + \frac{\partial f}{\partial y} \Delta y.$$

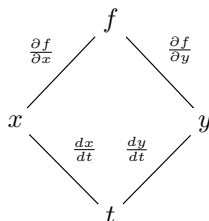
But since $\Delta x \approx \frac{dx}{dt} \Delta t$ and $\Delta y \approx \frac{dy}{dt} \Delta t$, so

$$\Delta f \approx \frac{\partial f}{\partial x} \frac{dx}{dt} \Delta t + \frac{\partial f}{\partial y} \frac{dy}{dt} \Delta t.$$

Dividing both sides by Δt and taking the limit $\Delta t \rightarrow 0$ yield the result

$$\frac{dz}{dt} = \frac{\partial f}{\partial x} \frac{dx}{dt} + \frac{\partial f}{\partial y} \frac{dy}{dt}.$$

This chain rule shows how df/dt is related to the derivatives of f , x , and y . The following tree diagram provides a good schematic way to keep track the above chain of dependence: f depends of x and y , which in turn depend on t .



For each top-to-bottom path, derivatives along the path are multiplied together. Then, the contributions from each paths are added up to form the relevant chain rule.

The chain rule can be visualized easily using tree diagram such that we do not have to memorize the chain rule formula, but just read from the tree diagram. More illustration in the subsequent examples.

EgTree1

Example 26.16. If $f = f(x, y)$ and $x = x(u, v)$, $y = y(u, v)$, then

$$\frac{\partial f}{\partial u} = \frac{\partial f}{\partial x} \frac{\partial x}{\partial u} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial u}, \quad \frac{\partial f}{\partial v} = \frac{\partial f}{\partial x} \frac{\partial x}{\partial v} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial v}.$$

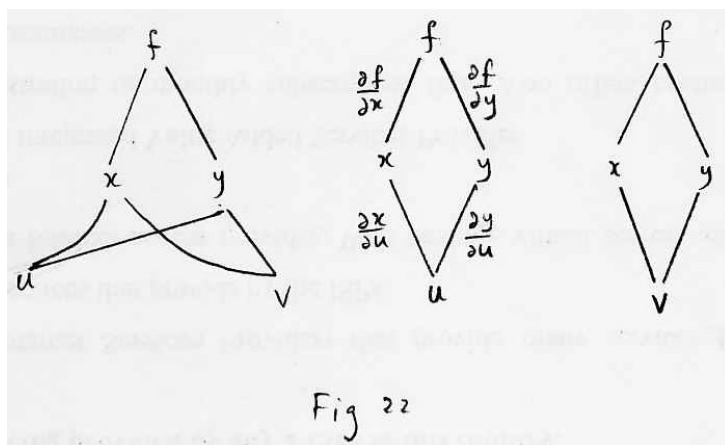


Fig 22 : Chain rule 2

□

EgTree2

Example 26.17. Calculate $\frac{\partial w}{\partial v}$ at $(u, v) = (0.5, 1)$ if

$$w = xy + yz + xz \quad \text{and} \quad \begin{aligned} x &= u + v, \\ y &= u - v, \\ z &= uv. \end{aligned}$$

Solution

$$\begin{aligned} \frac{\partial w}{\partial v} &= \frac{\partial w}{\partial x} \frac{\partial x}{\partial v} + \frac{\partial w}{\partial y} \frac{\partial y}{\partial v} + \frac{\partial w}{\partial z} \frac{\partial z}{\partial v} \\ &= (y + z)(1) + (x + z)(-1) + (y + x)(u) \\ &= (y - x) + u(x + y) \\ &= \left(-\frac{1}{2} - \frac{3}{2}\right) + \frac{1}{2} \left(\frac{3}{2} - \frac{1}{2}\right) \\ &= -\frac{3}{2} \end{aligned} \quad \left[\begin{aligned} u &= 1/2 \\ v &= 1 \\ x &= 3/2 \\ y &= -1/2 \\ z &= 1/2 \end{aligned} \right.$$

□

26.4.4 Implicit partial differentiation

If $F(x, y, z)$ is a given function of x , y and z , then the equation

$$F(x, y, z) = 0$$

may describe (implicitly) one or several functions z of x and y , i.e. $z(x, y)$. Thus, if $x^2 + y^2 + z^2 - 4 = 0$, then

$$z = \sqrt{4 - (x^2 + y^2)} \quad \text{or} \quad z = -\sqrt{4 - (x^2 + y^2)},$$

where both functions being defined for $x^2 + y^2 \leq 4$. Either function is said to be **implicitly defined** by the equation $x^2 + y^2 + z^2 - 4 = 0$. However notice that given $F(x, y, z) = 0$, one may not always be able to express z *explicitly* in terms of x and y . For instance in $\tan(xyz) + e^{xz} \ln(yz) = 0$. Similarly, an equation $F(p, q, r, z) = 0$ may define one or more implicit function z of p, q, r , and so on. Finding partial derivatives of implicit functions is our objective here.

The following results can be taken as a theorem.

$$\text{If } F(x, y) = 0, \text{ then } \frac{dy}{dx} = -\frac{F_x}{F_y} \quad (F_y \neq 0);$$

$$\text{if } F(x, y, z) = 0, \text{ then } \frac{\partial z}{\partial x} = -\frac{F_x}{F_z}, \quad \frac{\partial z}{\partial y} = -\frac{F_y}{F_z} \quad (F_z \neq 0).$$

Extension to four or more variables situations are obvious.

EgImplicit1

Example 26.18. Find dy/dx in $x^3 - 3xy^2 + y^3 = 1$.

Solution

$$\text{Define } F \text{ as } F(x, y) = x^3 - 3xy^2 + y^3 - 1 = 0,$$

$$\therefore F_x = 3x^2 - 3y^2 \quad \text{and} \quad F_y = -6xy + 3y^2.$$

$$\text{Using the above theorem, } \frac{dy}{dx} = -\frac{F_x}{F_y} = -\frac{3(x^2 - y^2)}{3(y^2 - 2xy)} = \frac{y^2 - x^2}{y^2 - 2xy}.$$

□

EgImplicit2

Example 26.19. Evaluate $\partial z / \partial x$ at the point (π, π, π) if $z = z(x, y)$ is implicitly defined by the equation $\sin(x + y) + \sin(y + z) + \sin(x + z) = 0$.

Solution

$$F(x, y, z) = \sin(x + y) + \sin(y + z) + \sin(x + z) = 0$$

$$F_x = \cos(x + y) + \cos(x + z)$$

$$F_z = \cos(y + z) + \cos(x + z)$$

$$\therefore \frac{\partial z}{\partial x} = -\frac{F_x}{F_z} = -\frac{\cos(x + y) + \cos(x + z)}{\cos(y + z) + \cos(x + z)}$$

$$\Rightarrow \left. \frac{\partial z}{\partial x} \right|_{(\pi, \pi, \pi)} = -\frac{\cos(2\pi) + \cos(2\pi)}{\cos(2\pi) + \cos(2\pi)} = -1$$

□

Chapter 27

Vector Field

Definition 27.1. A *vector field* on an open subset $D \subset \mathbb{R}^n$ is a (continuous and differentiable) mapping:

$$\mathbf{v} : D \rightarrow \mathbb{R}^n, \quad \mathbf{x} \mapsto \mathbf{v}(\mathbf{x}) = (v_1(\mathbf{x}), \dots, v_n(\mathbf{x})).$$

In lay man term, a vector field is a function that assigns to each point $\mathbf{x}$ in its domain D a unique vector $\mathbf{v}(\mathbf{x})$.

Example 27.1. Vector Fields:

1. Heat Flow
2. Static Electric Field, Gravitational Field, Electromagnetic Field
3. Fluid Flow

If to each point (x, y, z) of a region R in the 3- D space, there corresponds a vector $\mathbf{v}(x, y, z)$, then $\mathbf{v}$ is called a **vector function of position** or **vector point function** and we say that a **vector field** $\mathbf{v}$ has been defined in R .

Eg: 1) If the velocity at any point (x, y, z) within a moving fluid is known at a certain time, then a vector field is defined.

2) $\mathbf{v}(x, y, z) = x^2y\mathbf{i} - xz^3\mathbf{j} + yx^2z\mathbf{k}$ defines a vector field, or we write $\mathbf{v}(x, y, z) = v_1(x, y, z)\mathbf{i} + v_2(x, y, z)\mathbf{j} + v_3(x, y, z)\mathbf{k}$ in general. Here v_1, v_2, v_3 are the components for $\mathbf{v}$ in the (x, y, z) direction respectively.

27.1 Gradient, Divergence and Curl

The vector differential operator del, written ∇ , is defined by

$$\nabla = \frac{\partial}{\partial x} \mathbf{i} + \frac{\partial}{\partial y} \mathbf{j} + \frac{\partial}{\partial z} \mathbf{k} = \mathbf{i} \frac{\partial}{\partial x} + \mathbf{j} \frac{\partial}{\partial y} + \mathbf{k} \frac{\partial}{\partial z}$$

It possesses properties analogous to these of ordinary vectors.

The gradient

Let $\phi(x, y, z)$ defines a scalar field and is differentiable at each point (x, y, z) in a certain region of space. Then the gradient of ϕ , written $\nabla\phi$ or $\text{grad } \phi$, is define by

$$\nabla\phi = \left(\frac{\partial}{\partial x} \mathbf{i} + \frac{\partial}{\partial y} \mathbf{j} + \frac{\partial}{\partial z} \mathbf{k} \right) \phi = \frac{\partial\phi}{\partial x} \mathbf{i} + \frac{\partial\phi}{\partial y} \mathbf{j} + \frac{\partial\phi}{\partial z} \mathbf{k}$$

Note that $\nabla\phi$ defines a vector field. Gradient also know as gradient vector.

The directional derivative

The component of $\nabla\phi$ in the direction of a unit vector $\hat{a}$ is given by $\nabla\phi \cdot \hat{a}$ and is called the directional derivative of ϕ in the direction $\hat{a}$. Physically, this is the rate of change of ϕ at (x, y, z) in the direction $\hat{a}$.

The divergence

Let $\mathbf{v}(x, y, z) = v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k}$ defines a vector field and is differentiable at each point (x, y, z) in a certain region of space. Then the divergence of $\mathbf{v}$, written $\nabla \cdot \mathbf{v}$ or $\text{div } \mathbf{v}$, is defined by

$$\begin{aligned}\nabla \mathbf{v} &= \left(\frac{\partial}{\partial x} \mathbf{i} + \frac{\partial}{\partial y} \mathbf{j} + \frac{\partial}{\partial z} \mathbf{k} \right) \cdot (v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k}) \\ &= \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} + \frac{\partial v_3}{\partial z}\end{aligned}$$

Note the analogy with the dot product $A \cdot B = a_1b_1 + a_2b_2 + a_3b_3$.

Also note that $\nabla \cdot \mathbf{v} \neq \mathbf{v} \cdot \nabla$

The Curl

If $\mathbf{v}(x, y, z)$ is a differential vector field, then the curl or the rotation of $\mathbf{v}$, written $\nabla \times \mathbf{v}$ or $\text{curl } \mathbf{v}$, is defined by

$$\begin{aligned}\nabla \times \mathbf{v} &= \left(\frac{\partial}{\partial x} \mathbf{i} + \frac{\partial}{\partial y} \mathbf{j} + \frac{\partial}{\partial z} \mathbf{k} \right) \times (v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k}) \\ &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ v_1 & v_2 & v_3 \end{vmatrix} \\ &= \begin{vmatrix} \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ v_2 & v_3 \end{vmatrix} \mathbf{i} - \begin{vmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial z} \\ v_1 & v_3 \end{vmatrix} \mathbf{j} + \begin{vmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} \\ v_1 & v_2 \end{vmatrix} \mathbf{k} \\ &= \left(\frac{\partial v_3}{\partial y} - \frac{\partial v_2}{\partial z} \right) \mathbf{i} + \left(\frac{\partial v_1}{\partial z} - \frac{\partial v_3}{\partial x} \right) \mathbf{j} + \left(\frac{\partial v_2}{\partial x} - \frac{\partial v_1}{\partial y} \right) \mathbf{k}\end{aligned}$$

Note that in the expansion of the determinant, the operator $\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}$ must precede v_1, v_2, v_3 .

27.2 Formulae involving ∇

If $\mathbf{A}$ and $\mathbf{B}$ are differentiable vector functions, and ϕ and ψ are differentiable scalar functions of position (x, y, z) , then

1. $\nabla(\phi - \psi) = \nabla\phi - \nabla\psi$
2. $\nabla \cdot (\mathbf{A} + \mathbf{B}) = \nabla \cdot \mathbf{A} + \nabla \cdot \mathbf{B}$
3. $\nabla \times (\mathbf{A} + \mathbf{B}) = \nabla \times \mathbf{A} + \nabla \times \mathbf{B}$
4. $\nabla \cdot (\phi \mathbf{A}) = (\nabla \phi) \cdot \mathbf{A} + \phi(\nabla \cdot \mathbf{A})$
5. $\nabla \times (\phi \mathbf{A}) = (\nabla \phi) \times \mathbf{A} + \phi(\nabla \times \mathbf{A})$
6. $\nabla \cdot (\mathbf{A} \times \mathbf{B}) = \mathbf{B} \cdot (\nabla \times \mathbf{A}) - \mathbf{A} \cdot (\nabla \times \mathbf{B})$
7. $\nabla \times (\mathbf{A} \times \mathbf{B}) = (\mathbf{B} \cdot \nabla) \mathbf{A} - \mathbf{B}(\nabla \cdot \mathbf{A}) - (\mathbf{A} \cdot \nabla) \mathbf{B} + \mathbf{A}(\nabla \cdot \mathbf{B})$
8. $\nabla(\mathbf{A} \cdot \mathbf{B}) = (\mathbf{B} \cdot \nabla) \mathbf{A} + (\mathbf{A} \cdot \nabla) \mathbf{B} + \mathbf{B} \times (\nabla \times \mathbf{A}) + \mathbf{A} \times (\nabla \times \mathbf{B})$
9. $\nabla \cdot (\nabla \phi) = \nabla^2 \phi = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) \phi$ where ∇^2 is called the **Laplace Operator**.
10. $\nabla \times (\nabla \phi) = \mathbf{0}$
11. $\nabla \cdot (\nabla \times \mathbf{A}) = 0$
12. $\nabla \times (\nabla \times \mathbf{A}) = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A}$
13. $\nabla f(\phi) = \frac{df}{d\phi} \nabla \phi$
14. $\nabla \cdot (\mathbf{r} - \mathbf{a}) = 3$ $\mathbf{a} = \text{constant vector}$
15. $\nabla \times (\mathbf{r} - \mathbf{a}) = \mathbf{0}$ $\mathbf{r} = \text{radius vector}$
 $\quad \quad \quad = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$
16. $\nabla(|\mathbf{r} - \mathbf{a}|^n) = n|\mathbf{r} - \mathbf{a}|^{n-2}(\mathbf{r} - \mathbf{a})$

17. $\nabla(\mathbf{a} \cdot \mathbf{r}) = \mathbf{a}$
18. $\nabla \cdot (\nabla \phi \times \nabla \psi) = 0$
19. $\nabla r = \frac{1}{r} \mathbf{r} \quad r = \sqrt{x^2 + y^2 + z^2}$
20. $\nabla \cdot \mathbf{r} = 3 \quad \mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$
21. $\nabla \cdot \mathbf{r} = 0$
22. $\nabla r^n = n r^{n-2} \mathbf{r}$
23. $\nabla \cdot (r^n \mathbf{r}) = (n+3) r^n$
24. $\nabla \times (r^n \mathbf{r}) = \mathbf{0}$

Example 27.2. If $\mathbf{A}(x, y, z) = x^2 z \mathbf{i} - 2y^3 z^2 \mathbf{j} + xy^2 z \mathbf{k}$, $f(x, y, z) = xye^z$, (i) Find $\text{div}(\mathbf{A})$ at the point $(1, -1, 1)$

(ii) Find $(\mathbf{A} \cdot \nabla)f$

$$(i) \text{div}(\mathbf{A}) = \nabla \cdot \mathbf{A} = \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} \cdot \begin{pmatrix} x^2 z \\ -2y^3 z^2 \\ xy^2 z \end{pmatrix} = 2xz - 6y^2 z^2 + xy^2$$

$$\text{div}(\mathbf{A})|_{(1, -1, 1)} = 2xz - 6y^2 z^2 + xy^2 |_{(1, -1, 1)} = 2 - 6 + 1 = -3$$

$$(ii) \mathbf{A} \cdot \nabla = \begin{pmatrix} x^2 z \\ -2y^3 z^2 \\ xy^2 z \end{pmatrix} \cdot \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} = x^2 z \frac{\partial}{\partial x} - 2y^3 z^2 \frac{\partial}{\partial y} + xy^2 z \frac{\partial}{\partial z}$$

$$\begin{aligned} (\mathbf{A} \cdot \nabla)f &= x^2 z \frac{\partial f}{\partial x} - 2y^3 z^2 \frac{\partial f}{\partial y} + xy^2 z \frac{\partial f}{\partial z} \\ &= x^2 z (ye^z) - 2y^3 z^2 (xe^z) + xy^2 z (xye^z) \\ &= x^2 y z e^z - 2xy^3 z^2 e^z + x^2 y^3 z e^z \end{aligned}$$

Example 27.3. The vector field $\mathbf{F}(x, y, z) = xz^3 \mathbf{i} - 2x^2 y z \mathbf{j} + 2yz^4 \mathbf{k}$, find $\text{curl}(\mathbf{F})$ at $(1, -1, 1)$

$$\text{Curl}(\mathbf{F})|_{(1, -1, 1)} = \nabla \times \mathbf{F} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ xz^3 & -2x^2 y z & 2yz^4 \end{vmatrix} = \begin{pmatrix} 2z^4 + 2x^2 y \\ 3xz^2 \\ -4xyz \end{pmatrix} \bigg|_{(1, -1, 1)} = \begin{pmatrix} 0 \\ 3 \\ 4 \end{pmatrix} = 3\mathbf{j} + 4\mathbf{k}$$

Example 27.4. Show that $\nabla \cdot (r \mathbf{r}) = 4r$ if $\mathbf{r} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$; $r = |\mathbf{r}|$.

Solution. Use formula $\nabla \cdot (\phi \mathbf{A}) = (\nabla \phi) \cdot \mathbf{A} + \phi (\nabla \cdot \mathbf{A})$

$$\begin{aligned} \nabla \cdot (r \mathbf{r}) &= \mathbf{r} \cdot \nabla r + r (\nabla \cdot \mathbf{r}) \\ &= \frac{1}{r} \mathbf{r} \cdot \mathbf{r} + 3r \\ &= \frac{1}{r} (r^2) + 3r \quad (\mathbf{a} \cdot \mathbf{a} = |\mathbf{a}|^2 = a^2) \\ &= r + 3r \\ &= 4r \quad (\text{Q.E.D.}) \end{aligned}$$

$$\nabla r = \nabla (x^2 + y^2 + z^2)^{\frac{1}{2}}$$

$$\begin{pmatrix} \frac{\partial}{\partial x} & (x^2 + y^2 + z^2)^{\frac{1}{2}} \\ \frac{\partial}{\partial y} & (x^2 + y^2 + z^2)^{\frac{1}{2}} \\ \frac{\partial}{\partial z} & (x^2 + y^2 + z^2)^{\frac{1}{2}} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & (x^2 + y^2 + z^2)^{-\frac{1}{2}} (2x) \\ \frac{1}{2} & (x^2 + y^2 + z^2)^{-\frac{1}{2}} (2y) \\ \frac{1}{2} & (x^2 + y^2 + z^2)^{-\frac{1}{2}} (2z) \end{pmatrix}$$

$$= \frac{1}{\sqrt{x^2+y^2+z^2}} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \frac{1}{r} (\mathbf{r})$$

$$\nabla \cdot \mathbf{r} = \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = 1 + 1 + 1 = 3 \quad \square$$

27.3 Grad, Div, Laplacian, and Curl in Cartesian and Curvilinear Coordinate Systems

If f is a scalar function and $\mathbf{F} = F_1\mathbf{e}_1 + F_2\mathbf{e}_2 + F_3\mathbf{e}_3$ a vector field of orthogonal curvilinear coordinates u_1, u_2, u_3 , then

- $\text{grad } f = \nabla f = \frac{1}{h_1} \frac{\partial f}{\partial u_1} \mathbf{e}_1 + \frac{1}{h_2} \frac{\partial f}{\partial u_2} \mathbf{e}_2 + \frac{1}{h_3} \frac{\partial f}{\partial u_3} \mathbf{e}_3$
- $\text{div } \mathbf{F} = \nabla \cdot \mathbf{F} = \frac{1}{h_1 h_2 h_3} \left[\frac{\partial}{\partial u_1} (h_2 h_3 F_1) + \frac{\partial}{\partial u_2} (h_1 h_3 F_2) + \frac{\partial}{\partial u_3} (h_1 h_2 F_3) \right]$
- $\text{curl } \mathbf{F} = \nabla \times \mathbf{F} = \frac{1}{h_1 h_2 h_3} \begin{vmatrix} h_1 \mathbf{e}_1 & h_2 \mathbf{e}_2 & h_3 \mathbf{e}_3 \\ \frac{\partial}{\partial u_1} & \frac{\partial}{\partial u_2} & \frac{\partial}{\partial u_3} \\ h_1 F_1 & h_2 F_2 & h_3 F_3 \end{vmatrix}$
- Laplacian of $f = \nabla^2 f = \Delta f = \frac{1}{h_1 h_2 h_3} \left[\frac{\partial}{\partial u_1} \left(\frac{h_2 h_3}{h_1} \frac{\partial f}{\partial u_1} \right) + \frac{\partial}{\partial u_2} \left(\frac{h_3 h_1}{h_2} \frac{\partial f}{\partial u_2} \right) + \frac{\partial}{\partial u_3} \left(\frac{h_1 h_2}{h_3} \frac{\partial f}{\partial u_3} \right) \right]$

Remark 27.5. If $h_1 = h_2 = h_3 = 1$ and $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) = (\mathbf{i}, \mathbf{j}, \mathbf{k})$, these reduce to the usual expressions in rectangular coordinates where $(u_1, u_2, u_3) = (x, y, z)$.

The *physical interpretation* of $\text{curl } \mathbf{F}$:

Let the vector field $\mathbf{F}$ represents the velocity field in fluid flow. Particles near (x, y, z) tend to rotate about the axis that points in the direction of $\text{curl } \mathbf{F}$ and $\|\text{curl } \mathbf{F}\|$ is a measure of how quickly the particles move around the axis.

If $\text{curl } \mathbf{F} = \mathbf{0}$ at a point P , then the fluid is free from rotations at P and $\mathbf{F}$ is said to be *ir-rotational*.

The physical interpretation of $\text{div } \mathbf{F}$:

Let $\mathbf{F}$ be the velocity of a fluid, then $\text{div } \mathbf{F}$ represents the net rate of change (with respect to time) of the mass of fluid flowing from the point (x, y, z) per unit volume. That is, $\text{div } \mathbf{F}$ measures the tendency of the fluid to diverge from the point (x, y, z) ($\text{div } \mathbf{F} > 0$) or accumulate toward (x, y, z) ($\text{div } \mathbf{F} < 0$).

If $\text{div } \mathbf{F} = 0$, then $\mathbf{F}$ is said to be *incompressible*.

27.4 Vector Fields in Cylindrical and Spherical Coordinate Systems

27.4.1 Transformation of Coordinates

Let the rectangular coordinates (x, y, z) of any point be expressed as functions of (u_1, u_2, u_3) so that

$$x = x(u_1, u_2, u_3), \quad y = y(u_1, u_2, u_3), \quad z = z(u_1, u_2, u_3). \quad (27.1) \quad \text{eqn:1.1}$$

Suppose that [eqn:1.1](#) (27.1) can be solved for u_1, u_2, u_3 in terms of x, y, z , i.e.,

$$u_1 = u_1(x, y, z), \quad u_2 = u_2(x, y, z), \quad u_3 = u_3(x, y, z). \quad (27.2) \quad \text{eqn:1.2}$$

The functions of [eqn:1.1](#) (27.1) and [eqn:1.2](#) (27.2) are assumed to be single-valued and to have continuous derivatives so that the correspondence between (x, y, z) and (u_1, u_2, u_3) is unique.

Given a point P with rectangular coordinates (x, y, z) we can, from [eqn:1.2](#) (27.2) associate a unique set of coordinates (u_1, u_2, u_3) called the *curvilinear coordinates* of P . The sets of equations [eqn:1.1](#) (27.1) or [eqn:1.2](#) (27.2) defines a *transformation of coordinates*.

27.4.2 Orthogonal Curvilinear Coordinates

The surfaces $u_1 = c_1$, $u_2 = c_2$, $u_3 = c_3$, where c_1, c_2, c_3 are constants, are called *coordinate surfaces* and each pair of these surfaces intersect in curves called *coordinate curves* or *lines*.

If the coordinate surfaces intersect at right angles, then the curvilinear system is called *orthogonal*.

The u_1, u_2 , and u_3 coordinate curves of a curvilinear system are analogous to the x, y and z coordinate axes of a rectangular system.

27.4.3 Unit Vectors in Curvilinear Systems

Let $\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ be the position vector of a point P . Then (27.1) can be written $\mathbf{r} = \mathbf{r}(u_1, u_2, u_3)$.

A *tangent vector* to the u_1 curve at P (for which u_2 and u_3 are constants) is $\frac{\partial \mathbf{r}}{\partial u_1}$. Then a unit tangent vector in this direction is $\mathbf{e}_1 = \frac{\partial \mathbf{r}}{\partial u_1} / \left\| \frac{\partial \mathbf{r}}{\partial u_1} \right\|$ so that $\frac{\partial \mathbf{r}}{\partial u_1} = h_1 \mathbf{e}_1$ where $h_1 = \left\| \frac{\partial \mathbf{r}}{\partial u_1} \right\|$.

Similarly, if $\mathbf{e}_2$ and $\mathbf{e}_3$ are unit tangent vectors to the u_2 and u_3 curves at P respectively, then $\frac{\partial \mathbf{r}}{\partial u_2} = h_2 \mathbf{e}_2$ and $\frac{\partial \mathbf{r}}{\partial u_3} = h_3 \mathbf{e}_3$ where $h_2 = \left\| \frac{\partial \mathbf{r}}{\partial u_2} \right\|$ and $h_3 = \left\| \frac{\partial \mathbf{r}}{\partial u_3} \right\|$.

The quantities h_1, h_2, h_3 are called *scale factors*.

The unit vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ are in the directions of increasing u_1, u_2, u_3 , respectively.

Since the gradient vector

$$\nabla u_i = \frac{\partial u_i}{\partial x} \mathbf{i} + \frac{\partial u_i}{\partial y} \mathbf{j} + \frac{\partial u_i}{\partial z} \mathbf{k}$$

is normal to the surface $u_i = c_i$, a unit vector in this direction is given by $\mathbf{E}_i = \frac{\nabla u_i}{\|\nabla u_i\|}$.

Thus, at each point P of a curvilinear system, there exist, in general, two sets of unit vectors, $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ tangent to the coordinate curves and $\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3$ normal to the coordinate surfaces. These two sets become identical if and only if the curvilinear system is orthogonal.

Both sets are analogous to the $\mathbf{i}, \mathbf{j}, \mathbf{k}$ unit vectors in the rectangular coordinates but are unlike them in that they may change directions from point to point. A vector $\mathbf{A}$ can be represented in terms of the unit base vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ in the form

$$\mathbf{A} = A_1 \mathbf{e}_1 + A_2 \mathbf{e}_2 + A_3 \mathbf{e}_3.$$

27.4.4 Arc Length and Volume Elements

From $\mathbf{r} = \mathbf{r}(u_1, u_2, u_3)$, we have

$$d\mathbf{r} = \frac{\partial \mathbf{r}}{\partial u_1} du_1 + \frac{\partial \mathbf{r}}{\partial u_2} du_2 + \frac{\partial \mathbf{r}}{\partial u_3} du_3 = h_1 du_1 \mathbf{e}_1 + h_2 du_2 \mathbf{e}_2 + h_3 du_3 \mathbf{e}_3.$$

Then the differential of arc length ds is determined from

$$ds^2 = d\mathbf{r} \cdot d\mathbf{r}.$$

For orthogonal systems, $\mathbf{e}_i \cdot \mathbf{e}_j = 0$ for $i \neq j$ so that

$$ds^2 = d\mathbf{r} \cdot d\mathbf{r} = h_1^2 du_1^2 + h_2^2 du_2^2 + h_3^2 du_3^2.$$

The volume element for an orthogonal system is given by

$$dV = |(h_1 du_1 \mathbf{e}_1) \cdot (h_2 du_2 \mathbf{e}_2) \times (h_3 du_3 \mathbf{e}_3)| = h_1 h_2 h_3 du_1 du_2 du_3$$

since $|\mathbf{e}_1 \cdot \mathbf{e}_2 \times \mathbf{e}_3| = 1$.

27.4.5 Special Orthogonal Coordinate Systems

In *cylindrical coordinates* $(u_1, u_2, u_3) = (r, \theta, z)$, $r \geq 0$, $0 \leq \theta < 2\pi$, $z \in \mathbb{R}$:

- Transformation of Coordinates:
$$\begin{cases} x = r \cos \theta, \\ y = r \sin \theta, \\ z = z. \end{cases} \longleftrightarrow \begin{cases} r = \sqrt{x^2 + y^2}, \\ \theta = \tan^{-1} \frac{y}{x}, \\ z = z. \end{cases}$$
- Representation of a vector: $\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k} = r \cos \theta \mathbf{i} + r \sin \theta \mathbf{j} + z\mathbf{k}$

- Scale factors: $h_r = \left\| \frac{\partial \mathbf{r}}{\partial u_1} \right\| = \sqrt{\sin^2 \theta + \cos^2 \theta} = 1, h_\theta = r, h_z = 1$
- Arc length: $ds^2 = dr^2 + r^2 d\theta^2 + dz^2$
- Volume element: $dV = h_1 h_2 h_3 du_1 du_2 du_3 = r dz dr d\theta$

Example 27.6. The rectangular coordinates of the point with cylindrical coordinates $(r, \theta, z) = (4, \pi/3, 6)$ is $(2, 2\sqrt{3}, 6)$.

In *spherical coordinates* $(u_1, u_2, u_3) = (\rho, \phi, \theta)$, $\rho \geq 0$, $0 \leq \phi < 2\pi$, $0 \leq \theta \leq \pi$ (in many mathematical books, (ρ, θ, ϕ)):

- Transformation of Coordinates:
$$\begin{cases} x = \rho \sin \theta \cos \phi, \\ y = \rho \sin \theta \sin \phi, \\ z = \rho \cos \theta. \end{cases} \longleftrightarrow \begin{cases} \rho = \sqrt{x^2 + y^2 + z^2}, \\ \phi = \tan^{-1} \frac{y}{x}, \\ \theta = \tan^{-1} \frac{z}{\sqrt{x^2 + y^2}}. \end{cases}$$
- Scale factors: $h_\rho = 1, h_\phi = \rho \sin \phi, h_\theta = \rho$
- Arc length: $ds^2 = d\rho^2 + \rho^2 \sin^2 \theta d\phi^2 + \rho^2 d\theta^2$
- Volume element: $dV = \rho^2 \sin \theta d\rho d\phi d\theta$.

Here

- $\rho = \|\overrightarrow{OP}\|$, the distance between O and P ,
- ϕ is the angle between the vertical projection of $\overrightarrow{OP}$ onto the xy -plane and the positive x -axis,
- θ is the angle between $\overrightarrow{OP}$ and the positive z -axis.

Example 27.7. The cylindrical coordinates of the point with rectangular coordinates $(3, -3, \sqrt{6})$ is $(2\sqrt{6}, 7\pi/4, \pi/3)$.

Example 27.8. Equations of some surfaces in cylindrical and spherical coordinates:

	cone	cylinder	sphere	paraboloid
rectangular	$z = \sqrt{x^2 + y^2}$	$x^2 + y^2 = 1$	$x^2 + y^2 + z^2 = 1$	$z = x^2 + y^2$
cylindrical	$z = r$	$r = 1$	$z^2 = 1 - r^2$	$z = r^2$
spherical	$\phi = \pi/4$	$\rho = \csc \phi$	$\rho = 1$	$\rho = \cos \phi \csc^2 \phi$

27.4.6 The gradient vector

The **gradient** of $f(x, y, z)$, denoted by ∇f , is defined (in Cartesian representation) by

$$\nabla f = \frac{\partial f}{\partial x} \mathbf{i} + \frac{\partial f}{\partial y} \mathbf{j} + \frac{\partial f}{\partial z} \mathbf{k}. \quad (27.3) \quad \text{grad}_f$$

∇f , sometimes also denoted as $\text{grad } f$, is a vector quantity and therefore also known as **the gradient vector** of f . The ‘grad’ symbol ∇ (pronounced ‘del’) is a vector differential operator defined by

$$\nabla = \frac{\partial}{\partial x} \mathbf{i} + \frac{\partial}{\partial y} \mathbf{j} + \frac{\partial}{\partial z} \mathbf{k}. \quad (27.4) \quad \text{grad}$$

EgGrad1

Example 27.9. Let $f(x, y) = x^2 + 3y^4$, then $\nabla f = f_x \mathbf{i} + f_y \mathbf{j} = 2x \mathbf{i} + 12y^3 \mathbf{j}$. Substituting $x = 2$ and $y = 1$ gives the gradient at the specific point, $\nabla f(2, 1) = 4 \mathbf{i} + 12 \mathbf{j}$.

□

Geometric properties of the gradient vector

Suppose f is a differentiable function at the point (x_0, y_0, z_0) and $\nabla f(x_0, y_0, z_0) \neq \vec{0}$, then

1. The direction of $\nabla f(x_0, y_0, z_0)$ is
 - perpendicular to the level surface of f through the point (x_0, y_0, z_0) ;
 - perpendicular to the level curve of $f(x, y)$ through the point (x_0, y_0) ;
 - the direction in which f increases most rapidly.
2. The magnitude of $\text{grad } f$, $|\nabla f|$ is the maximum rate of change of f at the point (x_0, y_0, z_0) .

EgGrad2

Example 27.10. Find a vector perpendicular to the surface, $S : z = 4 - x^2 - 2y^2$ at the point $(1, -1, 1)$.

Solution

We may define a three-variable function $F(x, y, z)$ based on the equation of S , such that a level surface of F is S itself. Recall that the level surfaces of F are described by $F = \text{constants}$, i.e. letting F equal constants. Thus we define

$$F(x, y, z) = z - 4 + x^2 + 2y^2$$

so that S is represented by

$$F(x, y, z) = 0.$$

The gradient vector is calculated:

$$\nabla F = 2x\mathbf{i} + 4y\mathbf{j} + \mathbf{k}.$$

Thus at the point $(1, -1, 1)$, $\nabla F(1, -1, 1) = 2\mathbf{i} - 4\mathbf{j} + \mathbf{k}$. Now, $\nabla F(1, -1, 1)$ is the vector we are seeking since according to the properties of gradient: $\nabla F \perp (F = \text{constants})$. $\square$

EgGrad3

Example 27.11. Suppose the surface of a mountain is described by the equation $g(x, y) = 10 - (2x^2 + 3y^2)$. A hiker, starting from the point $(1, 1, 5)$, wish to climb uphill toward the peak. Find the direction of steepest ascent which leads to the peak in the shortest distance.

Solution

Applying the property “ ∇f points toward the direction in which f increases the fastest”, one understand at once that the required direction is just the direction of the gradient vector of g at $(1, 1)$.

$$\nabla g(x, y) = -4x\mathbf{i} - 6y\mathbf{j} \Rightarrow \nabla g(1, 1) = -4\mathbf{i} - 6\mathbf{j}$$

The vector $\nabla g(1, 1)$ lies in the xy -plane, pointing like a compass in the direction where g increases most rapidly. Also, the vector $\nabla g(1, 1)$ lies directly under the hiker's path of steepest ascent.

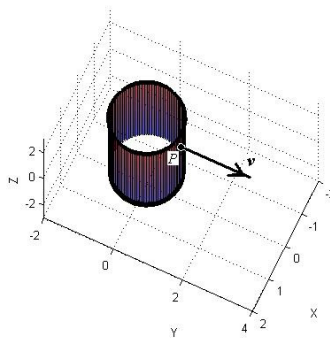


Fig 23:
The surface and its level curves;
path of steepest ascent and its projection on to xy -plane (i.e. the $\nabla g(1, 1)$)

$\square$

EgGrad4

Example 27.12. Find a vector $\mathbf{v}$ normal to the surface of the cylinder $x^2 + y^2 = 1$ at the point $P(0, 1, 1)$.

Solution

The cylinder $x^2 + y^2 = 1$ is a level surface of the function $f(x, y, z) = x^2 + y^2$ by letting $f = 1(\text{constant})$. Thus, a vector normal to the cylinder is given by the gradient of f (which is perpendicular to its level surfaces):

$$\nabla f = 2x\mathbf{i} + 2y\mathbf{j},$$

and at point P :

$$\mathbf{v} = \nabla f(0, 1, 1) = 2\mathbf{j}.$$

Notice $\mathbf{v}$ is horizontal, pointing toward the increasing $\mathbf{j}$ direction. This is what we expect since the cylinder stands vertically parallel with the z -axis.

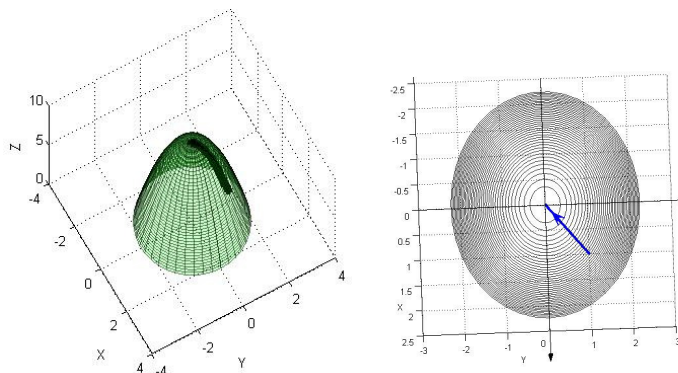


Fig 24:
The cylinder, point P and the vector $\mathbf{v}$

□

27.4.7 Directional derivative

In computing the partial derivative $\frac{\partial F}{\partial x}(x, y, z)$ at a point (x, y, z) , one considers the ration $\Delta F/\Delta x$. This is the ratio of the change in function F from (x, y, z) to $(x + \Delta x, y, z)$ to the change in x from x to $x + \Delta x$. Here only the values of F along a line parallel to the x -axis are considered. Similarly, $\frac{\partial F}{\partial y}$ and $\frac{\partial F}{\partial z}$ tell how F changes along the lines parallel to the y and z axes respectively.

However there is no reason to restrict our attention only to these three directions. Hence, one may define accordingly the directional derivatives of F in any given directions. Assume F changes ΔF moving a distance Δs in a given direction, then the **directional derivative** of F in the given direction is the limit of the ration $\Delta F/\Delta s$ as Δs approaches zero.

To be specific, let the direction in question be given by a nonzero vector $\mathbf{v} = v_x\mathbf{i} + v_y\mathbf{j} + v_z\mathbf{k}$ and denote the directional derivative of F in direction $\mathbf{v}$ at the point (x, y, z) by $D_{\mathbf{v}}F(x, y, z)$. A displacement from (x, y, z) in direction $\mathbf{v}$ corresponds to changes Δx , Δy , Δz is

$$(x + hv_x, y + hv_y, z + hv_z) - (x, y, z).$$

The changes above are assumed to be proportional to the components of $\mathbf{v}$, that is

$$\Delta x = hv_x, \quad \Delta y = hv_y, \quad \Delta z = hv_z,$$

where h is a positive scalar.

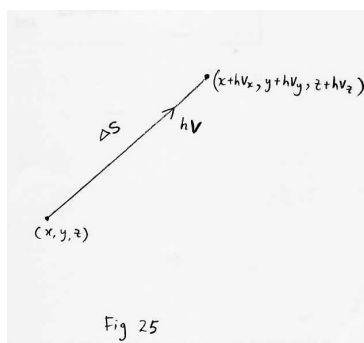


Fig 25: Displacement of Δs from (x, y, z)

The directional derivative of F in direction $\mathbf{v}$ at the point (x, y, z) , $D_{\mathbf{v}}F(x, y, z)$ is defined by the limit

$$D_{\mathbf{v}}F(x, y, z) = \lim_{h \rightarrow 0} \frac{F(x + hv_x, y + hv_y, z + hv_z) - F(x, y, z)}{h|\mathbf{v}|}. \quad (27.5)$$

EqDirectionalD

Directional derivative and the gradient vector

In practice, one hardly evaluate $D_{\mathbf{v}}F(x, y, z)$ by its definition in limit as above. Instead, the following expression which relates the directional derivative to the gradient vector is more useful :

$$D_{\hat{\mathbf{v}}}F(x, y, z) = \nabla F \cdot \hat{\mathbf{v}}, \quad (27.6)$$

where $D_{\hat{\mathbf{v}}}F$ is the directional derivative of F in the direction of the unit vector $\hat{\mathbf{v}}$ and the ‘hat’ in $\hat{\mathbf{v}}$ means unit vector. Equation (27.6) can be regarded as the generalization of partial derivatives, thus if $\hat{\mathbf{v}} = \mathbf{i}$, then $D_{\hat{\mathbf{v}}}F = \partial F / \partial x = F_x$. Similarly, $D_{\mathbf{j}}F = F_y$ and $D_{\mathbf{k}}F = F_z$.

EgDvF1

Example 27.13. Let $f(x, y, z) = x \sin z + x^2 y$, find the directional derivative of f at the point $P(1, 2, 0)$ in the direction of the vector $\mathbf{v} = 2\mathbf{i} - 2\mathbf{j} + \mathbf{k}$.

Solution

$$\begin{aligned} \nabla f(x, y, z) &= f_x \mathbf{i} + f_y \mathbf{j} + f_z \mathbf{k} = (\sin z + 2xy)\mathbf{i} + x^2 \mathbf{j} + x \cos z \mathbf{k} \\ \Rightarrow \nabla f(1, 2, 0) &= 4\mathbf{i} + \mathbf{j} + \mathbf{k} \\ \therefore D_{\hat{\mathbf{v}}}f(1, 2, 0) &= \nabla f(1, 2, 0) \cdot \hat{\mathbf{v}} = 7/3 \end{aligned} \quad \left[\begin{array}{l} |\mathbf{v}| = \sqrt{4 + 4 + 1} = 3 \\ \hat{\mathbf{v}} = \frac{\mathbf{v}}{|\mathbf{v}|} = \frac{1}{3}(2\mathbf{i} - 2\mathbf{j} + \mathbf{k}) \end{array} \right] \quad \square$$

27.5 Unconstrained optimization: Local extrema and saddle point

The concept and determination of local maximum, local minimum and inflection point has been discuss in calculus of single variable functions. Here, we consider the analogous questions for functions of two or more variables. The notions of a local maximum, a local minimum, a critical point and a saddle point are described and defined in the following.

Generally, a function has a **local maximum** (or **local minimum**) at a point where it takes on the largest (or the smallest) value in a small region around the point. Points at which the gradient of f is either $\vec{0}$ or undefined are called **critical points** of f . On the other hand, a function has a **saddle point** at a point if the value of the function becomes larger in some directions and smaller in other directions in a small region around the point. Note that a critical point can be either a local maximum, a local minimum or even a saddle point.

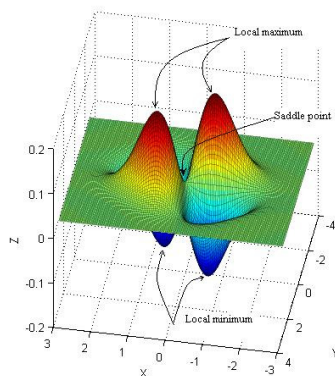


Fig 26: Local max, min and saddle point

More precisely, let $P(x, y)$ and $P_0(x_0, y_0)$ be two points inside the domain in which the function $f(x, y)$ is defined, then f is said to have a

- **local maximum** at P_0 if $f(P) \leq f(P_0)$ for P sufficiently close to P_0 ;
- **local minimum** at P_0 if $f(P) \geq f(P_0)$ for P sufficiently close to P_0 ;
- **critical point** at P_0 if $\nabla f(P_0) = \vec{0}$ or undefined.

Note that (assuming ∇f is defined everywhere) the condition for a critical point is

$$\nabla f = f_x \mathbf{i} + f_y \mathbf{j} + f_z \mathbf{k} = \vec{0},$$

which means all the partial derivative of f equal zero. Hence for a two-variable function, $f = f(x, y)$, the equivalent condition for a critical point is

$$f_x = f_y = 0.$$

On the other hand, if P_0 is a critical point of f and within a sufficiently small distance of P_0 , there exist two points P_1 and P_2 such that

$$f(P_1) > f(P_0) \text{ and } f(P_2) < f(P_0),$$

then we say f has a **saddle point** at P_0 .

27.5.1 The second derivative test for two-variable functions

Here we consider how to determine whether a critical point is a local maximum, local minimum or saddle point using the second derivative test. Our emphasis is on functions of two variables.

Consider $f = f(x, y)$ is a two-variable function having continuous first and second order partial derivatives. Suppose (x_0, y_0) is a point for which $f_x(x_0, y_0)$ and $f_y(x_0, y_0)$ are 0 (or equivalent to say $\nabla f = \vec{0}$), i.e. (x_0, y_0) is a critical point. Let the **discriminant**

$$D = f_{xx}(x_0, y_0)f_{yy}(x_0, y_0) - (f_{xy}(x_0, y_0))^2.$$

- If $D > 0$ and $f_{xx} > 0$, then f has a local minimum at (x_0, y_0) .
- If $D > 0$ and $f_{xx} < 0$, then f has a local maximum at (x_0, y_0) .
- If $D < 0$, then f has a saddle point at (x_0, y_0) .
- If $D = 0$, then no conclusion can be made.

MinMaxEg1

Example 27.14. Find the critical points of $f(x, y) = x^3 + 3xy^2 - 3x^2 - 3y^2 + 4$ and identify whether each such point gives a local maximum, a local minimum or a saddle point.

Solution

We start by evaluating all the partial derivatives that we need, sooner or later:

$$\begin{aligned} f_x &= 3x^2 + 3y^2 - 6x = 3(x^2 - 2x + y^2), \\ f_y &= 6xy = 6y = 6y(x - 1), \\ f_{xx} &= 6x - 6 = 6(x - 1), \\ f_{yy} &= 6x - 6 = 6(x - 1), \\ f_{xy} &= 6y \Rightarrow (f_{xy})^2 = 36y^2. \end{aligned}$$

Letting $f_x = 0$ and $f_y = 0$ give

$$x^2 - 2x + y^2 = 0 \Rightarrow x^2 - 2x = -y^2, \quad (a1)$$

$$6y(x - 1) = 0 \Rightarrow y = 0 \text{ or } x = 1. \quad (a2)$$

From (a2), for the case when $y = 0$ is substituted into (a1), then

$$\begin{aligned} x^2 - 2x = 0 &\Rightarrow x(x - 2) = 0 \Rightarrow x = 0, 2 \\ &\Rightarrow (0, 0), (2, 0) \text{ are critical points.} \end{aligned}$$

From (a2), for the case when $x = 1$ is substituted into (a1), then

$$\begin{aligned} -y^2 = -1 &\Rightarrow y = 1, -1 \\ &\Rightarrow (1, 1), (1, -1) \text{ are critical points.} \end{aligned}$$

So, there are four critical points: $(0, 0)$, $(2, 0)$, $(1, 1)$ and $(1, -1)$. The discriminant is

$$D = f_{xx}f_{yy} - (f_{xy})^2 = 36(x - 1)^2 - 36y^2 = 36[(x - 1)^2 - y^2].$$

From the second derivative test:

- ▷ $D(0, 0) = 36[(-1)^2 - 0] = 36 > 0$; $f_{xx}(0, 0) = 6(0 - 1) < 0 \Rightarrow (0, 0)$ is a local maximum.
- ▷ $D(2, 0) = 36[1 - 0] = 36 > 0$; $f_{xx}(2, 0) = 6(2 - 1) > 0 \Rightarrow (2, 0)$ is a local minimum.
- ▷ $D(1, 1) = -36 < 0 \Rightarrow (1, 1)$ is a saddle point.
- ▷ $D(1, -1) = -36 < 0 \Rightarrow (1, -1)$ is also a saddle point.

□

Example 27.15. Locate all the local extrema and saddle points of $f(x, y) = 4xy - x^4 - y^4$.

Solution

To find the critical points, we solve $f_x = f_y = 0$ to yield

$$f_x = 4y - 4x^3 = 0 \Rightarrow y = x^3, \quad (a1)$$

$$f_y = 4x - 4y^3 = 0 \Rightarrow x = y^3. \quad (a2)$$

Substituting (a1) into (a2) gives

$$x - x^9 = x(1 - x^8) = 0$$

$$\Rightarrow x = 0 \text{ or } x^8 = 1$$

$$\Rightarrow x = 0 \text{ or } x = 1, -1$$

Case 1: when $x = 0$, from (a1), one has $y = 0$, so $(0, 0)$ is a critical point.

Case 2: when $x = 1$, from (a1), one has $y = 1$, so $(1, 1)$ is also a critical point.

Case 3: when $x = -1$, from (a1), one has $y = -1$, which means $(-1, -1)$ is another critical point.

So, the three critical points of f are $(0, 0)$, $(1, 1)$ and $(-1, -1)$. Next, we use the second derivative test to determine the nature of these critical points. The discriminant is

$$D(x, y) = f_{xx}f_{yy} - f_{xy}^2 = (-12x^2)(-12y^2) - 4^2 = 144x^2y^2 - 16.$$

Then,

$$\triangleright D(0, 0) = -16 < 0 \Rightarrow (0, 0) \text{ is a saddle point.}$$

$$\triangleright D(1, 1) = 144 - 16 > 0; f_{xx}(1, 1) = -12 < 0 \Rightarrow (1, 1) \text{ is a local maximum.}$$

$$\triangleright D(-1, -1) = 144 - 16 > 0; f_{xx}(-1, -1) = -12 < 0 \Rightarrow (-1, -1) \text{ is also a local maximum.}$$

The graph of f is shown in **Fig 27**.

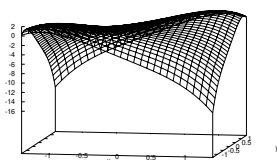


Fig 27: Graph of $f(x, y) = 4xy - x^4 - y^4$.

□

In problems of finding the critical points of a function, there can be quite some amount of algebra to handle especially when the function has many critical points.

27.6 Constrained optimization: The Lagrange multipliers method

In calculus of multivariable functions, problems of maximizing or minimizing a function where the variables are related by one or more equations, termed “constrains” or “side conditions”, are of considerable importance. Most real optimization problems are constrained by some kind of external conditions. For instance, a firm wanting to optimize its production must be subject to its limited budget. As a second example, the problem of finding the radius of the largest possible sphere inscribable in the ellipsoid $x^2 + 2y^2 + 3z^2 = 6$ is equivalent to minimizing the function $f = x^2 + y^2 + z^2$ under the side condition: $x^2 + 2y^2 + 3z^2 = 6$.

In handling such problems, it is sometimes possible to eliminate some of the variables by substituting from the constrain into the subjective function. This eventually reduce the problem to an ordinary maximum and minimum problem like those considered in single variable calculus. However, the method of Lagrange multipliers works when this substitution procedure is not feasible. The Lagrange multipliers method treats the variables in a more symmetrical way and often more convenient. The method is summarized below.

To find the critical points of $f(x, y)$ subject to the constrain $g(x, y) = 0$, solve the system of equations

$$\nabla f(x, y) = \lambda \nabla g(x, y) \text{ and } g(x, y) = 0, \quad (27.7) \quad \text{EqLagrangel}$$

where the parameter λ is known as the **Lagrange multiplier**.

The method is illustrated by the following examples.

EqLag1

Example 27.16. Find the local extrema of the function $f(x, y) = xy$ subject to the constraint $x^2 + 4y^2 = 8$.

Solution

We recognize the objective function is

$$f(x, y) = xy$$

and the constraint is

$$g(x, y) = x^2 + 4y^2 - 8 = 0.$$

Calculating ∇f and ∇g yield

$$\nabla f = y\mathbf{i} - x\mathbf{j} \text{ and } \nabla g = 2x\mathbf{i} + 8y\mathbf{j}.$$

Then $\nabla f = \lambda \nabla g$ gives

$$y = 2\lambda x, \quad (\text{a1})$$

$$x = 8\lambda y. \quad (\text{a2})$$

Dividing (a1) by (a2) gives

$$\frac{y}{x} = \frac{x}{4y} \Rightarrow 4y^2 = x^2 \Rightarrow x = \pm 2y. \quad (\text{a3})$$

From the constraint, we know

$$x^2 + 4y^2 = 8. \quad (\text{a4})$$

Substituting (a3) into (a4):

$$\begin{aligned} 4y^2 + 4y^2 &= 8 \\ \Rightarrow 8y^2 - 8 &= 0 \\ \Rightarrow 8(y^2 - 1) &= 8(y + 1)(y - 1) = 0 \\ \Rightarrow y &= \pm 1 \end{aligned}$$

From (a3), when $y = 1$, we have $x = \pm 2 \Rightarrow$ the critical points are $A(2, 1)$ and $B(-2, 1)$.

Similarly, when $y = -1$, we have $x = \mp 2 \Rightarrow$ the critical points are $C(-2, -1)$ and $D(2, -1)$.

Checking the value of f at these points:

$$\begin{aligned} \Rightarrow f(A) &= 2(1) = 2 = f(C), \\ \Rightarrow f(B) &= -2(1) = -2 = f(D). \end{aligned}$$

We conclude that under the constraint, f has two local maxima: $A(2, 1)$, $C(-2, -1)$ and two local minima: $B(-2, 1)$, $D(2, -1)$. Note the Lagrange multiplier λ is not useful here, despite it may be solved by substituting (a3) into either (a1) or (a2).

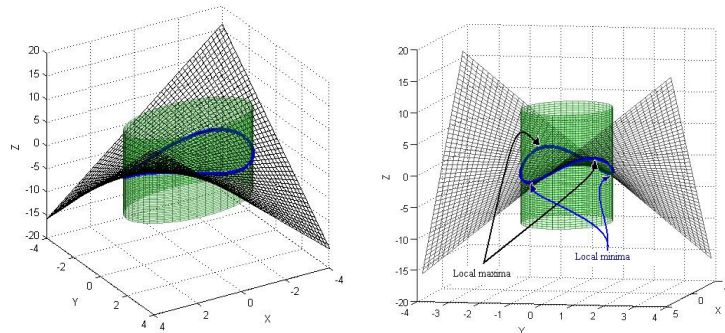


Fig 28: The intersection of $f(x, y) = xy$ and $x^2 + 4y^2 = 8$.

□

Example 27.17. Find the minimum of $f(x, y, z) = x^2 + y^2 + z^2$ subject to the constraint $x + 3y - 2z = 12$. Assume the critical point obtained do give a minimum.

Solution

Denote the quantity (or function) to be optimized as f and the constraint as g so that $f = x^2 + y^2 + z^2$ and $g = x + 3y - 2z - 12 = 0$. Calculating ∇f , ∇g and then forming $\nabla f = \lambda \nabla g$ yield

$$2x = \lambda, \quad (\text{a1})$$

$$2y = 3\lambda, \quad (\text{a2})$$

$$2z = -2\lambda, \quad (\text{a3})$$

$$x + 3y - 2z = 12. \quad (\text{a4})$$

Note that the last equation above comes from the constraint itself. There are four equations for us to solve for four unknowns. From (a1), (a2) and (a3), we straight away obtain

$$x = \lambda/2, \quad (\text{b1})$$

$$y = 3\lambda/2, \quad (\text{b2})$$

$$z = -\lambda. \quad (\text{b3})$$

Substituting all the (b1), (b2), (b3) into (a4) gives

$$\begin{aligned} \lambda \left(\frac{1}{2} + \frac{9}{2} + 2 \right) &= 12 \Rightarrow \lambda = \frac{12}{7}. \\ \Rightarrow x &= \frac{6}{7}, y = \frac{18}{7}, z = -\frac{12}{7}. \end{aligned}$$

Therefore, we conclude that under the constraint, the minimum of f occurs at point $P(\frac{6}{7}, \frac{18}{7}, -\frac{12}{7})$ and its minimum value is $f(P) = 72/7$. $\square$

Example 27.18. An open-top rectangular box is to have a volume of 32m^3 . Use the method of Lagrange multipliers to find its dimensions so that the amount of material used is minimized.

Solution

Assuming the box is having dimensions x, y, z as its length, width and height, respectively. The quantity to be optimized here is the material constituting the box which can be regarded as its surface area. This is denoted by f , and from simple geometry we know

$$f = xy + 2xz + 2yz.$$

This represents the base, the front and back, and the left and right surfaces of the box. The constraint here is the volume of the box to be denoted by g , thus

$$g = xyz - 32 = 0.$$

After computing ∇f and ∇g , one sees the components of $\nabla f = \lambda \nabla g$ give

$$y + 2z = \lambda yz,$$

$$x + 2z = \lambda xz,$$

$$2x + 2y = \lambda xy.$$

Multiplying these three equations by x, y and z respectively and knowing $xyz = 32$ from the constraint, we have

$$xy + 2xz = 32\lambda, \quad (\text{a1})$$

$$xy + 2yz = 32\lambda, \quad (\text{a2})$$

$$2xz + 2yz = 32\lambda. \quad (\text{a3})$$

Letting (a1)=(a2) gives

$$xy + 2xz = xy + 2yz \Rightarrow x = y. \quad (\text{b1})$$

Letting (a1)=(a3) gives

$$\begin{aligned}xy + 2xz &= 2xz + 2yz \\ \Rightarrow xy &= 2yz \text{ (after cancelling } 2xz\text{)} \\ \Rightarrow y^2 &= 2yz \text{ (using } x = y \text{ from (b1))} \\ \Rightarrow y &= 2z\end{aligned}\tag{b2}$$

Similarly, equalling (a2) to (a3) yields

$$\begin{aligned}xy + 2yz &= 2xz + 2yz \\ \Rightarrow x^2 &= 2xz \text{ (cancelling } 2yz \text{ and using (b1))} \\ \Rightarrow x &= 2z\end{aligned}\tag{b3}$$

Now, (b1) and (b2) are substituted into the constraint $xyz = 32$, so

$$xyz = y \cdot y \cdot \frac{y}{2} = 32 \Rightarrow y^3 = 64 \Rightarrow y = 4.$$

From (b1), (b2) and (b3), one straight away obtain $x = 4$ and $z = 2$. Therefore the optimal dimensions of the box in question are length = width = 4m and height = 2m.

□

The Lagrange multiplier method extends at once to situation in which f, g are functions of three or more variables, and to situation where there are two or more constraints. For example, to optimize $f(x, y, z)$ subject to the constraints $g(x, y, z) = 0$ and $h(x, y, z) = 0$, one solves the system of equations

$$\nabla f = \lambda \nabla g + \mu \nabla h, \quad g(x, y, z) = 0, \quad h(x, y, z) = 0$$

where the parameters λ, μ are the Lagrange multipliers. In this case, one has to solve for five unknowns x, y, z, λ, μ from a system of five (nonlinear) equations. One might notice as the number of variables and the number of constraints increase, handling the algebra can be a very tedious task although the Lagrange multipliers method can still be applied conveniently.

Chapter 28

Multiple Integral

28.1 Double Integrals

Definition 28.1. The *double integral* of f over a closed region R is defined as

$$\lim_{m,n \rightarrow \infty} \sum_{i=1}^m \sum_{j=1}^n f(x_i, y_j) \Delta x_i \Delta y_j$$

if the limit exists. If the limit exists, then f is said to be *integrable* over R and we denote this limit as $\iint_R f(x, y) dA$ where $dA = dx dy$ or $dy dx$.

Some Applications Of Double Integrals

- If $f(x, y) \geq 0$ and f is continuous on the rectangle R , then the volume of the solid that lies above R and under the surface $z = f(x, y)$ is given by

$$V = \iint_R f(x, y) dA.$$

- In particular, if $f(x, y) = 1$ in R , then $\iint_R f(x, y) dA = A(R)$, the area of R .

28.1.1 Double Integrals over Rectangular Regions

Theorem 28.1. Let $R = \{(x, y) : a \leq x \leq b, c \leq y \leq d\}$ be a closed rectangular region. If f is continuous on R , then

$$\iint_R f(x, y) dA = \int_c^d \int_a^b f(x, y) dx dy = \int_a^b \int_c^d f(x, y) dy dx.$$

Example 28.2. Evaluate the iterated integral $I = \int_0^3 \int_1^2 (4x + 6y) dx dy$.

Solution. $I = \int_0^3 [2x^2 + 6xy]_1^2 dy = [6y + 3y^2]_0^3 = 45.$ □

Example 28.3. Find the volume V of the solid under the plane $x + y + z = 4$ and over the rectangular region $R : 0 \leq x \leq 2, 0 \leq y \leq 1$ in the xy -plane.

Solution. $V = \int_0^2 \int_0^1 z dy dx = \int_0^2 \int_0^1 (4 - x - y) dy dx = 5.$ □

28.1.2 Double Integrals over Non-rectangular Regions

Theorem 28.4. Let f be continuous in the region R .

- If $R = \{(x, y) : a \leq x \leq b, g(x) \leq y \leq h(x)\}$, then

$$\iint_R f(x, y) dA = \int_a^b \int_{g(x)}^{h(x)} f(x, y) dy dx.$$

- If $R = \{(x, y) : g(y) \leq x \leq h(y), c \leq y \leq d\}$, then

$$\iint_R f(x, y) dA = \int_c^d \int_{g(y)}^{h(y)} f(x, y) dx dy.$$

Example 28.5. Evaluate $I = \iint_R 6xy dA$ over the region R enclosed between the curves $y = \sqrt{x}$, $2y = x$, $x = 2$ and $x = 4$.

Solution. $I = \int_2^4 \int_{x/2}^{\sqrt{x}} 6xy dy dx = 11$ □

Example 28.6. Use double integral to find the volume of the tetrahedron bounded by the coordinate planes and the plane $3x + 6y + 4z = 12$.

Solution. $V = \int_0^4 \int_0^{2-x/2} \frac{12-3x-6y}{4} dy dx = 4.$ □

Example 28.7. Find the volume of the solid that lies under the paraboloid $z = x^2 + y^2$ and above the region bounded by the x -axis, the y -axis, and the line $x + y = 1$.

Solution. $V = \int_0^1 \int_0^{1-x} (x^2 + y^2) dy dx = \frac{1}{6}.$ □

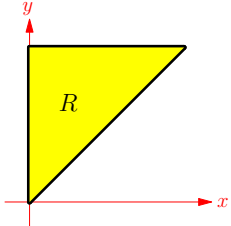
Example 28.8. Find the volume, V , of the solid that lies under the plane $2x + 2y + z = 18$ and above the triangular region R bounded by the lines $y = x$, $y = 2x$, and the line $x = 2$.

Solution. $V = \iint_R (18 - 2x - 2y) dA = \int_0^2 \int_x^{2x} (18 - 2x - 2y) dy dx = \int_0^2 [18y - 2xy - y^2]_x^{2x} dx = \int_0^2 (18x - 5x^2) dx = \left[9x^2 - \frac{5}{3}x^3 \right]_0^2 = \frac{68}{3}$ □

Example 28.9. Sketch the region of integration for the integral

$$I = \int_0^\pi \int_x^\pi \frac{\sin y}{y} dy dx$$

and write an equivalent integral with the order of integration reversed. Then evaluate the integral.

Solution.  $I = \int_0^\pi \int_0^y \frac{\sin y}{y} dx dy = \int_0^\pi \frac{\sin y}{y} (y - 0) dy = [-\cos y]_0^\pi = \cos 0 - \cos \pi = 2.$ □

Example 28.10. Evaluate the double integral $I = \int_0^1 \int_{y^2}^1 ye^{x^2} dx dy$ by reversing the order of integration.

Solution. The region of integration $R = \{(x, y) : y^2 \leq x \leq 1, 0 \leq y \leq 1\}$.

Describe R in another way : $0 \leq y \leq \sqrt{x}, 0 \leq x \leq 1$.

$$\text{Then } I = \int_0^1 \int_0^{\sqrt{x}} y e^{x^2} dy dx = \int_0^1 \left[\frac{y^2}{2} \right]_0^{\sqrt{x}} e^{x^2} dx = \int_0^1 \frac{x}{2} e^{x^2} dx = \left[\frac{e^{x^2}}{4} \right]_0^1 = \frac{e-1}{4}$$

□

28.1.3 Double Integrals in Polar Coordinates

Theorem 28.11. To evaluate a double integral using polar coordinates, we use

$$\iint_R f(x, y) dA = \iint_G f(r \cos \theta, r \sin \theta) r dr d\theta,$$

where G denotes the region of integration in polar coordinates.

Example 28.12. Use polar coordinates to evaluate the double integral $I = \int_{-1}^1 \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} (x^2 + y^2) dx dy$.

Solution. $I = \int_0^{2\pi} \int_0^1 r^2 r dr d\theta = 2\pi \times \frac{1}{4} = \pi/2$.

□

Example 28.13. Use polar coordinates to evaluate the double integral

$$I = \int_0^4 \int_0^{\sqrt{16-x^2}} \frac{dy dx}{(9+x^2+y^2)^{3/2}}.$$

Solution. $I = \int_0^{\pi/2} \int_0^4 \frac{1}{(9+r^2)^{3/2}} r dr d\theta = \frac{\pi}{2} \times \frac{2}{15} = \frac{\pi}{15}$.

□

Example 28.14. Use polar coordinates to evaluate the double integral

$$I = \int_0^1 \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} 6x^2 y dx dy.$$

Solution. Let R be the region of integration $\{(x, y) : -\sqrt{1-y^2} \leq x \leq \sqrt{1-y^2}, 0 \leq y \leq 1\}$.

In polar coordinates, $R = \{(r, \theta) : 0 \leq r \leq 1, 0 \leq \theta \leq \pi\}$.

Using the substitution, $x = r \cos \theta$, $y = r \sin \theta$, $r = \sqrt{x^2 + y^2}$, $dA = r dr d\theta$.

$$\begin{aligned} I &= \int_0^\pi \int_0^1 6(r \cos \theta)^2 r \sin \theta r dr d\theta = \int_0^\pi \int_0^1 6r^4 \cos^2 \theta \sin \theta dr d\theta \\ &= \int_0^\pi \left[\frac{6}{5} r^5 \right]_0^1 \cos^2 \theta \sin \theta d\theta = \int_0^\pi \frac{6}{5} \cos^2 \theta \sin \theta d\theta = \left[-\frac{6}{15} \cos^3 \theta \right]_0^\pi = \frac{4}{5} \end{aligned}$$

□

Exercise 1. Let R be the annular region lying between the two circles $x^2 + y^2 = 1$ and $x^2 + y^2 = 9$. Evaluate the integral $\iint_R (x + 2y^2) dA$ Answer: 40π

28.2 Triple Integrals and Volume Integrals

Definition 28.2. The *triple integral* of f over a bounded solid region D is defined as

$$\lim_{m,n,p \rightarrow \infty} \sum_{i=1}^m \sum_{j=1}^n \sum_{k=1}^p f(x_i, y_j, z_k) \Delta x_i \Delta y_j \Delta z_k$$

provided the limit exists. If it exists, then f is said to be *integrable* over D . We will denote the limit as $\iiint_D f(x, y, z) dV$ where $dV = dx dy dz$ or $dx dz dy$ or ...

Remark 28.15. The *volume* of $D = \iiint_D dV$.

28.2.1 Triple Integrals in A Rectangular Coordinates

Theorem 28.16. *Fubini's Theorem* Let $B = \{(x, y, z) : a \leq x \leq b, c \leq y \leq d, g \leq z \leq h\}$ be a closed rectangular box. If f is continuous on R , then

$$\iiint_B f(x, y, z) dV = \int_g^h \int_c^d \int_a^b f(x, y, z) dx dy dz.$$

Note: The iterated integral on the right can be replaced by any of the five other iterated integrals by changing the order of integration.

Example 28.17. Evaluate $\iiint_B xy^2z^3 dV$ over the rectangular box $B : 0 \leq x \leq 3, -2 \leq y \leq 2, 0 \leq z \leq 1$.

Solution. It evaluates to $\int_0^3 x dx \int_{-2}^2 y^2 dy \int_0^1 z^3 dz = \frac{9}{2} \times \frac{16}{3} \times \frac{1}{4} = 6$. □

Example 28.18. Evaluate the following integral by changing its order of integration:

$$\int_0^4 \int_0^1 \int_{2y}^2 \frac{2 \cos(x^2)}{\sqrt{z}} dx dy dz$$

Solution. After changing the order of integration, we have

$$\int_0^4 \int_0^2 \left[\int_0^{x/2} dy \right] 2 \cos x^2 dx z^{-1/2} dz = [2\sqrt{z}]_0^4 \int_0^2 x \cos x^2 dx = 4 \times \frac{\sin 4}{2} = 2 \sin 4.$$

□

Theorem 28.19. Let R be a closed region in the xy -plane and let $g(x, y)$ and $h(x, y)$ be continuous functions such that $g(x, y) \leq h(x, y)$ for all $(x, y) \in R$. If f is integrable over the region $D = \{(x, y, z) : (x, y) \in R, g(x, y) \leq z \leq h(x, y)\}$, then

$$\iiint_D f(x, y, z) dV = \iint_R \int_{g(x, y)}^{h(x, y)} f(x, y, z) dz dA.$$

Remark 28.20. Determining The Limits Of Integration

To find the limits for a particular order of integration, it is advisable to first determine the innermost limits, which may be functions of the outer two variables.

After integrating f with respect to the innermost variable, we end up with a double integral over the projection of D onto the coordinate plane of the outer two variables.

Then, you can determine the remaining limits of integration by methods used for double integrals.

Example 28.21. Using a triple integral to find volume Find the volume of the solid formed by the intersection of the cylinder $y = x^2$ and the two planes given by $z = 0$ and $y + z = 4$.

Solution. $V = \iint_R \int_0^{4-y} dz dA = \int_{-2}^2 \int_{x^2}^4 (4-y) dy dx = \frac{256}{15}$. □

Example 28.22. Find the volume of the region in the first octant bounded above by the cylinder $z = 1 - y^2$ and lying between the vertical planes $x + y = 1$ and $x + y = 3$.

Solution. The volume $V = \iint_R \int_0^{1-y^2} dz dA = \int_0^1 \int_{1-y}^{3-y} (1-y^2) dx dy = \frac{4}{3}$
where $R : 1 - y \leq x \leq 3 - y, 0 \leq y \leq 1$. □

28.2.2 Triple Integrals in Cylindrical and Spherical Coordinates

Theorem 28.23. *Triple Integrals in Cylindrical Coordinates* The volume element in cylindrical coordinates is $dV = r dz dr d\theta$.

Example 28.24. Use cylindrical coordinates to find $\iiint_D z \, dV$ where D is the region bounded by $z = \sqrt{x^2 + y^2}$ and $z = x^2 + y^2$.

Solution.
$$\int_0^{2\pi} \int_0^1 \int_{r^2}^r z r dz dr d\theta = \frac{\pi}{20}.$$

□

Example 28.25. Using a triple integral to find volume Find the volume of the solid enclosed by the paraboloids $z = x^2 + y^2$ and $z = 18 - x^2 - y^2$.

Solution.
$$\iint_R \int_{x^2+y^2}^{18-x^2-y^2} dz dA = 81\pi$$

where $R = \{(x, y) : x^2 + y^2 \leq 9\}$.

□

Example 28.26. Find the volume of the region D bounded by the paraboloids $z = 36 - 3x^2 - 3y^2$ and $z = x^2 + y^2$.

Solution.
$$\int_0^{2\pi} \int_0^3 \int_{r^2}^{36-3r^2} r dz dr d\theta = 162\pi.$$

□

Theorem 28.27. *Triple Integrals in Spherical Coordinates* The volume element in spherical coordinates is $dV = \rho^2 \sin \theta d\rho d\phi d\theta$.

Example 28.28. Use spherical coordinates to find

$$\iiint_D (x^2 + y^2 + z^2)^2 \, dV$$

where D is the solid in the first octant bounded by the sphere $x^2 + y^2 + z^2 = 4$.

Solution.

$$I = \int_0^{\pi/2} \int_0^{\pi/2} \int_0^2 (\rho^2)^2 \rho^2 \sin \theta d\rho d\phi d\theta = \frac{64\pi}{7}.$$

□

Example 28.29. Use spherical coordinates to find the volume of the solid that lies above the cone $z = \sqrt{x^2 + y^2}$ and below the sphere $x^2 + y^2 + z^2 = z$.

Solution.
$$I = \int_0^{\pi/4} \int_0^{2\pi} \int_0^{\cos \theta} \rho^2 \sin \theta d\rho d\phi d\theta = \frac{\pi}{8}.$$

Note that from $\rho = \sqrt{x^2 + y^2 + z^2}$, $\cos \theta = \frac{z}{\sqrt{x^2 + y^2 + z^2}}$, we have

$$x^2 + y^2 + z^2 = z \Leftrightarrow \rho^2 = \rho \cos \theta \Rightarrow \rho = \cos \theta$$

and $x^2 + y^2 + z^2 = z$, $z = \sqrt{x^2 + y^2}$ implies

$$2z^2 = z \Rightarrow z = 0, 1/2.$$

□

Exercise 2. Use spherical coordinates to find the volume of the solid enclosed by the sphere $x^2 + y^2 + z^2 = 36$ and the planes $z = 0$ and $z = 3$ Answer: $144\pi - 45\pi = 99\pi$

Exercise 3. Use spherical coordinates to evaluate

$$\int_{-5}^5 \int_0^{\sqrt{25-x^2}} \int_0^{\sqrt{25-x^2-y^2}} \frac{2}{\sqrt{x^2 + y^2 + z^2}} \, dz dy dx.$$

..... Answer: 25π

28.2.3 Volume (Triple) Integrals

Consider a bounded solid region $\mathcal{V}$ in the xyz -space. Then the volume of a small element $\Delta\mathcal{V}$ is given by

$$\Delta V \approx \Delta x \Delta y \Delta z.$$

When Δx , Δy and $\Delta z \rightarrow 0$, $dV = dx dy dz$.

Definition 28.3. The *total volume* of a solid region $\mathcal{V}$:

$$V = \iiint_{\mathcal{V}} dV = \iiint_{\mathcal{V}} dx dy dz.$$

Theorem 28.30 (Volume Integral Over Boxes). *For a closed rectangle box, $\mathcal{V} = \{(x, y, z) : a \leq x \leq b, c \leq y \leq d, r \leq z \leq s\}$, we have*

$$\begin{aligned} \iiint_{\mathcal{V}} f(x, y, z) dV &= \int_a^b \int_{y=c}^d \int_{z=r}^s f(x, y, z) dz dy dx \\ &= \int_{y=c}^d \int_{x=a}^b \int_{z=r}^s f(x, y, z) dz dx dy \\ &= \vdots \end{aligned}$$

NOTE. ALL TOGETHER, THERE ARE SIX DIFFERENT TYPES.

Example 28.31. Evaluate $\iiint_{\mathcal{V}} x^2 y z dV$ over the rectangle box,

$$\mathcal{V} : 0 \leq x \leq 1, \quad 0 \leq y \leq 2, \quad 1 \leq z \leq 3.$$

Solution.

$$\begin{aligned} \int_0^1 \int_0^2 \int_1^3 x^2 y z dz dy dx &= \int_0^1 \int_0^2 x^2 y \left[\frac{1}{2} z^2 \right]_{z=1}^3 dy dx = \int_0^1 \int_0^2 4x^2 y dy dx \\ &= \int_0^1 2x^2 [y^2]_0^2 dx = \int_0^1 8x^2 dx = \frac{8}{3}. \end{aligned}$$

Or

$$\int_0^1 \int_0^2 \int_1^3 x^2 y z dz dy dx = \int_0^1 x^2 dx \int_0^2 y dy \int_1^3 z dz = \left[\frac{x^3}{3} \right]_0^1 \left[\frac{y^2}{2} \right]_0^2 \left[\frac{z^3}{3} \right]_1^3 = \frac{8}{3}.$$

□

Theorem 28.32 (Volume Integral Over General Region). NOTE. AGAIN, THERE ARE SIX DIFFERENT TYPES. BUT HERE ONLY DISCRIBE TWO OF THEM.

I. For $\mathcal{V} = \{(x, y, z) : (x, y) \text{ in } \mathcal{R}, h_1(x, y) \leq z \leq h_2(x, y)\}$;

$$\begin{aligned} \iiint_{\mathcal{V}} f(x, y, z) dV &= \iint_{\mathcal{R}} \int_{z=h_1(x, y)}^{h_2(x, y)} f(x, y, z) dz dA \\ &= \int_a^b \int_{y=g_1(x)}^{g_2(x)} \int_{z=h_1(x, y)}^{h_2(x, y)} f(x, y, z) dz dy dx. \end{aligned}$$

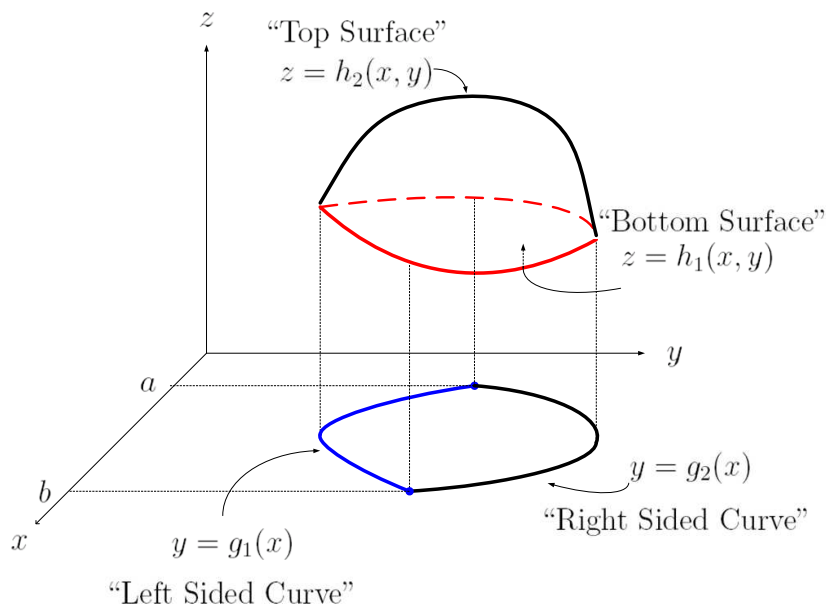
Here $\mathcal{R} = \{(x, y) : a \leq x \leq b, g_1(x) \leq y \leq g_2(x)\}$

II. For $\mathcal{V} = \{(x, y, z) : (x, z) \text{ in } \mathcal{R}, h_1(x, z) \leq y \leq h_2(x, z)\}$;

$$\begin{aligned} \iiint_{\mathcal{V}} f(x, y, z) dV &= \iint_{\mathcal{R}} \int_{y=h_1(x, z)}^{h_2(x, z)} f(x, y, z) dy dA \\ &= \int_a^b \int_{z=g_1(x)}^{g_2(x)} \int_{y=h_1(x, z)}^{h_2(x, z)} f(x, y, z) dy dz dx. \end{aligned}$$

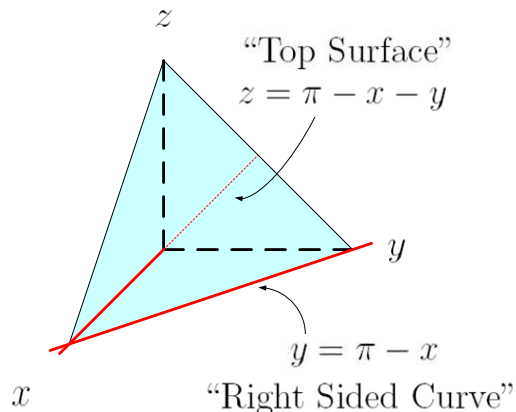
Here $\mathcal{R} = \{(x, z) : a \leq x \leq b, g_1(x) \leq z \leq g_2(x)\}$

$\vdots$



Example 28.33. Evaluate $\iiint_{\mathcal{V}} \sin(x+y+z) dV$ over the portion of the first octant cut off by the plane the plane $x+y+z=\pi$.

Solution. Step I. Sketch the surface.



Step II. Determine the solid region $\mathcal{V}$.

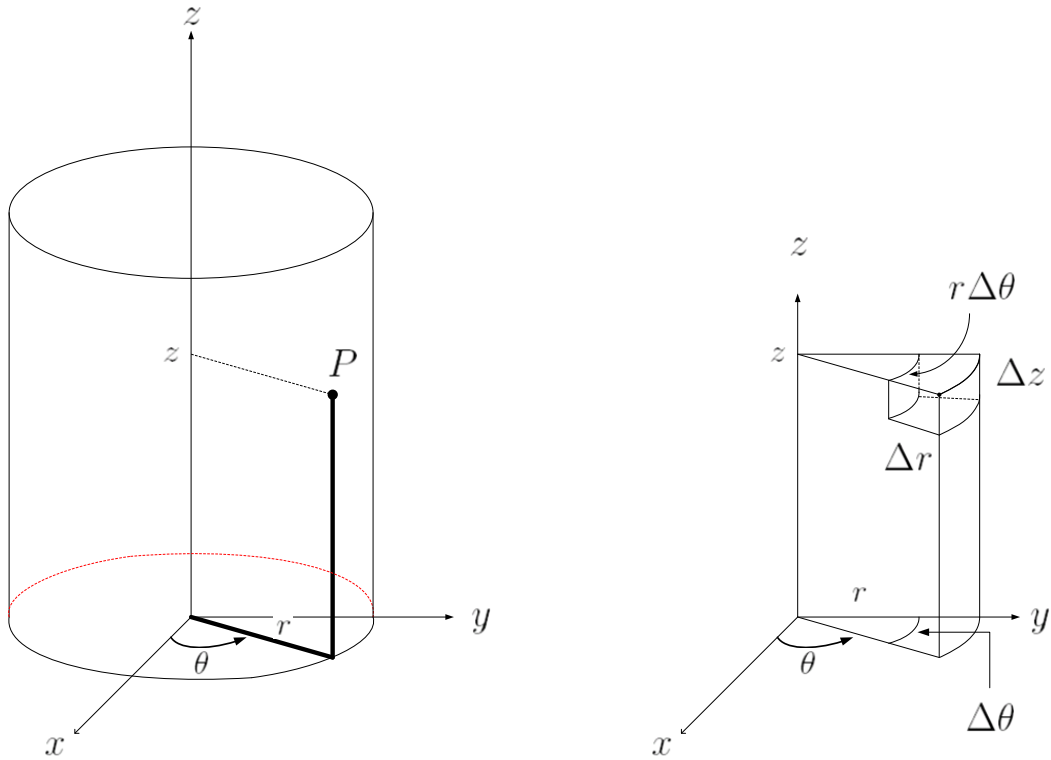
$$\mathcal{V} : 0 \leq x \leq \pi, \quad 0 \leq y \leq \pi - x, \quad 0 \leq z \leq \pi - y - x.$$

Step III.

$$\begin{aligned}
\iiint_{\mathcal{V}} \sin(x+y+z) dV &= \int_0^\pi \int_{y=0}^{\pi-x} \int_{z=0}^{\pi-y-x} \sin(x+y+z) dz dy dx \\
&= \int_0^\pi \int_{y=0}^{\pi-x} [-\cos(x+y+z)]_{z=0}^{\pi-y-x} dy dx \\
&= \int_0^\pi \int_{y=0}^{\pi-x} [1 + \cos(x+y)] dy dx \\
&= \int_0^\pi [y + \sin(x+y)]_{y=0}^{\pi-x} dx \\
&= \int_0^\pi [\pi - x - \sin x] dx = \frac{1}{2}\pi^2 - 2.
\end{aligned}$$

□

Cylindrical Coordinates System



A point $P(x, y, z)$ in space can be considered as having a position on a cylinder, i.e., by taking a transformation: $(r, \theta, z) \rightarrow (x, y, z)$, of the *cylindrical coordinates* as follows:

$$x = r \cos \theta, \quad y = r \sin \theta, \quad z = z;$$

where

$$r = \sqrt{x^2 + y^2}, \quad \tan \theta = \frac{y}{x},$$

From the above Figure, the volume of the small element is given by

$$\Delta V \approx (r \Delta \theta) \Delta r \Delta z.$$

When $\Delta \theta, \Delta r, \Delta z \rightarrow 0$, we have $dV = r dr d\theta dz$. So we have

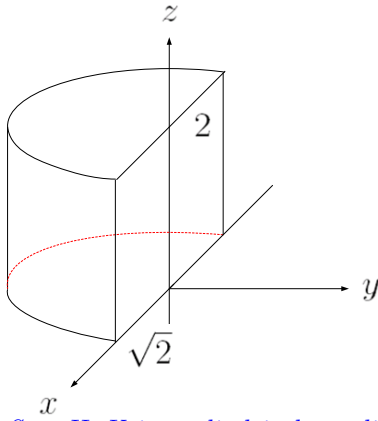
Formula 28.1 (Volume Form of Cylindrical Coordinates System). *For cylindrical coordinates:*

$$x = r \cos \theta, \quad y = r \sin \theta, \quad z = z; \quad (0 < r, \quad 0 < \theta \leq 2\pi, \quad -\infty < z)$$

we have $dV = r dr d\theta dz$.

Example 28.34. Evaluate $\iiint_{\mathcal{V}} \sqrt{z + x^2 + y^2} dV$ over the solid region $\mathcal{V}$ that bounded by the cylinder $x^2 + y^2 = 2$; the planes $z = 2$; placing above the xy -plane and on the left of xz -plane.

Solution. Step I.



Step II. Using cylindrical coordinates.

$$x = r \cos \theta, \quad y = r \sin \theta, \quad z = z; \quad (0 \leq r \leq \sqrt{2}, \quad \pi \leq \theta \leq 2\pi, \quad 0 \leq z \leq 2).$$

$$\Rightarrow dV = r dr d\theta dz.$$

Step III.

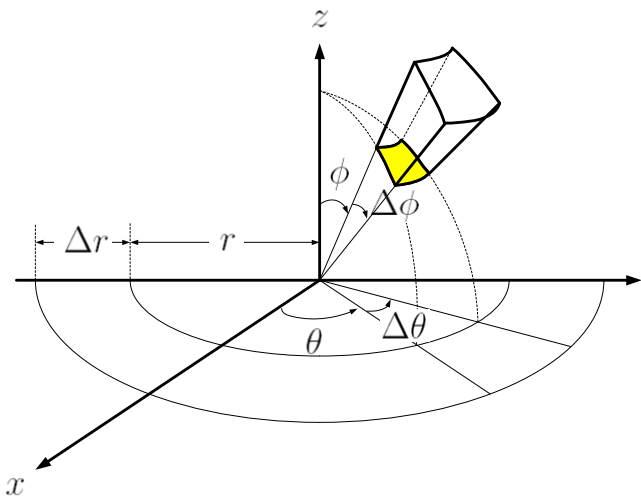
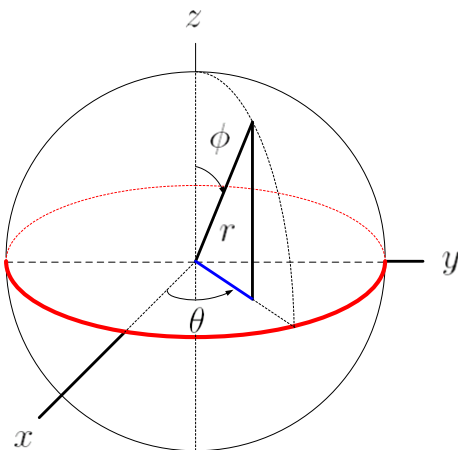
$$\begin{aligned} \therefore z + x^2 + y^2 &= z + r^2 \cos^2 \theta + r^2 \sin^2 \theta = z + r^2 \\ \therefore \iiint_{\mathcal{V}} \sqrt{z + x^2 + y^2} dV &= \int_{\pi}^{2\pi} \int_0^{\sqrt{2}} \int_0^2 \sqrt{z + r^2} r dr dz d\theta = \pi \int_0^2 \left[\frac{(z + r^2)^{3/2}}{3} \right]_{r=0}^{\sqrt{2}} dz \\ &= \pi \int_0^2 \frac{(z + r^2)^{3/2} - z^{3/2}}{3} dz = \frac{2\pi}{15} [2^5 - 2^{7/2}]. \end{aligned}$$

□

Spherical Coordinates System

For a point $P(x, y, z)$. The transformation : $(r, \phi, \theta) \rightarrow (x, y, z)$ of *spherical coordinates* is given as follows:

$$x = r \sin \phi \cos \theta, \quad y = r \sin \phi \sin \theta, \quad z = r \cos \phi.$$



The volume of the small element $\Delta \mathcal{V}$: $\Delta V \approx (r \Delta \phi \Delta r) r \sin \phi \Delta \theta$.

When $\Delta r, \Delta \phi, \Delta \theta \rightarrow 0$, we have $dV = r^2 \sin \phi d\phi d\theta dr$.

So we have

Formula 28.2 (Volume Form of Spherical Coordinates System). *For cylindrical coordinates:*

$$x = r \sin \phi \cos \theta, \quad y = r \sin \phi \sin \theta, \quad z = r \sin \phi; \quad (0 < r, \quad 0 < \phi \leq \pi, \quad 0 < \theta \leq 2\pi),$$

we have $dV = r^2 \sin \phi d\phi d\theta dr$.

Example 28.35. Evaluate $\iiint_{\mathcal{V}} x \, dV$ throughout the first octant of the sphere $x^2 + y^2 + z^2 = a^2$.

Solution. Step I.

[EXERCISE]

Step II. Using spherical polar coordinates.

$$x = r \sin \phi \cos \theta, \quad y = r \sin \phi \sin \theta, \quad z = r \cos \phi; \quad \left(0 \leq \phi \leq \frac{\pi}{2}, \quad 0 \leq \theta \leq \frac{\pi}{2}, \quad 0 \leq r \leq a\right).$$

$$\Rightarrow dV = r^2 \sin \phi \, dr \, d\theta \, d\phi.$$

Step III.

$$\begin{aligned} \therefore \iiint x \, dx \, dy \, dz &= \int_0^a \int_0^{\pi/2} \int_0^{\pi/2} r \sin \theta \cos \phi (r^2 \sin \theta \, dr \, d\theta \, d\phi) \\ &= \int_0^a r^3 \, dr \int_0^{\pi/2} \cos \phi \, d\phi \int_0^{\pi/2} \left(\frac{1}{2} - \frac{1}{2} \cos 2\theta\right) d\theta \\ &= \left(\frac{1}{4}a^4\right) \cdot (1) \cdot \left(\frac{1}{4}\pi\right) = \frac{1}{16}\pi a^4. \end{aligned}$$

□

28.3 Line Integrals

Definition 28.4. • A *plane curve* can be described by a pair of parametric equations

$$x = f(t), \quad y = g(t), \quad t \in I$$

where f and g are continuous functions on the interval I . The variable t is a *parameter* for the curve.

• A *space curve* can be described by a triple of parametric equations

$$x = f(t), \quad y = g(t), \quad z = h(t), \quad t \in I$$

where f, g and h are continuous functions on the interval I . A curve can also be specified by giving the position vector $\mathbf{r}(t)$ of a point $P = P(t) = (f(t), g(t), h(t))$. That is,

$$\mathbf{r}(t) = \overrightarrow{OP} = f(t)\mathbf{i} + g(t)\mathbf{j} + h(t)\mathbf{k}.$$

The tip of $\mathbf{r}(t)$ traces out the curve as t varies over I .

Example 28.36. The parametric equations for the line passing through the point (x_0, y_0, z_0) and parallel to the nonzero vector $\mathbf{v} = (a, b, c)$ are

$$x = x_0 + at, \quad y = y_0 + bt, \quad z = z_0 + ct.$$

Example 28.37.

- Parametrise the line segment joining the point $A(1, 4, 3)$ and $B(-2, 5, 6)$. [Note: The parametrization is not unique]
- Write down the corresponding vector equation for the line segment.

Solution.

- $\overrightarrow{AB} = (-3, 1, 3)$. The line is given by $x = 1 - 3t, y = 4 + t, z = 3 + 3t, 0 \leq t \leq 1$.
- $\mathbf{r}(t) = (1 - 3t)\mathbf{i} + (4 + t)\mathbf{j} + (3 + 3t)\mathbf{k}, 0 \leq t \leq 1$.

□

Example 28.38. The parametrization for the circle $x^2 + y^2 = a^2$ is

$$x = a \cos t, \quad y = a \sin t, \quad 0 \leq t \leq 2\pi.$$

Note that the circle is traced out *counterclockwise* as t increases from $t = 0$ to $t = 2\pi$.

Example 28.39. The parametrization of the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ is

$$x = a \cos t, \quad y = b \sin t, \quad 0 \leq t \leq 2\pi.$$

Note that the ellipse is traced out *counterclockwise* as t increases from $t = 0$ to $t = 2\pi$.

Example 28.40. Find parametric equations for the portion of the parabola $y = x^2 - x$ joining $A(-1, 2)$ and $B(3, 6)$, oriented from A to B .

Solution. $x = t, y = t^2 - t, -1 \leq t \leq 3.$ □

Example 28.41.

- Eliminate the parameter to find the Cartesian equation of the curve with parametric equations

$$x = 3t^2, \quad y = 5t + 2, \quad 0 \leq t \leq 2.$$

- Sketch the curve represented by the parametric equations. Indicate with an arrow the direction in which t increases.

Solution. $x = \frac{3}{25}(y - 2)^2, 2 \leq y \leq 12.$ □

Example 28.42. Find parametric equations for the line segment joining the points $A(2, 3, -4)$ and $B(-5, 0, 1)$. Write down the corresponding vector equation for the line segment.

Draw coordinate axes and sketch the line segment, indicate the direction of increasing t for your parametrisation.

Solution. $x = 2 - 7t, y = 3 - 3t, z = -4 + 5t, 0 \leq t \leq 1;$

$\mathbf{r} = (2 - 7t)\mathbf{i} + (3 - 3t)\mathbf{j} + (-4 + 5t)\mathbf{k}, 0 \leq t \leq 1;$

The line segment with this parametrization moves from A to B . □

Exercise 4. Find parametric equations for the portion of the parabola $x = y^2$ joining $A(4, -2)$ and $B(9, 3)$, oriented from A to B Answer: $x = t^2, y = t, -2 \leq t \leq 3.$

Definition 28.5. Let $\mathbf{r}(t)$ be the position vector of a curve C defined on an interval I . We call C a *smooth curve* or $\mathbf{r}(t)$ a *smooth function* if $\mathbf{r}'(t)$ is continuous and $\mathbf{r}'(t) \neq \mathbf{0}$ for any value of $t \in I$ (except possibly at any endpoints of I).

Note: The tangent vector $\mathbf{r}'(t)$ for a smooth curve varies “continuously” without abrupt changes in direction as t increases. We can think of a smooth curve as a curve with no “sharp corners” (called *cusps*).

Definition 28.6. A curve that is made up of a finite number of smooth curves is called *piecewise smooth*.

Definition 28.7. The *length* of a smooth curve $C : \mathbf{r}(t) = (x(t), y(t), z(t)), a \leq t \leq b$, that is traced exactly once as t increases from a to b is given by

$$L = \int_a^b \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2} dt = \int_a^b \|\mathbf{r}'(t)\| dt.$$

Definition 28.8. The *arc length parameter* s for the curve C with base point $P = P(a)$ is the function

$$s(t) = \int_a^t \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2} dt.$$

Note: By the Fundamental Theorem of Calculus,

$$\frac{ds}{dt} = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2}.$$

Definition 28.9. If f is defined on a smooth curve $C : x = x(t), y = y(t) \quad a \leq t \leq b$, then the *line integral of f along C* is

$$\int_C f(x, y) ds = \lim_{n \rightarrow \infty} \sum_{k=1}^n f(x_k^*, y_k^*) \Delta s_k$$

if the limit exists.

Theorem 28.43. If f is a continuous function, then

$$\int_C f(x, y) ds = \int_a^b f(x(t), y(t)) \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} dt.$$

- Recall that $ds = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} dt$.
- If $f(x, y) \geq 0$, then $\int_C f(x, y) ds$ represents the area of the “curved curtain” whose base is C and whose height above the point (x, y) is $f(x, y)$.

Example 28.44. Evaluate the line integral $\int_C x ds$, where C is the curve $x = t, y = t^2, \quad 0 \leq t \leq 1$.

Solution. $\int_C x ds = \int_0^1 t \sqrt{1 + 4t^2} dt = \frac{1}{12} \left[(1 + 4t^2)^{3/2} \right]_0^1 = \frac{\sqrt{125} - 1}{12}.$ □

Example 28.45. Evaluate $\int_C (2 + x^2 y) ds$, where C is the upper half of the circle $x^2 + y^2 = 9$.

Solution. A parametric form for C is $x = 3 \cos t, y = 3 \sin t, \quad 0 \leq t \leq \pi$.

Hence, $ds = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} dt = \sqrt{(-3 \sin t)^2 + (3 \cos t)^2} dt = 3 dt$ and

$$\int_C (2 + x^2 y) ds = \int_0^\pi (2 + (3 \cos t)^2 (3 \sin t)) 3 dt = 3 \left[2t - 9 \cos^3 t \right]_0^\pi = 6\pi + 54.$$
 □

Example 28.46. Integrate $f(x, y, z) = xy + y + z$ over the path

$$C : \mathbf{r}(t) = 2t\mathbf{i} + t\mathbf{j} + (2 - 2t)\mathbf{k}, \quad 0 \leq t \leq 1.$$

Solution. $\int_C (xy + y + z) ds = \int_0^1 (2t^2 - t + 2) 3 dt = 13/2.$ □

Theorem 28.47. If C is a piecewise-smooth curve; i.e., C is the union of a finite number of smooth curves $C_1, C_2, \dots, C_n$, where the end point of C_{i-1} is the initial point of C_i . Then

$$\int_C f(x, y) ds = \int_{C_1} f(x, y) ds + \int_{C_2} f(x, y) ds + \cdots + \int_{C_n} f(x, y) ds.$$

Example 28.48. Evaluate the line integral $I = \int_C (2 + x - yz) ds$, where C is the polygonal path consisting of the circular arc $x^2 + y^2 = 9, z = 2$ from $(3, 0, 2)$ to $(0, 3, 2)$ and the line segment from $(0, 3, 2)$ to $(3, 7, 2)$.

Solution.

- $C = C_1 + C_2$ where

$$C_1 : x = 3 \cos t, y = 3 \sin t, z = 2, \quad 0 \leq t \leq \frac{\pi}{2}$$

$$C_2 : x = 3t, y = 3 + 4t, z = 2, \quad 0 \leq t \leq 1.$$

- $I = I_1 + I_2$ where $I_1 = \int_{C_1} (2 + x - yz) ds$ and $I_2 = \int_{C_2} (2 + x - yz) ds$.

- On $C_1 : ds = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2} dt = \sqrt{(-3 \sin t)^2 + (3 \cos t)^2 + (0)^2} dt = 3dt$,
hence $I_1 = \int_0^{\pi/2} (2 + 3 \cos t - 2(3 \sin t))3dt = 3 \left[2t + 3 \sin t + 6 \cos t \right]_0^{\pi/2} = 3\pi - 9$.
- On $C_2 : ds = \sqrt{(3)^2 + (4)^2 + (0)^2} dt = 5dt$,
hence $I_2 = \int_0^1 (2 + 3t - 2(3 + 4t))5dt = 5 \left[-4t - \frac{5}{2}t^2 \right]_0^1 = -\frac{65}{2}$.
- $I = (3\pi - 9) + (-65/2) = 3\pi - \frac{83}{2}$.

□

Definition 28.10. The line integrals of f with respect to x and y are defined by

$$\int_C f(x, y) dx = \lim_{n \rightarrow \infty} \sum_{k=1}^n f(x_k^*, y_k^*) \Delta x_k \text{ and } \int_C f(x, y) dy = \lim_{n \rightarrow \infty} \sum_{k=1}^n f(x_k^*, y_k^*) \Delta y_k$$

provide the limits exist.

Note : If $x = x(t), y = y(t), a \leq t \leq b$ then $dx = x'(t)dt, dy = y'(t)dt$ and

$$\int_C f(x, y) dx = \int_a^b f(x(t), y(t)) x'(t) dt, \quad \int_C f(x, y) dy = \int_a^b f(x(t), y(t)) y'(t) dt.$$

Example 28.49. Evaluate the line integral $\int_C xy dx + (x - y) dy$, where C consists of the line segments from $(0, 0)$ to $(2, 0)$ and from $(2, 0)$ to $(3, 2)$.

Solution.

$$\int_C xy dx + (x - y) dy = \int_2^3 (2x^2 - 6x + 8) dx = 17/3$$

□

28.3.1 Line Integrals of Vector Fields

Definition 28.11. Consider a vector field on $\mathbb{R}^3$

$$\mathbf{F}(x, y, z) = M(x, y, z)\mathbf{i} + N(x, y, z)\mathbf{j} + P(x, y, z)\mathbf{k}.$$

$\mathbf{F}$ is *continuous* if its *component functions* M, N and P are continuous; *differentiable* if functions of M, N and P are differentiable, and so on.

If the partial derivatives of M, N and P all exist, then

- the *curl* of $\mathbf{F}$ is the vector field defined by

$$\text{curl } \mathbf{F} = \left(\frac{\partial P}{\partial y} - \frac{\partial N}{\partial z} \right) \mathbf{i} - \left(\frac{\partial P}{\partial x} - \frac{\partial M}{\partial z} \right) \mathbf{j} + \left(\frac{\partial N}{\partial x} - \frac{\partial M}{\partial y} \right) \mathbf{k}.$$

By introducing the differential operator $\nabla = \mathbf{i} \frac{\partial}{\partial x} + \mathbf{j} \frac{\partial}{\partial y} + \mathbf{k} \frac{\partial}{\partial z}$, we may write $\text{curl } \mathbf{F} = \nabla \times \mathbf{F}$.

- The *divergence* of $\mathbf{F}$ is the function defined by

$$\text{div } \mathbf{F} = \frac{\partial M}{\partial x} + \frac{\partial N}{\partial y} + \frac{\partial P}{\partial z}.$$

By introducing the differential operator $\nabla = \mathbf{i} \frac{\partial}{\partial x} + \mathbf{j} \frac{\partial}{\partial y} + \mathbf{k} \frac{\partial}{\partial z}$, we may write $\text{div } \mathbf{F} = \nabla \cdot \mathbf{F}$ and

$$\text{curl } \mathbf{F} = \nabla \times \mathbf{F} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ M & N & P \end{vmatrix}.$$

Example 28.50. Sketch the vector field $\mathbf{F}(x, y) = -\frac{1}{2}y\mathbf{i} + \frac{1}{2}x\mathbf{j}$.

Solution. $\mathbf{F} \cdot \mathbf{r} = 0$ and $\|\mathbf{F}\| = \frac{1}{2}\|\mathbf{r}\|$ implies that $\mathbf{F}$ is tangent to a circle centred at $(0, 0)$ and has length equal to $\frac{1}{2}$ the radius of that circle. □

Example 28.51. Suppose a spherical object of mass M is centred at the origin. Derive the formula for the gravitational field of force $\mathbf{F}(x, y, z)$ exerted by the mass on an object of mass m located at a point $P(x, y, z)$ in space. Then sketch this field.

Solution. $\mathbf{F} = -\frac{GMm}{\|\mathbf{r}\|^3}\mathbf{r}$. □

Exercise 5. Let $\mathbf{F} = 2xy\mathbf{i} - 5y^3z\mathbf{j} + 4xz\mathbf{k}$. Find the divergence and the curl of $\mathbf{F}$ Answer: $2y - 15y^2z + 4x$; $5y^3\mathbf{i} - 4z\mathbf{j} - 2x\mathbf{k}$

Exercise 6. Let $\mathbf{F} = 2yz\mathbf{i} - 5y^3x\mathbf{j} + 4yz^2\mathbf{k}$. Find the divergence and the curl of $\mathbf{F}$ Answer: $-15y^2x + 8zy$; $4z^2\mathbf{i} + 2y\mathbf{j} - (5y^3 + 2z)\mathbf{k}$

Definition 28.12. Let $C : x = x(t), y = y(t), z = z(t), a \leq t \leq b$ be a smooth curve with unit tangent vector $\mathbf{T}$. Then the *work done* by a force $\mathbf{F}$ in moving a particle along C from $t = a$ to $t = b$ is

$$W = \int_C \mathbf{F} \cdot \mathbf{T} \, ds.$$

Remark 28.52. Different ways to write the work integral:

- $W = \int_a^b \mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t) dt = \int_C \mathbf{F} \cdot d\mathbf{r}$
- $W = \int_C Mdx + Ndy + Pdz$ if $\mathbf{F}(x, y, z) = M(x, y, z)\mathbf{i} + N(x, y, z)\mathbf{j} + P(x, y, z)\mathbf{k}$.

Example 28.53. Find the work done by the force field $\mathbf{F}(x, y) = x^2\mathbf{i} - xy\mathbf{j}$ in moving a particle counterclockwise along the quarter-circle $\mathbf{r}(t) = \cos t\mathbf{i} + \sin t\mathbf{j}, 0 \leq t \leq \pi/2$.

Solution. $\int_0^{\pi/2} -2\cos^2 t \sin t \, dt = -\frac{2}{3}$ □

Example 28.54. Find the work done by the force field $\mathbf{F}(x, y, z) = y\mathbf{i} + z\mathbf{j} - x\mathbf{k}$ in moving a particle along the twisted cubic

$$C : x = t, y = t^2, z = t^3$$

from $(0, 0, 0)$ to $(1, 1, 1)$.

Solution. $\int_C \mathbf{F} \cdot d\mathbf{r} = \int_C ydx + zdy - xdz = \int_0^1 (t^2 + 2t^4 - 3t^3) dt = -\frac{1}{60}$. □

Example 28.55. Calculate the work done by $\mathbf{F}(x, y, z) = \mathbf{i} - y\mathbf{j} + xyz\mathbf{k}$ in moving a particle from $(0, 0, 0)$ to $(1, -1, 1)$ along the curve of intersection of the cylinder $y = -x^2$ and the plane $z = x$.

Solution. Let C be the curve of intersection.

Method 1

We can describe C using the parametric equations $x = t, y = -t^2, z = t$.

At $(0, 0, 0) : t = 0$ while at $(1, -1, 1) : t = 1$.

On $C :$

$$\mathbf{F} = \mathbf{i} + t^2\mathbf{j} - t^4\mathbf{k}$$

$$\frac{d\mathbf{r}}{dt} = \frac{dx}{dt}\mathbf{i} + \frac{dy}{dt}\mathbf{j} + \frac{dz}{dt}\mathbf{k} = \mathbf{i} - 2t\mathbf{j} + \mathbf{k}$$

$$\mathbf{F} \cdot \frac{d\mathbf{r}}{dt} = 1 - 2t^3 - t^4$$

$$\text{The work done} = \int_C \mathbf{F} \cdot d\mathbf{r} = \int_0^1 \mathbf{F} \cdot \frac{d\mathbf{r}}{dt} dt = \int_0^1 (1 - 2t^3 - t^4) dt = \frac{3}{10}$$

Method 2

$$y = -x^2, z = x \Rightarrow dy = -2xdx, dz = dx$$

$$\text{The work done} = \int_C \mathbf{F} \cdot d\mathbf{r} = \int_C dx - ydy + xyzdz$$

$$= \int_0^1 dx - (-x^2)(-2xdx) + x(-x^2)(x)dx = \int_0^1 (1 - 2x^3 - x^4)dx = \frac{3}{10} \quad \square$$

Exercise 7. Find the work done by the force $\mathbf{F} = -3xz\mathbf{i} + 2y\mathbf{j} + 4\mathbf{k}$ in the displacement along the line $x = -2z, y = 3z$ from $(4, -6, -2)$ to $(0, 0, 0)$ Answer: -60 .

Example 28.56. Assume a smooth curve C is described by the vector function $\mathbf{r}(t)$ for $a \leq t \leq b$. Let acceleration, velocity, and speed be given by $\mathbf{a} = \frac{d\mathbf{v}}{dt}$, $\mathbf{v} = \frac{d\mathbf{r}}{dt}$, and $v = \|\mathbf{v}\|$, respectively. Using Newton's second law $\mathbf{F} = m\mathbf{a}$, in the absence of friction, the work done by $\mathbf{F}$ in moving a particle of constant mass m from point A at $t = a$ to point B at $t = b$ is the same as the change in kinetic energy:

$$\int_C \mathbf{F} \cdot d\mathbf{r} = \int_C m\mathbf{a} \cdot d\mathbf{r} = m \int_C \frac{d\mathbf{v}}{dt} \cdot \mathbf{v} dt = m \int_C \frac{1}{2} \frac{d}{dt} v^2 dt = \frac{m}{2} [v(b)^2 - v(a)^2] = K(B) - K(A).$$

where $v^2 = \mathbf{v} \cdot \mathbf{v}$ and $\frac{d\mathbf{v} \cdot \mathbf{v}}{dt} = 2 \frac{d\mathbf{v}}{dt} \cdot \mathbf{v}$.

28.4 Surface Integrals

28.4.1 Parametric Surfaces

Definition 28.13. Let x, y , and z be functions of u and v that are continuous on a domain D in the uv -plane. The set of points (x, y, z) given by

$$\mathbf{r}(u, v) = x(u, v)\mathbf{i} + y(u, v)\mathbf{j} + z(u, v)\mathbf{k}$$

is called a *parametric surface*. The equations

$$x = x(u, v), y = y(u, v), z = z(u, v)$$

are called *parametric equations* for the surface. The variables u and v are the *parameters*, and D is the *parameter domain*.

Example 28.57. Identify and sketch the parametric surface S given by

$$\mathbf{r}(u, v) = 3 \cos u \mathbf{i} + 3 \sin u \mathbf{j} + v \mathbf{k}$$

where $0 \leq u \leq 2\pi$ and $0 \leq v \leq 4$.

Solution. S is the surface $x^2 + y^2 = 3^2, 0 \leq z \leq 4$; it is a right circular cylinder of height 4. The radius of the cylinder is 3, and the z -axis forms the axis of the cylinder. $\square$

Example 28.58. Identify and sketch the parametric surface S given by

$$\mathbf{r}(\phi, \theta) = \rho \sin \theta \cos \phi \mathbf{i} + \rho \sin \theta \sin \phi \mathbf{j} + \rho \cos \theta \mathbf{k}$$

where $0 \leq \theta \leq \pi$ and $0 \leq \phi \leq 2\pi$.

Solution. S is the surface $x^2 + y^2 + z^2 = \rho^2$; it is a sphere of radius ρ and centre $(0, 0, 0)$. $\square$

Example 28.59. Write the plane $x + 2y - 3z = 4$ in parametric form $\mathbf{r}(u, v)$.

Solution. $x = 4 - 2u + 3v, y = u, z = v$ or $\mathbf{r}(u, v) = (4 - 2u + 3v)\mathbf{i} + u\mathbf{j} + v\mathbf{k}$. $\square$

Example 28.60. Find a vector function whose graph is the surface $z = f(x, y)$.

Solution. The parametric representation of the surface can be written as $\mathbf{r}(u, v) = u\mathbf{i} + v\mathbf{j} + f(u, v)\mathbf{k}$ where $x = u, y = v$, or simply as $\mathbf{r}(x, y) = x\mathbf{i} + y\mathbf{j} + f(x, y)\mathbf{k}$. $\square$

Definition 28.14. Let S be a parametric surface given by

$$\mathbf{r}(u, v) = x(u, v)\mathbf{i} + y(u, v)\mathbf{j} + z(u, v)\mathbf{k}$$

over an open region D in the uv -plane such that x, y , and z have continuous partial derivatives on D . The *partial derivatives* of $\mathbf{r}$ with respect to u and v are defined as

$$\mathbf{r}_u = \frac{\partial x}{\partial u}(u, v)\mathbf{i} + \frac{\partial y}{\partial u}(u, v)\mathbf{j} + \frac{\partial z}{\partial u}(u, v)\mathbf{k}$$

and

$$\mathbf{r}_v = \frac{\partial x}{\partial v}(u, v)\mathbf{i} + \frac{\partial y}{\partial v}(u, v)\mathbf{j} + \frac{\partial z}{\partial v}(u, v)\mathbf{k}.$$

Let (u_0, v_0) be a point in D . A normal vector at the point

$$(x_0, y_0, z_0) = (x(u_0, v_0), y(u_0, v_0), z(u_0, v_0))$$

is given by

$$\mathbf{N} = \mathbf{r}_u \times \mathbf{r}_v = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} & \frac{\partial z}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} & \frac{\partial z}{\partial v} \end{vmatrix}.$$

Definition 28.15. Let S be a parametric surface $\mathbf{r}(u, v) = x(u, v)\mathbf{i} + y(u, v)\mathbf{j} + z(u, v)\mathbf{k}$ defined on a region D in the uv -plane.

S is *smooth* if $\mathbf{r}_u$ and $\mathbf{r}_v$ are continuous and $\mathbf{r}_u \times \mathbf{r}_v$ is not $\mathbf{0}$ for any (u, v) in D .

S is *piecewise smooth* if it consists of a finite number of smooth sections.

Remark 28.61. For a smooth surface, the *tangent plane* is the plane that contains the tangent vector $\mathbf{r}_u$ and $\mathbf{r}_v$, and the vector $\mathbf{r}_u \times \mathbf{r}_v$ is a normal vector to the plane.

Definition 28.16. Let S be a smooth parametric surface

$$\mathbf{r}(u, v) = x(u, v)\mathbf{i} + y(u, v)\mathbf{j} + z(u, v)\mathbf{k}$$

defined over an open region D in the uv -plane. Let (u_0, v_0) be a point in D . A normal vector at the point

$$(x_0, y_0, z_0) = (x(u_0, v_0), y(u_0, v_0), z(u_0, v_0))$$

is given by

$$\mathbf{N} = \mathbf{r}_u \times \mathbf{r}_v = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} & \frac{\partial z}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} & \frac{\partial z}{\partial v} \end{vmatrix}.$$

Example 28.62. Find the tangent plane to the surface $\mathbf{r}(u, v) = u\mathbf{i} + v\mathbf{j} + (u^2 + v^2)\mathbf{k}$ at the point $(1, 2, 3)$.

Solution. $\mathbf{r}_u = \mathbf{i} + 2u\mathbf{k}, \mathbf{r}_v = \mathbf{j} + 2v\mathbf{k}, \mathbf{r}_u \times \mathbf{r}_v = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & 0 & 2u \\ 0 & 1 & 2v \end{vmatrix} = -2u\mathbf{i} - 2v\mathbf{j} + \mathbf{k}; (x, y, z) = (u, v, u^2 + v^2) = (1, 2, 3) \Rightarrow (u, v) = (1, 2); \mathbf{r}_u \times \mathbf{r}_v = -2\mathbf{i} - 4\mathbf{j} + \mathbf{k}; -2x - 4y + z = -2(1) - 4(2) + 3 = -7 \quad \square$

28.4.2 Surface Area and Surface Integrals

Definition 28.17. Area of a Parametric Surface

If a smooth parametric surface S is given by the equation

$$\mathbf{r}(u, v) = x(u, v)\mathbf{i} + y(u, v)\mathbf{j} + z(u, v)\mathbf{k}, \quad (u, v) \in D$$

and S is covered just once as (u, v) varies over D , then the *surface area* of S is

$$A(S) = \iint_D \|\mathbf{r}_u \times \mathbf{r}_v\| dA$$

$$\text{where } \mathbf{r}_u = \frac{\partial x}{\partial u} \mathbf{i} + \frac{\partial y}{\partial u} \mathbf{j} + \frac{\partial z}{\partial u} \mathbf{k} \text{ and } \mathbf{r}_v = \frac{\partial x}{\partial v} \mathbf{i} + \frac{\partial y}{\partial v} \mathbf{j} + \frac{\partial z}{\partial v} \mathbf{k}.$$

Remark 28.63. For a surface S given by $z = f(x, y)$, we can parametrise the surface using the vector equation

$$\mathbf{r}(x, y) = x\mathbf{i} + y\mathbf{j} + f(x, y)\mathbf{k}$$

defined over the region R in the xy -plane. So $\mathbf{r}_x = \mathbf{i} + f_x\mathbf{k}$, $\mathbf{r}_y = \mathbf{j} + f_y\mathbf{k}$ and $\mathbf{r}_x \times \mathbf{r}_y = -f_x\mathbf{i} - f_y\mathbf{j} + \mathbf{k}$.

Thus we have $\|\mathbf{r}_x \times \mathbf{r}_y\| = \sqrt{f_x^2 + f_y^2 + 1}$ and

$$A(S) = \iint_R \|\mathbf{r}_x \times \mathbf{r}_y\| dA = \iint_R \sqrt{f_x^2 + f_y^2 + 1} dA.$$

Example 28.64. Find the area of the sphere of radius a given by

$$\mathbf{r}(\phi, \theta) = a \sin \theta \cos \phi \mathbf{i} + a \sin \theta \sin \phi \mathbf{j} + a \cos \theta \mathbf{k}$$

where the parameter domain is $D = \{(\phi, \theta) \mid 0 \leq \phi \leq 2\pi, 0 \leq \theta \leq \pi\}$.

Solution. $\mathbf{r}_\phi = -a \sin \theta \sin \phi \mathbf{i} + a \sin \theta \cos \phi \mathbf{j}$, $\mathbf{r}_\theta = a \cos \theta \cos \phi \mathbf{i} + a \cos \theta \sin \phi \mathbf{j} - a \sin \theta \mathbf{k}$ and $\mathbf{r}_\phi \times \mathbf{r}_\theta =$

$$\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ -a \sin \theta \sin \phi & a \sin \theta \cos \phi & 0 \\ a \cos \theta \cos \phi & a \cos \theta \sin \phi & -a \sin \theta \end{vmatrix} = -a^2 \sin^2 \theta \cos \phi \mathbf{i} - a^2 \sin^2 \theta \sin \phi \mathbf{j} - a^2 \sin \theta \cos \theta \mathbf{k}.$$

$$\begin{aligned} \text{Thus we have } \|\mathbf{r}_\phi \times \mathbf{r}_\theta\| &= \sqrt{a^4 \sin^4 \theta \cos^2 \phi + a^4 \sin^4 \theta \sin^2 \phi + a^4 \sin^2 \theta \cos^2 \theta} = \sqrt{a^4 \sin^4 \theta + a^4 \sin^2 \theta \cos^2 \theta} = \\ &= \sqrt{a^4 \sin^2 \theta} = a^2 \sin \theta; A(S) = \iint_D \|\mathbf{r}_\phi \times \mathbf{r}_\theta\| dA = \iint_D a^2 \sin \theta dA = \int_0^\pi \int_0^{2\pi} a^2 \sin \theta d\phi d\theta = 4\pi a^2. \end{aligned}$$

□

Example 28.65. Find the area of the surface S if it is the part of the plane $x + y + z = 1$ that lies in the first octant.

Solution. Let $f(x, y) = 1 - x - y$ and $R = \{(x, y) : 0 \leq y \leq 1 - x, 0 \leq x \leq 1\}$.

$$\Rightarrow A(S) = \iint_R \sqrt{3} dA = \sqrt{3}A(R) = \sqrt{3}/2.$$

□

Example 28.66. Find the surface area of the portion of the paraboloid $z = 4 - x^2 - y^2$ lying above the xy -plane.

Solution. The area of the surface S is $A(S) = \iint_R \sqrt{1 + f_x^2 + f_y^2} dA$.

$$f(x, y) = 4 - x^2 - y^2 \Rightarrow f_x = -2x, f_y = -2y \Rightarrow A(S) = \iint_R \sqrt{1 + 4x^2 + 4y^2} dA$$

The vertical projection of S onto the xy -plane is the circular region $R = \{(x, y) : x^2 + y^2 \leq 4\}$. In polar coordinates, $R = \{(r, \theta) : 0 \leq r \leq 2, 0 \leq \theta \leq 2\pi\}$.

Using the substitution $x^2 + y^2 = r^2$, $dA = r dr d\theta$,

$$A(S) = \int_0^{2\pi} \int_0^2 \sqrt{1 + 4r^2} r dr d\theta = 2\pi \left[\frac{1}{12} (1 + 4r^2)^{3/2} \right]_0^2 = \pi(17^{3/2} - 1)/6$$

□

Definition 28.18. If a smooth parametric surface S is given by the equation

$$\mathbf{r}(u, v) = x(u, v)\mathbf{i} + y(u, v)\mathbf{j} + z(u, v)\mathbf{k}, \quad (u, v) \in D$$

and S is covered just once as (u, v) varies over D , then the *surface integral* of $g(x, y, z)$ over S is

$$\iint_S g(x, y, z) dS = \iint_D g(x(u, v), y(u, v), z(u, v)) \|\mathbf{r}_u \times \mathbf{r}_v\| dA.$$

Remark 28.67. If S is the surface $z = f(x, y)$ and R is its vertical projection onto the xy -plane, then

$$\iint_S g(x, y, z) dS = \iint_R g(x, y, f(x, y)) \sqrt{1 + f_x^2 + f_y^2} dA.$$

Example 28.68. A curved lamina is the portion of the paraboloid $z = x^2 + y^2$ below the plane $z = 1$ and has a density $\sigma(x, y, z) = 6\sqrt{1 + 4z}$. Find the mass of the lamina.

$$\text{Solution. } M = \iint_S \sigma(x, y, z) \, dS = \iint_R 6\sqrt{1+4(x^2+y^2)}\sqrt{1+(2x)^2+(2y)^2} \, dA = \int_0^{2\pi} \int_0^1 6(1+4r^2)r \, dr \, d\theta = 18\pi. \quad \square$$

Example 28.69. Let S be the portion of the cone $z = 4 - 2\sqrt{x^2 + y^2}$ between $z = 0$ and $z = 4$. Evaluate the surface integral $\iint_S 3z\sqrt{5x^2 + 5y^2} \, dS$.

Solution.

$$S : z = 4 - 2\sqrt{x^2 + y^2}, \quad 0 \leq z \leq 4$$

$$z_x = -\frac{2x}{\sqrt{x^2 + y^2}}, \quad z_y = -\frac{2y}{\sqrt{x^2 + y^2}}$$

$$dS = \sqrt{1 + z_x^2 + z_y^2} \, dA = \sqrt{1 + \frac{4x^2}{x^2 + y^2} + \frac{4y^2}{x^2 + y^2}} \, dA$$

$$I = \iint_S 3z\sqrt{5x^2 + 5y^2} \, dS = \iint_R 3(4 - 2\sqrt{x^2 + y^2})\sqrt{5x^2 + 5y^2}\sqrt{1 + \frac{4x^2}{x^2 + y^2} + \frac{4y^2}{x^2 + y^2}} \, dA$$

$$= \iint_R 3(4 - 2\sqrt{x^2 + y^2})(5)\sqrt{x^2 + y^2} \, dA$$

Using polar coordinates, then $R : 0 \leq r \leq 2, 0 \leq \theta \leq 2\pi$ and

$$I = \int_0^{2\pi} \int_0^2 15(4 - 2r)r \cdot r \, dr \, d\theta = \int_0^{2\pi} \int_0^2 15(4r^2 - 2r^3) \, dr \, d\theta = 80\pi. \quad \square$$

Definition 28.19. Oriented Surfaces

A surface S is *orientable* if it is possible to choose a unit normal vector $\mathbf{n}$ at each point (x, y, z) on S (except possibly at any boundary points) in such a way that $\mathbf{n}$ varies continuously (have no abrupt changes in direction) as we traverse along any curve on S . Once $\mathbf{n}$ has been chosen, we say that we have *oriented* the surface.

Note: A surface can be oriented only if it has 2 sides, the process of orientation consists essentially in choosing which side we will call “positive” and which “negative”. (If the surface is closed, then it is natural to speak of the “inside” and “outside”.)

Example 28.70. The 2 possible orientations of a sphere are by inward unit normals or outward unit normals.

Example 28.71. A Mobius strip is a non-orientable surface since it has only 1 side.

Note : We will exclude the non-orientable surfaces from further consideration. Hence forth, whenever we say “surface”, we mean an “orientable surface”.

Definition 28.20. If $\mathbf{F}$ is a continuous vector field defined on an oriented surface S with unit normal vector $\mathbf{n}$, then the *surface integral of $\mathbf{F}$ over S* is $\iint_S \mathbf{F} \cdot \mathbf{n} \, dS$. This integral is also called the *flux* of $\mathbf{F}$ across S .

Remark 28.72.

- Recall that the gradient vector of a scalar function $h(x, y, z)$ at a point $P(x_0, y_0, z_0)$ is the vector

$$\nabla h(x_0, y_0, z_0) = h_x(x_0, y_0, z_0)\mathbf{i} + h_y(x_0, y_0, z_0)\mathbf{j} + h_z(x_0, y_0, z_0)\mathbf{k}.$$

- At any point (x_0, y_0, z_0) in the domain of $h(x, y, z)$, $\nabla h(x_0, y_0, z_0)$ is orthogonal to the surface $h(x, y, z) = h(x_0, y_0, z_0)$.
- If S is given by the equation $G(x, y, z) = 0$, then $\mathbf{n}$ may be taken to be one of the two unit vectors $\pm \frac{\nabla G}{\|\nabla G\|}$ depending on which one gives the preferred direction.

In particular, if S is given by $z = f(x, y)$, then by letting $G(x, y, z) = z - f(x, y)$, and we can orient S using the unit normal vectors

$$\mathbf{n} = \frac{-f_x\mathbf{i} - f_y\mathbf{j} + \mathbf{k}}{\sqrt{1 + f_x^2 + f_y^2}} \quad \text{or} \quad -\mathbf{n}.$$

Since the $\mathbf{k}$ -component of $\mathbf{n}$ ($-\mathbf{n}$) is positive (negative), $\mathbf{n}$ will give S an upward (downward) orientation.

- If S is given in parametric form by

$$\mathbf{r}(u, v) = x(u, v)\mathbf{i} + y(u, v)\mathbf{j} + z(u, v)\mathbf{k},$$

then

$$\mathbf{n} = \pm \frac{\mathbf{r}_u \times \mathbf{r}_v}{\|\mathbf{r}_u \times \mathbf{r}_v\|}.$$

$$\text{If } \mathbf{n} = \frac{\mathbf{r}_u \times \mathbf{r}_v}{\|\mathbf{r}_u \times \mathbf{r}_v\|}, \text{ then } \iint_S \mathbf{F} \cdot \mathbf{n} dS = \iint_D \mathbf{F} \cdot \frac{\mathbf{r}_u \times \mathbf{r}_v}{\|\mathbf{r}_u \times \mathbf{r}_v\|} \|\mathbf{r}_u \times \mathbf{r}_v\| dA = \iint_D \mathbf{F} \cdot (\mathbf{r}_u \times \mathbf{r}_v) dA.$$

Example 28.73. Evaluate the flux integral $I = \iint_S \mathbf{F} \cdot \mathbf{n} dS$ for the vector field $\mathbf{F} = -y\mathbf{i} + x\mathbf{j} + 3z\mathbf{k}$, where S is the hemisphere $z = \sqrt{16 - x^2 - y^2}$ with upward orientation.

Solution. Method 1: $z = \sqrt{16 - x^2 - y^2} \Rightarrow G(x, y, z) = x^2 + y^2 + z^2 = 16 \Rightarrow \mathbf{n} = \nabla G / \|\nabla G\|$
 $\Rightarrow \mathbf{F} \cdot \mathbf{n} = 6z^2 / \sqrt{4x^2 + 4y^2 + 4z^2} = \frac{3}{4}(16 - x^2 - y^2)$
 $\Rightarrow I = \int_0^{2\pi} \int_0^4 2r(16 - r^2)^{1/2} dr d\theta = 128\pi.$

Method 2: $\mathbf{r}(\phi, \theta) = 4 \sin \theta \cos \phi \mathbf{i} + 4 \sin \theta \sin \phi \mathbf{j} + 4 \cos \theta \mathbf{k}$, $0 \leq \theta \leq \pi/2$, $0 \leq \phi \leq 2\pi$,

$$\mathbf{r}_\phi \times \mathbf{r}_\theta = 16 \sin \theta (\sin \theta \cos \phi \mathbf{i} + \sin \theta \sin \phi \mathbf{j} + \cos \theta \mathbf{k}),$$

$$\mathbf{F} = -4 \sin \theta \sin \phi \mathbf{i} + 4 \sin \theta \cos \phi \mathbf{j} + 12 \cos \theta \mathbf{k},$$

$$\mathbf{F} \cdot (\mathbf{r}_\phi \times \mathbf{r}_\theta) = 64 \sin \theta (-\sin^2 \theta \sin \phi \cos \phi + \sin^2 \theta \sin \phi \cos \phi + 3 \cos^2 \theta),$$

and hence

$$I = \int_0^{2\pi} \int_0^{\pi/2} 192 \sin \theta \cos^2 \theta d\theta d\phi = 2\pi [-64 \cos^3 \theta]_0^{\pi/2} = \frac{128}{\pi}.$$

□

Example 28.74. Some Physical Examples of Surface Integrals

- If at any point (x, y, z) on a surface S , $f(x, y, z)$ gives the rate of flow of heat per unit area at that point, in units, say, of calories per second per square centimetre, then $\iint_S f(x, y, z) dS$ gives the number of calories per second flowing across the surface.
- If we assume a steady-state temperature distribution, where $T(x, y, z)$ denotes the temperature at each point in space, and if the region we consider is filled with homogeneous material having coefficient of thermal conductivity κ , then according to Fourier's Law,

$$\mathbf{Q} = -k \nabla T$$

gives, at each point in space, the direction in which the heat is flowing.

The magnitude of $\mathbf{Q}$ gives the rate of flow of heat per unit area across an area perpendicular to $\mathbf{Q}$. More generally, the scalar component of $\mathbf{Q}$ in the direction of a unit vector $\mathbf{n}$, namely $\mathbf{Q} \cdot \mathbf{n}$, gives the number of calories per unit time and per unit area crossing an element of area perpendicular to $\mathbf{n}$. It follows that $\iint_S (-k \nabla T) \cdot \mathbf{n} dS$ gives the number of calories per second flowing across the surface S .

Note. The reason for the negative sign is the the temperature gradient ∇T points in the direction of maximum rate of increase of temperature, whereas heat flows in the opposite direction, from hot to cold.

- If $\mathbf{v}$ denotes the velocity field of a fluid and μ its density, then $\iint_S \mu \mathbf{v} \cdot \mathbf{n} dS$ gives the rate of flow of liquid across the surface, expressed as mass per unit time.

28.5 Green's Theorem, Stokes' Theorem and Divergence Theorem

The theorems in this section are all the extension of Fundamental Theorem of Calculus to two dimensional spaces (Green's Theorem) and three dimensional spaces (Stokes' Theorem, Divergence Theorem).

28.5.1 Green's Theorem in the Plane

Definition 28.21. A plane curve $C : \mathbf{r} = \mathbf{r}(t)$ ($a \leq t \leq b$) has an *initial end point* at $\mathbf{r}(a)$ and a *final end point* at $\mathbf{r}(b)$. C is said to be *simple* if it does not intersect itself anywhere between its endpoints.

Example 28.75. A circle is a simple curve but the figure 8 is not.

Theorem 28.76. *Green's Theorem*

Let C be a positively oriented, piecewise-smooth simple closed curve in the plane and let D be the region bounded by C . If P and Q have continuous partial derivatives on an open region that contains D , then

$$\int_C Pdx + Qdy = \iint_D \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA$$

Remarks

- (Other formulas for Green's Theorem) Let $\mathbf{F} = P\mathbf{i} + Q\mathbf{j}$. Then

$$1. \int_C Pdx + Qdy = \iint_D (\text{curl } \mathbf{F}) \cdot \mathbf{k} dA$$

$$2. \int_C -Qdx + Pdy = \iint_D \left(\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} \right) dA = \iint_D \text{div } \mathbf{F} dA$$

- Green's Theorem gives the following formulas for the area of D :

$$A = \oint_C xdy = -\oint_C ydx = \frac{1}{2} \oint_C xdy - ydx$$

Example 28.77. Evaluate the line integral $\int_C (3y - e^{\sin x})dx + (7x + \sqrt{y^4 + 1})dy$, where C is the circle $x^2 + y^2 = 9$.

Solution. $\iint_R (7 - 3)dA = 36\pi$ □

Example 28.78. Verify that Green's Theorem is true for the line integral

$$\int_C xydx + x^2dy,$$

where C is the triangle with vertices $(0, 0)$, $(1, 0)$, and $(1, 2)$.

Solution. $\int_C xydx + x^2dy = \int_0^1 \int_0^{2x} xdydx = 2/3.$ □

Example 28.79. Find the area enclosed by the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$.

Solution. $\frac{1}{2} \oint_C xdy - ydx = \frac{1}{2} \int_0^{2\pi} a \cos t b \cos t dt - b \sin t a (-\sin t) dt = \pi ab$ □

Example 28.80. Verify that Green's Theorem is true for the line integral

$$\int_C y^2dx + x^2dy,$$

where C is the boundary curve of the region R lying between $y = x$ and $4x = y^2$.

Solution. Let $I_1 = \int_C Pdx + Qdy = \int_C y^2dx + x^2dy$ and $I_2 = \iint_R \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA = \iint_R (2x - 2y)dA.$

We have to verify that $I_1 = I_2$.

Calculate $I_2 = \iint_R (2x - 2y)dA$:

$$R : y \leq x \leq \frac{y^2}{4}, \quad 0 \leq y \leq 4$$

$$I_2 = \int_0^4 \int_{y^2/4}^y (2x - 2y) dx dy = \int_0^4 \left[x^2 - 2yx \right]_{y^2/4}^y dy = \int_0^4 \left(-\frac{y^4}{16} + \frac{y^3}{2} - y^2 \right) dy = \left[-\frac{y^5}{80} + \frac{y^4}{8} - \frac{y^3}{3} \right]_0^4 = -\frac{32}{15}.$$

Calculate $I_1 = \int_C y^2 dx + x^2 dy$:

$C = C_1 + C_2$

$C_1 : y = x$ from $x = 0$ to $x = 4$

On $C_1 : dy = dx$

$$\int_{C_1} y^2 dx + x^2 dy = \int_0^4 x^2 dx + x^2 dx = \frac{128}{3}$$

$C_2 : 4x = y^2$ from $y = 4$ to $y = 0$

On $C_2 : 4dx = 2ydy \Rightarrow dx = \frac{1}{2}ydy$

$$\int_{C_2} y^2 dx + x^2 dy = \int_4^0 y^2 \left(\frac{1}{2}ydy \right) + (y^2/4)^2 dy = \frac{-224}{5}.$$

$$I_1 = \int_{C_1} + \int_{C_2} = \frac{128}{3} + \frac{-224}{5} = -\frac{32}{15} = I_2 \text{ as expected.} \quad \square$$

Exercise 8. Verify that Green's Theorem is true for the line integral

$$\oint_C -5y^3 x dx + 5x^3 y dy,$$

where C is the positively oriented boundary curve of the region $R = \{(x, y) : x^2 + y^2 \leq 2, y \geq -x\}$.
Answer: 8

Green's Theorem can be extended to apply to regions with holes, i.e., regions that are not simply-connected.

Exercise 9. If $\mathbf{F}(x, y) = (-y\mathbf{i} + x\mathbf{j})/(x^2 + y^2)$, show that $\int_C \mathbf{F} \cdot d\mathbf{r} = 2\pi$ for every simple closed path that encloses the origin.

Answer: Note that $\int_C \mathbf{F} \cdot d\mathbf{r} = \int_{C'} \mathbf{F} \cdot d\mathbf{r}$ where C' is a counterclockwise-oriented circle with centre at the origin and radius a where a is chosen to be small enough that C' lies entirely inside C .

Exercise 10. Use Green's Theorem to evaluate the line integral along the given positively oriented curve:

$$\int_C (x^3 - y^3) dx + (x^3 + y^3) dy,$$

where C is the boundary of the region between the circles $x^2 + y^2 = 1$ and $x^2 + y^2 = 9$. . Answer: 120π

28.5.2 Stokes' Theorem

Definition 28.22. A surface $S : G(x, y, z) = C$ is *smooth* if the normal vector ∇G is continuous and never vanishes on S . (It has no "corners".)

It is said to be *piecewise smooth* if it consists of a finite number of smooth parts "joined together". For example, the surface of a sphere is smooth whereas the surface of a cube is piecewise smooth.

Theorem 28.81. Stokes' Theorem

Let S be an oriented piecewise-smooth surface that is bounded by a simple, closed, piecewise-smooth boundary curve C with positive orientation. If $\mathbf{F}$ is a continuously differentiable vector field on an open set that contains S , then

$$\oint_C \mathbf{F} \cdot d\mathbf{r} = \iint_S \text{curl} \mathbf{F} \cdot \mathbf{n} \, dS.$$

Example 28.82. Use Stokes' Theorem to compute the line integral $\oint_C \mathbf{F} \cdot d\mathbf{r}$, where $\mathbf{F}(x, y, z) = 4y\mathbf{i} - 3z\mathbf{j} + x\mathbf{k}$ and C is the triangle having vertices at $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$, counterclockwise when viewed from above.

$$\text{Solution. } \nabla \times \mathbf{F} = (3, -1, -4), \mathbf{n} = \frac{1}{\sqrt{3}}(1, 1, 1), dS = \sqrt{3}dA \Rightarrow I = \iint_R -2dA = -1. \quad \square$$

Remark 28.83. If S_1 and S_2 are oriented surfaces with the same oriented boundary curve C and both satisfy the hypotheses of Stokes' Theorem, then

$$\iint_{S_1} \text{curl} \mathbf{F} \cdot \mathbf{n} \, dS = \oint_C \mathbf{F} \cdot d\mathbf{r} = \iint_{S_2} \text{curl} \mathbf{F} \cdot \mathbf{n} \, dS.$$

This fact is useful when it is difficult to integrate over one surface but easy to integrate over the other.

Example 28.84. Let S be the part of the paraboloid $z = 5 - x^2 - y^2$ that lies above the plane $z = 1$, oriented upward. Verify Stokes' Theorem for the vector field $\mathbf{F} = 2z\mathbf{i} + 3x\mathbf{j} + 5y\mathbf{k}$.

$$\text{Solution. } \oint_C \mathbf{F} \cdot d\mathbf{r} = \oint_C 2zdx + 3xdy + 5ydz = \int_0^{2\pi} (-4 \sin t + 12 \cos^2 t) dt = 12\pi$$

$$\text{From } \nabla \times \mathbf{F} = (5, 2, 3) \text{ and } S = \{(x, y, z) : z = 1, x^2 + y^2 \leq 4\},$$

$$\iint_S \nabla \times \mathbf{F} \cdot \mathbf{n} dS = \int_0^{2\pi} \int_0^2 (5, 2, 3) \cdot (0, 0, 1) r dr d\theta = 3(\pi \times 2^2) = 12\pi. \quad \square$$

Example 28.85. Let $\mathbf{F} = (z - y)\mathbf{i} + (x - z)\mathbf{j} + (x - y)\mathbf{k}$.

Use Stokes' Theorem to evaluate the surface integral $I = \iint_S \text{curl} \mathbf{F} \cdot \mathbf{n} \, dS$, where S is the hemisphere $z = \sqrt{1 - x^2 - y^2}$.

Solution. By Stokes' Theorem,

$$I = \oint_C \mathbf{F} \cdot d\mathbf{r}$$

where C is the circle $x^2 + y^2 = 1, z = 0$.

$$C : x = \cos t, y = \sin t, z = 0, \quad 0 \leq t \leq 2\pi$$

$$\mathbf{r} = x\mathbf{i} + y\mathbf{j} + 0\mathbf{k} \Rightarrow \mathbf{r}' = -\sin t\mathbf{i} + \cos t\mathbf{j}$$

$$\mathbf{F} = (z - y)\mathbf{i} + (x - z)\mathbf{j} + (x - y)\mathbf{k} = (0 - \sin t)\mathbf{i} + (\cos t - 0)\mathbf{j} + (\cos t - \sin t)\mathbf{k}$$

$$\mathbf{F} \cdot \mathbf{r}' = \sin^2 t + \cos^2 t = 1$$

$$I = \int_0^{2\pi} \mathbf{F} \cdot \mathbf{r}' dt = \int_0^{2\pi} dt = 2\pi \quad \square$$

28.5.3 Gauss' Divergence Theorem

Theorem 28.86. *Gauss' Divergence Theorem*

Let S be a sectionally smooth surface. Let V be the solid enclosed by S . Let $\mathbf{F}$ be a vector field defined on S . If the component functions of $\mathbf{F}$ have continuous first derivatives in V , then

$$\oiint_S \mathbf{F} \cdot \mathbf{n} \, dS = \iiint_V \text{div} \mathbf{F} \, dV$$

where $\mathbf{n}$ is the outward unit normal.

Example 28.87. Let $\mathbf{F} = 3xy\mathbf{i} + y^2\mathbf{j} - x^2y^4\mathbf{k}$, and S be the surface of the tetrahedron with vertices $(0, 0, 0)$, $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. Use the divergence theorem to evaluate the surface integral $\iint_S \mathbf{F} \cdot \mathbf{n} \, dS$, where $\mathbf{n}$ is the outward-pointing unit normal.

$$\text{Solution. } \int_0^1 \int_0^{1-x} \int_0^{1-x-y} 5ydzdydx = \frac{5}{24}. \quad \square$$

Example 28.88. Verify the divergence theorem for $\mathbf{F} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ over the sphere $x^2 + y^2 + z^2 = a^2$.

$$\text{Solution. } \iiint_V \text{div} \mathbf{F} dV = \iiint_V 3dV = 4\pi a^3.$$

$$\iint_S \mathbf{F} \cdot \mathbf{n} dS = \iint_S adS = 4\pi a^3. \quad \square$$

Example 28.89. Verify the divergence theorem for $\mathbf{F} = x^2\mathbf{i} + y^2\mathbf{j} + z^2\mathbf{k}$ over the unit cube bounded by $x = 1, y = 1, z = 1$, and the coordinate planes.

Solution. Let $I_1 = \iint_S \mathbf{F} \cdot \mathbf{n} dS$ and $I_2 = \iiint_V \operatorname{div} \mathbf{F} dV$.

We have to verify that $I_1 = I_2$.

$$(a) \operatorname{div} \mathbf{F} = \frac{\partial}{\partial x}(x^2) + \frac{\partial}{\partial y}(y^2) + \frac{\partial}{\partial z}(z^2) = 2x + 2y + 2z$$

$$I_2 = \iiint_V (2x + 2y + 2z) dV = \int_0^1 \int_0^1 \int_0^1 (2x + 2y + 2z) dz dy dx = 3.$$

$$(b) I_1 = \int_{S_1} \mathbf{F} \cdot \mathbf{n} dS + \cdots + \int_{S_6} \mathbf{F} \cdot \mathbf{n} dS,$$

$S = S_1 + \cdots + S_6$, where $S_1, \dots, S_6$ are the 6 faces of the cube.

On $S_1 : x = 0, 0 \leq y \leq 1, 0 \leq z \leq 1$, we have

$\mathbf{n} =$ outward unit normal to $S_1 = -\mathbf{i}$

$$\mathbf{F} = x^2\mathbf{i} + y^2\mathbf{j} + z^2\mathbf{k} = y^2\mathbf{j} + z^2\mathbf{k}$$

$$\mathbf{F} \cdot \mathbf{n} = 0$$

$$\text{Hence, } J_1 = \iint_{S_1} \mathbf{F} \cdot \mathbf{n} dS = 0$$

On $S_2 : x = 1, 0 \leq y \leq 1, 0 \leq z \leq 1$, we have

$\mathbf{n} =$ outward unit normal to $S_2 = \mathbf{i}$

$$\mathbf{F} = \mathbf{i} + y^2\mathbf{j} + z^2\mathbf{k}$$

$$\mathbf{F} \cdot \mathbf{n} = 1$$

$$\text{Hence, } J_2 = \iint_{S_2} dS = \text{area of } S_2 = 1.$$

The remaining calculations are summarised in the following table:

S	$S_1 : x = 0$	$S_2 : x = 1$	$S_3 : y = 0$	$S_4 : y = 1$
$\mathbf{n}$	$-\mathbf{i}$	$\mathbf{i}$	$-\mathbf{j}$	$\mathbf{j}$
$\mathbf{F}$	$y^2\mathbf{j} + z^2\mathbf{k}$	$\mathbf{i} + y^2\mathbf{j} + z^2\mathbf{k}$	$x^2\mathbf{i} + z^2\mathbf{k}$	$x^2\mathbf{i} + \mathbf{j} + z^2\mathbf{k}$
$\mathbf{F} \cdot \mathbf{n}$	0	1	0	1
$\int_S \mathbf{F} \cdot \mathbf{n} dS$	0	1	0	1

$\therefore I_1 = 0 + 1 + 0 + 1 + 0 + 1 = 3 = I_2$ as expected. $\square$

28.6 Potential Functions, Path Independence, Conservative Fields

28.6.1 Path Independence and Conservative Fields

Definition 28.23. A set D of points on a xy -plane is called an *open set* if for every point P in D , there is a disk with centre P that lies entirely in D .

D is *connected* if any two points in D can be joined by a path that lies in D .

Definition 28.24. Conservative Fields

A vector field $\mathbf{F}$ is called a **conservative field** if $\mathbf{F} = \nabla\phi$ for some (differentiable) scalar field ϕ .

ϕ is called a **potential function** of $\mathbf{F}$.

Example 28.90. The vector field $\mathbf{F} = 2x\mathbf{i} + y\mathbf{j}$ is conservative since $\phi(x, y) = x^2 + \frac{y^2}{2}$ is a potential function of $\mathbf{F}$ i.e., $\mathbf{F} = \nabla\phi$

Example 28.91. The gravitational field $\mathbf{F} = -\frac{GMm}{\|\mathbf{r}\|^3} \mathbf{r}$ is conservative since $f(x, y, z) = \frac{GMm}{\sqrt{x^2 + y^2 + z^2}}$ is a potential function of $\mathbf{F}$ i.e., $\mathbf{F} = \nabla f$.

Theorem 28.92. Fundamental Theorem for Line Integrals

Let C be a piecewise smooth curve given by the vector function $\mathbf{r}(t)$, $a \leq t \leq b$, which begins at $\mathbf{r}(a)$ and ends at $\mathbf{r}(b)$. Let f be a differentiable function of 2 or 3 variables whose gradient vector $\nabla\phi$ is continuous on C . Then

$$\int_C \nabla\phi \cdot d\mathbf{r} = \phi(\mathbf{r}(b)) - \phi(\mathbf{r}(a)).$$

Proof.

$$\int_C \nabla\phi \cdot d\mathbf{r} = \int_a^b \nabla\phi \cdot \mathbf{r}'(t)dt = \int_a^b \left(\frac{\partial\phi}{\partial x} \frac{dx}{dt} + \frac{\partial\phi}{\partial y} \frac{dy}{dt} + \frac{\partial\phi}{\partial z} \frac{dz}{dt} \right) dt = \int_a^b \frac{d}{dt} \phi(\mathbf{r}(t)) dt = \phi(\mathbf{r}(b)) - \phi(\mathbf{r}(a)).$$

□

Example 28.93. Find the work done by the conservative field

$$\mathbf{F} = yz\mathbf{i} + xz\mathbf{j} + xy\mathbf{k} = \nabla(xyz)$$

along any smooth curve C from $A(-1, 3, 9)$ to $B(1, 6, -4)$.

Solution. $\phi(B) - \phi(A) = (1)(6)(-4) - (-1)(3)(9) = 3.$

□

Definition 28.25. Independence of Path

Let $\mathbf{F}$ be a continuous vector field with domain D . If

$$\varphi(C) = \int_C \mathbf{F} \cdot d\mathbf{r}$$

is a constant for any path C in D that have the same initial and terminal points, then we say that the line integral $\int_C \mathbf{F} \cdot d\mathbf{r}$ is *independent of path* and the field $\mathbf{F}$ is *conservative* on D .

Theorem 28.94. Independence of Path Theorem Suppose $\mathbf{F}$ is a vector field that is continuous on an open connected region D . Then $\int_C \mathbf{F} \cdot d\mathbf{r}$ is independent of path in D if and only if $\mathbf{F} = \nabla\phi$ for some scalar function ϕ ; i.e., if and only if $\mathbf{F}$ is conservative on D .

Theorem 28.95. The following statements are equivalent:

- The field $\mathbf{F}$ is conservative on D . That is, $\mathbf{F} = \nabla\phi$ for some scalar function ϕ .
- $\int_C \mathbf{F} \cdot d\mathbf{r}$ is independent of path.
- $\int_C \mathbf{F} \cdot d\mathbf{r} = 0$ around every closed path in D .
- $\nabla \times \mathbf{F} = \mathbf{0}$ implies $\mathbf{F} = -\nabla V$.

Remark 28.96. The above theorem tells us that

- the work done by a conservative field such as the gravitational or the electric field around a closed path is zero.
- the work done by a conservative field from a point A to another point B depends only A and B and not the path joining them.
- It is convenient to evaluate the work done integral by using the Fundamental Theorem for Line Integrals. But
 - how to tell if $\mathbf{F}$ is conservative?
 - how to find a potential function f if $\mathbf{F}$ is indeed conservative?

We need another concept to introduce a test for being conservative.

Definition 28.26. A *simply-connected region* in the plane is a connected region D such that every simple closed curve in D encloses only points that are in D . Intuitively speaking, a simply-connected region cannot contain any holes and cannot consist of two or more disjoint pieces.

Theorem 28.97. *The Component Test for Conservative Fields*

- Let $\mathbf{F}(x, y, z) = M(x, y, z)\mathbf{i} + N(x, y, z)\mathbf{j} + P(x, y, z)\mathbf{k}$ where M , N , and P have continuous first partial derivatives in an open simply connected region D . Then $\mathbf{F}$ is conservative if and only if $\nabla \times \mathbf{F} = \mathbf{0}$ in D if and only if

$$\frac{\partial P}{\partial y} = \frac{\partial N}{\partial z}, \quad \frac{\partial M}{\partial z} = \frac{\partial P}{\partial x}, \quad \frac{\partial N}{\partial x} = \frac{\partial M}{\partial y}.$$

- $\mathbf{F}(x, y) = M(x, y)\mathbf{i} + N(x, y)\mathbf{j}$ is conservative if and only if

$$\frac{\partial N}{\partial x} = \frac{\partial M}{\partial y}.$$

Note : To find a potential function f for $\mathbf{F}$, we solve the equation $\nabla\phi = \mathbf{F}$ for f . We do this by integrating the three equations

$$\frac{\partial\phi}{\partial x} = M, \quad \frac{\partial\phi}{\partial y} = N, \quad \frac{\partial\phi}{\partial z} = P.$$

Example 28.98. Show that $\mathbf{F} = (2x - 3)\mathbf{i} - z\mathbf{j} + \cos z\mathbf{k}$ is not conservative.

Solution. $\nabla \times \mathbf{F} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ 2x - 3 & -z & \cos z \end{vmatrix} = \mathbf{i} \neq \mathbf{0} \Rightarrow \mathbf{F} \text{ is not conservative.}$ □

Example 28.99. (a) Show that $\mathbf{F}(x, y, z) = (2x + 2y)\mathbf{i} + (2x - z^2)\mathbf{j} - 2yz\mathbf{k}$ is conservative.

(b) Find its potential function ϕ .

(c) Hence, use ϕ to find the work done by $\mathbf{F}$ along any smooth curve C joining the point $A(-1, 1, 0)$ to $B(2, 3, 0)$.

Solution. (a) $\nabla \times \mathbf{F} = (-2z + 2z)\mathbf{i} - (0 - 0)\mathbf{j} + (2 - 2)\mathbf{k} = \mathbf{0} \Rightarrow \mathbf{F}$ is conservative
 (b) $\phi(x, y, z) = x^2 + 2xy - yz^2 + C$
 (c) $\text{WD} = \phi(B) - \phi(A) = 17$ □

Example 28.100. Finding a Potential Function ϕ for $\mathbf{F}$

- Show that $\mathbf{F}(x, y) = 2xy\mathbf{i} + (x^2 - y)\mathbf{j}$ is conservative.
- Find its potential function ϕ .

Solution. 1. Let $M = 2xy$ and $N = x^2 - y$ be the components of $\mathbf{F}$.
 $\frac{\partial N}{\partial x} = 2x = \frac{\partial M}{\partial y} \Rightarrow \mathbf{F}$ is conservative
 2. Solve $\mathbf{F} = \nabla\phi(x, y)$ i.e. $2xy\mathbf{i} + (x^2 - y)\mathbf{j} = \phi_x\mathbf{i} + \phi_y\mathbf{j}$. This gives us the system

$$\begin{cases} \phi_x = 2xy & (1) \\ \phi_y = x^2 - y & (2) \end{cases}$$

 Integrating (1) wrt x , we obtain $\phi = x^2y + g(y)$ where $g(y)$ is a function of y .
 Differentiate the above equation wrt y and compare with (2), we have $\phi_y = x^2 + g'(y) = x^2 - y$
 i.e., $g'(y) = -y$ or $g(y) = -\frac{y^2}{2} + C$ where C is a constant.
 Hence, $\phi(x, y) = x^2y - \frac{y^2}{2} + C$ where C is a constant. □

Example 28.101. Finding a Potential Function ϕ for $\mathbf{F}$

1. Show that $I = \int_C 2xydx + (x^2 + z^2)dy + 2yzdz$ is independent of any path C between $(-1, 0, 1)$ and $(2, 3, 4)$.
2. Evaluate I .

Solution.

1. Let $\mathbf{F} = 2xy\mathbf{i} + (x^2 + z^2)\mathbf{j} + 2yz\mathbf{k}$.

Then $\nabla \times \mathbf{F} = \cdots = \mathbf{0} \Rightarrow \mathbf{F}$ is conservative.

First, we solve $\mathbf{F} = \nabla\phi(x, y, z)$ for ϕ .

That is, we have to solve the system

$$\mathbf{F} = \nabla\phi(x, y, z) \Leftrightarrow 2xy\mathbf{i} + (x^2 + z^2)\mathbf{j} + 2yz\mathbf{k} = \phi_x\mathbf{i} + \phi_y\mathbf{j} + \phi_z\mathbf{k}.$$

$$\begin{cases} \phi_x = 2xy & (1) \\ \phi_y = x^2 + z^2 & (2) \\ \phi_z = 2yz & (3) \end{cases}$$

Integrating (1) wrt x , we obtain $\phi = x^2y + g(y, z)$ where $g(y, z)$ is a function of y and z .

Differentiate the above equation wrt y and compare with (2), we have $\phi_y = x^2 + \frac{\partial g}{\partial y} = x^2 + z^2$

i.e., $\frac{\partial g}{\partial y} = z^2$ or $g(y, z) = z^2y + h(z)$ where $h(z)$ is a function of z .

Differentiate $\phi = x^2y + z^2y + h(z)$ wrt z and compare with (3), we have $\phi_z = 2yz + h'(z) = 2yz$ i.e., $h'(z) = 0$ or $h(z) = C$ where C is a constant.

Hence, $\phi(x, y, z) = x^2y + z^2y + C$ where C is a constant.

2. $\mathbf{F} = \nabla\phi(x, y, z) \Rightarrow I = \phi(-1, 0, 1) - \phi(2, 3, 4) = 0 - 60 = -60$.

□

Exercise 11. The inverse square law of gravitational attraction between two masses m_1 and m_2 is given by $\mathbf{F} = -\frac{Gm_1m_2}{\|\mathbf{r}\|^3}\mathbf{r}$, where $\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$.

- Show that $\mathbf{F}$ is conservative.
- Find a potential function for $\mathbf{F}$.

..... Answer: $\frac{Gm_1m_2}{\|\mathbf{r}\|}$.

Exercise 12. A particle in the plane is attracted to the origin with a force $\mathbf{F} = \|\mathbf{r}\|^n\mathbf{r}$, where n is a positive integer and $\mathbf{r} = x\mathbf{i} + y\mathbf{j}$ is the position vector of the particle.

- Show that $\mathbf{F}$ is conservative.
- Find the work done by $\mathbf{F}$ in moving the particle from (x_1, y_1) to (x_2, y_2) .

..... Answer: $\frac{1}{n+2} \left[(x_2^2 + y_2^2)^{(n+2)/2} - (x_1^2 + y_1^2)^{(n+2)/2} \right]$.

Part V

Differential Equations

Chapter 29

First Order Equations

The theory of *ordinary differential equation (ODE)* is important in engineering system modelling [Logan, 2006, Chapter 1]. The knowledge from this subject is also useful in the solution of boundary value problem and eigenfunction expansion in *Partial Differential Equation*.

In practise, numerical methods are used to solve problems modelling in ODE. Combining with the knowledge of numerical methods, many physical behaviour of engineering systems can be estimated numerically.

The main references for ordinary differential equations are Boyce and DiPrima [2012], O'Neil [2012], Zill [2017], James [2015], Kreyszig [2011], Boole [1859], Goursat [1904], Jordan and Smith [2007], Brauer and Nohel [1969], Chicone [1999], Robinson [2004], Hubbard and West [1991], Hubbard and West [1995], Soare et al. [2007], and Hartman [1982], etc.

29.1 Definitions

What is an Ordinary Differential Equation?

To answer this question, we would need to answer: “what is an equation?” An *equation* is a *formula* involving the “equality sign”. There are two types of equality:

$$\sin^2 x + \cos^2 x = 1, \quad 3x^4 - 7x + 1 = 0$$

where the left is true for all real number x while the right is true for only some real number x . The left equation is normally called an *identity*, while the right equation is called an *algebraic equation* (because it involves only polynomial).

Algebraic equations do not have “differentiation” in the formula. If we introduce “derivatives” into an equation, we will be getting ordinary differential equation which is defined below.

Definition 29.1 (First Order ODE). A *first order (implicit) ODE* is an equation involving the variable x , a function y which changes with respect to x and a derivative $y' = \frac{dy}{dx}$:

$$F(x, y, y') = 0. \tag{29.1}$$

e:ode:ord1

Here the partial derivative $F_{y'}(x, y, y') \neq 0$.

Definition 29.2 (Higher order ODE). An *ODE of order n* (or *n th order (implicit) ODE*) is an equation involving various derivatives of y to the n th order derivative $y^{(n)} = \frac{d^n y}{dx^n}$:

$$F(x, y, y', \dots, y^{(n)}) = 0. \tag{29.2}$$

e:ode:ordn

Here the partial derivative $F_{y^{(n)}}(x, y, y', \dots, y^{(n)}) \neq 0$.

Sometimes, it is possible to extract the highest order term to one side of the equation and the ODE in this form is said to be in an explicit form.

Definition 29.3. An ODE (29.1) is called *explicit* if it can be written in terms of a real-valued function f of two variables such that

$$\frac{dy}{dx} = f(x, y). \quad (29.3)$$

An n th order ODE (29.2) is said to be *explicit* if it can be written as

$$\frac{d^n y}{dx^n} = f\left(x, y, \frac{dy}{dx}, \dots, \frac{d^{n-1}y}{dx^{n-1}}\right). \quad (29.4)$$

Trying to solve an ODE is much more complicated compare to solving an algebraic equation. This is because the solutions of an algebraic function are just numbers. For example, the solution of a quadratic equation are two numbers:

$$ax^2 + bx + c = 0 \Rightarrow x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

where a, b, c are constants numbers.

But the solutions of an ODE are “functions”! Let us define the concept of (real-valued) solution for the first order ODE (29.1) and n th order ODE (29.2) properly as follows.

Definition 29.4. A (real-valued) function y is called the (*real-valued*) *solution* of an ODE (29.1) if y is a function over some interval $I \subset \mathbb{R}$ such that for all $x \in I$, $y'(x)$ exists and

$$F\left(x, y(x), \frac{dy}{dx}(x)\right) = 0.$$

A (real-valued) function y is called the (*real-valued*) *solution* of an ODE (29.2) if y is a function over some interval $I \subset \mathbb{R}$ such that for all $x \in I$, $y'(x), \dots, y^{(n)}(x)$ exist (i.e. y is n times differentiable) and

$$F\left(x, y(x), \frac{dy}{dx}(x), \dots, \frac{d^n y}{dx^n}(x)\right) = 0.$$

Definition 29.5. A solution of an n th order ODE (29.2) that contains n arbitrary constants (or parameters) is called the *general solution* of the equation (so the general solution of a first order ODE only consists of an arbitrary constant).

If the arbitrary constants in the general solution are assigned specific values, the result is called a *particular solution* of the equation.

A solution that cannot be obtained from the general solution for any choice of its arbitrary constants is called a *singular solution*.

eg:koayh12

Example 29.1. Show that the function $y = xe^x$ is a solution of the second order ODE $y'' - 2y' + y = 0$ on the interval $(-\infty, \infty)$.

Analysis: How do we show that the function is a solution to an ODE? We need to substitute the function into the ODE and show that the left-hand-side in (29.2) is zero.

Proof: By substituting $y = xe^x$ into the ODE, we have

$$y'' - 2y' + y = \frac{d^2}{dx^2}(xe^x) - 2\frac{d}{dx}(xe^x) + xe^x.$$

Now we calculate

$$\frac{d}{dx}(xe^x) = xe^x + e^x, \quad \frac{d^2}{dx^2}(xe^x) = xe^x + 2e^x,$$

and substituting back in the above formula

$$y'' - 2y' + y = (xe^x + 2e^x) - 2(xe^x + e^x) + xe^x = 0.$$

By definition, y is a solution of the ODE.

eg:koayh14

Example 29.2. Show that the function $y = e^{-x}$ is **not** a solution of the second order ODE $y'' - 2y' + y = 0$ on the interval $(-\infty, \infty)$.

Analysis: To show a function is **not** a solution to an ODE, we need to substitute the function into the ODE and show that the left-hand-side in (29.2) is **not** zero.

Proof: By substituting $y = e^{-x}$ into the ODE, we have

$$y'' - 2y' + y = \frac{d^2}{dx^2}(e^{-x}) - 2\frac{d}{dx}(e^{-x}) + e^{-x} = e^{-x} + 2e^{-x} + e^{-x} = 4e^{-x} \neq 0.$$

By definition, y is **not** a solution of the ODE.

Let C be an arbitrary constant throughout this topic.

eg:koayh11

Example 29.3. Show that $y = Ce^{3x}$ is the general solution of the ODE $y' = 3y$.

Proof: To apply the definition, we need to rewrite the ODE in the form

$$y' - 3y = 0$$

Now we substitute $y = Ce^{3x}$ into the left-hand-side of the ODE:

$$y' - 3y = C(3e^{3x}) - 3(Ce^{3x}) = 0.$$

By definition, y is a solution. It is a general solution because C is an arbitrary constant and the ODE is of order 1.

eg:koayh13

Example 29.4. Show that a one-parameter family of solutions $y = \left(\frac{x^2}{4} + C\right)^2$, $C > 0$, is a general solution of the ODE

$$y' = xy^{1/2}. \quad (*)$$

This ODE has a singular solution. Can you find the singular solution?

Solution. Since

$$y' - xy^{1/2} = 2\left(\frac{x^2}{4} + C\right)\left(\frac{2x}{4}\right) - x\left(\left(\frac{x^2}{4} + C\right)^2\right)^{1/2} = x\left(\frac{x^2}{4} + C\right) - x\left(\frac{x^2}{4} + C\right) = 0$$

By definition, y is a general solution of (*).

Clearly, $y = 0$ is also a solution of the DE. Since $y = 0$ cannot be obtained from the family for any choice of C , it is a singular solution of the ODE (*). $\square$

In Sections 29.2 to 29.4, we will investigate special types of first order ODEs in which the general solutions can be obtained using integration techniques.

29.2 Separable Equations

ord1:separable

If a first order ODE (29.3) has the form

$$\frac{dy}{dx} = f(x)$$

then the general solution can be obtained by taking “anti-derivative”

$$y = \int f(x)dx + C.$$

This is called a “*quadrature*” solution.

Example 29.5. Find the general solution of $\frac{dy}{dx} = \sin x$ and a specific solution such that $y(0) = 0$.

Solution: The general solution is

$$y = \int \sin x dx = -\cos x + C.$$

The constraint $y(0) = 0$ means that when $x = 0$,

$$y(0) = -\cos 0 + C = 0 \Rightarrow C = 1.$$

Hence, $y = 1 - \cos x$ is a “specific solution” to the ODE subjected to $y(0) = 0$.

If a first order ODE (29.3) has the form

$$\frac{dy}{dx} = f(x)g(y) \quad (29.5)$$

we call it a *separable ODE*. It has a general solution of the form by finding the “anti-derivative”:

$$\int \frac{1}{g(y)} dy = \int f(x) dx + C.$$

Example 29.6. Find the general solution of $\frac{dy}{dx} = \frac{e^{-x}}{y}$.

Solution: After the “separation of variables”, we obtain $\int y dy = \int e^{-x} dx$, which gives an “implicit” general solution

$$\frac{y^2}{2} = -e^{-x} + C.$$

By some algebraic operation, we can obtain an explicit general solution:

$$y = \pm \sqrt{2C - 2e^{-x}}, \quad x \geq \ln \frac{1}{C}, \quad C > 0.$$

Example 29.7 (Final Exam Sept 2016, Q1(a)). Find the general solution for the following separable equation

$$\frac{dy}{dx} = \frac{3+y}{1-x^2}, \quad 0 < x < 1.$$

Then, find the particular solution in which $y(\frac{1}{2}) = 0$. Write both (the general and particular) solutions in explicit analytic form. (8 marks)

Solution. By separation, we have

$$\frac{dy}{3+y} = \frac{dx}{1-x^2} \quad [1 \text{ mark}]$$

$$\Rightarrow \ln|3+y| = \int \left[\frac{\frac{1}{2}}{1+x} + \frac{\frac{1}{2}}{1-x} \right] dx = \frac{1}{2} \ln|1+x| - \frac{1}{2} \ln|1-x| + C \quad [4 \text{ marks}]$$

$$\Rightarrow y = \tilde{C} \sqrt{\frac{1+x}{1-x}} - 3 \quad [1 \text{ mark}]$$

$$\text{When } x = \frac{1}{2}, y = 0 = \tilde{C} \sqrt{\frac{\frac{3}{2}}{\frac{1}{2}}} - 3 \Rightarrow \tilde{C} = \sqrt{3} \dots\dots\dots [1 \text{ mark}]$$

$$\text{The particular solution is } y = \sqrt{\frac{3(1+x)}{1-x}} - 3. \dots\dots\dots [1 \text{ mark}] \quad \square$$

Sometimes, it may be necessary to apply some algebra (long division and partial fractions) or trigonometry techniques to reduce an ODE to a separable equation as illustrated below.

Example 29.8 (Final Exam Sept 2014, Q1(a)). Find the general solution (in explicit form) of the homogeneous equation

$$(x^2 - x)y' + 2(x^2 - x + 1)y = 0. \quad (6 \text{ marks})$$

Solution: Rewriting it as separable equation: $\frac{dy}{y} = \frac{2(x^2 - x + 1)}{x - x^2} dx \dots\dots\dots [1 \text{ mark}]$

Integrating both sides lead to

$$\begin{aligned} \ln|y| &= - \int \left[2 + \frac{2}{x(x-1)} \right] dx + C = -2x - \int \left[\frac{2}{x-1} - \frac{2}{x} \right] dx + C \\ &= -2x - 2 \ln|x-1| + 2 \ln|x| + C \end{aligned} \quad [3 \text{ marks}]$$

$$\text{Hence the general solution is } y = \frac{\tilde{C}x^2}{(x-1)^2 e^{2x}}, \quad \tilde{C} = e^C. \dots\dots\dots [2 \text{ marks}]$$

Example 29.9 (Final Exam Sept 2016, Q1(b)). Find the general solution for the following nonlinear ordinary differential equation

$$\frac{dy}{dx} + \sin \frac{x+y}{2} = \sin \frac{x-y}{2}.$$

Write your answer in analytic form.

(5 marks)

Solution. $\frac{dy}{dx} = \sin \frac{x-y}{2} - \sin \frac{x+y}{2} = 2 \cos \frac{\frac{x-y}{2} + \frac{x+y}{2}}{2} \sin \frac{\frac{x-y}{2} - \frac{x+y}{2}}{2}$ [2 marks]

$$= 2 \cos \frac{x}{2} \sin \frac{-y}{2} = -2 \cos \frac{x}{2} \sin \frac{y}{2}$$
 [1 mark]
$$\Rightarrow \int \frac{dy}{\sin \frac{y}{2}} = -2 \int \cos \frac{x}{2} dx \Rightarrow 2 \ln \tan \frac{y}{4} = -4 \sin \frac{x}{2} + C$$
 [2 marks]

□

29.3 Linear Equations and Integrating Factors

We say that an algebraic equation is “linear” when it has the form $ax = b$. Similarly, we say that a first order ODE (29.1) is “linear” when it has the following form:

$$\frac{dy}{dx} + P(x)y = Q(x). \quad (29.6)$$

Definition 29.6. When $Q(x) \neq 0$, the ODE (29.6) is called a *non-homogeneous linear ODE*. When $Q(x) = 0$, (29.6) is called a *homogeneous linear ODE*.

For a homogeneous linear ODE

$$\frac{dy}{dx} + P(x)y = 0, \quad (29.7)$$

we can rewrite it as a separable ODE and by applying the technique from the previous section, we can find the solution in implicit form:

$$\ln y = - \int P(x) dx + C$$

where C is an arbitrary constant. We can rewrite the general solution in explicit form:

$$y = \tilde{C} \exp \left\{ - \int P(x) dx \right\}, \quad \tilde{C} = e^C.$$

Example 29.10 (Radioactive decay). The decay of radioactive materials can be described by the first order ODE

$$\frac{dN}{dt} + kN = 0$$

with initial condition $N(0) = N_0$. Find the general solution and determine the particular solution that satisfies the initial condition.

Solution: By using separation of variables and taking anti-derivative, the general solution is

$$N = C \exp \left\{ - \int k dt \right\} = C e^{-kt}.$$

Form the initial condition $N(0) = N_0$, we obtain the particular solution:

$$N(0) = C = N_0 \Rightarrow \boxed{N = N_0 e^{-kt}}.$$

Example 29.11. Given a PDE $u_t - k u_{xx} = 0$ with boundary conditions $u(0, t) = u(L, t) = 0$ where $u = u(x, t)$; $k > 0$ and $L > 0$ are constants. When John solved this problem using a technique called “separation of variable”, he obtained many first order ODEs of the form

$$T'_n + k \left(\frac{n\pi}{L} \right)^2 T_n = 0, \quad n = 1, 2, \dots$$

where $T_n = T_n(t)$. You are requested to help John solve the above ODEs, i.e. find the general solution for all ODEs.

Solution. For each $n = 1, 2, \dots$, by using separation of variables, the ODE has a general solution

$$T_n = C_n e^{-(\frac{n\pi}{L})^2 kt}.$$

□

However, finding a solution for the non-homogeneous linear ODE (29.6) is more difficult. A new technique is required. Mathematicians found out that by multiplying an *integrating factor* $\alpha(x) = \exp(\int P(x)dx)$ to the right-hand and left-hand sides of (29.6), we obtain

$$\underbrace{\exp \int P(x)dx}_u \cdot \underbrace{\frac{dy}{dx}}_{v'} + \underbrace{\exp \int P(x)dx \cdot P(x)}_{u'} \underbrace{y}_v = Q(x) \exp \int P(x)dx. \quad (29.8)$$

Then, the *product rule for differentiation* from Calculus could be applied to reduce the left-hand-side of (29.8) to

$$\frac{d}{dx} \left[\exp \int P(x)dx \cdot y \right] = Q(x) \exp \int P(x)dx.$$

Denoting $\exp(\int P(x)dx)$ by $\alpha(x)$ and taking “anti-derivative”, we obtain the general solution

$$\alpha(x)y = \int \alpha(x)Q(x)dx \Rightarrow y = \frac{1}{\alpha(x)} \int \alpha(x)Q(x)dx.$$

An alternative solution is to apply the variation of parameters, which is used in Samuel S. Holland [1990, §1.3].

Let us investigate the integrating factor with a few examples.

Example 29.12. Find the general solution for the ODE:

$$\frac{1}{x} \frac{dy}{dx} - \frac{2y}{x^2} = x \cos x, \quad x > 0.$$

Solution: This can be transformed to non-homogeneous linear equation

$$\frac{dy}{dx} + \underbrace{\left(-\frac{2}{x}\right)}_{P(x)} y = \underbrace{x^2 \cos x}_{Q(x)}. \quad (*)$$

By using integrating factor

$$\alpha(x) = \exp \int \left(-\frac{2}{x}\right) dx = \exp(-2 \ln x) = \frac{1}{x^2},$$

the ODE (*) can be reduced and hence we obtain the general solution

$$\frac{1}{x^2} y = \int \frac{1}{x^2} x^2 \cos x dx \Rightarrow y = x^2 \int \cos x dx = x^2 (\sin x + C).$$

Example 29.13. Given a non-homogeneous PDE $u_t - k u_{xx} = \sinh(x)$ with boundary conditions $u(0, t) = u(L, t) = 0$ where $u = u(x, t)$; $k > 0$ and $L > 0$ are constants. When John solved this problem using a technique called “eigenfunction expansion” (found in the syllabus of UECM3003), he obtained many first order ODEs of the form

$$T'_n + k \left(\frac{n\pi}{L}\right)^2 T_n = \frac{2(-1)^{n+1} n\pi \sinh(L)}{L^2 + (n\pi)^2}, \quad n = 1, 2, \dots$$

where $T_n = T_n(t)$. You are requested to help John solve the above ODEs.

Solution: For each $n = 1, 2, \dots$, by using integrating factor method, we can reduce the ODEs to

$$\frac{d}{dt} \left[e^{(\frac{n\pi}{L})^2 kt} T_n \right] = e^{(\frac{n\pi}{L})^2 kt} \frac{2(-1)^{n+1} n\pi \sinh(L)}{L^2 + (n\pi)^2}$$

Integrating both sides, we have the implicit solution

$$e^{(\frac{n\pi}{L})^2 kt} T_n = \frac{1}{(\frac{n\pi}{L})^2 k} e^{(\frac{n\pi}{L})^2 kt} \frac{2(-1)^{n+1} n\pi \sinh(L)}{L^2 + (n\pi)^2} + C$$

and it can be written in explicit form

$$T_n = \frac{2L^2(-1)^{n+1} \sinh(L)}{k\pi(L^2 + (n\pi)^2)} + Ce^{-(\frac{n\pi}{L})^2 kt}.$$

Example 29.14 (Final Exam Sept 2014, Q1(a)). Write the general solution of the following first ODE in explicit form:

$$(1 + x^2)y' + xy = (1 + x^2)^{5/2}. \quad (7 \text{ marks})$$

[Remark: During the 2016 final exam, there were quite a few students mistakenly regarded this as the Bernoulli equation (29.15). This is wrong because the right-hand-side does not have the term y^α .]

Solution. Rewriting the first order linear non-homogeneous ODE in normal form:

$$y' + \frac{x}{1+x^2}y = (1+x^2)^{3/2}. \quad [1 \text{ mark}]$$

We then calculate the integrating factor

$$\alpha(x) = \exp \int \frac{x}{1+x^2} dx = \exp \left\{ \frac{1}{2} \ln |1+x^2| \right\} = \sqrt{1+x^2}. \quad [2 \text{ marks}]$$

Multiplying the integrating factor the ODE in normal form, we have

$$(\sqrt{1+x^2} \cdot y)' = (1+x^2)^2 = 1 + 2x^2 + x^4. \quad [2 \text{ marks}]$$

Taking integration, we obtain the general solution in explicit form

$$y = \frac{1}{\sqrt{1+x^2}} \left[x + \frac{2x^3}{3} + \frac{x^5}{5} + C \right]. \quad [2 \text{ marks}]$$

□

When an initial condition is imposed to the first order ODE, it will have a particular solution, which can normally be found from the general solution, as demonstrated from the following example.

Example 29.15. Find the solution for the ODE

$$x - xy - y' = 0, \quad y(0) = -1.$$

Solution. After rewriting the ODE to a non-homogeneous linear equation

$$y' + xy = x, \quad (*)$$

we can calculate the integrating factor

$$\alpha(x) = \exp \int x dx = \exp\left(\frac{x^2}{2}\right).$$

Multiplying the integrating factor to both sides of (*) and taking anti-derivative, we obtain the general solution

$$y = \exp\left(-\frac{x^2}{2}\right) \int x \exp\left(\frac{x^2}{2}\right) dx = \exp\left(-\frac{x^2}{2}\right) \left(\exp\left(\frac{x^2}{2}\right) + C \right) = 1 + C \exp\left(-\frac{x^2}{2}\right)$$

When $x = 0$, $y(0) = 1 + C = -1$. Hence $C = -2$ and we have the solution

$$y = 1 - 2 \exp\left(-\frac{x^2}{2}\right).$$

□

29.4 Exact Equations and Integrating Factors

In this section, we consider the first order ODE (29.1) of a more general form compare to the previous two sections:

$$M(x, y) + N(x, y) \frac{dy}{dx} = 0 \quad (29.9)$$

e:ode:MN

where $N(x, y) \neq 0$. It is equivalent to (29.3).

Using the knowledge of *total differential* from *vector calculus* (Topic 4 in *UECM2063 Calculus III*), if there *exists* a function $F(x, y)$ such that its total differential satisfies

$$dF(x, y) = \frac{\partial F}{\partial x} dx + \frac{\partial F}{\partial y} dy = M(x, y) dx + N(x, y) dy$$

then (29.9) can be solved easily:

$$M(x, y) + N(x, y) \frac{dy}{dx} = \frac{\partial F}{\partial x} + \frac{\partial F}{\partial y} \frac{dy}{dx} = \frac{d}{dx} F(x, y(x)) = 0 \Rightarrow F(x, y(x)) = C.$$

Definition 29.7. The expression $M(x, y)dx + N(x, y)dy$ is called a *differential form*. It is said to be *exact* in a rectangle $D : [a, b] \times [c, d]$ where a, b, c and d are constants, if there is a function $F(x, y)$ such that

$$\frac{\partial F}{\partial x} = M(x, y) \quad \text{and} \quad \frac{\partial F}{\partial y} = N(x, y)$$

for all $(x, y) \in D$.

Definition 29.8. If $M(x, y)dx + N(x, y)dy$ is an exact differential form, then the “form” equation

$$M(x, y)dx + N(x, y)dy = 0 \tag{29.10}$$

is called an *exact equation*.

Let us look at an example of exact equation.

Example 29.16. Show that $(2x + y^2)dx + (2xy)dy = 0$ is an exact equation.

Proof: Consider $F(x, y) = x^2 + xy^2$. Since

$$\frac{\partial F}{\partial x} = 2x + y^2, \quad \frac{\partial F}{\partial y} = 2xy,$$

By definition 29.8, the “form” equation is an exact equation.

Example 29.17. Find the general solution of $\frac{dy}{dx} = -\frac{2x + y^2}{2xy}$.

Analysis: Expanding the right-hand-side

$$\frac{dy}{dx} = -\left(\frac{1}{y} + \frac{y}{x}\right) \Rightarrow \frac{dy}{dx} + \frac{1}{x}y = -\frac{1}{y} \neq Q(x)$$

we cannot reduced the ODE to either of the simple forms in Sections 29.2 and 29.3. So we have to think of other methods.

Solution: Writing the ODE in differential form

$$2xydy = -(2x + y^2)dx \Rightarrow 2xydy + (2x + y^2)dx = 0 \Rightarrow d(x^2 + xy^2) = 0.$$

The general $y(x)$ satisfies the equation $x^2 + xy^2 = C$.

How can we find the function $F(x, y)$? Does it even exists in general? The theorem (from Vector Calculus) below says that a differential form in a domain has to satisfy an equality for it to be exact.

Theorem 29.18 (Exactness Theorem). *Suppose the first partial derivatives of $M(x, y)$ and $N(x, y)$ are continuous in rectangle D , then $M(x, y)dx + N(x, y)dy = 0$ is an exact equation in D if and only if*

$$\frac{\partial M}{\partial y}(x, y) = \frac{\partial N}{\partial x}(x, y), \quad \forall (x, y) \in D.$$

Let us apply this theorem to study the following example.

Example 29.19. Show that $(\sin x + x^2 e^y - 1)y' + (y \cos x + 2xe^y) = 0$ is exact and find the general solution.

Solution: This can be transformed to a differential form equation

$$\underbrace{(y \cos x + 2xe^y)}_{M(x,y)} dx + \underbrace{(\sin x + x^2 e^y - 1)}_{N(x,y)} dy = 0.$$

Since M and N are continuous in any rectangle in $\mathbb{R}^2$ and

$$\frac{\partial M}{\partial y} = \cos x + 2xe^y, \quad \frac{\partial N}{\partial x} = \cos x + 2xe^y,$$

we have $\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x}$. By Exactness Theorem, the differential form equation is an exact equation. Hence, there exist a function $F(x, y)$ such that

$$dF = \frac{\partial F}{\partial x} dx + \frac{\partial F}{\partial y} dy = M dx + N dy. \Rightarrow \frac{\partial F}{\partial x} = y \cos x + 2xe^y \Rightarrow F(x, y) = y \sin x + x^2 e^y + g(y)$$

and

$$\frac{\partial F}{\partial y} = \sin x + x^2 e^y + g'(y) = \sin x + x^2 e^y - 1 \Rightarrow g'(y) = -1 \Rightarrow g(y) = -y.$$

Hence we have a solution

$$F(x, y) = y \sin x + x^2 e^y - y = C.$$

Example 29.20 (Final Exam Sept 2013, Q1(a)). Show that the equation

$$(\sin y + y \cos x)dx + (\sin x + x \cos y)dy = 0$$

is exact and obtain its solution.

(12 marks)

Solution: To show that the equation is exact, we need to find $F(x, y)$ such that

$$\frac{\partial F}{\partial x} = \sin y + y \cos x \quad \text{and} \quad \frac{\partial F}{\partial y} = \sin x + x \cos y.$$

From the first equation, we have

$$F = x \sin y + y \sin x + f(y)$$

From the second equation, we have

$$\frac{\partial}{\partial y}(x \sin y + y \sin x + f(y)) = x \cos y + \sin x + f'(y) = \sin x + x \cos y \Rightarrow f'(y) = 0 \Rightarrow f(y) = C.$$

Therefore the equation is exact and the implicit solution is

$$x \sin y + y \sin x = C.$$

Note that it is “theoretically” always possible to convert the “form” equation (29.10) ^{le:exact-egn} which is not exact into an exact equation by multiplying an *integrating factor* provided that the differential form is differentiable. However, finding integrating factor can be a difficult task.

Definition 29.9. If (29.10) ^{le:exact-egn} is not exact but there is a function $\mu(x, y)$ such that

$$\mu(x, y)M(x, y)dx + \mu(x, y)N(x, y)dy = 0$$

is exact, then the function $\mu(x, y)$ is called an *integrating factor* of the ODE (29.9) ^{le:ode:MN} or the differential form equation (29.10) ^{le:exact-egn}.

We may be able to determine $\mu(x, y)$ from the equation

$$\frac{\partial}{\partial y}(\mu M) = \frac{\partial}{\partial x}(\mu N).$$

as demonstrated in the following equation.

Example 29.21. Find the general solution of

$$(3xy + y^2)dx + (x^2 + xy)dy = 0. \quad (29.11)$$

e:exmp:1

Solution: Let $M = 3xy + y^2$ and $N = x^2 + xy$. $\frac{\partial M}{\partial y} = 3x + 2y$, $\frac{\partial N}{\partial x} = 2x + y$ and $\frac{\partial M}{\partial y} \neq \frac{\partial N}{\partial x}$.

We will try to transform it to exact differential equation by multiplying integrating factor $\mu(x, y)$:

$$\mu(3xy + y^2)dx + \mu(x^2 + xy)dy = 0.$$

We try to find a μ such that

$$\frac{\partial}{\partial y}(\mu(3xy + y^2)) = \frac{\partial}{\partial x}(\mu(x^2 + xy)) \Rightarrow (3xy + y^2)\frac{\partial \mu}{\partial y} + \mu(3x + 2y) = (x^2 + xy)\frac{\partial \mu}{\partial x} + \mu(2x + y)$$

This is a first order PDE, which is difficult to solve. However, we can try to assume that μ is only a function of x so that $\frac{\partial \mu}{\partial y} = 0$ and $\frac{\partial \mu}{\partial x} = \frac{d\mu}{dx}$, which implies that

$$(x^2 + xy)\frac{d\mu}{dx} = \mu(3x + 2y) - \mu(2x + y) \Rightarrow \frac{d\mu}{dx} = \frac{\mu(x + y)}{x(x + y)} = \frac{\mu}{x}.$$

Solving it gives $\mu = x$.

Hence multiplying (29.11) by x , it can be rewrite in exact form:

$$(3x^2y + xy^2)dx + (x^3 + x^2y)dy = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy = df(x, y) = 0.$$

Since

$$\frac{\partial f}{\partial x} = 3x^2y + xy^2 \Rightarrow f(x, y) = x^3y + \frac{1}{2}x^2y^2 + g(y)$$

and

$$\frac{\partial f}{\partial y} = x^3 + x^2y + g'(y) = x^3 + x^2y$$

this implies $g(y) = -C$ and the general solution is

$$x^3y + \frac{1}{2}x^2y^2 = C.$$

Remark 29.22. The integrating factor μ is in general not unique. For example, in Example 29.21, $\mu = \frac{1}{xy(2x + y)}$ is also a possible integrating factor.

29.5 Change of Variables (or Symmetry)

The types of first order ODE and solution techniques studied in Sections 29.2 to 29.4 are already known in the middle of 19th centuries. Rota [1997] suggested that the technique of integrating factors (Section 29.4) be removed from ODE because it is useless. Never in his life has he heard of anyone solving a first order differential equation by finding an integrating factor. We still mention integrating factor here because it is an interesting calculus trick used to solve first order ODE.

An important theory in modern ODE which can explain why the techniques in Sections 29.2 and 29.3 work is related the notion of *symmetry*! Bluman and Anco [2002, p. 292] points out that in general, an integrating factor of an ODE has no obvious relation to any underlying invariance or geometrical motion (i.e. the symmetry of the ODE). This explains why Rota [1997] suggested the removal of integrating factors.

“Symmetry” means the “one-to-one” transformation of variables which preserves the ODE (29.1) or (29.2). It is related to Lie groups and Lie algebra which was discussed in Hydon [2000] and Bluman and Kumei [1989]. When apply to (29.3), it goes like:

- Assume that $\mathbf{w} = \xi(x, y)\frac{\partial}{\partial x} + \phi(x, y)\frac{\partial}{\partial y}$ is the generator element of a single-parameter invariant group G of (29.3).
- At a point (x_0, y_0) , if $\mathbf{w}(x_0, y_0) \neq \mathbf{0}$, then at a neighbourhood of (x_0, y_0) , there is a transform (difficult to find in general) $u = \eta(x, y)$, $v = \zeta(x, y)$ such that $\mathbf{w} = \frac{\partial}{\partial v}$ and (29.3) can be transformed to

$$\frac{dv}{du} = h(u)$$

which has a “quadrature” solution by taking “anti-derivative”:

$$w = \int h(u)du + C$$

Since the “symmetry” theory is very abstract, we will skip it and just illustrate the idea of symmetry very roughly with an example below.

Example 29.23. Find the general solution for the first order ODE

$$\frac{dy}{dx} = \frac{y^3 + x^2y - x - y}{x^3 + xy^2 - x + y}. \quad (29.12)$$

Solution: For this ODE, it is possible to find a transform $x = r \cos \theta$, $y = r \sin \theta$ which transforms (29.12) to a separable equation:

$$\frac{dr}{d\theta} = r(1 - r^2)$$

which has a general solution:

$$\int \frac{dr}{r(1-r)(1+r)} = \theta + C$$

By using partial fraction $\frac{1}{r(1-r)(1+r)} = \frac{1}{r} + \frac{1}{2(1-r)} - \frac{1}{2(1+r)}$, we have

$$\ln r - \frac{1}{2} \ln(1-r) - \frac{1}{2} \ln(1+r) = \theta + C$$

or in the original xy -coordinate:

$$\ln \sqrt{x^2 + y^2} - \frac{1}{2} \ln(1 - \sqrt{x^2 + y^2}) - \frac{1}{2} \ln(1 + \sqrt{x^2 + y^2}) = \arctan \frac{y}{x} + C$$

We now investigate a few types of first order ODE which have been investigated during the 18th and 19th centuries using the change of variables (and which have some typical symmetrical properties).

Definition 29.10. An explicit first order ODE (29.3) having the form

$$y' = f\left(\frac{y}{x}\right). \quad (29.13)$$

is called a *homogeneous equation*.

This type of ODE can be solved by a change of variable $u = \frac{y}{x}$, which reduces (29.13) to a separable equation (29.5):

$$\frac{dy}{dx} = u + x \frac{du}{dx} = f(u) \Rightarrow \frac{du}{dx} = \frac{f(u) - u}{x}$$

because $y = x \cdot u$.

Example 29.24 (Peter O’Neil, p40). Find the general solution of $xy' = \frac{y^2}{x} + y$.

Solution. Rewriting this equation as

$$y' = \left(\frac{y}{x}\right)^2 + \frac{y}{x}. \quad (29.14)$$

Let $u = \frac{y}{x}$, then $y = ux$, $y' = u + xu'$ and (29.14) becomes

$$u + xu' = u^2 + u \Rightarrow u' = \frac{u^2}{x} \Rightarrow -\frac{1}{u} = \ln|x| + C \Rightarrow u = \frac{y}{x} = -\frac{1}{\ln|x| + C}$$

Hence the solution is

$$y = -\frac{x}{\ln|x| + C}.$$

□

Definition 29.11. A first order ODE (29.1) having the form

$$y' + P(x)y = R(x)y^\alpha, \quad \alpha \neq 0, 1 \quad (29.15)$$

is called a *Bernoulli equation*.

According to [Boole, 1859, p. 37], this type of ODE can be solved by rearranging (29.15) to

$$y^{-\alpha}y' + P(x)y^{-\alpha+1} = R(x).$$

Multiplying it by $1 - \alpha$ and let $u = y^{1-\alpha}$, we obtain a first order non-homogeneous equation

$$u' + (1 - \alpha)P(x)u = (1 - \alpha)R(x)$$

which can be solved using the techniques mentioned previously and a “quadrature solution” (i.e. a solution in integral form) could be obtained.

$$y^{1-\alpha} = e^{-(1-\alpha) \int P(x)dx} \left(\int e^{(1-\alpha) \int P(x)dx} (1 - \alpha)R(x)dx + C \right).$$

Example 29.25. [Schroers, 2011, p. 12] Find the general solution of the ODE $y' + y = y^4$.

Solution: Let $u = y^{1-4}$. Then the ODE can be transformed to

$$\frac{du}{dx} - 3u = -3 \Rightarrow \frac{d}{dx}(e^{-3x}u) = -3e^{-3x} \Rightarrow e^{-3x}u = e^{-3x} + C \Rightarrow y^{-3} = 1 + Ce^{3x}.$$

Example 29.26. Find the general solution of the following ODE:

$$y' + \frac{y}{x} = 3x^2y^3. \quad (29.16)$$

Solution. This is a Bernoulli equation with $\alpha = 3$, $P(x) = \frac{1}{x}$ and $R(x) = 3x^2$. Let $v = y^{-2}$, (29.16) can be transformed to

$$v' + \frac{-2}{x}v = (-2)3x^2$$

which has solutions

$$v = x^2 \int \frac{1}{x^2}(-6x^2)dx = x^2(-6x + C) = Cx^2 - 6x^3.$$

where the integrator is $e^{\int \frac{-2}{x}dx} = \frac{1}{x^2}$. Hence the solution for (29.16) is

$$y^2 = \frac{1}{Cx^2 - 6x^3}.$$

□

Definition 29.12. A first order ODE (29.1) having the form

$$y' = P(x)y^2 + Q(x)y + R(x). \quad (29.17)$$

is called a *Riccati equation*.

This type of ODE does not have analytic solutions in general according to Liouville (1841). For example, when $P(x) = 1$, $Q(x) = 0$ and $R(x) = x^2$, we have

$$y' = y^2 + x^2$$

which does not have an analytical solution. However, if a specific solution $S(x)$ is known, then (29.17) transformed to a linear equation (29.6) by $y = S(x) + \frac{1}{z}$:

$$S'(x) + \left(\frac{1}{z}\right)' = P(x)(S(x))^2 + \frac{2P(x)S(x)}{z} + \frac{P(x)}{z^2} + Q(x)S(x) + \frac{Q(x)}{z} + R(x)$$

after cancelling terms related to $S(x)$, we have

$$-\frac{1}{z^2}z' = \frac{2P(x)S(x)}{z} + \frac{P(x)}{z^2} + \frac{Q(x)}{z} \Rightarrow z' + (2P(x)S(x) - Q(x))z = -P(x).$$

Example 29.27. Show that $S(x) = 4$ is a solution of the ODE

$$y' = \frac{1}{2x}y^2 - \frac{1}{x}y - \frac{4}{x} \quad (29.18)$$

e:exp:5

and find its general solution.

Solution. Since $S'(x) = 0$ and $\frac{4^2}{2x} - \frac{2}{x} - \frac{4}{x} = \frac{8-4-4}{x} = 0$, $S(x) = 4$ is a solution by definition.

Let $y = 4 + \frac{1}{z}$. Then $y' = -\frac{1}{z^2}z'$ and (29.18) can be written as

$$-\frac{1}{z^2}z' = \frac{1}{2x}\left[16 + \frac{8}{z} + \frac{1}{z^2}\right] - \frac{1}{x}\left[4 + \frac{1}{z}\right] - \frac{4}{x}.$$

Simplifying it gives

$$z' + \frac{3}{x}z = -\frac{1}{2x}.$$

It has an integrating factor $e^{\int \frac{3}{x}dx} = x^3$ and the solution for z is

$$z = \frac{1}{x^3} \int x^3 \left(-\frac{1}{2x}\right) dx = \frac{1}{x^3} \left(C - \frac{x^3}{6}\right) = \frac{C - x^3}{6x^3}.$$

The general solution for y is

$$y = 4 + \frac{1}{z} = 4 + \frac{6x^3}{C - x^3}.$$

□

29.6 Picard's Existence and Uniqueness Theorem

Until now, we find that we are only solve specific first order equations. Will any given ODE have a solution? The answer is obviously “no” since we know that the following algebraic equation

$$x^2 + 1 = 0$$

does not have a real solution since $x^2 \geq 0$. Therefore, we also know that the following ODE does not have any real-valued solution:

$$(y')^2 + 1 = 0.$$

In the previous sections, we have only discussed the techniques to find general solutions. Is it possible to find a unique solution?

To answer this question, let us investigate a variation of Example 29.3 below, in which we add a constraint.

Example 29.28. Find a (unique) particular solution of

$$y' = 3y$$

which satisfies the constraint:

$$y(0) = 4.$$

Solution: From Example 29.3, we know that $y = Ce^{3x}$ is the general solution of the ODE. From the constraint

$$y(0) = 4 \Rightarrow Ce^0 = 4 \Rightarrow C = 4.$$

Therefore $y = 4e^{3x}$ is a particular solution of the ODE. This solution is unique because the solution process gives a unique $C = 4$.

The constraint mentioned in the above example 29.28 is called the *initial condition* in mathematical and engineering community. It is a constraint which describes the initial physical state of an engineering system.

Let us provide a formal definition of the initial value problem related to the addition of this constraint.

Definition 29.13. An *initial value problem (IVP)* for a first order ODE is formed by either an ODE (29.1) or an ODE (29.3) with an *initial condition* $y(x_0) = y_0$:

$$y' = f(x, y), \quad y(x_0) = y_0. \quad (29.19)$$

e:ode:ord1:ivp

Definition 29.14. An *initial-value problem (IVP)* for an n th-order ODE is either an ODE (29.2) or an ODE (29.4) with *initial conditions*

$$y(x_0) = y_0, \quad y'(x_0) = y_1, \quad \dots, \quad y^{(n-1)}(x_0) = y_{n-1},$$

where $x_0 \in I$ and $y_0, y_1, \dots, y_{n-1}$ are given constants (or initial states of a system).

Imposing an initial condition does not mean that a solution will have a solution. Let us look at two examples.

Example 29.29. [Schroers, 2011, p. 16] Show that the IVP below does not have a solution.

$$\frac{dy}{dx} = \begin{cases} 1, & x \geq 0, \\ 0, & x < 0, \end{cases} \quad y(0) = 0.$$

Proof: Suppose a continuous and differentiable function $y(x)$ exists over $\mathbb{R}$. By restricting to $x \geq 0$, we learn that

$$y(x) = x.$$

By restricting to $x < 0$, we learn that

$$y(x) = C$$

and since $y(x)$ must be continuous and differentiable, take $x \rightarrow 0^-$, then $y(0^-) = 0$ and this implies that $C = 0$.

Since y is differentiable at $x = 0$, we have

$$0 = \lim_{x \rightarrow 0^-} \frac{0 - 0}{x - 0} = \lim_{x \rightarrow 0^-} \frac{y(x) - 0}{x - 0} = y'(0) = \lim_{x \rightarrow 0^+} \frac{y(x) - 0}{x - 0} = \lim_{x \rightarrow 0^+} \frac{x - 0}{x - 0} = 1.$$

This is a contradiction.

Example 29.30. Show that the following initial value problem does not have a solution:

$$y' = \tan x, \quad y\left(\frac{\pi}{2}\right) = 0.$$

Proof: Suppose a solution exists, then the ODE has a general solution

$$y(x) = -\ln |\cos x| + C$$

When $x = \frac{\pi}{2}$, $\cos \frac{\pi}{2} = 0$. The right hand side is an infinity while the left hand side is 0 (the requirement of the initial condition). A contradiction. Hence the initial problem has no solution.

Even if there is a solution when initial condition is given, it does not mean that it is the only solution, there may be multiple solutions as shown below.

Example 29.31. Show that $y_1(x) = 0$ and $y_2(x) = \begin{cases} 0, & x \leq 2, \\ (x-2)^2, & x \geq 2 \end{cases}$ are the solutions to the ODE

$$y' = 2\sqrt{y}, \quad y(0) = 0. \quad (29.20)$$

Proof: Both y_1 and y_2 satisfy $y_1(0) = y_2(0) = 0$.

LHS of ODE: $y_1'(x) = 0$; RHS of ODE: $2\sqrt{y_1(x)} = 2 \times 0 = 0 = \text{LHS}$.

Hence, $y_2(x)$ is a solution of the ODE (29.20).

LHS of ODE: $y_2'(x) = \begin{cases} 0, & x \leq 2, \\ 2(x-2), & x \geq 2 \end{cases}$;

RHS of ODE: $2\sqrt{y_2(x)} = 2 \begin{cases} 0, & x \leq 2, \\ (x-2), & x \geq 2 \end{cases}$.

Hence, $y_2(x)$ is a solution of the ODE (29.20).

So, when does the ODE (29.3) have a unique solution? Engineering and numerical problems require that an ODE gives a unique solution so that it is possible to compute and approximate the state of the actual problem. Charles Emile Picard (1856–1941) and Ernst Lindelöf (1870–1946) discovered that under fairly mild assumption on f (which is satisfied by most engineering and science ODE models), IVPs have unique solutions locally [Schroers, 2011, p. 14].

Theorem 29.32 (Picard(-Lindelöf Existence and Uniqueness) Theorem). *Let the functions f in (29.3) and $\frac{\partial f}{\partial y}$ be continuous in some domain $D = \{(x, y) : a < x < b, c < y < d\}$ containing the point (x_0, y_0) . Then, in some interval $x_0 - h < x < x_0 + h$ contained in $a < x < b$, there exists a unique solution $y = \phi(x)$ of the initial value problem (29.19).*

This version of the Picard-Lindelöf Theorem is useful in practice. It tells us that the IVPs of most differentiable ODEs have unique solution. However, one can replace the “continuity requirement” of the partial differentiation $\frac{\partial f}{\partial y}$ by Lipschitz condition with respect to the second argument in f . The proof of the theorem proceeds via the re-formulation of the IVP (29.19) in terms of an integral equation:

$$y(x) = y(x_0) + \int_{x_0}^x f(t, y(t)) dt$$

and the application of a contraction mapping theorem.

29.7 Formulation and Solution of Application Specific Problems

So far, we have been discussing theory. In this section, we will try to look at a few simple engineering examples — the mixture problem and population dynamics.

Example 29.33 (Mixture Problem, Zill [2017]). Initially 50 kg of salt is dissolved in a large tank holding 300 litres of water. A brine solution is pumped into the tank at the rate of 3 litres per minute, and the well-stirred solution is then pumped out at the same rate. If the concentration of the solution entering is 2 kg per gallon, determine the amount of salt in the tank at any time. How much salt is present after (a) 50 minutes? (b) a long time?

Solution: Let $A(t)$ be the amount of salt (in kg) in the tank at any time. The net rate at which $A(t)$ changes is given by

$$\frac{dA}{dt} = \left(\begin{array}{c} \text{rate of} \\ \text{substance entering} \end{array} \right) - \left(\begin{array}{c} \text{rate of} \\ \text{substance leaving} \end{array} \right) = R_1 - R_2.$$

Now the rate at which the salt enters the tank is, in kg per minute,

$$\begin{aligned} \underbrace{R_1}_{\text{lb/min}} &= \underbrace{(\text{Concentration In})}_{\text{lb/gal}} \underbrace{(\text{Flow Rate In})}_{\text{gal/min}} \\ &= (3 \text{ gal/min}) \cdot (2 \text{ lb/gal}) = 6 \text{ lb/min} \end{aligned}$$

whereas the rate at which salt is leaving is

$$\begin{aligned} \underbrace{R_2}_{\text{lb/min}} &= \underbrace{(\text{Concentration Out})}_{\text{lb/gal}} \underbrace{(\text{Flow Rate Out})}_{\text{gal/min}} \\ &= (3 \text{ gal/min}) \cdot \left(\frac{A}{300} \text{ lb/gal} \right) = \frac{A}{100} \text{ lb/min}. \end{aligned}$$

Thus, the IVP that describes the change in the amount of salt is

$$\frac{dA}{dt} = 6 - \frac{A}{100}, \quad A(0) = 50.$$

Solving the IVP, we obtain a particular solution

$$A(t) = 600 - 550e^{-t/100}.$$

(a). At $t = 50$, $A(50) = 266.41$ kg;

- (b). After a long time, $A(\infty) = \lim_{t \rightarrow \infty} A(t) = 600$, the amount of salt in the solution is $(300 \text{ gal}) \cdot (2 \text{ lb/gal}) = 600 \text{ lb}$, which is the amount of salt needed to make the concentration of salt in the tank equal to the inflow salt concentration.

Remark 29.34. The basic principle we will use in solving Example 29.33 is the *balance law*:

$$\text{Net Rate of Change} = \text{Rate of Inflow} - \text{Rate of Outflow}.$$

In Example 29.33, we have

$$\boxed{\text{rate of change of salt in the tank}} = \boxed{\text{rate at which salt enters the tank}} - \boxed{\text{rate at which salt leaves the tank}}$$

Let $A(t)$ = amount of salt (kg) in the tank at time t (minutes)

$V(t)$ = volume of solution (litres) in the tank at t

$c_i(t)$ = inflow salt concentration (kg/litres) at t

$c_o(t) = \frac{A(t)}{V(t)}$, outflow salt concentration (kg/litres) at t

$r_i(t)$ = inflow rate (gal/min) at t

$r_o(t)$ = outflow rate (gal/min) at time t

Then

$$\boxed{\text{rate of change of salt in the tank}} = \frac{dA}{dt}$$

$$\boxed{\text{rate at which salt enters the tank}} = r_i(t)c_i(t) \text{ (The units are: lb/min = gal/min} \times \text{lb/gal.)}$$

$$\boxed{\text{rate at which salt leaves the tank}} = r_o(t)c_o(t) \text{ (Note: unlike } c_i(t), c_o(t) \text{ is unknown)}$$

Example 29.35 (Unlimited Population Growth). The rate of growth of a population of fruit flies is proportional to the size of the population at any time. If there are 200 flies at the end of the third day and 800 at the end of the fifth day, how many were there in the beginning?

Solution: Let $P(t)$ be the population of fruit flies at day t . Given

$$\text{rate of growth} = kP(t) \Rightarrow \frac{d}{dt}P(t) = kP(t).$$

This is a separable equation and the general solution is

$$P(t) = Ce^{kt}$$

Given that

$$P(3) = Ce^{3k} = 200, \quad P(5) = Ce^{5k} = 800,$$

we can obtain

$$\frac{P(5)}{P(3)} = e^{2k} = 4 \Rightarrow e^k = 2$$

and

$$Ce^{3k} = C(e^k)^3 = 200 \Rightarrow C \cdot 2^3 = 200 \Rightarrow C = \frac{200}{8} = 25 \text{ flies at the beginning.}$$

Example 29.36 (Final Exam Sept 2014, Q2(a)). Assume that the rate of growth of a population of fruit flies is proportional to the size of the population, $P(t)$, at any time t . If there are 200 flies at the end of the third day and 900 at the end of the seventh day, how many fruit flies were there approximately in the beginning? (6 marks)

Solution. The growth model is $\frac{d}{dt}P(t) = kP(t)$ [1 mark]

The general solution is $P(t) = Ce^{kt}$ [1 mark]

Given that

$$P(3) = Ce^{3k} = 200, \quad P(7) = Ce^{7k} = 900, \quad \text{[1 mark]}$$

we can obtain

$$\frac{P(7)}{P(3)} = e^{4k} = 4.5 \Rightarrow k = \frac{\ln 4.5}{4} \approx 0.37602 \quad [1 \text{ mark}]$$

There were approximately

$$Ce^{3k} = 200 \Rightarrow C = \frac{200}{e^{3 \ln(4.5)/4}} \approx 64.73 \text{ flies at the beginning.} \quad [2 \text{ marks}]$$

□

The model is fairly good for small populations in large environments. For example, the first few spots of mold on a piece of bread or the first European settlers in the United States.

However, as time goes on, the model predicts that the population will continue to grow without any limits, and obviously, this cannot happen in the real world. Consequently, a more accurate model over a longer time scale need to take into account for the fact that populations exist in a finite amount of space and with limited resources.

To adjust the exponential growth population model to account for a limited environment and limited resources, we add the assumptions:

- If the population P is small, the rate of growth of the population is proportional to its size:

$$\frac{dP}{dt} \approx kP$$

- If the population is too large to be supported by its environment and resources, the population will decrease. That is, the rate of growth is negative:

$$P > L \Rightarrow \frac{dP}{dt} < 0$$

These conditions are met for the logistic population model:

Example 29.37 (Logistic Population Model). A more realistic model for population growth is given by the *logistic equation*

$$\frac{dP}{dt} = k \left(1 - \frac{P}{L} \right) P$$

where $k > 0$ and $L > 0$ are constants and P is the population at time t . This is called the *logistic population model with growth rate k and carrying capacity L* . Find a particular solution of this model if the initial condition is $P(0) = P_0$.

Solution: By applying the technique of separation of variables, the solution for this equation can be found to be the *logistic function*

$$P(t) = \frac{LP_0}{P_0 + (L - P_0)e^{-kt}}$$

Note that $\lim_{t \rightarrow \infty} P(t) = L$, the carrying capacity of the population. This is the limiting value of the population, beyond which it cannot grow.

Remark 29.38. The logistic equation can also be used to model the spread of information or an infectious disease. In the case of disease, if M denotes the total number of people in the population under consideration and $P(t)$ is the number of infected people at time t , then the differential equation $\frac{dP}{dt} = k(M - P)P$ states that the rate of growth of the disease is proportional to the product of the number of people who are infected and the number who are not infected. Clearly, this ODE is in fact the logistic equation:

$$\frac{dP}{dt} = kM \left(1 - \frac{P}{M} \right) P.$$

Example 29.39 (Final Exam Sept 2016, Q2(b)). The logistic population model below

$$\frac{dP}{dt} = k \left(1 - \frac{P}{L} \right) P$$

where the *growth rate* k and the *carrying capacity* L are positive constants and P is the population at time t . Find the general solution of the logistic population model. Determine the carrying capacity L given that the population of the United States of America in 1800, 1850 and 1900 was 5.3, 23.1 and 76 million people respectively. Write your answers to *at least three decimal places*. (10 marks)

Solution. By separating variables in the ODE,

$$\frac{dP}{(1 - \frac{P}{L})P} = kdt \Rightarrow \int \left[\frac{1}{L - P} + \frac{1}{P} \right] dP = kt + C \quad [2 \text{ marks}]$$

We further simplify it to

$$\ln |P| - \ln |L - P| = kt + C \Rightarrow \frac{P}{L - P} = Ce^{kt} \Rightarrow P = \frac{LCe^{kt}}{1 + Ce^{kt}} = \frac{LC}{C + e^{-kt}} \quad [4 \text{ marks}]$$

By a change of variable $t \rightarrow t + 1800$, the given conditions become

$$\begin{cases} P(0) = \frac{LC}{C+1} = 5.3 \\ P(50) = \frac{LC}{C + e^{-50k}} = 23.1 \\ P(100) = \frac{LC}{C + e^{-100k}} = 76 \end{cases} \Rightarrow \begin{cases} e^{-50k} = \frac{-17.8C + 5.3}{23.1} \\ e^{-100k} = \frac{-70.7C + 5.3}{76} \end{cases} \quad [1 \text{ mark}]$$

which reduces to the following quadratic equation

$$\left(\frac{-17.8C + 5.3}{23.1} \right)^2 = \frac{-70.7C + 5.3}{76} \Rightarrow 316.84C + 307.7177C - 9.1223 = 0 \quad [1 \text{ mark}]$$

By using scientific calculator, we obtain

$$C = -1, \quad 0.02879142 \quad [1 \text{ mark}]$$

Note that $C > 0$. Therefore,

$$L = 189.382619 \text{ million.} \quad [1 \text{ mark}]$$

□

Chapter 30

Second Order Equations

Many science and engineering problems can be modelled by second order ODEs. The most famous second order ODE is the Newton's second law of motion (for one-dimensional system):

$$m \frac{d^2 x}{dt^2} = F(t, x, \frac{dx}{dt})$$

which governs the change of position x of a particle of mass m moving under the influence of a force F [Braun, 1993] at time t . The method of separation of variables (refer to the course *UECM3003 Partial Differential Equations*) also reduces linear second order PDE problems to second order ODEs.

An ODE need not always use the symbol y and x as can be observed in classical mechanics textbooks such as Strauch [2009, Chapter 2]. However, for the discussion of this topic, we will use the notation y and x for the second order non-homogeneous linear ODE in *normal form* (or *explicit form*):

$$\frac{d^2 y}{dx^2} + p(x) \frac{dy}{dx} + q(x)y = g(x). \quad (30.1)$$

Why do we limit ourselves to just (30.1)? From a mathematical point of view, we have a generalised linear theory for (30.1) which is introduced in Section 30.1. From an engineering point of view, the essence of linear vibrations and oscillations are captured by two special cases of (30.1) which are discussed in Sections 30.2 and 30.3.

30.1 Linear Equations and Fundamental Solutions

A general n th order linear ODE has the form

$$a_0(x) \frac{d^n y}{dx^n} + a_1(x) \frac{d^{n-1} y}{dx^{n-1}} + \cdots + a_{n-1}(x) \frac{dy}{dx} + a_n(x)y = b(x). \quad (30.2)$$

where $a_0, a_1, \dots, a_{n-1}, a_n, b$ are functions of x . If $a_0(x) \neq 0$, (30.2) can be written in normal form

$$\underbrace{y^{(n)} + p_1(x)y^{(n-1)} + \cdots + p_{n-1}(x)y' + p_n(x)y}_{L[y]} = g(x), \quad (30.3)$$

where $p_i(x) = a_i(x)/a_0(x)$, $i = 1, \dots, n$ and $g(x) = b(x)/a_0(x)$.

Definition 30.1. An ODE of order n is said to be *linear* if it can be written in the implicit form (30.2) or the explicit form (30.3). An ODE which is not linear is said to be *nonlinear*.

If we further impose initial conditions

$$y(x_0) = y_0, \quad y'(x_0) = y_1, \quad \dots, \quad y^{(n-1)}(x_0) = y_{n-1} \quad (30.4)$$

to (30.3), we obtain an *initial value problem*.

Definition 30.2. The functions $a_i(x)$ in (30.2) or the function $p_i(x)$ in (30.3) are called the *coefficients* of the linear ODE. The function $b(x)$ (or $g(x)$) is called the *nonhomogeneous term*.

Definition 30.3. The ODE (30.3) is said to be *homogeneous* if $g(x)$ is identically 0 and *nonhomogeneous* otherwise. When the coefficients are constant functions, the linear ODE is said to have *constant coefficients*.

A homogeneous linear ODE of order n has the normal form

$$y^{(n)} + p_1(x)y^{(n-1)} + \cdots + p_{n-1}(x)y' + p_n(x)y = 0. \quad (30.5)$$

The definition of solution in Topic 1 is valid for general ODE of order n as well as (30.3). However, when (30.3) is homogeneous, we have a special particular solution stated below.

Definition 30.4. The particular solution $y = 0$ of a homogeneous ODE (30.3) and is called a *trivial solution*.

All possible solutions of the linear equation (30.3) can be obtained from its general solution, so linear equations have no singular solutions.

Theorem 30.1 (Existence and Uniqueness). *If the functions $p_1, p_2, \dots, p_n$ and g in ODE (30.3) are continuous on the open interval I , then there exists exactly one solution $y = \varphi(x)$ of the differential equation that also satisfies the initial conditions (30.4). This solution exists throughout the interval I .*

Example 30.2. Show that $y = C_1e^x + C_2e^{-x} + C_3\sin x + C_4\cos x$ is a general solution of

$$y^{(4)} - y = 0. \quad (30.6)$$

Proof: To show that y is a general solution, we need to first confirm that when y is substituted into (30.6), the equality is satisfied:

$$\begin{aligned} y' &= C_1e^x - C_2e^{-x} + C_3\cos x - C_4\sin x \\ y'' &= C_1e^x + C_2e^{-x} - C_3\sin x - C_4\cos x \\ y''' &= C_1e^x - C_2e^{-x} - C_3\cos x + C_4\sin x \\ y^{(4)} &= C_1e^x + C_2e^{-x} + C_3\sin x + C_4\cos x \end{aligned}$$

Subtracting $y^{(4)}$ with y , we obtain 0. Hence, y is a solution of (30.6). It is a general solution because (30.6) is ODE of order 4 and y has four arbitrary constants!

We can relate Example 30.2 with *UECM1313 Fundamentals of Linear Algebra* through the notion of linear operator L associated with (30.3):

$$L = \frac{d^{(n)}}{dx^{(n)}} + p_1(x)\frac{d^{(n-1)}}{dx^{(n-1)}} + \cdots + p_{n-1}(x)\frac{d}{dx} + p_n(x). \quad (30.7)$$

Then we can build the following comparison table:

Subject	Linear Algebra	Linear ODE
Linear Equation	$A\mathbf{x} = \mathbf{b}$	$L[y](x) = g(x)$
Homogeneous	Vector $\mathbf{b} = \mathbf{0}$	Function $g(x) = 0$
Linear property	$A(k\mathbf{x}) = kA\mathbf{x}$ $A(\mathbf{x} + \mathbf{y}) = A\mathbf{x} + A\mathbf{y}$	$L[ky_1] = kL[y_1]$ $L[y_1 + y_2] = L[y_1] + L[y_2]$
Example	$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 4 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\left\{ \frac{d^4}{dx^4} - 1 \right\} y = 0$ (Taken from Example eg:ode:ord4:linear 30.2)
Particular solutions	$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$ are two linearly independent solutions	e^x and $\sin x$ are two linearly independent solutions
Linear combinations of particular solutions	$C_1 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + C_2 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} C_1 \\ 0 \\ 0 \\ C_2 \end{bmatrix}$ is also a solution	$C_1 e^x + C_2 \sin x$ is also a solution

From the table, we can see that if we have two particular solutions, multiply these particular solutions with arbitrary constants and add them together, we still get a solution of the homogeneous linear equation. This is related to the *linear property* of the operator A and L in the above table.

Let us investigate a nonlinear second order ODE to demonstrate why linear property is special.

Example 30.3. [Braun, 1993, p. 131] Show that $y = 1/x$ is a solution of

$$y''(x) - 2x(y(x))^4 = 0$$

but $y = 2/x$ is not a solution of the ODE.

Proof: When $y = 1/x$,

$$\left(\frac{1}{x}\right)'' - 2x\left(\frac{1}{x}\right)^4 = \left(\frac{-1}{x^2}\right)' - \frac{2}{x^3} = \frac{2}{x^3} - \frac{2}{x^3} = 0.$$

Hence, $y = 1/x$ is a solution. However,

$$\left(\frac{2}{x}\right)'' - 2x\left(\frac{2}{x}\right)^4 = \left(\frac{-2}{x^2}\right)' - \frac{32}{x^3} = \frac{4}{x^3} - \frac{32}{x^3} = -\frac{28}{x^3} \neq 0.$$

Therefore $y = 2/x$ is not a solution.

So how does it compare to Example [30.2](#)? For nonlinear equation, we cannot simply multiply a number and get another solution but in Example [30.2](#), if we know that $y = e^x$ is a solution, we know that $y = 2e^x$ is also a solution.

In fact, any to solutions of (30.5) can form a new solution easily with the following principle.

Theorem 30.4 (Principle of Superposition). *If y_1 and y_2 are two solutions of the linear differential equation (30.5) then the linear combination $c_1 y_1 + c_2 y_2$ is also a solution for any values of the constants c_1 and c_2 .*

We normally expect the equation $a_0 x^n + a_1 x^{n-1} + \dots + a_n = 0$ to have n roots (may be the same) in the complex plane $\mathbb{C}$. We would also expect that the equation (30.5) to have n solutions. When these n solutions are linearly independent, we call it fundamental as defined below.

Definition 30.5 (Fundamental Set of Solutions). Any set $\{y_1, y_2, \dots, y_n\}$ of n linearly independent solutions of the homogeneous linear n -th order ODE (30.5) on an interval I is said to be a *fundamental set of solutions* on the interval I .

The notion of “linearly independent” and “linearly dependent” can be viewed as a generalisation from linear algebra where functions are regarded as a generalised vector.

Definition 30.6 (Linear Dependency). A set of functions $y_1(x), y_2(x), \dots, y_n(x)$ is said to be *linearly dependent* on an interval I if there **exist** constants $c_1, c_2, \dots, c_n$, not all zero, such that $c_1 y_1(x) + c_2 y_2(x) + \dots + c_n y_n(x) = 0$ for every x in the interval. If the set of functions is not linearly dependent on the interval, it is said to be *linearly independent*.

eg:lin-dep

Example 30.5. Show that $y_1 = \cos 2x$ and $y_2 = \frac{1}{2} - \sin^2 x$ are linearly dependent (over $\mathbb{R}$).

Proof: There **exist** constants $c_1 = 1$ and $c_2 = -2$ such that

$$c_1 y_1 + c_2 y_2 = y_1 - 2y_2 = \cos 2x - 2\left(\frac{1}{2} - \sin^2 x\right) = 0.$$

By definition, y_1 and y_2 are linearly dependent.

How do we know the values of c_1 and c_2 ? It is obtained by assuming $c_1, c_2 \in \mathbb{R}$ and solving

$$\begin{aligned} c_1 y_1 + c_2 y_2 &= c_1 \cos 2x + c_2 \left(\frac{1}{2} - \sin^2 x\right) = 0 \\ \Rightarrow c_1 \cos 2x + c_2 \left(\frac{1}{2} - \frac{1 - \cos 2x}{2}\right) &= c_1 \cos 2x + c_2 \frac{\cos 2x}{2} = \left(c_1 + \frac{c_2}{2}\right) \cos 2x = 0 \\ \Rightarrow c_1 + \frac{c_2}{2} &= 0, \quad \text{for } x \in \mathbb{R}. \end{aligned}$$

eg:lin-indep

Example 30.6. Show that $y_1 = x$ and $y_2 = x^2$ are linearly independent.

Solution. For any $c_1, c_2 \in \mathbb{R}$,

$$c_1 x + c_2 x^2 = 0 \Rightarrow c_1 = c_2 = 0 \quad (\text{coefficients of zero polynomial}).$$

By definition, y_1 and y_2 are linear independent. □

To determine the linear dependency by definition as shown can be very tedious. We first introduce the concept of Wronskian and state a theorem which provide a test for linear dependency.

def:wronskian

Definition 30.7. Given n functions $y_1, \dots, y_n$, the *Wronskian* of the functions is defined as

$$W[y_1, y_2, \dots, y_n] = \begin{vmatrix} y_1 & y_2 & \cdots & y_n \\ y_1' & y_2' & \cdots & y_n' \\ \vdots & \vdots & \cdots & \vdots \\ y_1^{(n-1)} & y_2^{(n-1)} & \cdots & y_n^{(n-1)} \end{vmatrix}.$$

Theorem 30.7 (Wronskian Test). *The differentiable functions y_1 and y_2 are linearly independent iff the Wronskian of the functions is not equal to zero, i.e.*

$$W[y_1, y_2](x) = \begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix} = y_1 y_2' - y_1' y_2 \neq 0.$$

Example 30.8. Use the Wronskian test to determine the linear dependency of the functions in (a) Example 30.5 and (b) Example 30.6.

Solution: (a) The Wronskian of Example [30.5](#) is [eg:lin-dep](#)

$$\begin{aligned} W[y_1, y_2] &= \begin{vmatrix} \cos 2x & \frac{1}{2} - \sin^2 x \\ -2 \sin 2x & -2 \sin x \cos x \end{vmatrix} \\ &= -2 \cos 2x \sin x \cos x + 2 \sin 2x \left(\frac{1}{2} - \sin^2 x \right) \\ &= -\cos 2x \sin 2x + 2 \sin 2x \left(\frac{1}{2} - \frac{1 - \cos 2x}{2} \right) = -\cos 2x \sin 2x + 2 \sin 2x \left(\frac{\cos 2x}{2} \right) = 0 \end{aligned}$$

The Wronskian test result indicates that y_1 and y_2 are linear dependent.

(b) The Wronskian of Example [30.6](#) is [eg:lin-indep](#)

$$W[y_1, y_2] = \begin{vmatrix} x & x^2 \\ 1 & 2x \end{vmatrix} = 2x^2 - x^2 = x^2 \neq 0.$$

The Wronskian test indicates that y_1 and y_2 are linear independent.

Example 30.9. Determine whether the functions $y_1 = e^x$, $y_2 = e^{-x}$ and $y_3 = \cosh x$ are linearly dependent or not by computing its Wronskian?

Solution. Note that $\cosh x = \frac{e^x + e^{-x}}{2}$ and $\sinh x = \frac{e^x - e^{-x}}{2}$, therefore

$$W[y_1, y_2, y_3] = \begin{vmatrix} e^x & e^{-x} & \cosh x \\ e^x & -e^{-x} & \sinh x \\ e^x & e^{-x} & \cosh x \end{vmatrix} = 0.$$

Hence, they are linearly dependent. □

As we can observe from the above examples, the Wronskian may not be easy to simplify. However, it allows us to characterise the “fundamental set of solutions” (see Definition [30.5](#)). [def:fund-set](#)

Theorem 30.10. Assume that the functions $p_1, p_2, \dots, p_n$ in the homogeneous ODE [\(30.5\)](#) are [le:ode:linear:n:homog](#) continuous on the open interval I and the functions $y_1, y_2, \dots, y_n$ are solutions of [\(30.5\)](#). If [le:ode:linear:n:homog](#)

$$W[y_1, y_2, \dots, y_n](x) \neq 0$$

for at least one point in I , then every solution can be expressed as a linear combination of the fundamental set of solutions $y_1, y_2, \dots, y_n$.

Since this topic is about the analysis of second order linear ODE, we will state a few specialised theorems associated with [\(30.1\)](#). [le:ode:ord2:linear](#)

Theorem 30.11 (Abel’s Theorem). If y_1 and y_2 are two solutions of linear ODE [\(30.1\)](#) with $g(x) = 0$ where p and q are continuous on an open interval I , then the Wronskian is given by [le:ode:ord2:linear](#)

$$W[y_1, y_2](x) = c \exp\left(-\int p(x) dx\right) \quad (\text{Abel’s formula})$$

where c is a certain constant that depends on y_1 and y_2 but not on x . Further, $W[y_1, y_2](x)$ is either zero for all $x \in I$ (if $c = 0$) or else is never zero in I (if $c \neq 0$).

A proof is given in Tenenbaum and Pollard [1985, 1963, Lesson 64].

Theorem 30.12. The general solution of the non-homogeneous equation [\(30.1\)](#) can be written in the form [le:ode:ord2:linear](#)

$$y = C_1 y_1(x) + C_2 y_2(x) + y_p(x),$$

where y_1 and y_2 are a fundamental set of solutions of the corresponding homogeneous equation C_1 and C_2 are arbitrary constants and y_p is some specific solution of the non-homogeneous equation [\(30.1\)](#). [le:ode:ord2:linear](#)

In other words, the general solution of non-homogeneous ODE consists of two parts $y = y_h + y_p$. The homogeneous solution y_h is the solution for the homogeneous equation with $g(x) = 0$ and y_p is a particular solution (30.1).

30.2 Homogeneous Equations with Constant Coefficients

Homogeneous linear ODE with constant coefficients is the type of ODE in which solutions can be easily found theoretically.

Definition 30.8. A homogeneous linear ODE with constant coefficients is a homogeneous linear ODE (30.5) where all the coefficients are constants:

$$a_0 y^{(n)} + a_1 y^{(n-1)} + \cdots + a_{n-1} y' + a_n y = 0. \quad (30.8)$$

Mathematicians found that with a suitable λ (some text books use the notation ω , r or m), the function $y = e^{\lambda x}$ is a solution of the equation (30.8). The values of λ can be determined by the following equation:

$$a_0 \lambda^n + a_1 \lambda^{n-1} + \cdots + a_{n-1} \lambda + a_n = 0. \quad (30.9)$$

Definition 30.9. The equation (30.9) is called the *characteristic equation* or *auxiliary equation* of the homogeneous linear ODE (30.8).

Let $p(\lambda) = a_0 \lambda^n + a_1 \lambda^{n-1} + \cdots + a_{n-1} \lambda + a_n$. $p(\lambda)$ is called the *characteristic polynomial*.

Since $p(\lambda)$ is a polynomial and according to Fundamental Theorem of Algebra (UECM3033 Complex Analysis), $p(\lambda)$ can be factorised into

$$(x - \lambda_1)^{m_1} (x - \lambda_2)^{m_2} \cdots (x - \lambda_k)^{m_k}$$

where λ_i are the (complex) roots of $p(\lambda)$ and $m_1 + \cdots + m_k = n$ (the degree of the characteristic polynomial or the order of the homogeneous linear ODE).

Definition 30.10. The values λ_i are called the *characteristic values* with *multiplicity* m_i of the ODE (30.8).

Theorem 30.13. If the multiplicity m_i are all one, then the general solution (30.8) is

$$y = C_1 e^{\lambda_1 x} + C_2 e^{\lambda_2 x} + \cdots + C_n e^{\lambda_n x}.$$

We will be discussing a few examples where the multiplicities are all one.

Example 30.14 (Final Exam Sept 2014, Q1(b)). Find the general solution of the homogeneous linear ODE $y'' - y' - 6y = 0$. (5 marks)

Solution. The characteristic equation is $\lambda^2 - \lambda - 6 = 0$ [2 marks]
 The characteristic values are $\lambda = -2, 3$ [1 mark]
 The general solution of the ODE is $y = C_1 e^{-2x} + C_2 e^{3x}$ [2 marks] $\square$

Example 30.15 (UECM3003 PDE). When solving a one-dimensional heat equation problem on a tiny rod of length $L > 0$, by applying the “separation of variable”, one will obtain the following ODE problem

$$\frac{d^2 X}{dx^2} + \lambda X = 0, \quad X(0) = X(L) = 0$$

where λ is an arbitrary constant. Find all the solutions of the above ODE problem for various different values of λ .

Solution. This is a linear ODE with constant coefficients, it has the following characteristic equation

$$\omega^2 + \lambda = 0 \quad (*)$$

- When $\lambda < 0$, the corresponding general solution is

$$(*) \Rightarrow \omega = \pm\sqrt{-\lambda} \Rightarrow X(x) = Be^{\sqrt{-\lambda}x} + Ce^{-\sqrt{-\lambda}x} \quad (30.10)$$

where B and C are arbitrary constants. From $X(0) = X(L) = 0$, we have

$$\begin{aligned} X(0) &= B + C = 0 \\ X(L) &= Be^{\sqrt{-\lambda}L} + Ce^{-\sqrt{-\lambda}L} = 0 \end{aligned}$$

Since $\sqrt{-\lambda}L \neq 0$,

$$\begin{vmatrix} 1 & 1 \\ e^{\sqrt{-\lambda}L} & e^{-\sqrt{-\lambda}L} \end{vmatrix} = e^{-\sqrt{-\lambda}L} - e^{\sqrt{-\lambda}L} = -2\sinh(\sqrt{-\lambda}L) \neq 0$$

The simultaneous equations above lead to $B = C = 0$ which implies that there is only one solution $X = 0$.

- When $\lambda = 0$, the general solution is

$$(*) \Rightarrow \omega = 0 \Rightarrow X(x) = Bx + C.$$

From $X(0) = X(L) = 0$, we have

$$\begin{aligned} X(0) &= C = 0 \\ X(L) &= BL + C = 0 \Rightarrow B = 0. \end{aligned}$$

Hence, there is only one solution $X(x) = 0$.

- When $\lambda > 0$, the general solution is

$$(*) \Rightarrow \omega = \pm\sqrt{\lambda}i \Rightarrow X(x) = B\sin(\sqrt{\lambda}x) + C\cos(\sqrt{\lambda}x).$$

From $X(0) = X(L) = 0$, we have

$$\begin{aligned} X(0) &= C = 0, \\ X(L) &= B\sin(\sqrt{\lambda}L) + C\cos(\sqrt{\lambda}L) = 0 \end{aligned}$$

From this simultaneous equations, we obtain

$$B\sin(\sqrt{\lambda}L) = 0$$

If $B = 0$, then we obtain a solution $X(x) = 0$.

If $B \neq 0$, then $\sin(\sqrt{\lambda}L) = 0 \Rightarrow \sqrt{\lambda}L = n\pi \Rightarrow \lambda = \left(\frac{n\pi}{L}\right)^2$ where $n = \pm 1, \pm 2, \pm 3, \dots$.

Note that n cannot be zero because $\lambda > 0$.

Therefore, when $\lambda = \left(\frac{n\pi}{L}\right)^2$, $n = 1, 2, \dots$, we have many solutions:

$$X(x) = B\sin \frac{n\pi x}{L}.$$

and when $\lambda > 0$ and $\lambda \neq \left(\frac{n\pi}{L}\right)^2$, $n = 1, 2, \dots$, we have only one solution $X = 0$.

□

Example 30.16 (Final Exam Sept 2016, Q2(a)). A spring, for which $k = 48$ g/cm, hangs in a vertical position with its upper end fixed. A mass, weighing $M = 0.5$ g, is attached to the lower end. After coming to rest, the mass is pulled down 2 cm and released. The mass is in a medium with a resistance equals to $64.375 \frac{dx}{dt}$. If the ODE that describes the motion of the mass is

$$M \frac{d^2x}{dt^2} = -kx - 64.375 \frac{dx}{dt}$$

where x is the position of centre of mass. Determine the motion of the mass by solving the above initial

value problem. Is the motion oscillatory? What is the final state? (10 marks)

Solution. Rearranging the equation, we obtain a homogeneous equation:

$$0.5 \frac{d^2 x}{dt^2} + 64.375 \frac{dx}{dt} + 48x = 0 \quad [1 \text{ mark}]$$

The characteristic equation is

$$0.5\lambda^2 + 64.375\lambda + 48 = 0. \quad [2 \text{ marks}]$$

By using hand calculation and scientific calculator, we obtain

$$\lambda = \frac{-64.375 \pm \sqrt{64.375^2 - 4 \times 0.5 \times 48}}{2 \times 0.5} = -0.75, -128 \quad [2 \text{ marks}]$$

Hence, the general solution is

$$x = C_1 e^{-128t} + C_2 e^{-0.75t} \quad [1 \text{ mark}]$$

From the initial conditions

$$\begin{aligned} x(0) &= C_1 + C_2 = 2 \\ x'(0) &= -128C_1 - 0.75C_2 = 0 \end{aligned} \quad [2 \text{ marks}]$$

Solving them gives $C_1 = -\frac{6}{509}$, $C_2 = \frac{1024}{509}$ and the motion of the position is

$$x(t) = \frac{1024}{509} e^{-0.75t} - \frac{6}{509} e^{-128t}. \quad [1 \text{ mark}]$$

This motion is not oscillatory and tends to rest at equilibrium again. [1 mark] □

Now we will be discussing the case when *some* multiplicity are not one. Consider a second order homogeneous linear ODE

$$ay'' + by' + cy = 0 \quad (30.11)$$

when one of the multiplicity $m_i > 1$. From its characteristic equation

$$a\lambda^2 + b\lambda + c = 0,$$

if $m_i > 1$, then $m_i = 2$ and $b^2 - 4ac = 0$ and we have a repeated root

$$\lambda = -\frac{b}{2a}.$$

The general solution is

$$y = (C_1 + C_2 x) e^{\lambda x}. \quad (30.12)$$

In general, the solution for characteristic values with multiplicity m greater than 1 has a corresponding solution $(C_1 + \dots + C_{m-1} x^{m-1}) e^{\lambda x}$.

Example 30.17. Find the general solution of the following linear ODE:

$$\frac{d^4 y}{dx^4} - 5 \frac{d^3 y}{dx^3} + 9 \frac{d^2 y}{dx^2} - 7 \frac{dy}{dx} + 2y = 0.$$

[**Hint:** If your calculator can handle polynomials up to degree 4, you should be able to solve it easily. Otherwise, you need to practise your algebra skill on polynomial factorisation together with the calculator that can handle polynomials up to degree 3.]

Solution. Since it is linear homogeneous with constant coefficients, the characteristic equation is

$$\lambda^4 - 5\lambda^3 + 9\lambda^2 - 7\lambda + 2 = 0$$

By looking at the coefficient 2 corresponding to degree 0, we guess the possible integer roots are ± 2 and ± 1 . Putting $\lambda = 1$ into the characteristic polynomial, we find that it is 0 and so it is a root. By long division, we obtain

$$(\lambda - 1)(\lambda^3 - 4\lambda^2 + 5\lambda - 2) = 0$$

Now by using calculator to do factorisation, we should obtain

$$(\lambda - 1)^3(\lambda - 2) = 0$$

Hence $\lambda = 1$ has a multiplicity $m_1 = 3$ and $\lambda = 2$ has multiplicity $m_2 = 1$. The general solution is

$$y = (C_1 + C_2x + C_3x^2)e^x + C_4e^{2x}.$$

□

30.3 Non-Homogeneous Equations with Constant Coefficients & Method of Undetermined Coefficients

Let us turn our discussion to the generalisation of homogeneous ODE (30.8) in the previous section to non-homogeneous form.

Definition 30.11. A non-homogeneous linear ODE with constant coefficients is a non-homogeneous linear ODE (30.2) where all the coefficients are constants:

$$a_0y^{(n)} + a_1y^{(n-1)} + \cdots + a_{n-1}y' + a_ny = b(x). \quad (30.13)$$

There is no general solution method to obtain a solution except by trial and error. To find the general solution of equation (30.13), we first find the general solution y_h of the corresponding homogeneous ODE of (30.13) and then try to guess a particular solution y_p from the form of non-homogeneous term $b(x)$ with the method of undetermined coefficients. If the trial function is part of the homogeneous solution y_h , we could multiply it by x and try again until we find a particular solution which is not part of the homogeneous solution. The general solution of (30.13) is the sum $y = y_h + y_p$ according to Theorem 30.12. The following shows some trial y_p for different type of functions.

Functions	Initial trial / guess for y_p
e^{ax}	de^{ax}
$\cos \beta x, \sin \beta x$	$d_1 \cos \beta x + d_2 \sin \beta x$
$a_0 + a_1x + a_2x^2 + \cdots$	$d_0 + d_1x + d_2x^2 + \cdots$

Let us investigate how to find y_h and y_p with a few of the following second order examples. Higher order examples are similar but the calculations are more tedious.

Example 30.18. Find the general solution of

$$y'' + 2y' - 3y = 4e^{2x} + 8e^x. \quad (30.14)$$

Solution: The characteristic equation is

$$\lambda^2 + 2\lambda - 3 = 0.$$

Hence $\lambda = -3, 1$ and the general solution of the corresponding homogeneous of (30.14) is

$$y_h = C_1e^{-3x} + C_2e^x$$

where C_1 and C_2 are arbitrary constants.

Remark: y_h is NOT a solution of (30.14) but a solution of $y'' + 2y' - 3y = 0$.

Looking at the right-hand-side, we find that it involves e^x , which is a special case of the y_h . We mentioned that the trial function needs to be multiply by x .

Therefore, the guess of the particular solution has the form $y_p = Ae^{2x} + Bxe^x$. Substituting into the non-homogeneous ODE (30.14), we have

$$A(2^2 + 2 \times 2 - 3)e^{2x} + B(2e^x + xe^x + 2(e^x + xe^x) - 3(xe^x)) = 4e^{2x} + 8e^x.$$

After simplification, we can determine the coefficient from $5A = 4$ and $4B = 8$ to be $A = \frac{4}{5}$ and

$$B = 2.$$

Hence, the general solution of (30.14) is

$$y = C_1e^{-3x} + C_2e^x + \frac{4}{5}e^{2x} + 2xe^x.$$

Remark: We must be careful to choose the right guess y_d . A wrong guess will lead to the LHS $\neq$ RHS.

Example 30.19. Solve

$$y'' - 4y = 8x^2 - 2x. \quad (30.15)$$

e:nonhomog:con

Solution. The corresponding homogeneous ODE has a characteristic equation $\lambda^2 - 4 = 0$. The characteristic values are $\lambda = \pm 2$ and the general solution of the homogeneous ODE is $y_h = C_1 e^{2x} + C_2 e^{-2x}$.

Looking at the right-hand-side $8x^2 - 2x$, we guess that the form of the particular solution is $y_p = d_1 x^2 + d_2 x + d_3$. By substituting into the non-homogeneous ODE (30.15), we have

$$y_d'' - 4y_d = 2d_1 - 4(d_1 x^2 + d_2 x + d_3) = (-4d_1)x^2 + (-4d_2)x + (2d_1 - 4d_3) = 8x^2 - 2x.$$

Comparing coefficients gives a simultaneous equation:

$$\begin{aligned} -4d_1 &= 8, \\ -4d_2 &= -2, \\ 2d_1 - 4d_3 &= 0. \end{aligned}$$

Solving it by hand or with calculator gives $d_1 = -2$, $d_2 = \frac{1}{2}$ and $d_3 = -1$.

Hence, the general solution of (30.15) is

$$y = C_1 e^{2x} + C_2 e^{-2x} - 2x^2 + \frac{1}{2}x - 1.$$

□

The values of the characteristic equation may not always be nice as illustrated by the following example.

Example 30.20. Solve $y'' + 7y' - 12y = x + \cos 3x$.

Solution. The characteristic equation is $\lambda^2 + 7\lambda - 12 = 0$, $\lambda = \frac{-7}{2} \pm \frac{\sqrt{97}}{2}$.

The homogeneous solution is

$$y_h = C_1 e^{(\frac{-7}{2} + \frac{\sqrt{97}}{2})x} + C_2 e^{(\frac{-7}{2} - \frac{\sqrt{97}}{2})x}.$$

Let

$$y_p = Ax + B + d_1 \cos 3x + d_2 \sin 3x.$$

Substituting into the non-homogeneous ODE, we have

$$7A - 12Ax - 12B + (-9d_1 + 21d_2 - 12d_1) \cos 3x + (-9d_2 - 21d_1 - 12d_2) \sin 3x = x + \cos 3x.$$

Solving it gives

$$A = -\frac{1}{12}, \quad B = \frac{-7}{144}, \quad d_1 = -\frac{1}{42}, \quad d_2 = \frac{1}{42}.$$

The general solution is

$$y = c_1 e^{(\frac{-7}{2} + \frac{\sqrt{97}}{2})x} + c_2 e^{(\frac{-7}{2} - \frac{\sqrt{97}}{2})x} - \frac{1}{12}x - \frac{7}{144} - \frac{1}{42} \cos 3x + \frac{1}{42} \sin 3x.$$

□

Example 30.21 (Final Exam Sept 2014, Q3(a)). A simple series circuit has an inductor of L , a capacitor of C and a resistor of R . The initial charge on the capacitor is zero. If a 12 volt battery is connected to the circuit, and the circuit is closed at $t = 0$, then the current i is described by the following ODE:

$$L \frac{di}{dt} + \frac{1}{C} \int_0^t i dt + Ri = 12.$$

Given that $L=1$ henry, $C = 10^{-6}$ farads, $R=1000$ ohms and the current i is related to the charge q in the series circuit by

$$i = \frac{dq}{dt}.$$

1. If the current at time $t = 0$ is zero, derive an **initial value problem** of a linear second order ODE that models the change of charge q of the given circuit with respect to time t .

(2 marks)

Solution. By substituting $i = \frac{dq}{dt}$ into the equation, we have the ODE:

$$\frac{d^2q}{dt^2} + 1000000(q - q(0)) + 1000\frac{dq}{dt} = 12 \quad [1 \text{ mark}]$$

and the initial conditions $q(0) = 0$, $q'(0) = i(0) = 0$ [1 mark] □

2. Find the solution of the initial value problem in (i) and show that the charge on the capacitor 1 second later is

$$q(1) = \frac{12}{10^6} \left[1 - e^{-500} \left(\cos 500\sqrt{3} + \frac{1}{\sqrt{3}} \sin 500\sqrt{3} \right) \right].$$

(8 marks)

Solution. This is a non-homogeneous linear ODE with constant coefficients. It is easy to find the particular solution $q_p = \frac{12}{1000,000}$ [1 mark]

and the general solution can be calculated as follows:

• $\lambda^2 + 1000\lambda + 1000000 = 0$ [1 mark]

• $\lambda = \frac{-1000 \pm \sqrt{3 \times 1000000}i}{2} = -500 \pm 500\sqrt{3}i$ [1 mark]

• $q = q_h + q_p = \frac{12}{1000,000} + e^{-500t}(C_1 \cos 500\sqrt{3}t + C_2 \sin 500\sqrt{3}t)$ [1 mark]

The solution can be calculated as follows:

• $q(0) = 0 \Rightarrow C_1 = -\frac{12}{1000,000}$ [1 mark]

• $q'(0) = 0 \Rightarrow C_1(-500) + C_2(500\sqrt{3}) = 0 \Rightarrow C_2 = \frac{1}{\sqrt{3}}C_1$ [1 mark]

The solution of the initial value problem is

$$q = \frac{12}{1000,000} \left[1 - e^{-500t}(\cos 500\sqrt{3}t + \frac{1}{\sqrt{3}} \sin 500\sqrt{3}t) \right] \quad [1.5 \text{ marks}]$$

When $t = 1$, we have [0.5 mark]

$$q(1) = \frac{12}{1000,000} \left[1 - e^{-500}(\cos 500\sqrt{3} + \frac{1}{\sqrt{3}} \sin 500\sqrt{3}) \right].$$

□

3. Find the steady state charge of the circuit. (1 mark)

Solution. Steady state charge is $\lim_{t \rightarrow \infty} q(t) = 1.2 \times 10^{-5}$ [1 mark] □

30.4 Homogeneous Linear Equation & Reduction of Order

Now, let us investigate the homogeneous linear equation (30.2) in which some or all of the **coefficients are not constant** and $b(x) = 0$.

A special type of homogeneous linear equation, which can be transformed to a homogeneous linear ODE with constant coefficients by a particular transform, is stated below.

Definition 30.12. [Zwillinger, 1998, p. 281] A homogeneous linear ODE of order n of the form

$$a_0x^n y^{(n)} + a_1x^{n-1}y^{(n-1)} + \cdots + a_{n-1}xy' + a_ny = 0 \quad (30.16)$$

e:ode:euler

is called an *Euler equation*.

Let $x = e^t$, then $\frac{dx}{dt} = e^t = x$ and

$$\frac{dy}{dx} = \frac{dy}{dt} \cdot \frac{dt}{dx} = \frac{1}{x} \frac{dy}{dt} \Rightarrow x \frac{dy}{dx} = \frac{dy}{dt}$$

$$\frac{d^2y}{dt^2} = \frac{d}{dt} \left[x \frac{dy}{dx} \right] = \frac{dx}{dt} \cdot \frac{dy}{dx} + x \frac{d^2y}{dx^2} \frac{dx}{dt} = x \frac{dy}{dx} + x^2 \frac{d^2y}{dx^2} \Rightarrow x^2 \frac{d^2y}{dx^2} = \frac{d^2y}{dt^2} - \frac{dy}{dt} = \frac{d}{dt} \left[\frac{d}{dt} - 1 \right] y$$

and in general

$$x^n \frac{d^n y}{dx^n} = \frac{d}{dt} \left[\frac{d}{dt} - 1 \right] \cdots \left[\frac{d}{dt} - (n-1) \right] y.$$

With the change of scale $x = e^t$, we can transform (30.16) to a homogeneous linear equation with constant coefficients

$$a_0 \frac{d}{dt} \left[\frac{d}{dt} - 1 \right] \cdots \left[\frac{d}{dt} - (n-1) \right] y + \cdots + a_{n-1} \frac{dy}{dt} + a_n y = 0$$

where the characteristic equation is

$$a_0 \lambda(\lambda-1) \cdots (\lambda-(n-1)) + \cdots + a_{n-1} \lambda + a_n = 0.$$

Example 30.22 (sci.math.symbolic, Ming Liu, 1995.7.3. Can Maple (do) more than MMA (Mathematica)?). Find a general solution for

$$y^{(4)} - \frac{4}{x^2} y'' + \frac{8}{x^3} y' - \frac{8}{x^4} y = 0.$$

Solution: Rewriting it as Euler equation:

$$x^4 y^{(4)} - 4x^2 y'' + 8xy' - 8y = 0 \quad (*)$$

Assume $x = e^t$ and $y = e^{\lambda t}$, then (*) becomes

$$\lambda(\lambda-1)(\lambda-2)(\lambda-3) - 4\lambda(\lambda-1) + 8\lambda - 8 = 0.$$

By extracting $\lambda-1$ from the terms and solving the polynomial of degree 3, we are able to obtain the following solution

$$\lambda = -1, 1, 2, 4.$$

Hence, the general solution is

$$y = \frac{C_1}{x} + C_1 x + C_3 x^2 + C_4 x^4.$$

Example 30.23. In solving the Laplace equation on a circular plate using the separation of variables, mathematicians obtain the following ODEs:

$$r^2 R'' + rR' - m^2 R = 0,$$

where $m = 0, 1, 2, \dots$. For each of these integers, determine the general solution of the ODE.

Solution. When $m = 0$, the general solution is found to be

$$R(r) = C_0 + D_0 \ln r$$

by applying the techniques from Topic 1. The characteristic equation is going to give only one characteristic solution $x^0 = 1$.

When $m = 1, 2, \dots$, we found the general solution of the ODE, which is an Euler equation, to be

$$R(r) = C_m r^m + D_m r^{-m}$$

from the characteristic equation $\lambda(\lambda-1) + \lambda - m^2 = 0$. □

The above example inspires us to investigate the roots of the characteristic equation of second order Euler Equation:

$$a_0 \lambda(\lambda-1) + a_1 \lambda + a_2 = 0$$

which can be categorised into three cases below:

Case 1: Two real roots λ_1 and λ_2 , the general solution has the form

$$y = C_1 |x|^{\lambda_1} + C_2 |x|^{\lambda_2}.$$

Case 2: Two complex roots $\lambda_{1,2} = \alpha \pm i\beta$, the general solution has the form

$$y = |x|^\alpha (C_1 \cos(\beta \ln |x|) + C_2 \sin(\beta \ln |x|)).$$

Case 3: Two repeating real roots $\lambda_1 = \lambda_2 = \lambda$, the general solution has the form

$$y = (C_1 + C_2 \ln |x|)|x|^\lambda.$$

Example 30.24. Solve the equation $x^2 y'' + 2xy' - 6y = 0$.

Solution: This is an Euler equation and the characteristic equation is

$$r(r-1) + 2r - 6 = 0$$

which gives the roots $r = -3, 2$ and the general solution

$$y = C_1 |x|^{-3} + C_2 |x|^2.$$

Example 30.25 (Final Exam April 2009, Q2(a)). Write down a fundamental set of solutions for $x^2 y'' + 2xy' - 6y = 0$, $x > 0$. (6 marks)

Solution. This is an Euler equation. The characteristic equation is

$$\lambda(\lambda-1) + 2\lambda - 6 = 0 \Rightarrow \lambda^2 + \lambda - 6 = 0 \Rightarrow \lambda = -3, 2.$$

The fundamental set of solutions are $\{x^{-3}, x^2\}$. □

Example 30.26. Find the general solution for $x^2 y'' - 5xy' + 9y = 0$

Solution. Let $y = |x|^\lambda$. Then we obtain

$$\lambda(\lambda-1) - 5\lambda + 9 = 0$$

which gives the roots $\lambda = 3, 3$ and the general solution

$$y = |x|^3 (C_1 + C_2 \ln |x|).$$

□

Example 30.27 (Final Exam Sept 2016, Q1(c)). Find the general solution for the following homogeneous equation:

$$x^2 y'' - 5xy' + 13y = 0. \quad (7 \text{ marks})$$

Solution. This is an Euler equation and it has the characteristic equation

$$\lambda(\lambda-1) - 5\lambda + 13 = 0 \Rightarrow \lambda^2 - 6\lambda + 13 = 0 \quad [3 \text{ marks}]$$

which gives the complex roots

$$\lambda = \frac{6 \pm \sqrt{6^2 - 4 \times 13}}{2} = 3 \pm 2i \quad [2 \text{ marks}]$$

and the general solution is

$$y = |x|^3 (C_1 \cos 2 \ln |x| + C_2 \sin 2 \ln |x|). \quad [2 \text{ marks}]$$

□

Example 30.28. Solve the initial value problem $x^2 y'' - 5xy' + 10y = 0$, $y(1) = 4$, $y'(1) = -6$.

Solution. Let $y = |x|^\lambda$. Then we obtain

$$\lambda(\lambda-1) - 5\lambda + 10 = 0$$

which gives the complex roots $\lambda = 3 \pm i$ and the general solution is

$$y = |x|^3 (C_1 \cos \ln |x| + C_2 \sin \ln |x|).$$

When $x = 1$,

$$y(1) = 4 = C_1, \quad y'(1) = 3 \times 4 + C_2 = -6 \Rightarrow C_2 = -18.$$

Therefore, the solution of the initial value problem is

$$y = x^3(4 \cos \ln x - 18 \sin \ln x).$$

□

If we have homogeneous linear equation of non-constant coefficients and non-Euler, we may not be able to solve it. However, if we know a particular solution y_1 of the homogeneous linear ODE of order n (normally by trial and error) and assume another solution is of the form $y_2 = vy_1$, we can derive the ODE for v , which is order $n - 1$. This technique is called the method of *reduction of order*.

In the case of order 2, after the reduction of order, we will get a first order equation, in which we can solve it using the techniques from Topic 1.

Consider an order 2 homogeneous linear ODE in normal form:

$$y'' + p(x)y' + q(x)y = 0 \quad (30.17)$$

with a particular solution $y_1 = y_1(x)$.

Assume $y = v \cdot y_1$ is another solution (just a guess), by substituting y into (30.17), we obtain

$$(v''y_1 + 2v'y_1' + vy_1'') + p(x)(v'y_1 + vy_1') + q(x)(vy_1) = 0.$$

Regrouping in terms of v , v' and v'' , we have

$$y_1 v'' + (2y_1' + py_1)v' + (\underbrace{y_1'' + py_1' + qy_1}_{y_1 \text{ is a particular solution}})v = 0.$$

The v -term vanishes because y_1 satisfies the ODE. Renaming $v' = u$ gives

$$y_1 u' + (2y_1' + py_1)u = 0 \Rightarrow u' + (2\frac{y_1'}{y_1} + p)u = 0$$

which is a linear first order ODE which could be solved by multiplying a suitable integrating factor. Thus the general solution of (30.17) should be $y = v \cdot y_1$.

Example 30.29. A solution is $y_1 = e^{-2x}$ is given, use the technique of reduction of order to solve $y'' + 4y' + 4y = 0$.

Solution: Assume the general solution is $y = vy_1$. Substituting this into the ODE, we have

$$\begin{aligned} & (vy_1'' + 2v'y_1' + v''y_1) + 4(vy_1' + v'y_1) + 4vy_1 = 0 \\ \Rightarrow & y_1 v'' + (2y_1' + 4y_1)v' + (y_1'' + 4y_1' + 4y_1)v = 0 \\ \Rightarrow & y_1 v'' + (2y_1' + 4y_1)v' = 0. \end{aligned}$$

Let $u = v'$ and $y_1' = -2e^{-2x}$. This implies

$$e^{-2x}u' + (-4e^{-2x} + 4e^{-2x})u = e^{-2x}u' = 0.$$

Solving it gives

$$u = C_1, \quad C_1 \text{ is a constant.}$$

Hence

$$v' = C_1 \Rightarrow v = C_1 x + C_2, \quad C_2 \text{ is another arbitrary constant.}$$

So the general solution of the linear ODE is

$$y = vy_1 = (C_1 x + C_2)e^{-2x}.$$

Remark: This is consistent with the general solution obtained using the technique mentioned in Section 30.2.

The above is a trivial. The following is a nontrivial example.

Example 30.30 (Dr Oon Shea Ming, Universiti Malaya). Find the general solution of

$$xy'' - 2y' - xy = 0.$$

Solution: This requires some observation, that is, if we can find something such that $y'' - y = 0$, we can have $xy'' - xy = 0$. Out of various elementary functions, $y = e^x$ seems to be a possible solution. Therefore, we assume $y_1 = (ax + b)e^x$ and

$$\begin{aligned} -xy_1 &= (-ax^2 - bx)e^x, \\ -2y_1' &= -2(ax + a + b)e^x, \\ xy_1'' &= x(ax + 2a + b)e^x. \end{aligned}$$

Summing up gives

$$0 = e^x(-ax^2 - bx - 2ax - 2a - 2b + ax^2 + 2ax + bx) = -2(a + b)e^x$$

which implies $a = -b$. Let us pick $a = 1$ and $b = -1$ and we have a particular solution $y_1 = (x - 1)e^x$ for the linear ODE.

Assume $y(x) = (x - 1)e^x v(x) = y_1(x)v(x)$ and substitute into the ODE, we obtain

$$x(y_1''v + 2y_1'v' + y_1v'') - 2(y_1'v + y_1v') - xy_1v = 0$$

which becomes after simplification:

$$2xy_1'v' + xy_1v'' - 2y_1v' = 2x^2e^xv' + (x^2 - x)e^xv'' - 2(x - 1)e^xv' = 0.$$

Let $u = v'$ and dividing by e^x , we have

$$(x^2 - x)u' + 2(x^2 - x + 1)u = 0 \Rightarrow u' + \frac{2(x^2 - x + 1)}{x^2 - x}u = 0.$$

The integrating factor is

$$\begin{aligned} \alpha(x) &= \exp \int \frac{2(x^2 - x + 1)}{x^2 - 1} dx = \exp \int \left[2 - \frac{2}{x} + \frac{2}{x - 1} \right] dx \\ &= \exp [2x - 2 \ln |x| + 2 \ln |x - 1|] = e^{2x} x^{-2} (x - 1)^2 = \frac{(x - 1)^2 e^{2x}}{x^2} \end{aligned}$$

Hence,

$$u \frac{(x - 1)^2 e^{2x}}{x^2} = C \Rightarrow v' = \frac{Cx^2}{(x - 1)^2} e^{-2x} \Rightarrow v = D - C \frac{(x + 1)e^{-2x}}{2x - 2}$$

and the general solution of the ODE is

$$y = (x - 1)e^x v = D(x - 1)e^x - \frac{C}{2}(x + 1)e^{-x}$$

where C and D are arbitrary constants.

30.5 Non-Homogeneous Linear Equation and Variation of Parameters

For the non-homogeneous second order equation (30.1) mentioned in page 591, we do not have general solution method. However, if we have the analytic solution $y_h = C_1 y_1(x) + C_2 y_2(x)$ of the corresponding homogeneous equation of (30.1), we may assume that the non-homogeneous linear equation has a similar solution by “variation of parameters”:

$$y_p = u_1(x)y_1(x) + u_2(x)y_2(x).$$

If we differentiate y_p and substitute inside the ODE, we obtain

$$\begin{aligned} y_p'' + p(x)y_p' + q(x)y_p &= u_1''y_1 + 2u_1'y_1' + u_1y_1'' + u_2''y_2 + 2u_2'y_2' + u_2y_2'' \\ &\quad + p(x)\{u_1'y_1 + u_1y_1' + u_2'y_2 + u_2y_2'\} + q(x)\{u_1y_1 + u_2y_2\} \\ &= u_1''y_1 + 2u_1'y_1' + u_2''y_2 + 2u_2'y_2' + p(x)\{u_1'y_1 + u_2'y_2\} = g(x). \end{aligned} \tag{30.18}$$

To simplify the above equation, we can choose u_1 and u_2 such that

$$u_1'y_1 + u_2'y_2 = 0. \tag{30.19}$$

Differentiating (30.19), we obtain

$$u_1''y_1 + u_1'y_1' + u_2''y_2 + u_2'y_2' = 0. \quad (30.20)$$

Substituting (30.19) and (30.20) into (30.18), we have

$$\underbrace{u_1''y_1 + u_1'y_1' + u_2''y_2 + u_2'y_2'}_{=0 \text{ because of (30.20)}} + (u_1'y_1' + u_2'y_2') + p(x) \cdot 0 = g(x). \quad (30.21)$$

Writing (30.19) and (30.21) in matrix form, we have

$$\begin{bmatrix} y_1 & y_2 \\ y_1' & y_2' \end{bmatrix} \begin{bmatrix} u_1' \\ u_2' \end{bmatrix} = \begin{bmatrix} 0 \\ g(x) \end{bmatrix}$$

and let $W[y_1, y_2]$ be the Wronskian of y_1 and y_2 (see Definition 30.7), we obtain a solution

$$u_1' = \frac{-y_2g(x)}{W[y_1, y_2]}, \quad u_2' = \frac{y_1g(x)}{W[y_1, y_2]}.$$

Taking integration (and ignoring the arbitrary constants) and we have

$$u_1(x) = \int \frac{-y_2(x)g(x)}{W[y_1, y_2]} dx \quad \text{and} \quad u_2(x) = \int \frac{y_1(x)g(x)}{W[y_1, y_2]} dx. \quad (30.22)$$

Example 30.31 (Final Exam April 2009, Q2(b)). Given that general solution to $y'' + p(x)y' + q(x)y = 0$ is of the form $y = c_1y_1 + c_2y_2$, and a particular solution to the non-homogeneous equation $y'' + p(x)y' + q(x)y = f(x)$ is $y_p = u_1y_1 + u_2y_2$ show that $u_1 = - \int \frac{f(x)y_2}{y_1y_2' - y_2y_1'} dx$. (9 marks)

A proof is contained in the above analysis.

Example 30.32. Find a general solution for $y'' + 2y' - 3y = 8e^x$.

Solution: The characteristic equation of the corresponding homogeneous equation is $\lambda^2 + 2\lambda - 3 = 0$. The characteristic values $\lambda = -3, 1$. The homogeneous solution is

$$y_h = C_1e^{-3x} + C_2e^x.$$

Note that $g(x) = 8e^x$. Let $y_1 = e^{-3x}$, $y_2 = e^x$ and

$$y_p = u_1y_1 + u_2y_2 = u_1e^{-3x} + u_2e^x.$$

The method of variation of parameter tell us that we need to calculate the Wronskian:

$$W[y_1, y_2] = \begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix} = \begin{vmatrix} e^{-3x} & e^x \\ -3e^{-3x} & e^x \end{vmatrix} = e^{-2x} + 3e^{-2x} = 4e^{-2x}$$

and then the functions u_1 and u_2 are

$$u_1 = \int \frac{-y_2g(x)}{W[y_1, y_2]} dx = \int \frac{-e^x \cdot 8e^x}{4e^{-2x}} dx = \frac{-2}{4}e^{4x} + C_1$$

$$u_2 = \int \frac{y_1g(x)}{W[y_1, y_2]} dx = \int \frac{e^{-3x} \cdot 8e^x}{4e^{-2x}} dx = 2x + C_2.$$

Hence, the general solution is

$$y = u_1y_1 + u_2y_2 = C_1e^{-3x} + C_2e^x - \frac{1}{2}e^x + 2xe^x.$$

Example 30.33. Find a general solution for $y'' - \frac{4}{x}y' + \frac{4}{x^2}y = x^2 + 1$.

Solution. Note that

$$y'' - \frac{4}{x}y' + \frac{4}{x^2}y = 0 \Rightarrow x^2y'' - 4xy' + 4y = 0$$

has a characteristic equation

$$\lambda(\lambda - 1) - 4\lambda + 4 = 0 \Rightarrow \lambda^2 - 5\lambda + 4 = 0$$

and the characteristic values are $\lambda = 1, 4$. Let $y_1 = x$, $y_2 = x^4$, $y_p = u_1y_1 + u_2y_2$ and $g(x) = x^2 + 1$. The Wronskian is

$$W = \begin{vmatrix} x & x^4 \\ 1 & 4x^3 \end{vmatrix} = 4x^4 - x^4 = 3x^4.$$

Then

$$u_1 = \int \frac{(-y_2)g(x)}{W} dx = - \int \frac{x^4(x^2 + 1)}{3x^4} dx = -\frac{1}{3} \int (x^2 + 1) dx = -\frac{1}{9}x^3 - \frac{1}{3}x + C_1$$

and

$$u_2 = \int \frac{y_1g(x)}{W} dx = \int \frac{x(x^2 + 1)}{3x^4} dx = \frac{1}{3} \int \left(\frac{1}{x} + \frac{1}{x^3}\right) dx = \frac{1}{3} \ln x - \frac{1}{6}x^{-2} + C_2.$$

Hence the general solution is

$$\begin{aligned} y &= u_1y_1 + u_2y_2 = \left(-\frac{1}{9}x^3 - \frac{1}{3}x + C_1\right)x + \left(\frac{1}{3} \ln x - \frac{1}{6}x^{-2} + C_2\right)x^4 \\ &= C_1x + C_2x^4 - \frac{1}{9}x^4 - \frac{1}{2}x^2 + \frac{1}{3}x^4 \ln x. \end{aligned}$$

□

Example 30.34 (Final Exam April 2009, Q2(c)). Find the general solution for $x^2y'' + 2xy' - 6y = 4x^2$, $x > 0$. (10 marks)

[**Warning:** To use the method variation of parameters, you must write the ODE in **normal form** to get the g , otherwise, you will get the wrong answer. It is possible to apply method of undetermined coefficients for this problem but you have guess it correctly. $y = Ax^2$ is not the right guess.]

Solution. The characteristic equation of the homogeneous Euler equation is

$$\lambda(\lambda - 1) + 2\lambda - 6 = 0 \Rightarrow (\lambda + 3)(\lambda - 2) = 0 \Rightarrow \lambda = -3, 2.$$

Therefore $y_1 = x^2$, $y_2 = x^{-3}$ and the Wronskian is

$$W = \begin{vmatrix} x^2 & x^{-3} \\ 2x & -3x^{-4} \end{vmatrix} = -3x^{-2} - 2x^{-2} = -\frac{5}{x^2}.$$

We need to write the ODE to normal form to get $g(x)$:

$$y'' + \frac{2}{x}y' - \frac{6}{x^2}y = \boxed{4 = g(x)}$$

Then

$$u_1 = \int \frac{(-x^{-3}) \cdot 4}{-5x^{-2}} dx = \frac{4}{5} \ln |x| + C_1, \quad u_2 = \int \frac{x^2 \cdot 4}{-5x^{-2}} dx = -\frac{4}{25}x^5 + C_2$$

Hence the general solution is

$$y = u_1y_1 + u_2y_2 = (C_1 + \frac{4}{5} \ln |x|)x^2 + (C_2 - \frac{4}{25}x^5)x^{-3} = C_1x^2 + C_2x^{-3} + \frac{4}{5}x^2 \ln |x| - \frac{4}{25}x^2.$$

□

Example 30.35 (Final Exam Sept 2013, Q1(b)). Find a particular solution of

$$x^2y'' - xy' + y = \frac{1}{x}, \quad x > 0.$$

Simplify your answer.

(13 marks)

[**Remark:** Unlike the previous example, it is possible to guess a “particular” solution by using method

of undetermined coefficients with $y = \frac{A}{x}$. It is easy to determine $A = \frac{1}{4}$, therefore, $y_p = \frac{1}{4x}$. I am not sure if this is acceptable for the lecturer setting this question, for me, it is OK.]

Solution. The roots of the characteristic equation of the homogeneous Euler equation is

$$\lambda(\lambda - 1) - \lambda + 1 = 0 \Rightarrow \lambda = 1.$$

Therefore $y_1 = x$, $y_2 = x \ln x$ and the Wronskian is

$$W = \begin{vmatrix} x & x \ln x \\ 1 & \ln x + 1 \end{vmatrix} = x \ln x + x - x \ln x = x.$$

From $y'' - \frac{1}{x}y' + \frac{1}{x^2}y = \frac{1}{x^3} = g(x)$, we have (using the change of variable $x = e^t$)

$$u_1 = \int \frac{-x \ln x \cdot x^{-3}}{x} dx = \frac{\ln x}{2x^2} + \frac{1}{4x^2}, \quad u_2 = \int \frac{x \cdot x^{-3}}{x} dx = -\frac{1}{2}x^{-2}$$

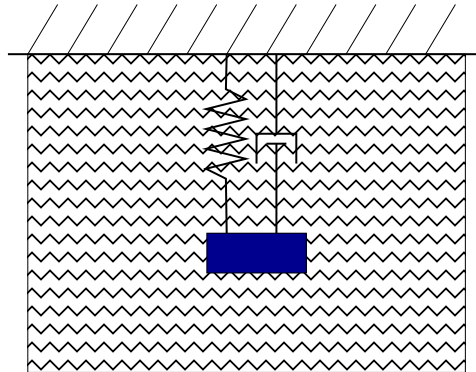
Therefore,

$$y = y_h + u_1 y_1 + u_2 y_2 = C_1 x + C_2 x \ln x + \frac{\ln x}{2x} + \frac{1}{4x} - \frac{\ln x}{2x}.$$

□

30.6 Mechanical and Electrical Vibrations

Consider the case where a small object of mass m is attached to an elastic spring of length ℓ , which is suspended from a rigid horizontal support.



The elastic spring has the property that if it is stretched or compressed a distance $\Delta\ell$ which is small compared to its natural length ℓ , then it will exert a restoring force of magnitude $k\Delta\ell$. The constant k is called the spring-constant, and is a measure of the stiffness of the spring. In addition, the mass and spring may be immersed in a medium, such as oil, which impedes the motion of an object through it. Engineers usually refer to such systems as *spring-mass-dashpot systems* [Braun, 1993, p. 165].

The analysis of mechanical vibration in spring-mass-dashpot systems is very important. The shock absorbers in our automobiles can be approximated by this system and we can learn its dynamics from the analysis of mechanical vibration.

Let y be the position of the position of the mass m and during equilibrium, $mg = k\Delta\ell$. If we use the symbol x to denote time (just here), the forces exerted on the mass m are

1. the weight of the mass, $F_{\text{weight}} = mg$;
2. the restoration force, $F_{\text{spring}} = -k(\Delta\ell + y)$;
3. the friction force, $F_{\text{friction}} = -\gamma y'$; and
4. the external force, $F_{\text{external}} = f(x)$.

By Newton's law,

$$F = ma \Rightarrow F_{\text{weight}} + F_{\text{spring}} + F_{\text{friction}} + F_{\text{external}} = my''$$

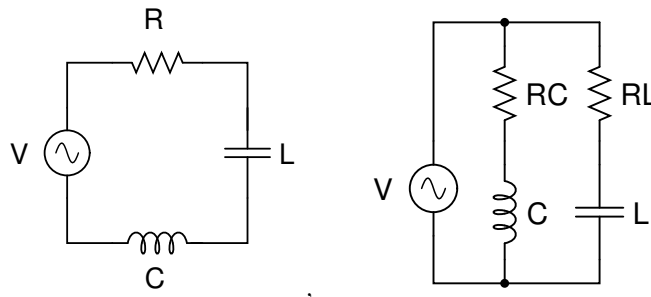
which simplifies to the following equation after the cancellation $mg = k\Delta\ell$:

$$my'' + \gamma y' + ky = f(x). \quad (30.23)$$

e:mech-vibr

When $\gamma = 0$, the system is called *undamped system*; when $f(x) = 0$, the system is called *free system*.

Electrical vibration or *electrical oscillation* is important in the analysis of the *resonance* in AC (alternating current) circuits. Resonance happens at a special frequency determined by the values of the resistance R , capacitance C , and inductance L . There are two kinds of RLC circuit: (a) series, on the left, and (b) parallel, on the right of the following figure:



For series RLC circuit, the condition of resonance is straightforward and it is characterised by minimum impedance and zero phase. However, for a parallel RLC circuit, the characterisation of resonance is much more complicated.

Let us investigate the series RLC circuit. The voltage drop across resistor, capacitor and inductor are iR , q/C and $L \frac{di}{dt}$ respectively. By Kirchhoff's Voltage Law, we have

$$iR + \frac{q}{C} + L \frac{di}{dt} = V(t).$$

Since the “charge” q is related to the “current” i by $i = \frac{dq}{dt}$, replacing i with $\frac{dq}{dt}$, we have second order ODE for the “charge” q [Xie, 2010]:

$$L \frac{d^2q}{dt^2} + R \frac{dq}{dt} + \frac{1}{C}q = V(t). \quad (30.24)$$

This equation is very similar to the equation (30.23) describing the mechanical system of a single degree-of-freedom.

In the following subsections, we investigate the behaviour of the vibrations for various values of parameters for (30.23) (the oscillation in (30.24) is similar where the variable is the current instead of position in (30.23)).

30.6.1 Free Vibration

When the system is undamped and there is no external force, i.e. $\gamma = 0$ and $f(x) = 0$, we have a *free vibration* or *simple harmonic motion*:

$$my'' + ky = 0.$$

This is a second order linear ODE with constant coefficients and the characteristic equation is

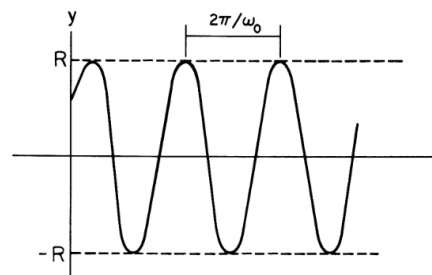
$$m\lambda^2 + k = 0 \Rightarrow \lambda = \pm i\sqrt{k/m}.$$

Let $\omega_0 = \sqrt{k/m}$, the general solution is

$$y = C_1 \cos \omega_0 x + C_2 \sin \omega_0 x = R \cos(\omega_0 x - \delta)$$

where $R = \sqrt{C_1^2 + C_2^2}$ and $\tan \delta = C_2/C_1$.

The position y can be regarded as an oscillation with amplitude R and phase δ with a *natural period* of $T_0 = \frac{2\pi}{\omega_0}$. The ω_0 is called the *angular frequency* of the oscillation. The value ω_0 is called the *natural frequency* of the oscillation. The sketch of this solution is shown on the right [Braun, 1993].



30.6.2 Damped Free Vibration

When the system is damped and there is no external force, i.e. $\gamma \neq 0$ and $f(x) = 0$, we have a *damped free vibration*:

$$my'' + \gamma y' + ky = 0. \quad (30.25)$$

The characteristic equation and characteristic values are

$$m\lambda^2 + \gamma\lambda + k = 0 \Rightarrow \lambda = \frac{-\gamma}{2m} \pm \frac{\sqrt{\gamma^2 - 4mk}}{2m}.$$

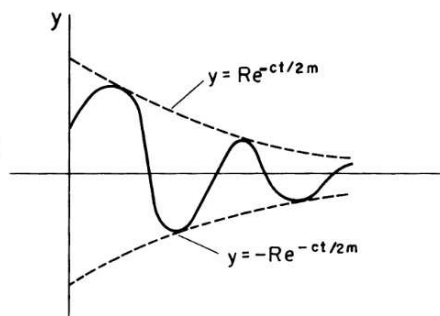
Based on the sign of $\gamma^2 - 4mk$, we have the following cases just like Section 30.2.

- **Case I: Underdamped** ($\gamma^2 < 4mk$)

Let $\beta = \frac{\sqrt{4mk - \gamma^2}}{2m}$. The solution of (30.25) is

$$y(x) = e^{-\gamma t/2m}(C_1 \cos \beta x + C_2 \sin \beta x) = Re^{-\gamma t/2m} \cos(\beta x - \delta).$$

Note that $y(x)$ is oscillatory and bounded by exponentially decaying curves $\pm Re^{-\gamma t/2m}$. The motion of $y(x)$ is not periodic. However, we could define the angular frequency β with which the mass oscillating up and down as quasi-frequency. Similarly, the quasi period is defined as $\frac{2\pi}{\beta}$. The sketch of this solution is shown on the right [Braun, 1993].

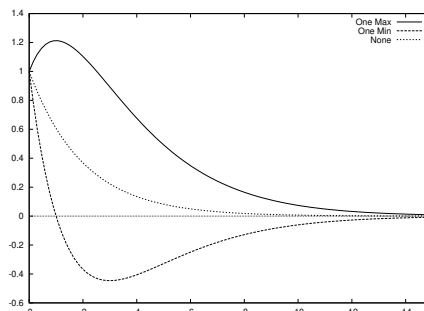


- **Case II: Critically Damped** ($\gamma^2 = 4mk$)

According to Section 30.2, the general solution is

$$y(x) = e^{-\gamma x/2m}(C_1 + C_2 x).$$

Note that $y(x)$ decays to zero since $\lim_{x \rightarrow \infty} y(x) = \lim_{x \rightarrow \infty} \frac{C_1 + C_2 x}{e^{\gamma x/2m}} = \lim_{x \rightarrow \infty} \frac{C_2}{(\gamma/2m)e^{\gamma x/2m}} = 0$. Examining the derivative of $y(x)$ shows that $y(x)$ can at most one maximum or minimum as shown on the right.



- **Case III: Overdamped** ($\gamma^2 > 4mk$)

According to Section 30.2, the general solution is

$$y(x) = C_1 e^{\lambda_1 x} + C_2 e^{\lambda_2 x}$$

where $\lambda_1 = \frac{-\gamma}{2m} + \frac{\sqrt{\gamma^2 - 4mk}}{2m}$ and $\lambda_2 = \frac{-\gamma}{2m} - \frac{\sqrt{\gamma^2 - 4mk}}{2m}$

Since $\gamma^2 > 4mk$ and $\gamma > \sqrt{\gamma^2 - 4mk}$ then λ_1 is negative and $\lambda_2 < \lambda_1 < 0$. This implies $y(x)$ dies off as $x \rightarrow \infty$. Examining the derivative of $y(x)$ shows that $y(x)$ can at most one maximum or minimum. Thus the sketches of $y(x)$ for overdamped case are qualitatively similar to those of critically damped motion.

30.6.3 Forced Free Vibration

When the system is undamped, i.e. $\gamma = 0$ and there is a external force $f(x) = F_0 \cos \omega x$, we have a *forced free vibration*:

$$my'' + ky = F_0 \cos \omega x. \quad (30.26)$$

We can use the method of undetermined coefficients from Section 30.3 to find $y(x)$. Let $y_p = d_1 \cos \omega x + d_2 \sin \omega x$, the

$$my_p'' + ky_p = -m\omega^2(d_1 \cos \omega x + d_2 \sin \omega x) + k(d_1 \cos \omega x + d_2 \sin \omega x) = F_0 \cos \omega x.$$

By comparing coefficients, we have

$$-m\omega^2 d_1 + kd_1 = F_0, \quad -m\omega^2 d_2 + kd_2 = 0.$$

When $k - m\omega^2 \neq 0$, we have $d_1 = \frac{F_0}{k - m\omega^2}$ and $d_2 = 0$ and the general solution is

$$y = C_1 \cos \sqrt{\frac{k}{m}}x + C_2 \sin \sqrt{\frac{k}{m}}x + \frac{F_0}{k - m\omega^2} \cos \omega x.$$

When $k - m\omega^2 = 0$, the $y_p = d_1 x \cos \omega x + d_2 x \sin \omega x$ according to the method of undetermined coefficients and

$$\begin{aligned} my_p'' + ky_p &= md_1(-2\omega \sin \omega x - \omega^2 x \cos \omega x) + md_2(2\omega \cos \omega x - \omega^2 x \sin \omega x) \\ &\quad + k(d_1 x \cos \omega x + d_2 x \sin \omega x) \\ &= 2m\omega(d_2 \cos \omega x - d_1 \sin \omega x) + (k - m\omega^2)d_1 x \cos \omega x + (k - m\omega^2)d_2 x \sin \omega x \\ &= F_0 \cos \omega x. \end{aligned}$$

By comparing coefficients, we have

$$2m\omega d_2 = F_0, \quad -2m\omega d_1 = 0 \Rightarrow d_1 = 0, \quad d_2 = \frac{F_0}{2m\omega}$$

and the general solution is

$$y = C_1 \cos \omega x + C_2 \sin \omega x + \frac{F_0}{2m\omega} x \sin \omega x, \quad \omega = \sqrt{\frac{k}{m}}.$$

The third term, represents an oscillation with increasing amplitude. Thus the forcing term $F_0 \cos \omega x$, if it is in resonance with the natural frequency $\omega_0 = \sqrt{k/m}$ of the system, will always cause unbounded oscillations. Such a phenomenon was responsible for the collapse of the Tacoma Narrows Bridge (at Puget Sound in the state of Washington, U.S. on the 1 July, 1940) and many other mechanical catastrophes [Braun, 1993, p. 171].

30.6.4 Damped Forced Vibration

When the system is damped, i.e. $\gamma \neq 0$ and there is an external force $f(x) = F_0 \cos \omega x$, we have a *damped forced vibration*:

$$my'' + \gamma y' + ky = F_0 \cos \omega x. \quad (30.27)$$

We can use the method of undetermined coefficients from Section 30.3 to find $y(x)$. First, we need to find the homogeneous solution x_h which is a solution of (30.25). Then we need to find a particular solution for (30.27) using the method of undetermined coefficients, which gives

$$x_p(t) = \frac{F_0}{(k - m\omega^2)^2 + \gamma^2 \omega^2} \left\{ (k - m\omega^2) \cos \omega x + \gamma \omega \sin \omega x \right\} = R \cos(\omega x - \delta)$$

where $\omega_0 = \sqrt{\frac{k}{m}}$, $\tan \delta = \frac{\gamma \omega}{m(\omega_0^2 - \omega^2)}$, and $R = \frac{F_0}{\sqrt{m^2(\omega_0^2 - \omega^2)^2 + \gamma^2 \omega^2}}$.

The amplitude R will be at its maximum when the driving frequency ω is closed to its natural frequency ω_0 , i.e. at $\omega_{\max}^2 = \omega_0^2(1 - \frac{\gamma^2}{4mk})$.

The response of amplitude R at driving frequencies closed to $\omega_{\max}$ is large and this phenomena is called *resonance*. The corresponding resonating amplitude is given by

$$R_{\max} = \frac{F_0}{\gamma \omega_0 \sqrt{1 - \frac{\gamma^2}{4mk}}} \approx \frac{F_0}{\gamma \omega_0} \left(1 + \frac{\gamma^2}{4mk}\right).$$

30.7 Working with Maxima

ode2() solves first and second order ODEs.

For example, (30.24) can be solved using

```
assume(notequal(R, 0));
ode2(L*diff(q(t),t,2)+R*diff(q(t),t)+q/C-V(t), q(t), t);
```

Chapter 31

Series Solutions of Linear Equations

In Topic 2, we learned the theory of general linear ODEs. It was found that constant coefficient linear ODEs are rather easy to solve by using characteristic equations and methods of undetermined parameters. On the other hand, variable coefficient linear ODEs are difficult to solve and most of them do not have quadrature solution [Smirnov, 1964a, p.45, item 45]. The second order variable coefficient linear ODEs appeared in the solution of second order linear partial differential equations (*UECM3003*) on *separable coordinate systems* using separation of variables technique [Morse and Feshbach, 1953a, Chapter 5]:

$$a(x)\frac{d^2y}{dx^2} + b(x)\frac{dy}{dx} + c(x)y = 0. \quad (31.1)$$

e:ode:ord2:homog

Here are some important examples arising from PDE [Temme, 1996a, 96]

- Hypergeometric differential equation

$$x(1-x)y'' + [c - (a+b+1)x]y' - aby = 0, \quad (31.2)$$

e:hypergeometric

- *Bessel's equation* of order $\nu \geq 0$ (a given parameter)

$$x^2y'' + xy' + (x^2 - \nu^2)y = 0, \quad (31.3)$$

e:bessel

- Legendre's equation

$$(1-x^2)y'' - 2xy' + \lambda y = 0, \quad -1 < x < 1 \quad (31.4)$$

e:legendre

where λ is a real constant. In PDE, $\lambda = \nu(\nu+1)$, and (31.4) is called the *Legendre's equation of order ν* .

- Hermite's equation of order λ

$$y'' - 2xy' + \lambda y = 0 \quad (31.5)$$

e:hermite

- Kummer differential equation

$$xy'' + (c-x)y' - ay = 0 \quad (31.6)$$

e:kummer

The possibility of approximating functions with Taylor series inspired mathematicians to solve the variable coefficient linear ODEs using general power series [Samuel S. Holland, 1990, §2.2]. Series solution is applicable to n -th order homogeneous linear ordinary differential equations, most frequently second order differential equations, to yield an infinite series expansion of the two independent solutions [Zwillinger, 1998].

Modern analysis of the series solutions were studied for the linear ODEs over complex domain $\mathbb{C}$, where the knowledge of Analyticity (Topic 4) and Residue Theorem (Topic 5) from *UECM3013: Complex Analysis* are required. Morse and Feshbach [1953a, Chapter 5] explains how complex integral can be more general than the series solution. The analytic theory of complex linear ODE is developed into an advanced theory of the global analysis and differential algebra. According to Kohno [1999], the global analysis of linear ODE (where (31.1) is a special case) is initiated by the works of G. G. Stokes and B. Riemann in 1857, the Airy function and Gauss hypergeometric function, which was found based on their “behaviour in large”, became very important in the characterisation of special functions [Kristensson, 2010], [Bell, 1968a].

For instance, Gauss hypergeometric function arises in all fields of mathematics, e.g., in number theory, in the theory of groups and in analysis itself. In a way, all power series are special or extended cases of the hypergeometric series.

Both the Airy function and the Gauss hypergeometric function are particular solutions of linear differential equations. In the theory of linear differential equations in $\mathbb{C}$, the solution of connection problems and Stokes phenomena, and the derivation of monodromy groups, are important problems. Temme [1996a] pointed out that special functions are often introduced as solutions of a linear homogeneous second order differential equation (31.1).

Due to the complexity of the theory of linear differential equations in $\mathbb{C}$, we will limit ourselves to the domain in $\mathbb{R}$. In Section 31.3, we will substitute an infinite series into a linear equation, then the different coefficients may be matched to obtain recurrences for the coefficients of the series. Solving these recurrences results in an explicit solution at a regular point. However, many second order ODE problems arising from physics contains singular points (Section 31.2), where the series expansion normally **diverges** except when the singular point is regular (Section 31.4). An English physicist Michael Berry said that “almost all series describing physics diverge!”

This topic is based on Boyce and DiPrima [2012] and Braun [1993, Chapter 2]. For advanced study of the series solutions, refer to Balser [1994].

31.1 Review of Power Series

In this section, we define the notion of power series and the associated operations, e.g., equality, addition, etc.

Definition 31.1. A *power series* at $x = x_0$ or in $x - x_0$ is an infinite series of the form

$$\sum_{n=0}^{\infty} a_n (x - x_0)^n \quad (31.7)$$

where x_0 is a point on $\mathbb{R}$ and the constants $a_0, a_1, \dots$ are called the (*real*) *coefficients* of the series and the point a is called the *centre* of the series.

Power series can be combined through the operations of equality, addition, multiplication, and division.

Definition 31.2. Consider two series $\sum_{n=0}^{\infty} a_n (x - x_0)^n$ and $\sum_{n=0}^{\infty} b_n (x - x_0)^n$. The two series are said

to be *equal* and denoted as $\sum_{n=0}^{\infty} a_n (x - x_0)^n = \sum_{n=0}^{\infty} b_n (x - x_0)^n$ if $a_n = b_n$ for all n . In particular, if

$\sum_{n=0}^{\infty} a_n (x - x_0)^n = 0$, then $a_n = 0$ for all n .

The *addition and subtraction* are defined as

$$\sum_{n=0}^{\infty} a_n (x - x_0)^n \pm \sum_{n=0}^{\infty} b_n (x - x_0)^n = \sum_{n=0}^{\infty} (a_n \pm b_n) (x - x_0)^n.$$

The *multiplication* is defined as

$$\left(\sum_{n=0}^{\infty} a_n (x - x_0)^n \right) \left(\sum_{n=0}^{\infty} b_n (x - x_0)^n \right) = \sum_{n=0}^{\infty} c_n (x - x_0)^n$$

where

$$c_n = a_0 b_n + a_1 b_{n-1} + a_2 b_{n-2} + \dots + a_n b_0 = \sum_{k=0}^n a_k b_{n-k}.$$

The *division* is defined as

$$\frac{f(x)}{g(x)} := \frac{\sum_{n=0}^{\infty} a_n (x - x_0)^n}{\sum_{n=0}^{\infty} b_n (x - x_0)^n} = \sum_{n=0}^{\infty} d_n (x - x_0)^n =: h(x)$$

where d_n can be determined recursively by solving the following linear equations (obtained by equating the coefficients of $f(x) = g(x)h(x)$):

$$\begin{aligned}
a_0 &= b_0 d_0 \Rightarrow d_0 = \frac{a_0}{b_0} \\
a_1 &= b_0 d_1 + b_1 d_0 \Rightarrow d_1 = \frac{a_1 - b_1 d_0}{b_0} \\
a_2 &= b_0 d_2 + b_1 d_1 + b_2 d_0 \Rightarrow d_2 = \frac{a_2 - b_1 d_1 - b_2 d_0}{b_0} \\
&\vdots
\end{aligned}$$

Before proceeding to the discussion of the convergence property of power series, we will recall the notion of convergence of number series from Calculus II.

Definition 31.3. A series $\sum_{n=0}^{\infty} a_n$ is said to be *convergent* if the limit $\lim_{N \rightarrow \infty} \sum_{n=0}^N a_n$ exists; otherwise, it is said to be *divergent*. A series is said to be *absolutely convergent* if the series $\sum_{n=0}^{\infty} |a_n|$ is convergent.

With this, we can define the notion of convergence for power series.

Definition 31.4. A power series (31.7) is said to be *converging* (or *is convergent*) at a point x if the limit

$$\lim_{N \rightarrow \infty} \sum_{n=0}^N a_n (x - x_0)^n$$

exists. A series is said to be *diverging* at a point x if it is not converging.

The *interval of convergence* of a power series is the set of all numbers for which the series converges.

A power series is said to be *absolutely converging* at a point x if

$$\lim_{N \rightarrow \infty} \sum_{n=0}^N |a_n (x - x_0)^n|$$

exists.

For a given power series (31.7), there are only 3 possibilities of the convergence behaviour:

1. The series converges only at $x = x_0$.
2. The series converges for all x .
3. There exists $R > 0$, called the *radius of convergence* of the power series, such that the series converges for $|x - x_0| < R$ and diverges for $|x - x_0| > R$. At $x - x_0 = \pm R$, the series may converge or diverge.

In case (a), we say $R = 0$; while in case (b), we say $R = \infty$.

If a power series converges at $x = x_1$, it converges absolutely for $|x - x_0| < |x_1 - x_0|$. If the power series diverges at $x = x_1$, it diverges for $|x - x_0| > |x_1 - x_0|$.

One of the most useful tests for the absolute convergence of a series is the ratio test stated below.

Theorem 31.1 (Ratio Test). If $a_n \neq 0$, and if

$$L = \lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right|,$$

then the power series $\sum_{n=0}^{\infty} a_n$ converges absolutely if $L < 1$, and diverges if $L > 1$.

If $L = 1$, the test is inconclusive and other test is required to determine whether the series converges or not.

Adapting the ratio test to power series, we obtain the following theorem.

Theorem 31.2 (Ratio Test for Power Series). *If*

$$\lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right| = L, \quad L \in \mathbb{R}$$

then the power series (31.7) converges absolutely for x in the range $|x - x_0| < \frac{1}{L}$ and diverges for x in the range $|x - x_0| > \frac{1}{L}$.

Given a power series, we can differentiate it formally as defined in Calculus II.

Definition 31.5. A power series defines a continuous function in its interval of convergence and can be differentiated (term-by term) any number of times:

$$\begin{aligned} \frac{d}{dx} \left[\sum_{n=0}^{\infty} a_n (x - x_0)^n \right] &= \sum_{n=1}^{\infty} n a_n (x - x_0)^{n-1}, \\ \frac{d^2}{dx^2} \left[\sum_{n=0}^{\infty} a_n (x - x_0)^n \right] &= \sum_{n=2}^{\infty} n(n-1) a_n (x - x_0)^{n-2} \end{aligned}$$

and so on.

A power series is an infinite summation, while a function is a mapping from an interval I to $\mathbb{R}$. An important question is: Does a power series actually represent a function at the point x_0 ?

Definition 31.6 (Analytic at a Point). A function f is *analytic at point x_0* if, in an open interval containing x_0 , it can be approximated exactly by a power series in $x - x_0$ with a positive radius of convergence.

Definition 31.7 (Analyticity). A function is said to be *analytic (over an open interval I)* if it is analytic at every point in I .

Mathematically, we are interested to know if under the “arithmetic” operations (addition, subtraction, etc.), the analytic property will be preserved. The following theorem provides a positive answer.

Theorem 31.3. *If functions f and g are analytic at point x_0 , then so are the functions $f \pm g$ and $f \cdot g$. If $g(x_0) \neq 0$, the quotient $\frac{f}{g}$ is analytic at x_0 .*

Let us look at two examples of analytic functions.

Example 31.4 (Polynomial functions are analytic over $\mathbb{R}$). A polynomial function $P_n(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ is analytic at any point $x_0 \in \mathbb{R}$ because it can always be written in the form $P_n(x) = b_0 + b_1(x - x_0) + b_2(x - x_0)^2 + \cdots + b_n(x - x_0)^n$.

Example 31.5 (Rational functions $\frac{P(x)}{Q(x)}$ are analytic over $\mathbb{R} \setminus \text{Zero}(Q(x))$). Let $\text{Zero}(Q(x)) := \{x_0 : Q(x_0) = 0\}$ and let $P(x)$ and $Q(x)$ be polynomials without a common factor. A rational function $\frac{P(x)}{Q(x)}$ is analytic except at $x_0 \in \text{Zero}(Q(x))$.

The following theorem provides a characterisation of analytic function using differentiation.

Theorem 31.6 (Taylor and Maclaurin Series Representations). *If f is analytic at $x = x_0$, then f can be represented by the Taylor series about x_0 :*

$$f(x) = \sum_{n=0}^{\infty} \underbrace{\frac{f^{(n)}(x_0)}{n!}}_{a_n} (x - x_0)^n.$$

When $x_0 = 0$, the series is also called the Maclaurin series.

According to this theorem, *analyticity* requires at least that f be infinitely differentiable at $x = x_0$. The following are two examples of non-analytic functions.

Example 31.7. The function $f(x) = |x - 2|$ is not analytic at $x = 2$ because $f'(2)$ does not exist.

Example 31.8. The function $f(x) = x^{7/3}$ is not analytic at $x = 0$ because $f'''(0)$ does not exist.

Many of the fundamental functions that we learned in Calculus are analytic.

Example 31.9. The functions e^x , $\sin x$, $\cos x$, $\sinh x$ and $\cosh x$ are analytic over $\mathbb{R}$, while $\frac{1}{1-x}$ is analytic over $\mathbb{R} \setminus \{1\}$ and $\ln x$ is analytic for $x > 0$. Some of the corresponding Maclaurin series are listed below:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots, \quad -\infty < x < \infty; \quad (31.8) \quad \text{e:exp}$$

$$\sin x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots, \quad -\infty < x < \infty; \quad (31.9) \quad \text{e:sin}$$

$$\cos x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots, \quad -\infty < x < \infty; \quad (31.10) \quad \text{e:cos}$$

$$\sinh x = \sum_{n=0}^{\infty} \frac{x^{2n+1}}{(2n+1)!} = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots, \quad -\infty < x < \infty; \quad (31.11) \quad \text{e:sinh}$$

$$\cosh x = \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!} = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots, \quad -\infty < x < \infty. \quad (31.12) \quad \text{e:cosh}$$

31.2 Ordinary Points and Singular Points

singular-points

To find the solutions of (31.1) using power series (31.7), we first rewrite (31.1) in *normal form* as mentioned in Topic 2:

$$\frac{d^2 y}{dx^2} + p(x) \frac{dy}{dx} + q(x)y = 0, \quad p(x) := \frac{b(x)}{a(x)}, \quad q(x) := \frac{c(x)}{a(x)}. \quad (31.13) \quad \text{e:ode:ord2:normal}$$

Then, we look at what happens if we substitute the power series (31.7) into (31.13):

$$\sum_{n=0}^{\infty} a_n n(n-1)(x-x_0)^{n-2} + p(x) \sum_{n=0}^{\infty} a_n n(x-x_0)^{n-1} + q(x) \sum_{n=0}^{\infty} a_n (x-x_0)^n = 0?$$

To determine the coefficients a_n , we need to identify the ordinary and singular points of $p(x)$ and $q(x)$ associated with (31.13) and the original homogeneous linear ODE (31.1), which do not have solutions in analytical form or quadrature form using the methods in Topic 2.

Definition 31.8. A point x_0 is said to be an *ordinary point* of the ODE (31.1) if both $p(x)$ and $q(x)$ in the normal form (31.13) are analytic at x_0 . A point that is not an ordinary point is called a *singular point*.

Definition 31.9. A singular point $x = x_0$ of the ODE (31.1) is said to be a *regular (singular) point* [Whittaker and Watson, 1920a] if the Taylor series of $(x-x_0)p(x)$ and $(x-x_0)^2 q(x)$ about x_0 converge in an open interval $(x_0 - h, x_0 + h)$, i.e. they are *analytic*, see Definition 31.7. A singular point that is not regular is called an *irregular singular point*.

Example 31.10. Determine the (a) ordinary points and (b) singular points (and the type of singularity) of the given differential equation $(x+1)y'' - x^2 y' + 3y = 0$.

Solution: (a) From the coefficient of y'' , we know that $x+1 \neq 0$ give $x \neq -1$. So $x \neq -1$ are ordinary points.

(b) We rewrite the ODE in normal form:

$$y'' - \frac{x^2}{x+1} y' + \frac{3}{x+1} y = 0, \quad p(x) = -\frac{x^2}{x+1}, \quad q(x) = \frac{3}{x+1}.$$

Since $p(x)$ and $q(x)$ are not defined for $x = -1$, $x = -1$ is the only singular point.

Since $(x+1)p(x) = -x^2$ is analytic at $x = -1$ and $(x+1)^2q(x) = 3(x+1)$ is analytic at $x = -1$, $x = -1$ is a regular singular point.

Example 31.11. Determine the singular points of the ODE $xy'' + x(1-x)^{-1}y' + (\sin x)y = 0$. Are the singular points regular or irregular? Justify your answer.

Solution. Writing the ODE in normal form:

$$y'' + \underbrace{\frac{1}{1-x}}_{p(x)} y' + \underbrace{\frac{\sin x}{x}}_{q(x)} y = 0.$$

- When $x = 1$, $p(x)$ is not analytic, hence $x = 1$ is a singular point. Since $(x-1)p(x) = -1$ and $(x-1)^2p(x)$ are analytic, hence $x = 1$ is a regular singular point.
- When $x = 0$, $q(x)$ is a “removable” singular point since $\lim_{x \rightarrow 0} \frac{\sin x}{x} = 1$ (refer to Calculus I). Therefore $x = 0$ is considered to be an ordinary point and all $x \neq 1$ are ordinary points.

□

Example 31.12. For the following ODEs, classify points as ordinary or singular points. Classify singular points further as regular or irregular singular points.

1. $x^2y'' + xy' + y = 0$ (an Euler equation)

Solution.

$x \neq 0$ are ordinary points because $p(x) = \frac{1}{x}$ and $q(x) = \frac{1}{x^2}$ are well-defined.

$x = 0$ is a regular singular point because $xp(x) = 1$ and $x^2q(x) = 1$ are analytic at $x = 0$. □

2. [Boyce, Ex5.4, P239,Q9] $(x+2)^2(x-1)y'' + 3(x-1)y' - 2(x+2)y = 0$

Solution.

$x \neq 1, -2$ are ordinary points because $p(x) = \frac{3}{(x+2)^2}$ and $q(x) = \frac{-2}{(x+2)(x-1)}$ are well-defined.

$x = 1$ is a regular singular point because $(x-1)p(x)$ and $(x-1)^2q(x)$ are analytic at $x = 1$.

$x = -2$ is an irregular singular point because $\lim_{x \rightarrow -2} (x+2)p(x) = \lim_{x \rightarrow -2} \frac{3}{x+2}$ does not exist, i.e. $(x+2)p(x)$ is not analytic at $x = -2$. □

Finally, let us investigate the ODEs which are closely related to physics.

Example 31.13. Find the ordinary and singular points of (a) Hypergeometric differential equation (31.2); (b) Bessel's equation of order ν (36.23); (c) Legendre's equation of order n (37.20); (d) Hermite's equation of order ν (42.11).

Solution:

(a) This ODE has regular singular points at 0, 1 and ∞ . A solution is the hypergeometric function.

(b) It is found in the solution to Laplace equation in cylindrical coordinates. $x = 0$ is a singular point, all the other points are ordinary points. $x = 0$ is a regular singular point because $xp(x) = 1$ is analytic at $x = 0$ and when the order $\nu \neq 0$, $x^2q(x) = x^2 - \nu^2$ is analytic at $x = 0$.

To see what happens when $x \rightarrow \infty$ one has to use a Möbius transformation, for example $x = 1/w$. After performing the algebra, we have

$$\frac{d^2f}{dw^2} + \frac{1}{w} \frac{df}{dw} + \left[\frac{1}{w^4} - \frac{\nu^2}{w^2} \right] f = 0.$$

Now $p(w) = 1/w$ has a pole of first order at $w = 0$. And $q(w)$ has a pole of fourth order at $w = 0$. Thus this equation has an irregular singularity $w = 0$ corresponding to x at ∞ . There is a basis for solutions of this differential equation that are Bessel functions.

(c) It is found in the solution of Laplace equation in spherical coordinates. $x = \pm 1$ and ∞ are regular singular points and all the other points are ordinary points.

(d) All points in $\mathbb{R}$ are ordinary points. This ODE has an irregular singularity at ∞ . Its solutions are Hermite polynomials.

One encounters this ordinary second order ODE in solving the one-dimensional time independent Schrödinger equation

$$E\psi = -\frac{\hbar^2}{2m} \frac{d^2\psi}{dx^2} + V(x)\psi$$

for a harmonic oscillator with the potential energy $V(x) = \frac{1}{2}m\omega^2x^2$.

31.3 Series Solutions Near an Ordinary Point

In this section, we will study the possibility of approximating $y(x)$ at an ordinary point x_0 (normally, we assume $x_0 = 0$) by the power series $y = \sum_{n=0}^{\infty} a_n(x - x_0)^n$. A power series approximation always exists at an ordinary point as stated by the following theorem.

Theorem 31.14 (Series Solution and Radii of Convergence). *If x_0 is an ordinary point of the homogeneous equation (31.1) or (31.13), then the general solution at $x = x_0$ can be written as*

$$y(x) = \sum_{n=0}^{\infty} a_n(x - x_0)^n = a_0y_1(x) + a_1y_2(x)$$

where a_0 and a_1 are arbitrary constants, and y_1 and y_2 are linearly independent series solutions that are analytic at x_0 . Furthermore, the radius of convergence for each of the series solutions y_1 and y_2 is at least as large as the minimum of the radii of convergence of the series of $p(x)$ and $q(x)$.

A series solution normally converges at least for $|x - x_0| < R$, where R is the distance from x_0 to the closest singular point.

Example 31.15 (Final Exam Sept 2014, Q1(d)). Find the ordinary points of the ODE $x^4(x - 3)y'' - (x - 2)y' + 3xy = 0$ and write down the form of the series solution that the ODE should have at an ordinary point. Do not actually find the solution. (5 marks)

Solution. For a point $x_0 \in \mathbb{R}$ to be ordinary, $x_0^4(x_0 - 3) \neq 0$ or $x_0 \neq 0, 3$ [3 marks]
At an ordinary point x_0 , the series solution should have the following form:

$$y(x) = \sum_{n=0}^{\infty} a_n(x - x_0)^n. \quad [2 \text{ marks}]$$

□

Let us look at how power series method works with two first order ODE examples below.

Example 31.16 (Homogeneous First Order ODE). Find a solution of $y' - y = 0$ in the form of a power series in x .

Solution: Note that $x_0 = 0$. So let $y = \sum_{n=0}^{\infty} a_nx^n$. By substituting into the ODE, we obtain

$$y' - y = \sum_{n=0}^{\infty} na_nx^{n-1} - \sum_{n=0}^{\infty} a_nx^n = \sum_{n=0}^{\infty} (n+1)a_{n+1}x^n - \sum_{n=0}^{\infty} a_nx^n = 0.$$

By equating the coefficients to zero, we obtain a recurrence relation

$$(n+1)a_{n+1} - a_n = 0 \Rightarrow a_{n+1} = \frac{a_n}{n+1} = \frac{a_{n-1}}{(n+1)n} = \cdots = \frac{a_0}{(n+1)!}.$$

A solution is $y = a_0 \sum_{n=0}^{\infty} \frac{x^n}{n!}$ or $y = a_0e^x$ where a_0 is an arbitrary constant.

Example 31.17 (Non-Homogeneous First Order ODE). Find the first four terms of the Maclaurin series solution of the IVP: $y' + e^xy = x^2$, $y(0) = 4$.

Solution: Expanding y, e^x at $x = 0$, we have

$$\sum_{n=0}^{\infty} n a_n x^{n-1} + \left(\sum_{n=0}^{\infty} \frac{x^n}{n!} \right) \left(\sum_{n=0}^{\infty} a_n x^n \right) = \sum_{n=0}^{\infty} (n+1) a_{n+1} x^n + \sum_{n=0}^{\infty} \left(\sum_{i=0}^n \frac{a_i}{(n-i)!} \right) x^n = x^2.$$

This gives the recurrence relation:

$$(n+1)a_{n+1} + \sum_{i=0}^n \frac{a_i}{(n-i)!} = \delta_{2n} = \begin{cases} 1, & n = 2, \\ 0, & n \neq 2. \end{cases}$$

This leads to

$$n = 0: a_1 + a_0 = 0$$

$$n = 1: 2a_2 + a_1 + a_0 = 0$$

$$n = 2: 3a_3 + a_2 + a_1 + \frac{1}{2}a_0 = 1$$

and so on. They can be reduced to give us a general solution

$$y = a_0 - a_0 x + 0x^2 + \left(\frac{1}{3} + \frac{1}{6}a_0 \right) x^3 + \dots$$

When $x = 0, y(0) = 4$. This gives $a_0 = 4$. Hence the solution in power series form is

$$y = 4 - 4x + x^3 + \frac{1}{12}x^4 + \dots$$

Applying power series method to solve first order ODE are cumbersome as illustrate above. We will normally apply the technique from Topic 1 to solve first ODE problem. Now, let us look at a very simple second order ODE example from Topic 2.

Example 31.18. Find the series solution for the ODE $y'' + y = 0$ at $x = 0$.

Solution: Expanding y in power series $\sum_{n=0}^{\infty} a_n x^n$ leads to

$$y'' + y = \sum_{n=0}^{\infty} (n+2)(n+1)a_{n+2}x^n + \sum_{n=0}^{\infty} a_n x^n = 0$$

and equating the coefficients to zero gives a recurrence relation

$$(n+2)(n+1)a_{n+2} + a_n = 0 \Rightarrow a_{n+2} = \frac{-1}{(n+2)(n+1)} a_n.$$

Hence, by (31.10) and (31.9), we have

$$\begin{aligned} y &= a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + a_5 x^5 + \dots \\ &= a_0 + a_1 x + \frac{-1}{2!} a_0 x^2 + \frac{-1}{3!} a_1 x^3 + \frac{a_0}{4!} x^4 + \frac{a_1}{5!} x^5 + \dots \\ &= a_0 \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \dots \right) + a_1 \left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots \right) \\ &= a_0 \cos x + a_1 \sin x \end{aligned}$$

as expected!

Example 31.19. Find the series solution for the ODE $y'' - y = 0$ at $x = 0$.

Solution. Expanding y in power series leads to

$$y'' - y = \sum_{n=0}^{\infty} (n+2)(n+1)a_{n+2}x^n - \sum_{n=0}^{\infty} a_n x^n = 0$$

and equating the coefficients to zero gives a recurrence relation

$$(n+2)(n+1)a_{n+2} - a_n = 0 \Rightarrow a_{n+2} = \frac{1}{(n+2)(n+1)} a_n.$$

When n is even, $a_{2n} = \frac{1}{(2n)!}a_0$ and when n is odd, $a_{2n+1} = \frac{1}{(2n+1)!}a_1$. Hence

$$y = \sum_{n=0}^{\infty} a_{2n}x^{2n} + \sum_{n=0}^{\infty} a_{2n+1}x^{2n+1} = \sum_{n=0}^{\infty} \frac{a_0}{(2n)!}x^{2n} + \sum_{n=0}^{\infty} \frac{a_1}{(2n+1)!}x^{2n+1} = a_0 \cosh x + a_1 \sinh x$$

according to (31.12) and (31.11) on page 616. $\square$

Mathematicians were able to link the fundamental set of solutions of $y'' + y = 0$ and $y'' - y = 0$, i.e. sine, cosine functions and hyperbolic sine, cosine functions, through complex analysis:

$$\begin{aligned}\cos x &= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!}x^{2n} = \sum_{n=0}^{\infty} \frac{1}{(2n)!}(ix)^{2n} = \cosh ix \\ \sin x &= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!}x^{2n+1} = \sum_{n=0}^{\infty} \frac{i^{2n}}{(2n+1)!}x^{2n+1} = -i \sum_{n=0}^{\infty} \frac{1}{(2n+1)!}(ix)^{2n+1} = -i \sinh ix.\end{aligned}$$

We mentioned in Topic 2 that the method of separation of variables can reduce a particular PDE to ODEs

$$y'' + \lambda y = 0. \quad (31.14)$$

With appropriate change of variable (which we mentioned in Topic 1, this technique also applies to Topic 2 as well as in this topic), it is possible to transform (31.14) to $y'' + y = 0$ when $\lambda > 0$ and to $y'' - y = 0$ when $\lambda < 0$. We will see this idea again when we try to solve Bessel equation.

Now, let us proceed to analyse a simple linear ODE where the coefficient is no longer constant.

Example 31.20 (Airy Equation, http://en.wikipedia.org/wiki/Airy_function). Find the series solution for the ODE $y'' - xy = 0$ about (a) $x = 0$ and (b) $x = 1$.

Solution. (a) Expanding y in power series at $x = 0$ leads to

$$\sum_{n=0}^{\infty} (n+2)(n+1)a_{n+2}x^n - \sum_{n=0}^{\infty} a_{n-1}x^n = 0$$

and the difference equation and recurrence relation is

$$(n+2)(n+1)a_{n+2} - a_{n-1} = 0 \Rightarrow a_{n+2} = \frac{a_{n-1}}{(n+2)(n+1)}, \quad n \geq 0.$$

This leads to the solution

$$\begin{aligned}y(x) &= a_0 + a_1x + 0 + \frac{1}{6}a_0x^3 + \frac{1}{12}a_1x^4 + 0 + \frac{a_0}{180}x^6 + \frac{a_1}{504}x^7 + \cdots \\ &= a_0\left(1 + \frac{1}{6}x^3 + \frac{1}{180}x^6 + \cdots\right) + a_1\left(x + \frac{1}{12}x^4 + \frac{1}{504}x^7 + \cdots\right) \\ &= a_0 \left[1 + \sum_{n=1}^{\infty} \frac{x^{3n}}{(3n)(3n-1)(3n-3)(3n-4) \cdots 6 \cdot 5 \cdot 3 \cdot 2} \right] \\ &\quad + a_1 \left[x + \sum_{n=1}^{\infty} \frac{x^{3n+1}}{(3n+1)(3n)(3n-2)(3n-3) \cdots 7 \cdot 6 \cdot 4 \cdot 3} \right].\end{aligned}$$

(b) Expanding y at $x = 1$, $y(x) = \sum_{n=0}^{\infty} a_n(x-1)^n$, let $u = x - 1$, the ODE can be transformed to

$$\frac{d^2y}{du^2} - (u+1)y = 0$$

expanded to

$$\sum_{n=0}^{\infty} (n+2)(n+1)a_{n+2}u^n - \sum_{n=0}^{\infty} a_nu^{n+1} - \sum_{n=0}^{\infty} a_nu^n = 0$$

which simplifies to by performing index shifting

$$\sum_{n=0}^{\infty} (n+2)(n+1)a_{n+2}u^n - \sum_{n=0}^{\infty} a_{n-1}u^n - \sum_{n=0}^{\infty} a_nu^n = 0$$

and the difference equation and recurrence relation is

$$(n+2)(n+1)a_{n+2} - a_{n-1} - a_n \Rightarrow a_{n+2} = \frac{a_n + a_{n-1}}{(n+2)(n+1)}.$$

This leads to the solution

$$\begin{aligned} y(x) &= a_0 + a_1u + \frac{1}{2}a_0u^2 + \left(\frac{1}{6}a_0 + \frac{1}{6}a_1\right)u^3 + \left(\frac{1}{12}a_1 + \frac{1}{24}a_0\right)u^4 + \cdots \\ &= a_0\left(1 + \frac{1}{2}u^2 + \frac{1}{6}u^3 + \frac{1}{24}u^4 + \cdots\right) + a_1\left(u + \frac{1}{6}u^3 + \frac{1}{12}u^4 + \cdots\right) \\ &= a_0\left(1 + \frac{1}{2}(x-1)^2 + \frac{1}{6}(x-1)^3 + \frac{1}{24}(x-1)^4 + \cdots\right) \\ &\quad + a_1\left(x-1 + \frac{1}{6}(x-1)^3 + \frac{1}{12}(x-1)^4 + \cdots\right). \end{aligned}$$

□

We will now analyse the power series at the ordinary point (note that by default, we assume the ordinary point $x_0 = 0$) of an important ODE in physics, the Legendre's equation (37.20).

Example 31.21 (Legendre's Equation, Al-Gwaiz [2008], p139). Use the power series method to solve the Legendre's equation (37.20). Explain what happens when $\lambda = m(m+1)$.

Solution. As $x = 0$ is an ordinary point of this ODE, Theorem 31.14 says that we can find two

linearly independent series solutions of the form $y = \sum_{n=0}^{\infty} a_n x^n$.

$$\text{Formally, } \begin{cases} y' = \sum_{n=0}^{\infty} n a_n x^{n-1}, \\ xy' = \sum_{n=0}^{\infty} n a_n x^n, \\ y'' = \sum_{n=0}^{\infty} n(n-1) a_n x^{n-2} = \sum_{n=0}^{\infty} (n+2)(n+1) a_{n+2} x^n \\ x^2 y'' = \sum_{n=0}^{\infty} n(n-1) a_n x^n. \end{cases}$$

Substitute them into the ODE, we get

$$\sum_{n=0}^{\infty} \left\{ (n+1)(n+2)a_{n+2} + [-n(n-1) - 2n + \lambda]a_n \right\} x^n = 0.$$

By equating the coefficients to zero, we obtain a recursion formula

$$a_{n+2} = \frac{n^2 + n - \lambda}{(n+1)(n+2)} a_n, \quad n \geq 0.$$

The general solution is

$$\begin{aligned} y &= a_0 \underbrace{\left[1 - \frac{\lambda}{2!}x^2 - \frac{\lambda(6-\lambda)}{4!}x^4 - \frac{\lambda(6-\lambda)(20-\lambda)}{6!}x^6 - \cdots \right]}_{y_1} \\ &\quad + a_1 \underbrace{\left[x + \frac{2-\lambda}{3!}x^3 + \frac{(2-\lambda)(12-\lambda)}{5!}x^5 + \frac{(2-\lambda)(12-\lambda)(30-\lambda)}{7!}x^7 + \cdots \right]}_{y_2}. \end{aligned}$$

If $\lambda = m(m+1)$, then the recursion formula becomes

$$a_{n+2} = \frac{n^2 + n - m^2 - m}{(n+1)(n+2)} a_n = \frac{(n-m)(n+m+1)}{(n+1)(n+2)} a_n, \quad n \geq 0.$$

When $n = m$, $a_{m+2} = 0$ and $a_{m+2k} = \cdots a_m = 0$ for $k \geq 0$. This means that one of the two solutions is a polynomial.

The general solution is

$$y = a_0 \left[1 - \frac{m(m+1)}{2!}x^2 + \frac{(m-2)m(m+1)(m+3)}{4!}x^4 + \cdots \right] + a_1 \left[x - \frac{(m-1)(m+2)}{3!}x^3 + \frac{(m-3)(m-1)(m+2)(m+4)}{5!}x^5 + \cdots \right].$$

□

From the solution, we learned that if $\lambda = m(m+1)$ for some nonnegative integer m , one of the linearly independent solutions is a polynomial of degree m . These polynomials, multiplied by some suitable constant, are called *Legendre polynomials*.

Definition 31.10. We define *Legendre polynomial* $P_m(x)$ as the polynomial solution of the Legendre equation (37.20) with $P_m(1) = 1$.

Remark 31.22. The only solutions of Legendre equation (37.20) that are bounded at $x = \pm 1$ are the Legendre polynomials.

When $\lambda = m(m+1)$ with $m = 0, 1, 2, 3, 4, 5$, we obtain the first few Legendre polynomials

1. When $\lambda = 0$, we have $y_1 = P_0(x) = 1$.
2. When $\lambda = 2$, we have $y_2 = P_1(x) = x$.
3. When $\lambda = 6$, we have $y_1 = P_2(x) = -\frac{1}{2}(1 - 3x^2)$.
4. When $\lambda = 12$, we have $y_2 = P_3(x) = -\frac{3}{2}(x - \frac{5}{3}x^3)$.
5. When $\lambda = 20$, we have $y_1 = P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3)$.
6. When $\lambda = 30$, we have $y_2 = P_5(x) = \frac{1}{8}(63x^5 - 70x^3 + 15x)$.

Example 31.23 (Final Exam Sept 2014, Q5(a)). Find the series solution of the ODE

$$(x^2 + 1)y'' + xy' - xy = 0$$

near $x = 0$. Write down the coefficients of the series solution up to and including that of the degree 5 term and organise the series solution into the form $C_1y_1 + C_2y_2$ where $\{y_1, y_2\}$ is a fundamental set of solutions. (14 marks)

Solution: Since $x = 0$ is an ordinary point. The series solution has the form

$$y = \sum_{n=0}^{\infty} a_n x^n. \quad [1 \text{ mark}]$$

Substituting into the ODE, we obtain

$$(x^2 + 1) \sum_{n=0}^{\infty} n(n-1)a_n x^{n-2} + x \sum_{n=0}^{\infty} na_n x^{n-1} - x \sum_{n=0}^{\infty} a_n x^n = 0. \quad [2 \text{ marks}]$$

Arranging it gives

$$2a_2 + 6a_3x + \sum_{n=2}^{\infty} (n+2)(n+1)a_{n+2}x^n + \sum_{n=2}^{\infty} n(n-1)a_n x^n + a_1x + \sum_{n=2}^{\infty} na_n x^n - a_0x - \sum_{n=2}^{\infty} a_{n-1}x^n = 0 \quad [3 \text{ marks}]$$

which simplifies to

$$2a_2 + (6a_3 + a_1 - a_0)x + \sum_{n=2}^{\infty} \left\{ (n+2)(n+1)a_{n+2} + n(n-1)a_n + na_n - a_{n-1} \right\} x^n = 0. \quad [1 \text{ mark}]$$

By equating the coefficients to zero, we obtain

$$\begin{aligned} 2a_2 &= 0, \\ 6a_3 + a_1 - a_0 &= 0, \\ (n+2)(n+1)a_{n+2} + n^2a_n - a_{n-1} &= 0, \quad n = 2, 3, \dots \end{aligned} \quad [2 \text{ marks}]$$

Hence,

$$a_{n+2} = \frac{a_{n-1} - n^2a_n}{(n+2)(n+1)}. \quad [1 \text{ mark}]$$

In particular, when $n = 0, 1, 2$, $a_2 = 0$, $a_3 = \frac{a_0 - a_1}{6}$, $a_4 = \frac{a_1 - 4a_2}{12} = \frac{a_1}{12}$ [1 mark]

When $n = 3$, $a_5 = \frac{a_2 - 9a_3}{20} = \frac{-9}{20} \cdot \frac{a_0 - a_1}{6} = \frac{-3}{40}(a_0 - a_1)$ [1 mark]

Hence the solution is

$$\begin{aligned} y &= a_0 + a_1x + \frac{a_0 - a_1}{6}x^3 + \frac{a_1}{12}x^4 - \frac{3}{40}(a_0 - a_1)x^5 + \dots \\ &= a_0 \underbrace{\left(1 + \frac{1}{6}x^3 - \frac{3}{40}x^5 + \dots\right)}_{y_1} + a_1 \underbrace{\left(x - \frac{1}{6}x^3 + \frac{1}{12}x^4 + \frac{3}{40}x^5 + \dots\right)}_{y_2}. \end{aligned} \quad [2 \text{ marks}]$$

Example 31.24 (Final Exam Sept 2016, Q3(a)). Show that the equation $(1 - x^2)y'' - 2xy' + \frac{3}{4}y = 0$ has the **fundamental system of solutions**

$$\left\{ 1 - \frac{3}{8}x^2 - \frac{21}{128}x^4 - \dots, \quad x + \frac{5}{24}x^3 + \frac{15}{128}x^5 + \dots \right\}$$

by using power series method.

(11 marks)

Solution. Substitute $y = \sum_{n=0}^{\infty} a_n x^n$ into the equation:

$$(1 - x^2) \sum_{n=0}^{\infty} a_n n(n-1)x^{n-2} - 2x \sum_{n=0}^{\infty} a_n n x^{n-1} + \frac{3}{4} \sum_{n=0}^{\infty} a_n x^n \quad [2 \text{ marks}]$$

By expanding the terms and adjusting index, we have

$$\sum_{n=0}^{\infty} a_{n+2}(n+1)(n+2)x^n + \sum_{n=0}^{\infty} \left[-a_n n(n-1) - 2a_n n + \frac{3}{4}a_n \right] x^n = 0 \quad [2 \text{ marks}]$$

From here, we obtain the recurrence relation

$$(n+1)(n+2)a_{n+2} + \left[-n(n-1) - 2n + \frac{3}{4} \right] a_n = 0 \Rightarrow a_{n+2} = \frac{n^2 + n - \frac{3}{4}}{(n+1)(n+2)} a_n$$

for $n = 0, 1, 2, \dots$ [2 marks]

For $n=2, 3, 4, 5$, we have

$$\begin{aligned} a_2 &= \frac{-\frac{3}{4}}{2} a_0 = -\frac{3}{8} a_0, & a_3 &= \frac{1 + 1 - \frac{3}{4}}{2 \cdot 3} a_1 = \frac{5}{24} a_1, \\ a_4 &= \frac{4 + 2 - \frac{3}{4}}{3 \cdot 4} \left(-\frac{3}{8} a_0\right) = -\frac{21}{128} a_0, & a_5 &= \frac{9 + 3 - \frac{3}{4}}{4 \cdot 5} \left(\frac{5}{24} a_1\right) = \frac{15}{128} a_1 \end{aligned} \quad [4 \text{ marks}]$$

The general solution is

$$y = a_0 + a_1x - \frac{3}{8}a_0x^2 + \frac{5}{24}a_1x^3 - \frac{21}{128}a_0x^4 + \frac{15}{128}a_1x^5 + \dots$$

which implies the two fundamental system of solutions above. [1 mark] □

Finally, we look at a nonhomogeneous problem of a second order ODE.

Example 31.25. Find the series solution for the ODE $y'' + x^2y' + 4y = 1 - x^2$.

Solution: Expanding at $x = 0$, we simplify to

$$\underbrace{\sum_{n=0}^{\infty} (n+2)(n+1)a_{n+2}x^n}_{y''} + \underbrace{\sum_{n=0}^{\infty} (n-1)a_{n-1}x^n}_{x^2y', a_{-1}=0} + \sum_{n=0}^{\infty} 4a_nx^n = 1 - x^2.$$

Comparing coefficients leads to

$$(n+2)(n+1)a_{n+2} + (n-1)a_{n-1} + 4a_n = \delta_{n,0} - \delta_{n,2}, \quad n \geq 0.$$

Solving for a_n leads to the solution

$$\begin{aligned} y &= a_0 + a_1x + \left(\frac{1}{2} - 2a_0\right)x^2 + \left(-\frac{2}{3}a_1\right)x^3 + \left(-\frac{3}{4} - 2a_0 - \frac{1}{4}a_1\right)x^4 + \dots \\ &= a_0 \underbrace{\left(1 - 2x^2 - 2x^4 - \dots\right)}_{y_1} + a_1 \underbrace{\left(x - \frac{2}{3}x^3 - \frac{1}{4}x^4 - \dots\right)}_{y_2} + \underbrace{\left(\frac{1}{2}x^2 - \frac{3}{4}x^4 - \dots\right)}_{y_p}. \end{aligned}$$

31.4 Series Solutions Near a Regular Singular Point

The solution near a regular singular point *cannot be characterised by a power series* (31.7). However, the solution at a regular singular point has a pattern (31.15) where the *method of Frobenius* can be applied.

Theorem 31.26 (Frobenius Theorem). *If x_0 is a regular singular point of (31.1), then there is at least one solution of the form of **Frobenius series**:*

$$y = (x - x_0)^r \sum_{n=0}^{\infty} a_n (x - x_0)^n = \sum_{n=0}^{\infty} a_n (x - x_0)^{n+r}, \quad a_n \neq 0 \quad (31.15)$$

where the number r is a constant that must be determined.

Suppose $x_0 = 0$ is a **regular singular point**. Let r_1 and r_2 be roots of the **indicial equation** and $r_1 > r_2$. Then

1. When $r_1 - r_2$ is not an integer, i.e. $r_1 - r_2 \notin \mathbb{Z}$,

$$y_1(x) = \sum_{n=0}^{\infty} a_n x^{n+r_1} \quad \text{and} \quad y_2(x) = \sum_{n=0}^{\infty} b_n x^{n+r_2}, \quad a_n, b_n \neq 0.$$

2. When $r_1 - r_2 = 0$,

$$y_1(x) = \sum_{n=0}^{\infty} a_n x^{n+r_1} \quad \text{and} \quad y_2(x) = y_1(x) \ln(x) + \sum_{n=1}^{\infty} b_n x^{n+r_1}.$$

3. When $r_1 - r_2$ is a non-zero integer, i.e. $r_1 - r_2 \in \mathbb{Z} \setminus \{0\}$,

$$y_1(x) = \sum_{n=0}^{\infty} a_n x^{n+r_1} \quad \text{and} \quad y_2(x) = k y_1(x) \ln(x) + \sum_{n=0}^{\infty} b_n x^{n+r_2}.$$

The mathematical theory behind Frobenius Theorem is again the complex analytic property of singular points! This was pointed out in Smirnov [1964b, item 97]. It is also possible for mathematicians to analyse the convergence of series solution (31.7) from a complex analytic point of view as illustrated by the following example [Tenenbaum and Pollard, 1985, 1963, Example 40.392].

Example 31.27. Find the interval of convergence of the series solution of

$$x^2 y'' + \frac{x}{1+x^2} y' - 3y = 0.$$

Solution: By definition, $p(x) = \frac{1}{x(1+x^2)}$ is singular at $x = 0$. So $x = 0$ is singular. It is regular because both $xp(x) = \frac{1}{1+x^2}$ and $x^2q(x) = -3$ are analytic at $x = 0$.

The other two regular singular points are $x = i$ and $x = -i$. They form the obstacles for the radius of convergence for the regular singular point $x = 0$ on the complex plane.

The interval of convergence is $-1 < x < 1$.

Example 31.28. Find the regular singular points of

$$(x+2)x^2y'' - xy' + (1+x)y = 0$$

and write down the series solution at $x = 0$.

Solution: Rewriting the equation as

$$y'' - \frac{1}{x(x+2)}y' + \frac{1+x}{x^2(x+2)}y = 0$$

we know that $x = 0$ and -2 are singular points, furthermore

$$xp(x) = x\left(\frac{-1}{x(x+2)}\right) = -\frac{1}{x+2} \text{ is analytic at } x = 0,$$

$$(x+2)p(x) = \frac{-1}{x} \text{ is analytic at } x = -2$$

and

$$x^2q(x) = x^2\left(\frac{1+x}{x^2(x+2)}\right) = \frac{1+x}{x+2} \text{ is analytic at } x = 0,$$

$$(x+2)q(x) = \frac{1+x}{x^2} \text{ is analytic at } x = -2.$$

The points $x = 0$ and -2 are regular singular points.

Therefore, we can apply Frobenius method to study the series at $x = 0$.

Let $y = \sum_{n=0}^{\infty} a_n x^{n+r}$. Then $y' = \sum_{n=0}^{\infty} (n+r)a_n x^{n+r-1}$, $y'' = \sum_{n=0}^{\infty} (n+r)(n+r-1)a_n x^{n+r-2}$.

Substitute into the original ODE, we have

$$(x+2) \sum_{n=0}^{\infty} (n+r)(n+r-1)a_n x^{n+r} - \sum_{n=0}^{\infty} (n+r)a_n x^{n+r} + (1+x) \sum_{n=0}^{\infty} a_n x^{n+r} = 0.$$

This can be simplified as follows

$$\begin{aligned} & \sum_{n=0}^{\infty} (n+r)(n+r-1)a_n x^{n+r+1} + \sum_{n=0}^{\infty} 2(n+r)(n+r-1)a_n x^{n+r} + \sum_{n=0}^{\infty} [-(n+r)a_n]x^{n+r} \\ & + \sum_{n=0}^{\infty} a_n x^{n+r} + \sum_{n=0}^{\infty} a_n x^{n+r+1} = 0 \\ \Rightarrow & \sum_{n=0}^{\infty} [(n+r)(n+r-1) + 1]a_n x^{n+r+1} + \sum_{n=0}^{\infty} [2(n+r)(n+r-1) - (n+r) + 1]a_n x^{n+r} = 0 \\ \Rightarrow & \sum_{n=0}^{\infty} [(n-1+r)(n-1+r-1) + 1]a_{n-1} x^{n-1+r+1} \quad (\text{let } a_{-1} = 0, \text{ performing index shifting}) \\ & + \sum_{n=0}^{\infty} [2(n+r)(n+r-1) - (n+r) + 1]a_n x^{n+r} = 0 \\ \Rightarrow & \sum_{n=0}^{\infty} \left\{ [(n+r-1)(n+r-2) + 1]a_{n-1} + [2(n+r)(n+r-1) - (n+r) + 1]a_n \right\} x^{n+r} = 0. \end{aligned}$$

Equating the coefficients to zero, we have

$$[(n+r-1)(n+r-2) + 1]a_{n-1} + [2(n+r)(n+r-1) - (n+r) + 1]a_n = 0, \quad n \geq 0.$$

Note that a_0 is arbitrary and when $n = 0$, we obtain the **indicial equation** and their roots:

$$0 + [2(0+r)(0+r-1) - (0+r) + 1]a_0 = 0 \Rightarrow 2r^2 - 3r + 1 = 0 \Rightarrow r_1 = \frac{1}{2}, r_2 = 1.$$

This is case 1, i.e. two different roots with $r_1 - r_2$ not an integer. The general solution is

$$y(x) = c_1 \sum_{n=0}^{\infty} a_n^{(1)} x^{n+\frac{1}{2}} + c_2 \sum_{n=0}^{\infty} a_n^{(2)} x^{n+1} = c_1 y_1(x) + c_2 y_2(x)$$

where

$$y_1(x) = \sqrt{x} \sum_{n=0}^{\infty} a_n^{(1)} x^n = \sqrt{x} \left(1 - \frac{3}{4}x + \frac{7}{32}x^2 - \frac{133}{1920}x^3 + \dots \right),$$

$$y_2(x) = x \sum_{n=0}^{\infty} a_n^{(2)} x^n = x - \frac{1}{3}x^2 + \frac{1}{10}x^3 - \frac{1}{30}x^4 - \dots,$$

$$\text{and } \begin{cases} a_0^{(1)} = 1, \\ a_n^{(1)} = -\frac{4n^2 - 8n + 7}{8n^2 - 4n} a_{n-1}, \quad n = 1, 2, \dots \end{cases} \quad \begin{cases} a_0^{(2)} = 1, \\ a_n^{(2)} = -\frac{n^2 - n + 1}{2n^2 + n} a_{n-1}, \quad n = 1, 2, \dots \end{cases}$$

Example 31.29. Use the method of Frobenius to find a solution of the ODE $2xy'' + y' - y = 0$.

Solution. Let $y = \sum_{n=0}^{\infty} a_n x^{n+r}$ and by substitution, we obtain

$$\begin{aligned} & \sum_{n=0}^{\infty} 2(n+r)(n+r-1)a_n x^{n+r-1} + \sum_{n=0}^{\infty} (n+r)a_n x^{n+r-1} - \sum_{n=0}^{\infty} a_n x^{n+r} = 0 \\ \Rightarrow & \sum_{n=0}^{\infty} \left[2(n+r)(n+r-1)a_n + (n+r)a_n - a_{n-1} \right] x^{n+r-1} = 0. \end{aligned}$$

Equating the coefficients to zero leads to a recurrence relation

$$(n+r)(2n+2r-2+1)a_n - a_{n-1} = 0, \quad n \geq 0 \quad (*)$$

and when $n = 0$, we obtain the indicial equation

$$a_{-1} = 0, \quad r(2r-1)a_0 - a_{-1} = 0 \Rightarrow r = 0, \quad \frac{1}{2}.$$

The difference of roots is not an integer (Case 1), therefore, no logarithmic function appears. When $r = 0$, (*) becomes

$$a_n = \frac{a_{n-1}}{n(2n-1)} = \dots = \frac{a_0}{n(n-1) \dots 1 \cdot (2n-1)(2n-3) \dots 1} = \frac{2^n a_0}{(2n)!} \Rightarrow y_1 = \sum_{n=0}^{\infty} \frac{2^n}{(2n)!} x^n$$

When $r = \frac{1}{2}$, (*) becomes

$$\begin{aligned} a_n &= \frac{a_{n-1}}{(n+\frac{1}{2})(2n)} = \frac{a_0}{(2n+1)(2n-1) \dots 3 \cdot n(n-1) \dots 1} = \frac{2^n a_0}{(2n+1)!} \\ \Rightarrow y_2 &= \sum_{n=0}^{\infty} \frac{2^n}{(2n+1)!} x^{n+\frac{1}{2}}. \end{aligned}$$

The functions y_1 and y_2 are two linearly independent solutions of the ODE. Hence, the general solution is

$$y = C_1 \sum_{n=0}^{\infty} \frac{2^n}{(2n)!} x^n + C_2 \sum_{n=0}^{\infty} \frac{2^n}{(2n+1)!} x^{n+1/2}.$$

□

Example 31.30 (Final Exam Sept 2016, Q3(a)). Consider the following differential equation

$$2x^2 y'' - xy' + (1-x^2)y = 0.$$

1. Show that $x = 0$ is a regular singular point. (3 marks)

Solution. Note that $p(x) = -\frac{1}{2x}$ and $q(x) = \frac{1-x^2}{2x^2}$ and $x = 0$ is a singular point of $p(x)$.
[1 mark]

Since both

$$xp(x) = -\frac{1}{2} \quad \text{and} \quad x^2q(x) = \frac{1-x^2}{2}$$

are analytic at $x = 0$, $x = 0$ is a regular singular point. [2 marks] $\square$

2. By applying the method of Frobenius, determine the roots of the indicial equation at $x = 0$.

(6 marks)

Solution. Let $y = \sum_{n=0}^{\infty} a_n x^{n+r}$. Substitute into the ODE and we have

$$2x^2 \sum_{n=0}^{\infty} a_n(n+r)(n+r-1)x^{n+r-2} - x \sum_{n=0}^{\infty} a_n(n+r)x^{n+r-1} + (1-x^2) \sum_{n=0}^{\infty} a_n x^{n+r} = 0. \quad [2 \text{ marks}]$$

Let $a_{-2} = a_{-1} = 0$. By appropriate change of index and simplification, we have

$$\begin{aligned} & \sum_{n=0}^{\infty} \left(2(n+r)(n+r-1) - (n+r) + 1 \right) a_n x^{n+r} - \sum_{n=0}^{\infty} a_n x^{n+r+2} = 0 \\ \Rightarrow & \sum_{n=0}^{\infty} \left[\left(2(n+r)(n+r-1) - (n+r) + 1 \right) a_n - a_{n-2} \right] x^{n+r} = 0. \end{aligned} \quad [2 \text{ marks}]$$

When $n = 0$, we obtain the indicial equation and its roots:

$$2r(r-1) - r + 1 = 0 \Rightarrow \boxed{2r^2 - 3r + 1 = 0} \Rightarrow r = \frac{1}{2}, 1. \quad [2 \text{ marks}]$$

$\square$

eg: case2:r1=r2

Example 31.31. Find the regular singular point of

$$x^2 y'' + 5xy' + (x+4)y = 0$$

and the series solution at the regular singular point.

Solution.

Similar to the previous example, we can verify that $x = 0$ is a regular singularity.

Again let $y = \sum_{n=0}^{\infty} a_n x^{n+r}$, the ODE can be transformed to

$$\begin{aligned} & \sum_{n=0}^{\infty} (n+r)(n+r-1) a_n x^{n+r} + \sum_{n=0}^{\infty} 5(n+r) a_n x^{n+r} + \sum_{n=0}^{\infty} (a_n x^{n+r+1} + 4a_n x^{n+r}) = 0 \\ \Rightarrow & \sum_{n=0}^{\infty} (n+r)(n+r-1) a_n x^{n+r} + \sum_{n=0}^{\infty} 5(n+r) a_n x^{n+r} \\ & + \sum_{n=0}^{\infty} a_{n-1} x^{n-1+r+1} + \sum_{n=0}^{\infty} 4a_n x^{n+r} = 0 \\ \Rightarrow & \sum_{n=0}^{\infty} \left[((n+r)(n+r-1) + 5(n+r) + 4) a_n + a_{n-1} \right] x^{n+r} = 0 \end{aligned}$$

This implies that

$$((n+r)(n+r-1) + 5(n+r) + 4) a_n + a_{n-1} = 0, \quad n \geq 0.$$

When $n = 0$, we have

$$(r(r-1) + 5r + 4) a_0 + a_{-1} = 0, \quad a_{-1} = 0.$$

For non-trivial solution, $a_0 \neq 0$, we have the indicial equation

$$r^2 + 4r + 4 = 0 \Rightarrow r = -2.$$

This is a repeated root case $r_1 - r_2 = 0$ and the general solution should have the form

$$y(x) = C_1 \underbrace{\sum_{n=0}^{\infty} a_n x^{n-2}}_{y_1} + C_2 \left[\underbrace{y_1 \ln x + \sum_{n=0}^{\infty} b_n x^{n-2}}_{y_2} \right]$$

where $\boxed{b_0 = 0}!$

From $r = -2$ and the recurrence relation in explicit form

$$a_n = \frac{-1}{(n-2)(n-3) + 5(n-2) + 4} a_{n-1}, \quad n \geq 1.$$

we obtain

$$y_1(x) = \frac{1}{x^2} \sum_{n=0}^{\infty} a_n x^n = \frac{1}{x^2} \left(1 - x + \frac{1}{4}x^2 - \frac{1}{36}x^3 + \cdots \right).$$

The coefficient b_n can be determined as follows:

$$\begin{aligned} y_2(x) &= y_1(x) \ln(x) + \sum_{n=0}^{\infty} b_n x^{n-2}, \\ y_2'(x) &= y_1'(x) \ln(x) + \frac{y_1(x)}{x} + \sum_{n=0}^{\infty} (n-2)b_n x^{n-3}, \\ y_2''(x) &= y_1''(x) \ln(x) + \frac{2y_1'(x)}{x} - \frac{y_1(x)}{x^2} + \sum_{n=0}^{\infty} (n-2)(n-3)b_n x^{n-4}. \end{aligned}$$

Substitute into the original ODE, we have

$$\begin{aligned} &x^2 \left[y_1'' \ln x + \frac{2}{x} y_1' - \frac{1}{x^2} y_1 + \sum_{n=0}^{\infty} (n-2)(n-3)b_n x^{n-4} \right] \\ &+ 5x \left[y_1' \ln x + \frac{1}{x} y_1 + \sum_{n=0}^{\infty} (n-2)b_n x^{n-3} \right] + (x+4) \left[y_1 \ln x + \sum_{n=0}^{\infty} b_n x^{n-2} \right] = 0 \\ \Rightarrow &\underbrace{[x^2 y_1'' + 5x y_1' + (x+4)y_1]}_{=0} \ln x + 2x y_1' - y_1 + \sum_{n=0}^{\infty} (n-2)(n-3)b_n x^{n-2} \\ &+ 5y_1 + \sum_{n=0}^{\infty} 5(n-2)b_n x^{n-2} + \sum_{n=0}^{\infty} b_n x^{n-1} + \sum_{n=0}^{\infty} 4b_n x^{n-2} = 0 \\ \Rightarrow &2x y_1' + 4y_1 + \sum_{n=0}^{\infty} \left[(n-2)(n-3) + 5(n-2) + 4 \right] b_n x^{n-2} + \sum_{n=0}^{\infty} b_n x^{n-1} = 0 \\ \Rightarrow &2x \sum_{n=0}^{\infty} (n-2)a_n x^{n-3} + 4 \sum_{n=0}^{\infty} a_n x^{n-2} \\ &+ \sum_{n=0}^{\infty} \left[(n-2)(n-3) + 5(n-2) + 4 \right] b_n x^{n-2} + \sum_{n=0}^{\infty} b_{n-1} x^{n-2} = 0 \\ \Rightarrow &\sum_{n=0}^{\infty} \left[(2(n-2)a_n + 4a_n) + ((n-2)(n-3) + 5(n-2) + 4)b_n + b_{n-1} \right] x^{n-2} = 0. \end{aligned}$$

This leads to a recurrence relation for the coefficient b_n :

$$2na_n + n^2 b_n + b_{n-1} = 0.$$

When $n = 1$, $b_0 = 0$, the recurrence relation gives $b_1 = 2$.

When $n = 2$, the recurrence relation gives $b_2 = -\frac{3}{4}$.

When $n = 3$, the recurrence relation gives $b_3 = \frac{11}{108}$.

The first three terms of y_2 can be written as

$$y_2 = y_1 \ln x + \frac{1}{x^2} \left(2x - \frac{3}{4}x^2 + \frac{11}{108}x^3 - \cdots \right).$$

□

Example 31.32. [Harper, 1999, p. 182] Consider the solution of the following ODE at $x = 0$:

$$xy'' + 2y' + xy = 0.$$

By use of Frobenius series method, obtain the (a) indicial equation and its solution(s), (b) recursion formula, and (c) general solution of the ODE.

Solution. The standard form of the ODE is $y'' + \frac{2}{x}y' + y = 0$.

The singularity at $x = 0$ is regular, hence the solution has the form $y(x) = \sum_{n=0}^{\infty} a_n x^{n+r}$. and

$$\underbrace{\sum_{n=0}^{\infty} (n+r)(n+r-1)a_n x^{n+r-1}}_{xy''} + \underbrace{\sum_{n=0}^{\infty} 2(n+r)a_n x^{n+r-1}}_{2y'} + \sum_{n=0}^{\infty} a_n x^{n+r+1} = 0$$

$$\Rightarrow \sum_{n=0}^{\infty} \left[(n+r)(n+r-1)a_n + 2(n+r)a_n + a_{n-2} \right] x^{n+r-1} = 0$$

where $a_{-2} = a_{-1} = 0$.

(a) By equating the coefficients to zero, we obtain

$$(n+r)(\underline{n+r-1}+2)a_n + a_{n-2} = 0, \quad n \geq 0.$$

When $n = 0$, we can obtain and solve the indicial equation

$$r(r+1)a_0 = 0 \Rightarrow \boxed{r^2 + r = 0} \Rightarrow r_1 = 0, \quad r_2 = -1.$$

This is case 3, where $r_1 - r_2 = 1$ is a nonzero integer and the general solution should have the form

$$y(x) = C_1 \underbrace{\sum_{n=0}^{\infty} a_n x^n}_{y_1} + C_2 \left[k y_1 \ln x + \underbrace{\sum_{n=0}^{\infty} b_n x^n}_{y_2} \right]$$

where k is a constant to be determined by substitution.

(b) We can determine y_1 using $r_1 = 0$. However, if we use $r = -1$, we will maintain a_1 as an arbitrary constant and be able to obtain the general solution directly. The recursion formula is

$$a_{n+2} = \frac{-a_n}{(n+1)(n+2)}, \quad 0, 1, 2, \dots$$

or more explicitly, for $j = 1, 2, \dots$,

$$a_2 = -\frac{a_0}{2!}, \quad a_4 = -\frac{a_2}{4 \cdot 3} = \frac{a_0}{4!}, \quad \dots, \quad a_{2j} = \frac{(-1)^j a_0}{(2j)!}$$

$$a_3 = -\frac{a_1}{3!}, \quad a_5 = -\frac{a_3}{5 \cdot 4} = \frac{a_1}{5!}, \quad \dots, \quad a_{2j+1} = \frac{(-1)^j a_1}{(2j+1)!}.$$

$$(c) \text{ The general solution is } y(x) = a_0 \underbrace{\sum_{j=0}^{\infty} \frac{(-1)^j}{(2j)!} x^{2j-1}}_{y_1} + a_1 \underbrace{\sum_{j=0}^{\infty} \frac{(-1)^j}{(2j+1)!} x^{2j}}_{y_2}. \quad \square$$

One very important linear ODE which occurs in the study of linear PDE in cylindrical coordinate is the Bessel equation (36.23).

Example 31.33 (Bessel Equation, Al-Gwaiz [2008]). Find a series solution for the Bessel's equation (36.23) of order $\nu \geq 0$ at the regular singular point $x = 0$.

Solution. By Frobenius Theorem, the ODE should have a series solution of the form

$$y = \sum_{n=0}^{\infty} a_n x^{n+r}.$$

Substitute the following

$$\begin{aligned}
\nu^2 y &= \sum_{n=0}^{\infty} \nu^2 a_n x^{n+r} = \nu^2 a_0 x^r + \nu^2 a_1 x^{r+1} + \sum_{n=2}^{\infty} \nu^2 a_n x^{n+r} \\
x^2 y &= \sum_{n=0}^{\infty} a_n x^{n+r+2} = \sum_{n=2}^{\infty} a_{n-2} x^{n+r} \quad (\text{by replacing } n \text{ with } n-2) \\
xy' &= \sum_{n=0}^{\infty} (n+r) a_n x^{n+r} = r a_0 x^r + (r+1) a_1 x^{r+1} + \sum_{n=2}^{\infty} (n+r) a_n x^{n+r} \\
x^2 y'' &= \sum_{n=0}^{\infty} (n+r)(n+r-1) a_n x^{n+r} \\
&= r(r-1) a_0 x^r + r(r+1) a_1 x^{r+1} + \sum_{n=2}^{\infty} (n+r)(n+r-1) a_n x^{n+r}
\end{aligned}$$

into the ODE, we obtain

$$\begin{aligned}
&[r(r-1) + r - \nu^2] a_0 x^r + [r(r+1) + (r+1) - \nu^2] a_1 x^{r+1} \\
&+ \sum_{n=2}^{\infty} \left[\{(n+r)(n+r-1) + (n+r) - \nu^2\} a_n + a_{n-2} \right] x^{n+r} = 0
\end{aligned}$$

By the identification of the coefficients to zero, we obtain

$$\begin{aligned}
x^r : [r(r-1) + r - \nu^2] a_0 &= 0, \\
x^{1+r} : [(r+1)^2 - \nu^2] a_1 &= 0, \dots \\
x^{n+r} : [(n+r)^2 - \nu^2] a_n + a_{n-2} &= 0, \quad n \geq 2.
\end{aligned}$$

Since a_0 is arbitrary (hence we can pick $a_0 \neq 0$), we obtain the indicial equation is

$$r(r-1) + r - \nu^2 = 0 \Rightarrow r^2 = \nu^2 \Rightarrow r = \pm \nu.$$

$$\text{When } r = \nu, a_n = \frac{a_{n-2}}{(\nu+n)^2 - \nu^2} = \frac{a_{n-2}}{2n\nu + n^2}, \quad n \geq 2.$$

Note that $a_{-1} = 0$ and a_0 is arbitrary. Therefore,

- When n is odd, $a_n = 0$.
- When n is even, we know from UECM1313 Discrete Maths, $n = 2j$ for some integer j and

$$a_2 = -\frac{a_0}{2^2(\nu+1)}, \quad a_4 = \frac{a_0}{2^4 2!(\nu+1)(\nu+2)}, \quad \dots, \quad a_{2j} = \frac{(-1)^j a_0}{2^{2j} j! (\nu+1)(\nu+2) \cdots (\nu+j)}.$$

Hence, a series solution is obtained:

$$y_1(x) = a_0 \sum_{j=0}^{\infty} \frac{(-1)^j}{2^{2j} j! (\nu+1)(\nu+2) \cdots (\nu+j)} x^{\nu+2j}$$

where a_0 is an arbitrary constant. □

To simplify the expression of the series solution of Bessel's equation, mathematicians introduced the *gamma function*:

$$\Gamma(\nu) = \int_0^{\infty} e^{-t} t^{\nu-1} dt, \quad \nu > 0$$

which has these properties: (1) $\Gamma(\nu+1) = \nu\Gamma(\nu)$; (2) $\Gamma(n+1) = n!$ for $n = 0, 1, 2, \dots$.

The *Bessel functions of the first kind of order ν* is defined as

$$J_{\nu}(x) := \sum_{j=0}^{\infty} \frac{(-1)^j}{j! \Gamma(\nu+j+1)} \left(\frac{x}{2}\right)^{\nu+2j} = \frac{y_1(x)}{2^{\nu} \Gamma(\nu+1)}$$

where the a_0 is taken as 1 in $y_1(x)$ in this case.

Finding the general solution for Bessel's equation when ν is an integer can be very time-consuming because it is similar to Example eg: case 3 31.32 (in which we were lucky that by choosing the right values of r_1 and r_2 we get the general solution). For Bessel's equation, it won't work and we have to go through a longer process like Example eg: case 2: r1=r2 31.31.

When ν is not an integer, then $r = -\nu$ will give a solution $J_{-\nu}(x)$ and the general solution is

$$y(x) = C_1 J_{\nu}(x) + C_2 J_{-\nu}(x).$$

The *Bessel function of the second kind of order ν* , defined as

$$Y_\nu(x) := \frac{\cos(\nu\pi)J_\nu(x) - J_{-\nu}(x)}{\sin(\nu\pi)},$$

is a second linearly independent solution of the Bessel's equation (36.23). With this, for any value of ν , the general solution of Bessel's equation on $(0, \infty)$ can be written as

$$y = C_1 J_\nu(x) + C_2 Y_\nu(x). \quad \square$$

Remark 31.34. $J_\nu(x)$ and $J_{-\nu}(x)$ are linearly dependent when ν is a positive integer.

The solutions of many other second order linear ODEs can be expressed in terms of Bessel's functions.

Example 31.35 (SJF Ex 4.5, Q5). Express the general solution of the following ODE

$$y'' + \frac{1}{x}y' + \left(1 - \frac{4}{x^2}\right)y = 0$$

in terms of Bessel's functions of the first kind, $J_\nu(x)$, and the second kind, $Y_\nu(x)$.

Solution: We identify this ODE to be Bessel's equation of order 2 ($\nu^2 = 4$), hence the general solution is

$$y = C_1 J_2(x) + C_2 Y_2(x).$$

Example 31.36. Use the change of variables $t = \lambda x$ to find the general solution of the *parametric Bessel equation of order ν*

$$x^2 y'' + xy' + (\lambda^2 x^2 - \nu^2)y = 0$$

where λ is a positive parameter. Express your answer in terms of the Bessel functions.

Solution: Let $t = \lambda x$ and we have

$$\frac{dy}{dx} = \frac{dy}{dt} \cdot \frac{dt}{dx} = \lambda \frac{dy}{dt}, \quad \frac{d^2y}{dx^2} = \frac{d}{dx} \left(\lambda \frac{dy}{dt} \right) = \lambda \frac{d^2y}{dt^2} \cdot \frac{dt}{dx} = \lambda^2 \frac{d^2y}{dt^2}.$$

Hence, the original equation becomes

$$x^2 \lambda^2 \frac{d^2y}{dt^2} + x \lambda \frac{dy}{dt} + (\lambda^2 x^2 - \nu^2)y = 0 \Rightarrow t^2 \frac{d^2y}{dt^2} + t \frac{dy}{dt} + (t^2 - \nu^2)y = 0.$$

The solution is

$$y = C_1 J_\nu(t) + C_2 Y_\nu(t) = C_1 J_\nu(\lambda x) + C_2 Y_\nu(\lambda x).$$

Example 31.37 (Peter O'Neil, 10e, p221, 5.4q9). Use the change of variables $y = x^{-n}u$ to find the general solution of the ODE

$$xy'' + (1 + 2n)y' + xy = 0.$$

Express your answer in terms of the Bessel functions.

Solution. Since

$$y' = -nx^{-n-1}u + x^{-n}u', \quad y'' = n(n+1)x^{-n-2}u - 2nx^{-n-1}u' + x^{-n}u''$$

the ODE becomes

$$n(n+1)x^{-n-1}u - 2nx^{-n}u' + x^{-n+1}u'' - n(2n+1)x^{-n-1}u + (2n+1)x^{-n}u' + x^{-n+1}u = 0$$

which simplifies to by multiplying x^{n-1} :

$$u'' + ((2n+1)x^{-1} - 2nx^{-1})u' + (n(n+1)x^{-2} - n(2n+1)x^{-2} + 1)u = \boxed{u'' + \frac{1}{x}u' + \frac{x^2 - n^2}{x^2}u = 0}.$$

Hence, the general solution for u is $u = C_1 J_\nu(x) + C_2 Y_\nu(x)$ and the general solution for the original ODE is

$$y = x^{-n}[C_1 J_n(x) + C_2 Y_n(x)]. \quad \square$$

Definition 31.11. A *modified Bessel equation of order ν* is

$$x^2 y'' + xy' + (-x^2 - \nu^2)y = 0. \quad (31.16)$$

e:messel

Here the range of x can either be $0 < x < L$, $0 < a < x < b < \infty$ or $0 < L < x$ depending on the engineering problem.

By a change of variable, it is possible to write the solutions of (36.33) as

$$y(x) = aJ_\nu(ix) + bJ_{-\nu}(ix) \quad \text{or} \quad y(x) = aJ_\nu(ix) + bY_\nu(ix).$$

They are not in real form. Mathematicians introduced a transform to turn $J_\nu(ix)$ and $Y_\nu(ix)$ into real functions:

$$I_\nu(x) = e^{-\frac{1}{2}\nu\pi i} J_\nu(e^{\frac{1}{2}\pi i} x) = \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(k + \nu + 1)} \left(\frac{x}{2}\right)^{2k+\nu}, \quad -\pi < \arg x \leq \frac{\pi}{2}$$

and

$$K_\nu(x) = \lim_{r \rightarrow \nu} \frac{I_r(x) \cos r\pi - I_{-r}(x)}{\sin r\pi}.$$

The functions I_ν and K_ν above are all the *modified Bessel functions of the first and second type/kind* respectively. We can regard them as the complementary to Bessel functions of the first and second type as follows [Abramowitz and Stegun, 1964]:

ODE	$y'' + y = 0$	$y'' - y = 0$	(36.23)	(36.33)
Fundamental Solution 1	$\cos x, \sin x$	$\cosh x, \sinh x$	$J_\nu(x), Y_\nu(x)$	$I_\nu(x), K_\nu(x)$
Fundamental Solution 2	e^{ix}, e^{-ix}	e^x, e^{-x}	$H_\nu^{(1)}(x), H_\nu^{(2)}(x)$	?

31.5 Comments on Convergence & Irregular Singular Points

So far, we have avoided the discussion of convergence. It is because Theorem 31.14 guarantees the convergence of the power series in an open interval consisting of an ordinary point x_0 . In the case of regular singular point x_0 , although it is infinite at x_0 , according to Theorem 31.26, the series is well defined. However, for irregular singular points, the series may diverge everywhere as demonstrated in the following example.

Example 31.38 (Coddington and Levinson [1984]). Find the formal power series solution at $x = 0$ for the ODE $x^2 y'' + (3x - 1)y' + y = 0$.

Solution: Assume $y = \sum_{n=0}^{\infty} a_n x^n$. Substituting into the equation gives

$$\begin{aligned} & x^2 \sum_{n=0}^{\infty} n(n-1)a_n x^{n-2} + (3x-1) \sum_{n=0}^{\infty} na_n x^{n-1} + \sum_{n=0}^{\infty} a_n x^n \\ &= \sum_{n=0}^{\infty} [n(n-1)a_n + 3na_n - (n+1)a_{n+1} + a_n] x^n = 0 \end{aligned}$$

Equating the coefficients to zero, we obtain

$$(n+1)a_{n+1} = (n^2 - n + 3n + 1)a_n \Rightarrow a_{n+1} = \frac{(n+1)^2}{n+1} a_n = (n+1)a_n = \cdots = (n+1)!a_0$$

Therefore the formal power series solution is $y = a_0 \sum_{n=0}^{\infty} n! x^n$ which is divergent for $x \neq 0$!

In the above example, 0 is an irregular singular point. In general, solutions expanded around an irregular singular points are distinctive in one aspect: they are usually in the form of an exponential function times a Frobenius series. Due to the factor of the exponential function, a solution near an

irregular singular point behaves very differently from that near a regular singular point. It may blow up exponentially, or vanish exponentially, or oscillate wildly.

Let analyse the irregular singular point with the first order homogeneous equation

$$y' + P(x)y = 0$$

at the origin. If $P(x)$ has a pole of order $k + 1$ at $x = 0$, then $P(x)$ has the Laurent series expansion

$$P(x) = \frac{b_{k+1}}{x^{k+1}} + \cdots + \frac{b_1}{x} + a_0 + a_1x + \cdots .$$

Therefore,

$$y(x) = Ce^{-\int P(x)dx} = C \exp \left[\frac{b_{k+1}}{kx^k} + \cdots + \frac{b_2}{x} \right] \cdot \underbrace{x^{-b_1} M(x)}_{\text{Frobenius series}} .$$

Series that are divergent may still be useful for the purpose of approximation. The study of divergent series has lead to the development of “asymptotic series”, which is a large subject, consult the references Murray [1984], Henrici [1977], and Balser [2000].

Chapter 32

Laplace Transform

Engineers (especially electrical engineers) find the Laplace transform [Schiff, 1999] to be a handy tool for solving the initial value problems of general linear ODEs (and systems of linear ODEs) of constant coefficients. In Section 32.1, Laplace transform is defined mathematically and it is used to perform some calculation. Section 32.2 discusses various rules or properties of Laplace transform and how these rules could be used to find Laplace transform of various combination of functions. Section 32.3 defines the inverse transform and apply it to solve various ODE problems. Applications of Laplace transform and some basic concepts in classical control theory will be discussed in later sections.

32.1 Definitions

In this section, a special kind of *linear transform* [Glaeske et al., 2006] called the *Laplace transform* is defined. It relates a set of causal functions to a set of complex-valued functions. The variable t is used instead of variable x (which was used in Topics 1 to 3) because in engineering, Laplace transform is mostly used to solve problems that changes with respect with time t .

Definition 32.1. A function f over $\mathbb{R}$ is called a *causal function* if $f(t)$ is zero when $t < 0$.

Definition 32.2. A *Heaviside step function* $H : \mathbb{R} \rightarrow \mathbb{R}$ is defined as $H(t) = \begin{cases} 0, & t < 0, \\ 1, & t \geq 0. \end{cases}$

We can construct a causal function f from any function g over $\mathbb{R}$ by multiplying it with the Heaviside step function H , i.e. $f(t) = g(t)H(t)$, for all $t \in \mathbb{R}$. If f is a causal function then $f(t) = H(t)f(t)$, for all $t \in \mathbb{R}$.

Definition 32.3. Let f be a causal function. The *Laplace transform* F of f is defined by the following improper integral:

$$F(s) = \mathcal{L}\{f(t)\} = \int_{0-}^{\infty} f(t)e^{-st} dt = \lim_{L \rightarrow \infty} \lim_{\varepsilon \rightarrow 0+} \int_{-\varepsilon}^L f(t)e^{-st} dt. \quad (32.1)$$

The function $K(s, t) = e^{-st}$ is called the *kernel* of the Laplace transform and the variable $s \in \mathbb{C}$ is called the *transformed variable*. The real part of s is required to be larger than some real number, usually 0 if not otherwise stated.

Since f is assumed causal, that means when $t < 0$, $f(t) = 0$ and the inverse Laplace transformation will return the “causal portion” of f , i.e. $\mathcal{L}^{-1}\{F(s)\} = f(t)H(t)$. If one wishes to retain the information in the region of $t < 0$ for a non-causal function, a bilateral Laplace transform $F_B(s) = \mathcal{L}_B\{f(t)\} = \int_{-\infty}^{\infty} f(t)e^{-st} dt$ may be used.

Not all functions have a Laplace transform, only functions that are of *exponential order* (defined below) have a Laplace transform.

def:exp-order

Definition 32.4. A piecewise-continuous function f is said to be of *exponential order* α_f if for every positive number A , there exists a real constants $K > 0$ and α_f such that $|f(t)| \leq Ke^{\alpha_f t}$ and if $c > \alpha_f$ then $\lim_{t \rightarrow \infty} \frac{f(t)}{e^{ct}} = 0$.

Functions that are of exponential order include power functions t^n , polynomial functions, exponential function a^t ($a > 1$), etc.

Theorem 32.1. (Existence Theorem, [Schiff, 1999, p. 26]) *If a function f is of exponential order α_f then the Laplace transform (32.1) exists for $\text{Re}(s) > \alpha_f$.*

Remark 32.2. The region $\text{Re}(s) \geq \alpha_f$ is known as the *region of convergence* of the Laplace transform [Goodwin et al., 2000, p. 66].

Remark 32.3. According to Doetsch [1974, Theorem 23.2], $F(s) \rightarrow 0$ as $\text{Re}(s) \rightarrow \infty$ whenever the Laplace transform exists. As a consequence, any function $F(s)$ without this behaviour, say $\frac{s-1}{s+1}$, $\frac{e^s}{s}$, or s^2 , cannot be the Laplace transform of any “classical” function f .

Let us look at how to calculate Laplace transform for causal functions by definition.

eg:by-defn

Example 32.4. Find the Laplace Transform for the functions (a) $f(t) = 1$ (or $H(t)$); and (b) $f(t) = t$.

Solution:

$$(a) \quad F(s) = \int_0^{\infty} 1 \cdot e^{-st} dt = \left[\frac{e^{-st}}{-s} \right]_0^{\infty} = \frac{1}{s}, \text{Re}(s) > 0.$$

Note that the Laplace transform for 1 and $H(t)$ are the same because Laplace only deals with functions (or in engineering terminology, signals) starting from time $t = 0$, i.e. causal functions. The two functions are the same for $t \geq 0$.

$$(b) \quad F(s) = \int_0^{\infty} t \cdot e^{-st} dt = \left[\frac{te^{-st}}{-s} \right]_0^{\infty} + \frac{1}{s} \left[\frac{e^{-st}}{(-s)} \right]_0^{\infty} = \frac{1}{s^2}, \text{Re}(s) > 0.$$

eg:by-defn2

Example 32.5. Find the Laplace Transform for the following functions:

(a) $f(t) = t^2$;	(c) $f(t) = \cos at$;	(e) $f(t) = H(t-a)$, $a > 0$;
(b) $f(t) = e^{at}$;	(d) $f(t) = \sin at$;	(f) $f(t) = \cosh at$.

Solution. (a) $F(s) = \int_0^{\infty} t^2 \cdot e^{-st} dt = \left[\frac{t^2 e^{-st}}{-s} - \frac{2te^{-st}}{s^2} - \frac{2e^{-st}}{s^3} \right]_0^{\infty} = \frac{2}{s^3}, \text{Re}(s) > 0.$

$$(b) \quad F(s) = \int_0^{\infty} e^{at} \cdot e^{-st} dt = \int_0^{\infty} e^{(a-s)t} dt = \left[\frac{e^{(a-s)t}}{a-s} \right]_0^{\infty} = \lim_{t \rightarrow +\infty} \frac{e^{(a-s)t}}{a-s} - \frac{1}{a-s} = \frac{1}{s-a}, \text{Re}(s) > a.$$

$$(c) \quad F(s) = \int_0^{\infty} \cos at \cdot e^{-st} dt = \left[\frac{e^{-st}}{s^2 + a^2} (-s \cos at + a \sin at) \right]_0^{\infty} = \frac{s}{s^2 + a^2}, \text{Re}(s) > 0.$$

$$(d) \quad F(s) = \int_0^{\infty} \sin at \cdot e^{-st} dt = \left[\frac{e^{-st}}{s^2 + a^2} (-s \sin at - a \cos at) \right]_0^{\infty} = \frac{a}{s^2 + a^2}, \text{Re}(s) > 0.$$

$$(e) \quad F(s) = \int_0^{\infty} H(t-a) \cdot e^{-st} dt = \left[\frac{e^{-st}}{-s} \right]_a^{\infty} = \frac{e^{-as}}{s}, \text{Re}(s) > 0.$$

$$(f) \quad F(s) = \int_0^{\infty} \cosh at \cdot e^{-st} dt = \int_0^{\infty} \frac{e^{at} + e^{-at}}{2} \cdot e^{-st} dt = \frac{1}{2} \left[\frac{e^{(a-s)t}}{a-s} + \frac{e^{(-a-s)t}}{-a-s} \right]_0^{\infty} = \frac{1}{2} \left[-\frac{1}{a-s} - \frac{1}{-a-s} \right] = \frac{s}{s^2 - a^2}, \text{Re}(s) > a.$$

□

32.2 Rules of Laplace Transform

sec:properties

Finding Laplace transform using definition (32.1) is time-consuming. To find Laplace transform faster, we will investigate its properties and then use them together with the Laplace transform in Examples 32.4 and 32.5 to evaluate the more complicated causal functions.

In the following theorems, functions f and g are assumed to be causal, piecewise-continuous with exponential order α_f and α_g respectively.

Theorem 32.6 (Linearity). *If a and b are constants then*

$$\mathcal{L}\{af(t) + bg(t)\} = a\mathcal{L}\{f(t)\} + b\mathcal{L}\{g(t)\}, \quad \operatorname{Re}(s) > \max\{\alpha_f, \alpha_g\}.$$

Example 32.7. Find $\mathcal{L}\{e^{3t} - t + 2 \cos t\}$.

Solution: $\mathcal{L}\{e^{3t} - t + 2 \cos t\} = \mathcal{L}\{e^{3t}\} - \mathcal{L}\{t\} + 2\mathcal{L}\{\cos t\} = \frac{1}{s-3} - \frac{1}{s^2} + \frac{2s}{s^2+1},$

The exponential order is $\operatorname{Re}(s) > \max\{3, 0, 0\} = 3$.

Example 32.8. Find $\mathcal{L}\{7 - 3t + 4 \sin 2t - 5e^{4t}\}$.

Solution. $\mathcal{L}\{7 - 3t + 4 \sin 2t - 5e^{4t}\} = 7\mathcal{L}\{1\} - 3\mathcal{L}\{t\} + 4\mathcal{L}\{\sin 2t\} - 5\mathcal{L}\{e^{4t}\} = \frac{7}{s} - \frac{3}{s^2} + \frac{8}{s^2+2^2} - \frac{5}{s-4},$

The exponential order is $\operatorname{Re}(s) > \max\{0, 0, 0, 4\} = 4$. □

Theorem 32.9 (Frequency shift/First shift/Exponential Modulation). *If $\mathcal{L}\{f(t)\} = F(s)$, $\operatorname{Re}(s) > \alpha_f$ then*

$$\mathcal{L}\{e^{at}f(t)\} = F(s-a), \quad \operatorname{Re}(s) > \alpha_f + \operatorname{Re}(a).$$

Example 32.10. Find $\mathcal{L}\{te^{-2t}\}$.

Solution: From Example 32.4 (b), we know $\mathcal{L}\{t\} = \frac{1}{s^2}$. $\therefore \mathcal{L}\{e^{-2t}t\} = \frac{1}{(s+2)^2}$, $\operatorname{Re}(s) > -2$.

Example 32.11. Find $\mathcal{L}\{e^{-3t} \sin 2t\}$.

Solution. By Example 32.5 (d), $\mathcal{L}\{\sin 2t\} = \frac{2}{s^2+2^2}$. $\therefore \mathcal{L}\{e^{-3t} \sin 2t\} = \frac{2}{(s+3)^2+2^2}$, $\operatorname{Re}(s) > -3$. □

Theorem 32.12 (Time shift/Second shift/Delay Theorem). *If constant $a > 0$ then*

$$\mathcal{L}\{f(t-a)H(t-a)\} = e^{-as}\mathcal{L}\{f(t)H(t)\}.$$

This property can be useful when we want to find the Laplace transform of a piecewise function $f(t)$. However, in the calculation, one needs to make use of

$$\mathcal{L}\{f(t)H(t-a)\} = e^{-as}\mathcal{L}\{f(t+a)H(t)\}, \quad (32.2) \quad \text{e:time-shift}$$

as demonstrated in the following example.

Example 32.13. Determine $\mathcal{L}\{f(t)\}$ when $f(t) = \begin{cases} t, & 0 \leq t \leq a, \\ 0, & \text{otherwise} \end{cases}$

Solution: To apply Laplace transform, $f(t)$ needs to be rewritten in terms of Heaviside step functions:

$$f(t) = tH(t) + (0-t)H(t-a) = tH(t) - tH(t-a).$$

$$F(s) = \mathcal{L}\{tH(t)\} - \mathcal{L}\{tH(t-a)\} = \mathcal{L}\{t\} - e^{-as}\mathcal{L}\{(t+a)H(t)\}$$

$$= \frac{1}{s^2} - e^{-as} \mathcal{L}\{t\} - e^{-as} \mathcal{L}\{a\} = \frac{1}{s^2} - e^{-as} \frac{1}{s^2} - e^{-as} \frac{a}{s}.$$

Example 32.14. Determine $\mathcal{L}\{f(t)\}$ when $f(t) = \begin{cases} 2t^2, & 0 \leq t < 3, \\ t+4, & 3 \leq t < 5, \\ 9, & t \geq 5. \end{cases}$

Solution. Rewriting $f(t)$ in terms of Heaviside step function, we have

$$\begin{aligned} f(t) &= 2t^2 H(t) + (t+4-2t^2)H(t-3) + (9-t-4)H(t-5) \\ &= 2t^2 H(t) + (t+4-2t^2)H(t-3) - (t-5)H(t-5) \\ F(s) &= \mathcal{L}\{2t^2 H(t) + (t+4-2t^2)H(t-3) - (t-5)H(t-5)\} \\ &= \mathcal{L}\{2t^2 H(t)\} + \mathcal{L}\{(t+4-2t^2)H(t-3)\} - \mathcal{L}\{(t-5)H(t-5)\} \\ &= \frac{4}{s^3} + e^{-3s} \mathcal{L}\{((t+3)+4-2(t+3)^2)H(t)\} - e^{-5s} \mathcal{L}\{tH(t)\} \\ &= \frac{4}{s^3} + e^{-3s} \mathcal{L}\{(t+7-2(t^2+6t+9))H(t)\} - e^{-5s} \frac{1}{s^2} \\ &= \frac{4}{s^3} + e^{-3s} \mathcal{L}\{(-2t^2-11t-11)H(t)\} - e^{-5s} \frac{1}{s^2} \\ &= \frac{4}{s^3} - e^{-3s} \left(\frac{4}{s^3} + \frac{11}{s^2} + \frac{11}{s} \right) - e^{-5s} \frac{1}{s^2} \end{aligned}$$

□

thm:times-tn

Theorem 32.15 (Derivative of Transform/Multiplication by t^n). If $\mathcal{L}\{f(t)\} = F(s)$ then

$$\frac{d^n}{ds^n}[F(s)] = \mathcal{L}\{(-1)^n t^n f(t)\}, \quad \mathcal{L}\{t^n f(t)\} = (-1)^n \frac{d^n}{ds^n}[F(s)].$$

Example 32.16. Find $\mathcal{L}\{t^n\}$.

Solution: By Example [32.4](#), we know that $\mathcal{L}\{1\} = \frac{1}{s}$. By Theorem [32.15](#),

$$\mathcal{L}\{t^n\} = \mathcal{L}\{t^n \cdot 1\} = (-1)^n \frac{d^n}{ds^n} \left(\frac{1}{s} \right) = (-1)^n (-1) \cdots (-n) s^{-(n+1)} = \frac{n!}{s^{n+1}}.$$

Example 32.17. Find (a) $\mathcal{L}\{t \sin 3t\}$ and (b) $\mathcal{L}\{t^2 e^t\}$.

Solution.

$$\begin{aligned} \text{(a) } \mathcal{L}\{\sin 3t\} &= \frac{3}{s^2 + 9}. \text{ So } \mathcal{L}\{t \sin 3t\} = (-1)^1 \frac{d}{ds} \left(\frac{3}{s^2 + 9} \right) = -\frac{-6s}{(s^2 + 9)^2} = \frac{6s}{(s^2 + 9)^2} \\ \text{(b) } \mathcal{L}\{e^t\} &= \frac{1}{s-1}. \text{ So } \mathcal{L}\{t^2 e^t\} = (-1)^2 \frac{d^2}{ds^2} \left(\frac{1}{s-1} \right) = \frac{2}{(s-1)^3}. \end{aligned}$$

□

thm:int-lt

Theorem 32.18 (Integration of Transform). If $\lim_{t \rightarrow 0^+} \frac{f(t)}{t}$ is finite and $\mathcal{L}\{f(t)\} = F(s)$ then

$$\mathcal{L}\left\{\frac{f(t)}{t}\right\} = \int_s^\infty F(\sigma) d\sigma.$$

eg:int-L

Example 32.19. Find $\mathcal{L}\left\{\frac{\sin t}{t}\right\}$.

Solution: We know that $\mathcal{L}\{\sin t\} = \frac{1}{1+s^2}$ and $\lim_{t \rightarrow 0^+} \frac{\sin t}{t} = 1$. Hence

$$\mathcal{L}\left\{\frac{\sin t}{t}\right\} = \int_s^\infty \frac{1}{1+\sigma^2} d\sigma = \frac{\pi}{2} - \tan^{-1} s = \tan^{-1} \frac{1}{s}$$

eg:int-L2

Example 32.20. Find $\mathcal{L}\left\{\frac{e^t - 1 - t}{t}\right\}$.

Solution. Note that by L'Hôpital rule, $\lim_{t \rightarrow 0} \frac{e^t - 1 - t}{t^2} = \lim_{t \rightarrow 0} \frac{e^t - 1}{2t} = \lim_{t \rightarrow 0} \frac{e^t}{2} = \frac{1}{2}$. By Theorem [32.18](#),

$$\mathcal{L}\{e^t - 1 - t\} = \frac{1}{s-1} - \frac{1}{s} - \frac{1}{s^2}$$

implies

$$\begin{aligned}\mathcal{L}\left\{\frac{e^t - 1 - t}{t}\right\} &= \int_s^\infty \left[\frac{1}{\sigma - 1} - \frac{1}{\sigma} - \frac{1}{\sigma^2} \right] d\sigma = \left[\ln(\sigma - 1) - \ln \sigma \right]_s^\infty + \left[\frac{1}{\sigma} \right]_s^\infty \\ &= \lim_{\sigma \rightarrow \infty} \ln \frac{\sigma - 1}{\sigma} - \ln \frac{s - 1}{s} + \frac{1}{\infty} - \frac{1}{s} = \ln \frac{s}{s - 1} - \frac{1}{s}.\end{aligned}$$

□

Theorem 32.21 (Time Differentiation). Suppose that f is continuous and having n -th order derivatives on $[0, \infty)$ and of exponential order α . If $\mathcal{L}\{f(t)\} = F(s)$ then

$$\mathcal{L}\left\{\frac{d^n}{dt^n}f(t)\right\} = s^n F(s) - s^{n-1}f(0) - s^{n-2}f'(0) - \dots - f^{(n-1)}(0).$$

In the case of $n = 1$, we have $\mathcal{L}\{f'(t)\} = sF(s) - f(0)$.

In the case of $n = 2$, we have $\mathcal{L}\{f''(t)\} = s^2F(s) - sf(0) - f'(0)$.

In the case of $n = 3$, we have $\mathcal{L}\{f'''(t)\} = s^3F(s) - s^2f(0) - sf'(0) - f''(0)$.

The time differentiation property is usually applied in solving linear ODEs shown in the following example and Section [32.4](#)

Example 32.22. Given $y'' + y = 0$, $y(0) = 0$, $y'(0) = 1$. Find $Y(s) = \mathcal{L}\{y\}$.

Solution: $Y(s)$ can be found using time differentiation property:

$$\mathcal{L}\{y''\} + \mathcal{L}\{y\} = \mathcal{L}\{0\} \Rightarrow s^2Y(s) - sy(0) - y'(0) + Y(s) = 0 \Rightarrow s^2Y(s) - 1 + Y(s) = 0.$$

Hence
$$Y(s) = \frac{1}{s^2 + 1}.$$

Example 32.23. Find $\mathcal{L}\{\sin^2 at\}$.

Solution. Let $f(t) = \sin^2 at$. Then

$$\begin{aligned}\mathcal{L}\{f'(t)\} &= \mathcal{L}\{f(t)\} - f(0) \Rightarrow \mathcal{L}\{2 \sin at \cdot \cos at \cdot a\} = s\mathcal{L}\{\sin^2 at\} - \sin^2 0 \\ &\Rightarrow \mathcal{L}\{a \sin 2at\} = \frac{2a^2}{s^2 + 4a^2} = s\mathcal{L}\{\sin^2 at\}.\end{aligned}$$

Therefore, $\mathcal{L}\{\sin^2 at\} = \frac{2a^2}{s(s^2 + 4a^2)}$. Or, $\mathcal{L}\{\sin^2 at\} = \mathcal{L}\left\{\frac{1 - \cos 2at}{2}\right\} = \frac{1}{2s} - \frac{s}{2((2a)^2 + s^2)}$. □

Theorem 32.24 (Time Integration). If $t > 0$ then

$$\mathcal{L}\left\{\int_0^t f(\tau)d\tau\right\} = \frac{1}{s}\mathcal{L}\{f(t)\}.$$

Example 32.25. Evaluate $\mathcal{L}\left\{\int_0^t \tau + \sin 2\tau d\tau\right\}$.

Solution: By applying Laplace Transform Time Integration, we obtain

$$\frac{1}{s}\mathcal{L}\{t + \sin 2t\} = \frac{1}{s}\left[\frac{1}{s^2} + \frac{2}{s^2 + 4}\right].$$

Remark 32.26. A comprehensive list of Laplace transforms was compiled in Prudnikov et al. [CRC1992].

32.3 Inverse Laplace Transform

The Laplace transform associates every piecewise-continuous function of exponential order with a complex function. The inverse Laplace transform is a process to determine what piecewise-continuous function f leads to a given complex function F .

Definition 32.5. Let $F(s)$ be the Laplace transform of a function $f(t)$, $t \geq 0$. The *inverse Laplace transform* of $F(s)$ is defined as the contour integral (Chapter 14)

$$f(t) = \mathcal{L}^{-1}\{F(s)\} = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} F(s)e^{st} ds \quad (32.3)$$

where $c > \alpha_f$.

Finding Laplace transform using formula (32.1) only requires the knowledge from Calculus but finding inverse Laplace transform using formula (32.3) requires the use of Residue Theorem as illustrated in <http://residuetheorem.com/2013/04/14/inverse-laplace-transform-part-i/>.

Since finding inverse Laplace transform with (32.3) is difficult, Engineers find the inverse Laplace transform of $F(s)$ by

- (1) checking if $F(s)$ can be found in the Laplace transform table;
- (2) using algebraic manipulations, especially *partial fractions*, to transform $F(s)$ into a combination of recognisable forms in the Laplace transform table;
- (3) utilising the properties of Laplace transform and Laplace transform table to find the inverse Laplace transform.

Some properties of inverse Laplace transform follow closely with the properties of Laplace transform in Section 32.2 as follows:

- Linearity: $\mathcal{L}^{-1}\{\alpha F(s) + \beta G(s)\} = \alpha \mathcal{L}^{-1}\{F(s)\} + \beta \mathcal{L}^{-1}\{G(s)\}$;
- Frequency Shift: $\mathcal{L}^{-1}\{F(s-a)\} = e^{at} \mathcal{L}^{-1}\{F(s)\}$;
- Time Shift: $\mathcal{L}^{-1}\{e^{-as}F(s)\} = \mathcal{L}^{-1}\{F(s)\} \Big|_{t \rightarrow t-a} = f(t-a)H(t-a)$.
- Multiplication by t^n : $\mathcal{L}^{-1}\left\{\frac{d^n}{ds^n}[F(s)]\right\} = (-1)^n t^n f(t)$.
- Time Integration: $\mathcal{L}^{-1}\left\{\frac{1}{s}F(s)\right\} = \int_0^t f(\tau) d\tau$.

Learning to find an inverse Laplace transform of a function is like learning to find an integration of a function. One needs practise a lot to build the intuition for using the appropriate techniques.

Example 32.27. Find the inverse Laplace transform for $F(s) = \frac{2}{s^2 + 36}$ (Hint: Table of Transform)

Solution. $f(t) = \mathcal{L}^{-1}\left\{\frac{2}{s^2 + 36}\right\} = \mathcal{L}^{-1}\left\{\frac{2}{6} \frac{6}{s^2 + 6^2}\right\} = \frac{1}{3} \mathcal{L}^{-1}\left\{\frac{6}{s^2 + 6^2}\right\} = \frac{1}{3} \sin 6t \quad \square$

Example 32.28. Find the inverse Laplace transform for $F(s) = \frac{1}{(s+3)(s-2)}$ (Hint: Partial Fraction)

Solution: $f(t) = \mathcal{L}^{-1}\left\{\frac{1}{(s+3)(s-2)}\right\} = \mathcal{L}^{-1}\left\{\frac{\frac{1}{5}}{s-2} + \frac{\frac{-1}{5}}{s+3}\right\} = \frac{1}{5}(e^{2t} - e^{-3t})$

Example 32.29. Find the inverse Laplace transform for $F(s) = \frac{s+1}{s^2(s^2+9)}$ (Hint: Partial Fraction)

Solution. $f(t) = \mathcal{L}^{-1}\left\{\frac{s+1}{s^2(s^2+9)}\right\} = \mathcal{L}^{-1}\left\{\frac{1}{9s} + \frac{1}{9s^2} - \frac{s+1}{9(s^2+9)}\right\} = \frac{1}{9} + \frac{t}{9} - \frac{1}{9} \cos 3t - \frac{1}{27} \sin 3t \quad \square$

Example 32.30. Find the inverse Laplace transform for $F(s) = \frac{1}{(s+2)^2}$ (Hint: Frequency shift)

Solution: $f(t) = \mathcal{L}^{-1}\left\{\frac{1}{(s+2)^2}\right\} = e^{-2t} \mathcal{L}^{-1}\left\{\frac{1}{s^2}\right\} = e^{-2t} t$.

Example 32.31. Find the inverse Laplace transform for $F(s) = \frac{2}{s^2 + 6s + 13}$ (Hint: Frequency shift)

$$\text{Solution. } f(t) = \mathcal{L}^{-1}\left\{\frac{2}{s^2 + 6s + 13}\right\} = \mathcal{L}^{-1}\left\{\frac{2}{(s+3)^2 + 4}\right\} = e^{-3t} \mathcal{L}^{-1}\left\{\frac{2}{s^2 + 2^2}\right\} = e^{-3t} \sin 2t \quad \square$$

Example 32.32. Find the inverse Laplace transform for $F(s) = \frac{s+7}{s^2 + 2s + 5}$ (Hint: Frequency shift)

$$\text{Solution. } f(t) = \mathcal{L}^{-1}\left\{\frac{s+7}{s^2 + 2s + 5}\right\} = \mathcal{L}^{-1}\left\{\frac{s+2}{(s+1)^2 + 2^2} + \frac{6}{2} \left(\frac{2}{(s+1)^2 + 2^2}\right)\right\} = e^{-t} \cos 2t + 3e^{-t} \sin 2t \quad \square$$

Example 32.33. Find the inverse Laplace transform for $F(s) = \frac{1}{(s+1)^2(s^2 + 4)}$. (Hint: Partial fraction and frequency shift)

$$\begin{aligned} \text{Solution. } f(t) &= \mathcal{L}^{-1}\left\{\frac{1}{(s+1)^2(s^2 + 4)}\right\} = \mathcal{L}^{-1}\left\{-\frac{2s+3}{25(s^2 + 4)} + \frac{2}{25(s+1)} + \frac{1}{5(s+1)^2}\right\} \\ &= -\frac{2}{25} \cos 2t - \frac{3}{50} \sin 2t + \frac{2}{25} e^{-t} + \frac{1}{5} e^{-t} t \end{aligned} \quad \square$$

Example 32.34. Find the inverse Laplace transform for $F(s) = \frac{4e^{-4s}}{s(s+2)}$ (Hint: Time shift)

Solution:

$$\begin{aligned} f(t) &= \mathcal{L}^{-1}\left\{e^{-4s} \left(\frac{4}{s(s+2)}\right)\right\} = \mathcal{L}^{-1}\left\{e^{-4s} \left(\frac{2}{s} - \frac{2}{s+2}\right)\right\} \\ &= \mathcal{L}^{-1}\left\{\frac{2}{s} - \frac{2}{s+2}\right\} \Big|_{t \rightarrow t-4} = [(2 - 2e^{-2t})H(t)]_{t \rightarrow t-4} = (2 - 2e^{-2(t-4)})H(t-4) \end{aligned}$$

Example 32.35. Find the inverse Laplace transform for $F(s) = \frac{e^{-\pi s}(s+3)}{s(s^2 + 1)}$ (Hint: Time shift)

$$\begin{aligned} \text{Solution. } f(t) &= \mathcal{L}^{-1}\left\{e^{-\pi s} \left(\frac{s+3}{s(s^2 + 1)}\right)\right\} = \mathcal{L}^{-1}\left\{e^{-\pi s} \left(\frac{3}{s} + \frac{1-3s}{s^2 + 1}\right)\right\} \\ &= \mathcal{L}^{-1}\left\{\frac{3}{s} + \frac{1}{s^2 + 1} - 3\left(\frac{s}{s^2 + 1}\right)\right\} \Big|_{t \rightarrow t-\pi} = [(3 + \sin t - 3 \cos t)H(t)]_{t \rightarrow t-\pi} \\ &= (3 + \sin(t-\pi) - 3 \cos(t-\pi))H(t-\pi) \end{aligned} \quad \square$$

Example 32.36. Find the inverse Laplace transform for $F(s) = \ln\left(\frac{s+a}{s+b}\right)$ where $a \neq b$.

Solution. First, note that $F(s) \rightarrow 0$ as $\text{Re}(s) \rightarrow \infty$. Therefore, the Laplace transform may exist. A formal calculation below suggests a possibility for $f(t)$:

$$\begin{aligned} F(s) &= \ln(s+a) - \ln(s+b) \Rightarrow F'(s) = \frac{1}{s+a} - \frac{1}{s+b} \\ &\Rightarrow \mathcal{L}^{-1}\{F'(s)\} = \mathcal{L}^{-1}\left\{\frac{1}{s+a} - \frac{1}{s+b}\right\} = e^{-at} - e^{-bt} \\ &\Rightarrow (-1)tf(t) = e^{-at} - e^{-bt} \\ &\Rightarrow f(t) = \frac{e^{-bt} - e^{-at}}{t} \end{aligned}$$

Since

$$\lim_{t \rightarrow 0^+} \frac{e^{-bt} - e^{-at}}{t} = \lim_{t \rightarrow 0^+} \frac{-be^{-bt} + ae^{-at}}{1} = a - b,$$

according to Theorem 32.18, F is the Laplace transform of f and therefore, f is the inverse Laplace transform of F . [thm: int-1t] [sec: impulse] □

The following question is a bit tough because the impulse function $\delta(t)$ stated in Section 32.6.1 is involved, i.e. we need to use $\mathcal{L}\{\delta(t)\} = 1$ in the following example.

Example 32.37 (Final Exam April 2009, Q4(a)). Find the inverse Laplace transform for $\frac{1}{s^2 + 3s - 4} + \frac{2}{s^2 - 6s + 12} + \tan^{-1} s$. (8 marks)

Solution.

$$\mathcal{L}^{-1}\left\{\frac{1}{s^2+3s-4}\right\} = \mathcal{L}^{-1}\left\{\frac{\frac{1}{-4-1}}{s+4} + \frac{\frac{1}{1+4}}{s-1}\right\} = \frac{1}{5}(e^t - e^{-4t}).$$

$$\mathcal{L}^{-1}\left\{\frac{2}{s^2-6s+12}\right\} = \mathcal{L}^{-1}\left\{\frac{2}{(s-3)^2+3}\right\} = \frac{2e^{3t}}{\sqrt{3}}\mathcal{L}^{-1}\left\{\frac{\sqrt{3}}{s^2+3}\right\} = 2e^{3t}\sin\sqrt{3}t.$$

According to Example [32.19](#),

$$\mathcal{L}\left\{\frac{\sin t}{t}\right\} = \frac{\pi}{2} - \tan^{-1}s \Rightarrow \mathcal{L}^{-1}\{\tan^{-1}s\} = \frac{\pi}{2}\delta(t) - \frac{\sin t}{t}.$$

Therefore,

$$\mathcal{L}^{-1}\left\{\frac{1}{s^2+3s-4} + \frac{2}{s^2-6s+12} + \tan^{-1}s\right\} = \frac{e^t - e^{-4t}}{5} + \frac{2}{\sqrt{3}}e^{3t}\sin\sqrt{3}t + \frac{\pi}{2}\delta(t) - \frac{\sin t}{t}.$$

□

Note that if $F(s)$ does not tend to zero as $\text{Re}(s) \rightarrow \infty$, f may not exist. The following example illustrates how engineers can misuse some of the Laplace transform rules and lead to mathematically invalid answer.

Example 32.38. Try to find the inverse Laplace transform for $F(s) = \ln s$.

Trial (by Engineers): $F'(s) = \frac{1}{s} \Rightarrow \mathcal{L}^{-1}\{F'(s)\} = \mathcal{L}^{-1}\left\{\frac{1}{s}\right\} = 1 \Rightarrow (-1)tf(t) = 1 \Rightarrow f(t) = \frac{-1}{t}.$

Why is the answer above **wrong**? Because, $f(t) = \frac{-1}{t}$ does not have a Laplace transform by Definition [32.3](#):

$$\mathcal{L}\left\{\frac{-1}{t}\right\} = \int_{0-}^{\infty} \frac{-1}{t}e^{-st}dt.$$

As $t \rightarrow 0$, $\frac{-1}{t}e^{-st} \rightarrow \infty$! So the above integral does not exist!

Note that $\ln s$ does not tend to zero as $\text{Re}(s) \rightarrow \infty$. So the properties in Section [32.2](#) do not work. A correct answer involving complicated mathematics is given in Doetsch [1974, p. 320]:

$$\mathcal{L}\left\{-\frac{1}{t}H(t)\right\} = \ln s + \gamma \Rightarrow \mathcal{L}^{-1}\{\ln s\} = -\gamma\delta(t) - \frac{1}{t}H(t)$$

where $\gamma = \lim_{n \rightarrow \infty} \left(\sum_{k=1}^n \frac{1}{k} - \ln n\right) = \int_1^{\infty} \left(\frac{1}{[x]} - \frac{1}{x}\right) dx$ is called the *Euler(-Mascheroni) constant*.

32.4 Applications in IVP of Solving Ordinary (Integro-)Differential Equation

In this section, we will investigate how engineers use Laplace transform to solve *linear-time-invariant (LTI)* (see https://en.wikipedia.org/wiki/Linear_time-invariant_theory) ODE and LTI integro-differential equations with a few simple examples. First, let us recall the “time differentiation” and “time integration” properties from Section [32.2](#) which are fundamental in the solution of LTI ODE and LTI integro-differential equations:

$$\mathcal{L}\left\{\frac{d^n}{dt^n}f(t)\right\} = s^n F(s) - s^{n-1}f(0) - s^{n-2}f'(0) - \dots - f^{(n-1)}(0),$$

$$\mathcal{L}\left\{\int_0^t f(\tau)d\tau\right\} = \frac{1}{s}F(s).$$

A typical example of LTI ODE with initial conditions is a second order ODE-IVP:

$$ay'' + by' + cy = u(t), \quad y(0) = y_0, \quad y'(0) = v_0$$

where a , b and c are constants. By taking the Laplace transform the ODE becomes

$$a[s^2Y(s) - sy_0 - v_0] + b[sY(s) - y_0] + cY(s) = U(s)$$

or

$$Y(s) = \frac{U(s) + ay_0s + av_0 + by_0}{as^2 + bs + c}.$$

By taking the inverse transform, engineers recover the solution $y(t)$ to the ODE-IVP.

In the following examples, Laplace transform and inverse Laplace transform are used in combination to find the analytic solution of various ODE-IVP. Note that y is causal and $t \geq 0$.

Example 32.39. Find the solution $y(t)$ of $y' - 4y = 1$, $y(0) = 1$.

Solution: We can solve the IVP by first taking Laplace transform, simplifying terms and taking inverse Laplace transform:

$$sY(s) - 1 - 4Y(s) = \frac{1}{s} \Rightarrow Y(s) = \frac{s+1}{s} \frac{1}{s-4} = \frac{5}{4(s-4)} - \frac{1}{4s}$$

$$y(t) = \frac{5}{4}e^{4t} - \frac{1}{4}$$

Example 32.40. [Johnson, 1997, p. 40] Solve the initial value problem

$$y'' + 3y' + 2y = 1 - t, \quad y(0) = 1, \quad y'(0) = -2.$$

Solution: Taking Laplace transform, we obtain

$$(s^2Y(s) - s - (-2)) + 3(sY(s) - 1) + 2Y(s) = \frac{1}{s} - \frac{1}{s^2}.$$

Rearranging them, we have

$$(s^2 + 3s + 2)Y(s) = s + 1 + \frac{1}{s} - \frac{1}{s^2} \Rightarrow Y(s) = \frac{s^3 + s^2 + s - 1}{s^2(s^2 + 3s + 2)}.$$

After a rather long calculation of partial fraction, we obtain

$$Y(s) = \frac{5}{4s} - \frac{1}{2s^2} - \frac{2}{s+1} + \frac{7}{4(s+2)}.$$

Taking the inverse Laplace transform, we obtain

$$y(t) = \frac{5}{4} - \frac{t}{2} - 2e^{-t} + \frac{7}{4}e^{-2t}.$$

Example 32.41. Find the solution $y(t)$ of $y'' + 4y = t$, $y(0) = 1$, $y'(0) = 0$.

Solution. $(s^2 + 4)Y(s) = \frac{1}{s^2} + s$

$$\Rightarrow Y(s) = \frac{s^3 + 1}{s^2} \frac{1}{s^2 + 4} = \frac{1}{4s^2} + \frac{s - \frac{1}{4}}{s^2 + 4} = \frac{1}{4s^2} + \frac{s}{s^2 + 2^2} - \frac{1}{8} \frac{2}{s^2 + 2^2}$$

$$\therefore y(t) = \frac{1}{4}t + \cos 2t - \frac{1}{8} \sin 2t.$$

□

Example 32.42. Find the solution $x(t)$ of $\frac{d^2x}{dt^2} + 5\frac{dx}{dt} + 6x = 2e^{-t}$, $x(0) = 1$, $x'(0) = 0$

Solution. $(s^2 + 5s + 6)X(s) = \frac{2}{s+1} + s + 5$

$$\Rightarrow X(s) = \frac{2 + s^2 + 6s + 5}{(s+1)(s+3)(s+2)} = -\frac{1}{s+3} + \frac{1}{s+2} + \frac{1}{s+1}$$

$$\therefore x(t) = -e^{-3t} + e^{-2t} + e^{-t}.$$

□

Example 32.43 (Final Exam Sept 2014, Q1(c)). Use Laplace transform to solve the initial value problem:

$$\frac{d^2y}{dt^2} + 4\frac{dy}{dt} + 3y = 2, \quad y(0) = 0, \quad y'(0) = 1. \quad (8 \text{ marks})$$

Solution. Let $Y = \mathcal{L}\{y\}$. Taking Laplace transform, we have

$$s^2Y - y'(0) - sy(0) + 4(sY - y(0)) + 3Y = \frac{2}{s} \quad [2 \text{ marks}]$$

Substituting the initial conditions, we obtain

$$s^2Y - 1 + 4sY + 3Y = \frac{2}{s} \quad [1 \text{ mark}]$$

$$\Rightarrow Y = \frac{1 + \frac{2}{s}}{s^2 + 4s + 3} = \frac{s + 2}{s(s+1)(s+3)} = \frac{2}{3s} - \frac{1}{2(s+1)} - \frac{1}{6(s+3)} \quad [1+3=4 \text{ marks}]$$

Taking inverse Laplace transform, we obtain the solution

$$y(t) = \frac{2}{3} - \frac{1}{2}e^{-t} - \frac{1}{6}e^{-3t}. \quad [1 \text{ mark}]$$

□

Example 32.44 (Final Exam Sept 2014, Q2(b)). Use Laplace transform to solve the initial value problem

$$y' + 3y = \sin^2 3t, \quad y(0) = 1. \quad (9 \text{ marks})$$

Solution. Taking Laplace transform to the left, we obtain

$$\mathcal{L}\{y' + 3y\} = sY - y(0) + 3Y = (s+3)Y - 1 \quad [2 \text{ marks}]$$

Taking the Laplace transform to the right, we obtain

$$\mathcal{L}\{e^{2t} \sin^2 3t\} = \mathcal{L}\left\{\frac{1 - \cos 6t}{2}\right\} = \frac{1}{2s} - \frac{s}{2(s^2 + 6^2)} \quad [2 \text{ marks}]$$

Hence,

$$Y = \frac{1}{s+3} \left[1 + \frac{1}{2s} - \frac{s}{2(s^2 + 6^2)} \right] = \frac{s^3 + 36s + 18}{s(s+3)(s^2 + 6^2)} \quad [3 \text{ marks}]$$

$$= \frac{1}{6s} + \frac{13}{15(s+3)} + \frac{-s-12}{30(s^2 + 6^2)}.$$

$$\text{Hence } y(t) = \frac{1}{6} + \frac{13}{15}e^{-3t} - \frac{1}{30} \cos 6t - \frac{1}{15} \sin 6t. \dots\dots\dots [2 \text{ marks}] \quad \square$$

Example 32.45. Find the solution $y(t)$ of $y''' + 5y'' + 17y' + 13y = 1$, $y(0) = y'(0) = 1$, $y''(0) = 0$

Solution: $s^3Y(s) - s^2 - s + 5(s^2Y(s) - s - 1) + 17(sY(s) - 1) + 13Y(s) = \frac{1}{s}$

$$\Rightarrow (s^3 + 5s^2 + 17s + 13)Y(s) = \frac{1}{s} + s^2 + s + 5s + 5 + 17 = s^2 + 6s + 22 + \frac{1}{s}$$

$$Y(s) = \frac{1}{13s} + \frac{8}{5(s+1)} - \frac{44s+7}{65(s^2+4s+13)} = \frac{1}{13s} + \frac{8}{5(s+1)} - \frac{1}{65} \frac{44(s+2) - 88 + 7}{(s+2)^2 + 9}$$

$$y(t) = \frac{1}{13} + \frac{8}{5}e^{-t} - e^{-2t} \left(\frac{44}{65} \cos 3t - \frac{27}{65} \sin 3t \right)$$

nal-2014.05q2b

Example 32.46 (Final Exam Sept 2014, Q2(b)). Use Laplace transform to solve the following initial value problem

$$y'' + 2y' + 2y = \cos 3x, \quad y(0) = 1, \quad y'(0) = 0.$$

[Hint: $\frac{s^3 + 2s^2 + 10s + 18}{(s^2 + 2s + 2)(s^2 + 9)} = \frac{92s + 166}{85(s^2 + 2s + 2)} - \frac{7s - 18}{85(s^2 + 9)}$] (8 marks)

Solution. By taking Laplace transform, we have

$$(s^2Y - s) + 2(sY - 1) + 2Y = \frac{s}{s^2 + 9} \quad [2 \text{ marks}]$$

This is simplified to

$$Y = \frac{s + 2 + \frac{s}{s^2 + 9}}{s^2 + 2s + 2} = \frac{s^3 + 2s^2 + 10s + 18}{(s^2 + 2s + 2)(s^2 + 9)} \quad [2 \text{ marks}]$$

Taking inverse Laplace transform

$$y = \mathcal{L}^{-1}\left\{ \frac{92s + 166}{85((s+1)^2 + 1)} - \frac{7s - 18}{85(s^2 + 9)} \right\} \quad [1 \text{ mark}]$$

$$= e^{-x} \mathcal{L}^{-1}\left\{ \frac{92s + 74}{85(s^2 + 1)} \right\} - \frac{7}{85} \cos 3x + \frac{6}{85} \sin 3x \quad [1+1=2 \text{ marks}]$$

$$= e^{-x} \left(\frac{92}{85} \cos x + \frac{74}{85} \sin x \right) - \frac{7}{85} \cos 3x + \frac{6}{85} \sin 3x$$

[1 mark]

□

Example 32.47. Find the solution of $\frac{dy}{dt} + 4 \int_0^t y(\tau) d\tau = \frac{t^2}{2}$.

Solution. $sY(s) - y(0) + 4 \frac{1}{s} Y(s) = \frac{1}{2} \frac{2!}{s^3} \Rightarrow \frac{s^2 + 4}{s} Y(s) = \frac{1}{s^3} + y(0)$
 $\Rightarrow Y(s) = \frac{s}{s^2 + 4} y(0) + \frac{1}{s^2(s^2 + 4)} = \frac{s}{s^2 + 4} y(0) + \frac{\frac{1}{4}}{s^2} - \frac{\frac{1}{4}}{s^2 + 4}$
 $y(t) = y(0) \cos 2t + \frac{1}{4} \left(t - \frac{1}{2} \sin 2t \right)$

□

32.5 Transfer Function and Convolution

The main reason engineers used Laplace transform is because it allows engineers to express *linear-time invariant engineering system* as the composition of various “*engineering*” blocks called transfer functions:



Figure: A transfer function “engineering block”.

For example, in analogue electrical networks, a cascading network may be viewed as a cascading of the above blocks (with two inputs): [Raut and Swamy, 2010, p. 33]. Concrete examples of application of block diagram are given in Giurgiutiu and Lyshevski [2009].

output is denoted by $y(t)$, it is called a continuous-time single-input-single-output (SISO) and is normally represented by a transfer function defined below.

Definition 32.6. Given a **linear time-invariant (LTI)** ODE:

$$a_n y^{(n)} + a_{n-1} y^{(n-1)} + \dots + a_2 y'' + a_1 y' + a_0 y = b_m u^{(m)} + \dots + b_1 u' + b_0 u \quad (32.4)$$

the **transfer function** of the ODE, denoted by $G(s)$, is defined as the *ratio* of the Laplace transform of the system **output**, $y(t)$ $Y(s) = \mathcal{L}\{y(t)\}$, to the Laplace transform of the system **input** $u(t)$, $U(s) = \mathcal{L}\{u(t)\}$:

$$G(s) := \frac{Y(s)}{U(s)} = \frac{b_m s^m + \dots + b_1 s + b_0}{a_n s^n + \dots + a_1 s + a_0} \quad (32.5)$$

under a *quiescent initial conditions* or *zero initial conditions*:

$$y(0) = y'(0) = \dots = y^{(n-1)}(0) = u(0) = \dots = u^{(m-1)}(0) = 0.$$

Definition 32.7. Equating the denominator in (32.5) to zero, i.e., $Q(s) = a_n s^n + \dots + a_1 s + a_0 = 0$ leads to the *characteristic equation* of the transfer function (32.5). From the factorisation of the numerator, $P(s) = (s - z_1)(s - z_2) \dots (s - z_m)$, we can obtain z_i , $i = 1, \dots, m$, they are called the *zeroes* of (32.5). From the factorisation of the denominator $Q(s) = (s - p_1)(s - p_2) \dots (s - p_n)$ we can obtain p_i , $i = 1, \dots, n$, they are called the *poles* of transfer function (32.5).

Example 32.48. Find the transfer function and characteristic equation for the linear system $x'' + \omega^2 x = f(t)$. Write the zeroes and poles of the transfer function.

Solution: The transfer function is

$$G(s) = \frac{X(s)}{F(s)} = \frac{1}{s^2 + \omega^2}$$

and characteristic equation is

$$s^2 + \omega^2 = 0.$$

There is no zero and the poles are $\pm\omega i$.

The poles of a transfer function is important because they characterise the stability of a LTI system.

Definition 32.8. A LTI system (33.18) (or (32.5)) is said to be *stable* if all the poles of (32.5), p_i , (i.e. the roots of the characteristic polynomial) have negative real part ($\text{Re}(p_i) < 0$).

Example 32.49. Find the transfer function and characteristic equation for the linear system $y'' + 12y' + 13y = u(t)$ and determine if it is stable.

Solution. The transfer function and characteristic equation are respectively

$$G(s) = \frac{Y(s)}{U(s)} = \frac{1}{s^2 + 12s + 13}, \quad s^2 + 12s + 13 = 0.$$

Since $s = \frac{-12 \pm \sqrt{12^2 - 4(13)}}{2} = -\sqrt{23} - 6, \sqrt{23} - 6 < 0$. Hence the system is stable. $\square$

Example 32.50 (Final Exam April 2009, Q4(b)). By using the Laplace transform, solve the following initial value problem.

$$9y'' + 3y' - 2y = \begin{cases} t, & 0 \leq t < 3, \\ 0, & t \geq 3, \end{cases} \quad \text{and} \quad y(0) = 0, \quad y'(0) = 1.$$

Is the solution stable?

(17 marks)

Solution. To solve this problem, we need to express the right hand side in terms of Heaviside step function:

$$9y'' + 3y' - 2y = tH(t) + (0 - t)H(t - 3).$$

Taking Laplace transform, we have

$$9(s^2Y - 1) + 3sY - 2Y = \frac{1}{s^2} + e^{-3s}\mathcal{L}\{-(t+3)H(t)\} = \frac{1}{s^2} - e^{-3s}\left[\frac{1}{s^2} + \frac{3}{s}\right].$$

This implies

$$(9s^2 + 3s - 2)Y = 9 + \frac{1}{s^2} - e^{-3s}\left[\frac{1}{s^2} + \frac{3}{s}\right] \Rightarrow Y = \frac{9s^2 + 1}{s^2(3s + 2)(3s - 1)} - e^{-3s}\frac{3s + 1}{s^2(3s + 2)(3s - 1)}.$$

By partial fraction, we have

$$Y = \frac{-\frac{3}{4}}{s} + \frac{-\frac{1}{2}}{s^2} + \frac{6}{3s - 1} + \frac{-\frac{15}{4}}{3s + 2} - e^{-3s}\left[\frac{-\frac{9}{4}}{s} + \frac{-\frac{1}{2}}{s^2} + \frac{\frac{3}{4}}{3s + 2} + \frac{6}{3s - 1}\right].$$

Taking inverse Laplace transform, we have

$$y(t) = -\frac{3}{4} - \frac{1}{2}t + 2e^{\frac{1}{3}t} - \frac{5}{4}e^{-\frac{2}{3}t} - \left(-\frac{9}{4} - \frac{1}{2}t + \frac{1}{4}e^{-\frac{2}{3}t} + 2e^{\frac{1}{3}t}\right)H(t)\Big|_{t \rightarrow t-3}.$$

Note that $2e^{\frac{1}{3}t} - 2e^{\frac{1}{3}(t-3)}H(t-3)$ cannot cancel after $t \geq 3$. Therefore, the solution is unstable because $y \rightarrow \infty$ as $t \rightarrow \infty$. Or, using the definition above, it is unstable because it has a pole which is positive $\frac{1}{3}$.

```
1 # Final Exam April 2009, Q4(b)
2 # Lecture Notes Example 4.5.6
3 from scipy import linspace, exp, zeros
4
5 def dx(x,t):
6     y = x[0]; z = x[1]
7     if 0<t<3:
8         return [z, (2*y-3*z + t)/9.0]
9     else:
10        return [z, (2*y-3*z)/9.0]
11
12 def exact(tarr):
```

```

13     xexact = zeros(tarr.shape)
14     for i in range(len(tarr)):
15         t = tarr[i]
16         xexact[i] = -0.75-0.5*t + 2*exp(t/3.0)-1.25*exp(-2*t/3.0)
17         if t >= 3:
18             xexact[i] -= -9.0/4-0.5*(t-3) + 0.25*exp(-2*(t-3)/3.0) + \
19                 2*exp((t-3)/3)
20     return xexact
21
22 from scipy.integrate import odeint
23 tarr = linspace(0,10,500+1)
24 x = odeint(dx, [0.0,1.0], tarr)
25 xexact = exact(tarr)
26 import pylab as p
27 for i in range(20):
28     print("t=", tarr[i], ":", x[i,0], "vs", xexact[i])
29 p.plot(tarr, x[:,0], tarr, xexact, '+')
30 p.show()

```

□

Using the notion of transfer function $G(s)$, the solution $y(t)$ of (33.18), which can be obtained by taking the inverse Laplace transform:

$$y(t) = \mathcal{L}^{-1}\{Y(s)\} = \mathcal{L}^{-1}\{G(s) \cdot U(s)\}.$$

What is the correspondence of “multiplication operation” from the “frequency domain” s back to the “time domain” t ? Mathematicians discovered that “convolution” is the corresponding “operation”.

def:convol

Definition 32.9. If f and g are piecewise-continuous causal functions, the *convolution* of functions f and g is defined by

$$(f * g)(t) = \left[\int_0^t f(\tau)g(t-\tau)d\tau \right] H(t). \quad (32.6)$$

e:convolution

Note that $f * g$ is a function of t and convolution is commutative, i.e. $f * g = g * f$.

The following example demonstrates how to perform convolution of two causal functions.

Example 32.51. Find the convolution for the functions $f(t) = \sin 2t H(t)$ and $g(t) = tH(t)$ by definition.

Solution. The calculation is given below:

$$\begin{aligned}
 (f * g)(t) &= \int_0^t \sin 2\tau H(\tau) \cdot (t-\tau)H(t-\tau)d\tau = \left(\int_0^t (t-\tau) \sin 2\tau d\tau \right) H(t) \\
 &= H(t) \left[\frac{-(t-\tau) \cos 2\tau}{2} - \frac{\sin 2\tau}{4} \right]_0^t = H(t) \left[\frac{t}{2} - \frac{\sin 2t}{4} \right]
 \end{aligned}$$

□

The following important theorem relates the operations of “multiplication” and “convolution” in the context of Laplace transform.

Theorem 32.52 (Convolution Theorem). If $\mathcal{L}\{f(t)\} = F(s)$ and $\mathcal{L}\{g(t)\} = G(s)$ then

$$\begin{aligned}
 \mathcal{L}\{f * g\} &= \mathcal{L}\{f(t)\}\mathcal{L}\{g(t)\} = F(s)G(s), \\
 \mathcal{L}^{-1}\{F(s)G(s)\} &= (f * g)(t).
 \end{aligned}$$

Example 32.53. Find $\mathcal{L}^{-1}\left\{\frac{1}{s^2(s+2)^2}\right\}$ using the convolution theorem.

Solution.

$$\begin{aligned}
 \mathcal{L}^{-1}\left\{\frac{1}{s^2(s+2)^2}\right\} &= \mathcal{L}^{-1}\left\{\frac{1}{s^2}\right\} * \mathcal{L}^{-1}\left\{\frac{1}{(s+2)^2}\right\} = t * (e^{-2t}t) = \int_0^t (t-\tau)e^{-2\tau}\tau d\tau \\
 &= \left[e^{-2\tau} \left(\frac{\tau^2 - t\tau}{2} - \frac{t-2\tau}{4} + \frac{1}{4} \right) \right]_0^t = e^{-2t} \left(\frac{1+t}{4} \right) - \frac{1-t}{4}.
 \end{aligned}$$

□

Example 32.54. Solve the integral equation $y(t) = 2t^2 + \int_0^t y(t - \tau)e^{-\tau} d\tau$.

Solution:

$$\begin{aligned} y(t) &= 2t^2 + y(t) * e^{-t} \\ \Rightarrow Y(s) &= 2\left(\frac{2!}{s^3}\right) + Y(s)\frac{1}{s+1} \Rightarrow Y(s) = \frac{s+1}{s}\left(\frac{4}{s^3}\right) = 2\left(\frac{2!}{s^3}\right) + \frac{4}{3!}\frac{3!}{s^4} \\ \Rightarrow y(t) &= 2t^2 + \frac{2}{3}t^3. \end{aligned}$$

In general, the convolution of functions is difficult to calculate, therefore, it is normally only used as a formal representation of the solution of an ODE with unknown (input) function $u(t)$ as illustrated in the following examples.

Example 32.55. By expressing the solution using the convolution notation, find the solution of the ODE

$$y'' + 9y = u(t), \quad y(0) = 3, \quad y'(0) = -1.$$

Solution:

$$\begin{aligned} s^2Y(s) - sy(0) - y'(0) + 9Y(s) &= U(s) \\ \Rightarrow (s^2 + 9)Y(s) &= U(s) + 3s - 1 \\ \Rightarrow Y(s) &= \frac{U(s) + 3s - 1}{s^2 + 9} = \frac{3s}{s^2 + 3^2} - \frac{1}{3}\frac{3}{s^2 + 3^2} + \frac{1}{3}\frac{3}{s^2 + 3^2}U(s) \\ \Rightarrow y(t) &= 3\cos 3t - \frac{1}{3}\sin 3t + \frac{1}{3}(\sin 3t) * u(t). \end{aligned}$$

Example 32.56 (Final Exam Sept 2016, Q4(a)). Use Laplace transform to solve the non-homogeneous equation

$$y' + 4y = \cos^5 3t, \quad y(0) = 0.$$

You do not have to actually compute the Laplace transform of $\cos^5 3t$. Leave the solution in convolution form. (4 marks)

Solution. Taking Laplace transform

$$(s + 4)Y = \mathcal{L}\{\cos^5 3t\} \Rightarrow Y = \frac{1}{s + 4}\mathcal{L}\{\cos^5 3t\}. \quad [1+2=3 \text{ marks}]$$

Taking inverse Laplace transform, we obtain the solution

$$y(t) = e^{-4t} * \cos^5 3t. \quad [1 \text{ mark}]$$

□

Example 32.57. Solve the ODE $y'' - 2y' - 8y = g(t)$, $y(0) = 1$, $y'(0) = 0$. Leave the solution in convolution form.

Solution.

$$\begin{aligned} s^2Y(s) - sy(0) - y'(0) - 2(sY(s) - y(0)) - 8Y(s) &= G(s) \\ \Rightarrow s^2Y(s) - s - 2(sY(s) - 1) - 8Y(s) &= G(s) \\ \Rightarrow Y(s) &= \frac{G(s) + s - 2}{s^2 - 2s - 8} = \frac{s - 2}{(s - 4)(s + 2)} + \frac{G(s)}{(s - 4)(s + 2)} \\ &= \frac{2}{3(s + 2)} + \frac{1}{3(s - 4)} + G(s) \left[\frac{1}{6} \left(\frac{1}{s - 4} - \frac{1}{s + 2} \right) \right] \\ \Rightarrow y(t) &= \frac{2}{3}e^{-2t} + \frac{1}{3}e^{4t} + \frac{1}{6}(e^{4t} - e^{-2t}) * g(t). \end{aligned}$$

□

“Transfer functions” are implemented in commercial software such as MATLAB/Simulink, ANSYS, Dymola, SimulationX, MapleSim, Wolfram SystemModeler, etc. However, there is only very limited implementation of transfer function in freely available software such as JModelica, OpenModelica, SciLab/Xcos. SciLab is only one which is still a bit user friendly and is freely available from <https://www.scilab.org/>. Unlike the commercial software such as MATLAB/Simulink, the Xcos has a terrible GUI based on the terrible Java/AWT widget and only the command line is usable.

eg:scilab

Example 32.58. Investigate how to represent the following transfer function in software such as Scilab (or MATLAB):

$$G(s) = \frac{5s + 10}{s^2 + 4s + 5}$$

Solution: In Scilab,

```
s = %s; // Define the complex number frequency parameter.
// Define the linear continuous transfer function
TF = syslin("c", 5*s + 10, s^2 + 4*s + 5);
disp(TF, "Transfer Function: ") // Wierd: LIFO
```

In MATLAB (or Octave with control package),

```
TF = tf([5 10], [1 4 5])
```

32.6 System Responses

For each system block (and the whole engineering system), it is important to know how the system responds to the input. This section investigates how a system responds to particular types of inputs. The system response and stability of a continuous-time LTI system [Åtröm and Murray, 2008, Section 2.2]:

$$\begin{aligned}\dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \\ \mathbf{y} &= \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u}\end{aligned}\tag{32.7}$$

e:lti-sys-ss

are associated with the poles and zeros of the transfer function $Y(s)/U(s)$ determined by [Chen, 2009]:

$$\begin{bmatrix} sX \\ Y \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} X \\ U \end{bmatrix}.\tag{32.8}$$

e:lti-sys-tf

Here A , B , C and D are constant matrices. The matrix A is called the *dynamics matrix*, the matrix B is called the *control matrix*, the matrix C is called the *sensor matrix* and the matrix D is called the *direct term*. The state variables $\mathbf{x}$ are normally not observable while the system output variables $\mathbf{y}$ are any system variable of interest.

The system response of the system (33.18) with a transfer function $G(s)$, characteristic equation $Q(s) = 0$, any given input $u(t)$ and any given initial conditions will be

$$y(t) = \mathcal{L}^{-1}\{Y(s)\} = \underbrace{\mathcal{L}^{-1}\{I(s)/Q(s)\}}_{\text{natural response}} + \underbrace{\mathcal{L}^{-1}\{G(s)U(s)\}}_{\text{forced response}}\tag{32.9}$$

e:lt:response

where “natural response” originates from the initial conditions.

For engineers to use transfer functions to characterise a system block, they extract the information of the “steady state” $\lim_{t \rightarrow \infty} y(t)$ and “initial state” $y(0)$ from $Y(s)$ by using the following two mathematical theorems.

Theorem 32.59 (Final-Value Theorem). *If $F(s) = \mathcal{L}\{f(t)\}$ and $f(t)$ has a limit at ∞ , then*

$$\lim_{t \rightarrow \infty} f(t) = \lim_{s \rightarrow 0} sF(s).$$

Theorem 32.60 (Initial-Value Theorem). *If $F(s) = \mathcal{L}\{f(t)\}$ and $f(t)$ has a limit at 0, then*

$$\lim_{t \rightarrow 0^+} f(t) = \lim_{s \rightarrow \infty} sF(s)$$

Engineers estimate three fundamental types of forced responses below using the initial- and final-value theorems:

1. Impulse response: This is important theoretically but violent in practice, i.e. using impulse input to determine the system response may cause the system to break down.

2. Step response: This is commonly use in practise because it is simple to quantify.
3. Frequency response: By using sinusoidal excitation (with different frequencies), the system's "steady-state" response with respect to the excitation is investigated.

A transfer function $G(s)$ can be derived for the system (32.8):

$$X = (sI - A)^{-1}BU \Rightarrow Y = C(sI - A)^{-1}BU + DU \Rightarrow G(s) = \frac{Y}{U} = C(sI - A)^{-1}B + D.$$

The final value theorem implies the step response of (32.8) leads to the DC gain matrix K : (32.8):

$$\lim_{t \rightarrow \infty} y(t) = \lim_{s \rightarrow 0} sY(s) = \lim_{s \rightarrow 0} sG(s) \cdot \frac{1}{s} = C(-A)^{-1}B + D =: K.$$

32.6.1 Impulse Function and Impulse Response

To model certain physical events mathematically, such as a sudden power surge caused by a bolt of lightning, or the dynamical effects of a hammer blow to a spring-mounted weight, it turns out that ordinary functions are ill suited for these purposes. An entirely new entity that is not a function at all [Schiff, 1999, p. 87], but something a mathematician called "generalised function" seems to be a better suited entity.

Definition 32.10. An *impulse function* or *Dirac delta function* $\delta(x)$ is a "generalised function" which satisfies

$$\int_{-\infty}^{\infty} \delta(t)f(t)dt = f(0).$$

for every function f over $\mathbb{R}$.

Let $\tau = t - a$ where $a \in \mathbb{R}$. By definition, we have

$$\int_{-\infty}^{\infty} \delta(t - a)f(t)dt = \int_{-\infty}^{\infty} \delta(\tau)f(\tau + a)d\tau = f(0 + a) = f(a).$$

This implies

$$\mathcal{L}\{\delta(t)\} = \int_{0^-}^{\infty} e^{-st}\delta(t)dt = \int_{-\infty}^{\infty} e^{-st}\delta(t)H(t)dt = e^0H(0) = 1.$$

Example 32.61. Find the Laplace transform of $\delta(t - a)$ where $a \geq 0$.

Solution. $\mathcal{L}\{\delta(t-a)\} = \int_{0^-}^{\infty} e^{-st}\delta(t-a)dt = \int_{-\infty}^{\infty} e^{-st}\delta(t-a)H(t)dt = e^{-sa}H(a) = e^{-as}$, $\text{Re}(s) > 0$. $\square$

Definition 32.11. For a linear system (33.18) or (32.5), the *impulse response* of the system is defined as the response (output) of the system to a Dirac delta function under zero initial conditions.

This means that $u = \delta(t)$ and (32.5) becomes

$$Y(s) = G(s)U(s) = G(s)\mathcal{L}\{\delta(t)\} = G(s).$$

This indicates that *from the impulse response, we can obtain the transfer function directly*.

Example 32.62. Investigate impulse response of the transfer function $G(s) = \frac{5s + 10}{s^2 + 4s + 5}$ using Scilab.

Solution: Mathematically, the impulse response is

$$y(t) = \mathcal{L}^{-1}\left\{\frac{5s + 10}{s^2 + 4s + 5}\right\} = \mathcal{L}^{-1}\left\{\frac{5(s + 2)}{(s + 2)^2 + 1}\right\} = e^{-2t}\mathcal{L}^{-1}\left\{\frac{5s}{s^2 + 1}\right\} = 5e^{-2t}\cos t.$$

The impulse cannot be calculated numerically but Scilab can handle it through simplification symbolically making it possible for numerical calculation:

```
exec("tfeg.sci")

// Impulse Response for a time range [0, 5]
t=0:0.01:5;
```

```

plot2d(t, csim('impulse',t,TF));
xlabel("Time [s]");           // Add a title and label axis
ylabel("y");
title("Impulse Response");
// Define a nice grid for the plot to make it easier to read
xgrid(1, 1, 10);

```

32.6.2 Step Function and Step Response

Although impulse response is useful theoretically, an engineering system normally cannot withstand an impulse. A step function may help us understand the system without destroying it.

Definition 32.12. A function f over $\mathbb{R}$ is called a *step function* (or *staircase function* if it can be written as

$$f: \mathbb{R} \rightarrow \mathbb{R}, \quad f(t) = \begin{cases} A_0, & -\infty < t < a_0, \\ A_1, & -a_0 < t < a_1, \\ \dots & \dots \\ A_n, & -a_{n-1} < t < \infty. \end{cases}$$

A *unit step function* is the Heaviside step function $H(t)$ defined in Definition [32.2](#). [def:heaviside](#)

Definition 32.13. Given a linear system [\(33.18\)](#) or [\(32.5\)](#), the *step response* of the system is defined as the output with respect to the input of unit step function under zero initial conditions. [e:ode:linearLT-TF](#)

According to [\(32.9\)](#), the output from a system [\(33.18\)](#) will normally consists of two parts. One part is input independent and only depends on the initial conditions, called the *zero input response*. The other part is directly related to the input function, called the *zero state response*. In control engineering, the zero input response is related to “transient response” while the zero state response is related to the “steady state”. The *transient response* of a system is that part of the output which is independent of the input. The *steady state* is the “additional” part of the output which arises due to the presence of the input. [Johnson, 1997, p. 60] [e:lt:response](#) [e:ode:linear](#)

In the following example, we will analyse the impulse response and step response of a second order ODE which is in normal form.

Example 32.63. Find the impulse response and step response for a linear system given by $y'' + 5y' + 6y = 5u$.

Solution. Taking Laplace transform, one arrives at

$$Y(s) = \frac{5}{s^2 + 5s + 6} U(s).$$

For impulse response, $u(t) = \delta(t)$, hence $U(s) = 1$.

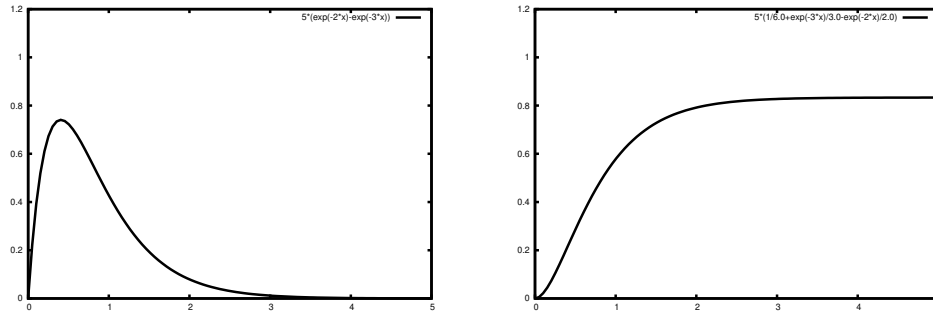
$$Y(s) = \frac{5}{s^2 + 5s + 6} = 5 \left[\frac{1}{s+2} - \frac{1}{s+3} \right] \Rightarrow y(t) = 5(e^{-2t} - e^{-3t})$$

For step response, $u(t) = H(t)$, hence $U(s) = \frac{1}{s}$.

$$Y(s) = \frac{5}{s^2 + 5s + 6} \left(\frac{1}{s} \right) = 5 \left[\frac{1}{3(s+3)} - \frac{1}{2(s+2)} + \frac{1}{6s} \right]$$

$$\therefore y(t) = 5 \left(\frac{1}{3} e^{-3t} - \frac{1}{2} e^{-2t} + \frac{1}{6} \right) \quad \square$$

Remark 32.64. The sketch of the impulse and step responses are given below:



The Scilab script to plot the step response is

```
s = %s;
TF = syslin("c", 5, s^2 + 5*s + 6);
t=0:0.01:5;
// 'csim' performs a general purpose continuous simulation function
// and can be used in various ways including step, impulse, u(t)
plot2d(t, csim('step', t, TF));
xlabel("Time [s]");
ylabel("y");
title("Step Response");
xgrid(1, 1, 10);
```

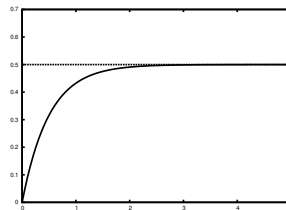
In general, the step response of a first order and second order linear systems are categorised as follows:

- First order system: $a_0 y' + a_1 y = H(t)$, $y(0) = 0$

The solution of this system is

$$y(t) = \mathcal{L}^{-1}\left\{\frac{1}{a_1 s} - \frac{a_0}{a_1(a_0 s + a_1)}\right\} = \frac{1}{a_1} - \frac{1}{a_1} e^{-(a_1/a_0)t}.$$

Consider $a_0 = 1$ and $a_1 = 2$, the step response is as follows:

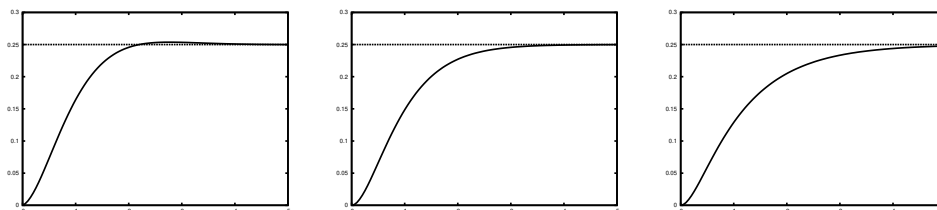


- Second order system: $a_0 y'' + a_1 y' + a_2 y = H(t)$, $y(0) = 0$

According to Topic 2, we can classify the system response and the form of the solution into

- $a_1^2 - 4a_0a_2 > 0$ (over-damped): $\frac{1}{a_2} - A_1 e^{-\lambda_1 t} - A_2 e^{-\lambda_2 t}$
- $a_1^2 - 4a_0a_2 = 0$ (critically-damped): $\frac{1}{a_2} - (A_1 t + A_2) e^{-\lambda t}$
- $a_1^2 - 4a_0a_2 < 0$ (under-damped): $\frac{1}{a_2} - e^{-\sigma t} (A_1 \cos \omega t + A_2 \sin \omega t)$.

Consider $a_0 = 1$ and $a_2 = 4$. The sketch of the response for $a_1 = 3$ (under-damped), $a_1 = 4$ (critically-damped) and $a_1 = 5$ (over-damped) are as follows:



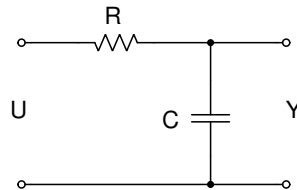
Engineers will look at the above category to determine the right transfer function to be used as a mathematical model for an engineering block.

32.6.3 Frequency Response

Many electrical systems need to operate for a given frequency range and to understand their operation, we need to investigate their frequency response.

Definition 32.14. Given a linear system (33.18) or (32.5), the *frequency response* of the system is defined as the output with respect to the input of sinusoidal signal under zero initial conditions.

We will demonstrate this with a low pass filter as follows.



The ODE that describes the relation between the input voltage u and the output voltage y can be derived as follows:

- Let $i = i(t)$ be the current. According to voltage analysis, we have

$$u(t) = R \cdot i(t) + y(t), \quad y(t) = \frac{1}{C} \int_{-\infty}^t i(t) dt.$$

- Taking derivative of y , we have $\frac{dy}{dt}(t) = \frac{1}{C} i(t)$
- The first ODE below is obtained: $u(t) = RC \frac{dy}{dt}(t) + y(t)$.

The frequency analysis can be carried by letting $u(t) = E_0 \sin \omega t$ where ω is the frequency. Under zero initial condition assumption, taking Laplace transform leads to

$$Y = \frac{1}{RCs + 1} \cdot \frac{E_0 \omega}{s^2 + \omega^2} = E_0 \left[\frac{\omega R^2 C^2}{(\omega^2 R^2 C^2 + 1)(RCs + 1)} - \frac{\omega RCs - \omega}{(\omega^2 R^2 C^2 + 1)(s^2 + \omega^2)} \right].$$

Taking inverse Laplace transform gives us the frequency response

$$y(t) = \frac{E_0 \omega RC}{(\omega^2 R^2 C^2 + 1)} \left[e^{-\frac{t}{RC}} - \cos \omega t + \frac{1}{\omega RC} \sin \omega t \right].$$

How would an engineer know if this is a low-pass filter (LPF)? An engineer recognises it as LPF from the coefficient $\frac{E_0 \omega RC}{(\omega^2 R^2 C^2 + 1)}$, it is small when ω is large, i.e. high frequency, so small frequency signals can pass through while high frequency signals cannot.

Instead of performing the above complicated analysis, in practice, engineers assume $s = i\omega$ and analyse the transfer function

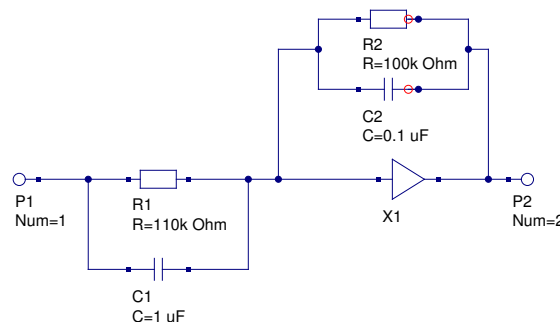
$$G(i\omega) = \frac{Y(i\omega)}{U(i\omega)} = \frac{1}{RCi\omega + 1}.$$

They are two ways to study the changes of transfer function with respect to the frequency ω . One way is by using the *Bode diagram* analysis, in which the *gain* (in dB) and *phase* (in degree) of the transfer function are plotted in pair. The other way is to plot the gain and phase on a plane.

The following example demonstrates how engineers model an electronic circuit.

eg:sussex

Example 32.65. The ‘control electronics’ for a magnetic suspension system has the following circuit system (drawn using Qucs).



Derive the circuit transfer function $Y(s)/U(s)$ and plot the Bode diagram using Scilab or Gnuplot.

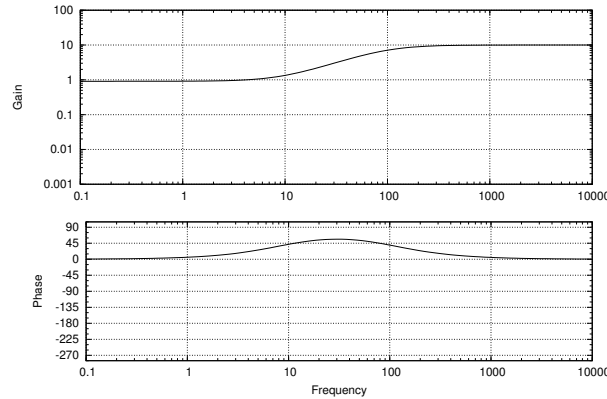
Solution: An ideal amplifier is normally assumed to have infinite gain (hence infinite resistance), therefore the input node to the amplifier is very closed to ground. Hence, the current analysis gives

$$\frac{U(s)}{Z_1(s)} = -\frac{Y(s)}{Z_2(s)} \Rightarrow \frac{Y(s)}{U(s)} = -\frac{Z_2(s)}{Z_1(s)} = -\frac{(\frac{1}{R_2} + sC_2)^{-1}}{(\frac{1}{R_1} + sC_1)^{-1}} = \frac{R_2(sC_1R_1 + 1)}{R_1(sC_2R_2 + 1)} = \frac{10(0.11s + 1)}{11(0.01s + 1)}.$$

To plot the Bode diagram, we need to let $s = i\omega$ and calculate amplitude and angle of the transfer function

$$\left| \frac{Y}{U}(i\omega) \right| = \left| \frac{10\sqrt{(0.11\omega)^2 + 1}}{11\sqrt{(0.01\omega)^2 + 1}} \right|, \quad \phi = \tan^{-1} 0.11\omega - \tan^{-1} 0.01\omega.$$

Using a complicated Gnuplot script, we can obtain a bode plot below:



For an engineer, he/she will recognise the circuit as a high pass filter. However, an engineer will prefer to obtain the above plot (which is easier) using Scilab:

```
s = %s;
TF = syslin("c", 10*(0.11*s + 1), 11*(0.01*s + 1));
f = 0.1:10000;           // Set-up the frequency range
bode(TF, f);             // Generate the Bode plot
title("Example 4.6.12"); // Add a title
```

A ubiquitous requirement in today's world of consumer electronics is to generate high performance, high frequency, reference clock signals. These signals are required for both digital and analogue systems with clock generation and communication standards being the most common applications. In the case of a digital system, the frequency of a clock signal may drift from its designed value in the presence of temperature and process variations, making a frequency compensation mechanism essential. For analogue systems, especially for wireless communication carrier signals, smarter and more accurate frequency adjustment approaches are required than for their digital counterparts [Ke, 2010].

Fundamentally, no matter whether designed for digital or analogues applications, a clock generation circuit must ensure the devices are working at their designed operational speed. Furthermore, besides the clock generation mechanism, many of these devices require their clock signal to be tuneable over a range of frequency bands. This requires the use of a *Voltage Controlled Oscillator* (VCO) and the controllability requirement demands the utilisation of the *Phase-Locked Loop* (PLL).

32.7 Formulation and Solution of Application-Specific Problems

Laplace transform is not very important in ODE theoretically because it lacks beautiful mathematical properties compared to Fourier transform, which can be generalised to *harmonic analysis* because of the group structure.

However, it is very important in *classical control engineering* because the transfer function obtain from Laplace transform is very convenient to handle using algebra. Virtually every system is underpinned by sophisticated *control systems* (a system block which is used to control and to stabilise the engineering system). Examples range from simple household products (temperature regulation in air-conditioners, thermostats in hot-water heaters, etc.), to more sophisticated systems, such as family car (which has hundreds of control loops), to large-scale systems, such as chemical plants, aircraft and manufacturing process. Beside industrial applications, “feedback” control mechanisms are central to the operations of

biological systems (important for medicine design), communication networks, national economies and even human interactions [Goodwin et al., 2000, p. 6].

During Second World War, Bode, Nyquist [Åtröm and Murray, 2008, Chapter 9], Nichols, Evans, etc. have pioneered simple graphical means for analysing single-output feedback control problems and these methods are now the mathematical contents of *classical control theory*.

In 1960's, the publication of work by Wiener, Kalman, etc. on optimal estimation and control promoted an alternate state space approach to control (some part of this theory will be investigated in Topic 5). This work allowed multivariable problems to be treated in a unified fashion. This set of developments is loosely termed *modern control theory*.

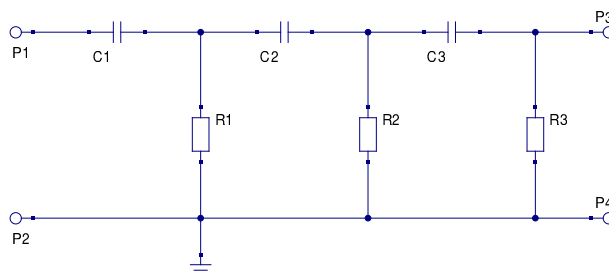
By the 1980's various approaches to control had reached a sophisticated level, and emphasis then shifted to other related issues, including the effect of model error on the performance of feedback controllers. This can be classified as the period of *robust control theory* [Goodwin et al., 2000, p. 9].

Engineers build a Laplace transform table to carry out the analysis of simple linear systems and their system responses. The modelling of engineering system is also based of a table of basic components. Note that the current $i = \frac{dq}{dt}$.

System	Physical Law	Kinetic energy	Potential energy	Dissipation	State
Mechanical	$\sum \text{Force} = 0$, $\sum \text{Moment} = 0$	Mass, $F = m\ddot{x}$	Spring, $F = kx$	Damper, $F = \gamma\dot{x}$	Displacement, x
Electrical	$\sum \text{Current} = 0$	Inductor, $V = L \frac{d^2q}{dt^2}$	Capacitor, $V = \frac{1}{C}q$	Resistor, $V = R \frac{dq}{dt}$	Charge (capacity), q
Thermal	$\sum \text{Energy} = \text{constant}$	None	Thermal Capacitor, $Q = C\Delta T$	Heat Transfer, $\Delta T = R\Phi$	Heat, Q
Hydraulic Fluid	$\sum \text{Mass} = \text{constant}$	Fluid Inertance, $\Delta p = L \frac{d^2v}{dt^2}$	Fluid Capacitor, $M = Cx\Delta p$	None	Flow, Q

With this, we can transform a system block or an engineering system to systems of ODE. The following simple example demonstrates this.

Example 32.66. Consider the following circuit, which is popular for phase shifting in oscillators.



Let $C_1 = C_2 = C_3 = C$ and $R_1 = R_2 = R_3 = R$. Formulate the systems of ODE which characterises the input u and output y .

Solution: Let voltages across the capacitors C_1, C_2, C_3 be x_1, x_2, x_3 . Then by Kirchorff voltage law, we have

$$u = x_1 + x_2 + x_3 + y \quad (*)$$

Note that “Voltage $x_j = q_j/C_j$ ” for capacitors $j = 1, 2, 3$, therefore

$$q_j = \frac{x_j}{C_j} \Rightarrow \frac{dq_j}{dt} = i_j = C_j \frac{dx_j}{dt}.$$

By Kirchorff current law,

$$\begin{aligned} C_1 \dot{x}_1 &= \frac{u - x_1}{R_1} + C_2 \dot{x}_2 \\ C_2 \dot{x}_2 &= \frac{u - x_1 - x_2}{R_2} + C_3 \dot{x}_3 \\ C_3 \dot{x}_3 &= \frac{u - x_1 - x_2 - x_3}{R_3} \end{aligned} \quad (\dagger)$$

By writing down (†) and (*) in matrix form, we obtain

$$\begin{bmatrix} C_1 & -C_2 & 0 \\ 0 & C_2 & -C_3 \\ 0 & 0 & C_3 \end{bmatrix} \dot{\mathbf{x}} = \begin{bmatrix} -1/R_1 & 0 & 0 \\ -1/R_2 & -1/R_2 & 0 \\ -1/R_3 & -1/R_3 & -1/R_3 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1/R_1 \\ 1/R_2 \\ 1/R_3 \end{bmatrix} u$$

$$y = [-1 \quad -1 \quad -1] \mathbf{x} + [1]u.$$

where $\mathbf{x} = [x_1, x_2, x_3]^T$.

After simplification taking inverse matrix, we obtain

$$\dot{\mathbf{x}} = \frac{1}{RC} \begin{bmatrix} -3 & -2 & -1 \\ -2 & -2 & -1 \\ -1 & -1 & -1 \end{bmatrix} \mathbf{x} + \frac{1}{RC} \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} u$$

$$y = [-1 \quad -1 \quad -1] \mathbf{x} + [1]u.$$

In general, many engineering linear systems that one encounters can be written in the matrix form similar to the above example:

$$\begin{cases} \dot{\mathbf{x}} = A\mathbf{x} + B\mathbf{u}, & \mathbf{x}(0) = \mathbf{x}_0 \\ \mathbf{y} = C\mathbf{x} + D\mathbf{u} \end{cases} \quad (32.10)$$

where $\mathbf{x} = (x_1(t), \dots, x_n(t))^T$ is the *system state* vector, $\mathbf{y} = (y_1(t), \dots, y_m(t))^T$ is the output vector and $\mathbf{u} = (u_1(t), \dots, u_k(t))^T$ is *control input* vector. A is a square matrix while B, C, D are appropriate rectangular matrices.

Taking Laplace transform on the ODE in (32.10), we have

$$s\mathbf{X}(s) - \mathbf{x}(0) = A\mathbf{X}(s) + B\mathbf{U}(s) \Rightarrow (s\mathbf{I} - A)\mathbf{X}(s) = \mathbf{x}(0) + B\mathbf{U}(s)$$

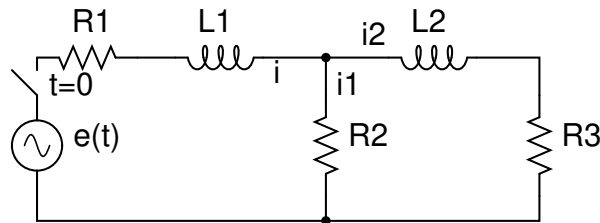
which can be solved by taking inverse Laplace transform

$$\begin{aligned} \mathbf{x}(t) &= \mathcal{L}^{-1}\{(s\mathbf{I} - A)^{-1}\}\mathbf{x}(0) + \mathcal{L}^{-1}\{(s\mathbf{I} - A)^{-1}B\mathbf{U}(s)\} \\ &= \underbrace{\Phi(t)\mathbf{x}(0)}_{\text{zero input response}} + \underbrace{\Phi(t) * B\mathbf{u}(t)}_{\text{zero state response}} = \underbrace{\Phi(t)\mathbf{x}(0)}_{\text{natural response}} + \underbrace{\int_0^t \Phi(t-\tau)B\mathbf{u}(\tau)d\tau}_{\text{forced response}} \end{aligned} \quad (32.11)$$

where $\Phi(t) := \mathcal{L}^{-1}\{(s\mathbf{I} - A)^{-1}\}$ is the *state transition matrix*.

In the following, we will solve a few initial value problem of systems of ODE.

Example 32.67. Using Kirchhoff's Law, it is possible to describe the following circuit (with $e(t) = 200V$, $R_1 = 20\Omega$, $L_1 = 0.5H$, $L_2 = 1H$, $R_2 = 8\Omega$ and $R_3 = 10\Omega$)



using the differential-algebraic equation:
$$\begin{cases} 20i(t) + 0.5i'(t) + 8i_1(t) = 200H(t), \\ i_2'(t) + 10i_2(t) - 8i_1(t) = 0, \\ i(t) - i_1(t) - i_2(t) = 0 \end{cases}$$

which can be simplified to the follow ODEs:
$$\begin{cases} 28i_1(t) + 20i_2(t) + 0.5i_1'(t) + 0.5i_2'(t) = 200H(t), \\ -8i_1(t) + i_2'(t) + 10i_2(t) = 0. \end{cases}$$

Solve the above ODE (write the numerical values in two decimal places) assuming zero initial conditions.

Solution: Taking Laplace transform we have

$$\begin{aligned} (s + 56)I_1 + (s + 40)I_2 &= \frac{400}{s}, \\ -8I_1 + (s + 10)I_2 &= 0. \end{aligned}$$

Solving it gives us

$$I_2 = \frac{3200}{(s^2 + 74)s + 880} = \frac{40}{11s} + \frac{1.22}{s + 59.11} - \frac{4.86}{s + 14.89}$$

$$I_1 = \frac{400s + 4000}{s^3 + 74s^2 + 880s} = \frac{50}{11s} - \frac{50s - 700}{11(s^2 + 74s + 880)} = \frac{50}{11s} - \frac{7.49}{s + 59.11} + \frac{2.98}{s + 14.89}.$$

Taking inverse Laplace transform, we have

$$i_2 \approx \frac{40}{11} + 1.22e^{-59.11t} - 4.86e^{-14.89t},$$

$$i_1 \approx \frac{50}{11} - 7.52e^{-59.11t} + 2.97e^{-14.89t}.$$

nal-2016.05q4c

Example 32.68 (Final Exam Sept 2016, Q4(c)). Use Laplace transform to find the solution $(x(t), y(t))^T$ of the system of ODEs:

$$\dot{x} = x + y + e^{-3t}$$

$$\dot{y} = x - y + 3e^{-t}$$

where $x(0) = 1, y(0) = -1$.

(8 marks)

Solution. Taking Laplace transform: $\begin{cases} sX - 1 = X + Y + \frac{1}{s+3} \\ sY + 1 = X - Y + \frac{3}{s+1} \end{cases}$ [2 marks]

Rearranging it to

$$\begin{bmatrix} s-1 & -1 \\ -1 & s+1 \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} 1 + \frac{1}{s+3} \\ \frac{3}{s+1} - 1 \end{bmatrix} = \begin{bmatrix} \frac{s+4}{s+3} \\ \frac{2-s}{s+1} \end{bmatrix} \quad [1 \text{ mark}]$$

and taking inverse, we have

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \frac{1}{s^2 - 2} \begin{bmatrix} s+1 & 1 \\ 1 & s-1 \end{bmatrix} \begin{bmatrix} \frac{s+4}{s+3} \\ \frac{2-s}{s+1} \end{bmatrix} \quad [1 \text{ mark}]$$

Expanding it leads to

$$X = \frac{s^3 + 5s^2 + 8s + 10}{(s^2 - 2)(s+3)(s+1)} = \frac{\frac{5\sqrt{8}+20}{\sqrt{8}(1+\sqrt{2})(3+\sqrt{2})}}{s - \sqrt{2}} + \frac{\frac{5\sqrt{8}-20}{\sqrt{8}(1-\sqrt{2})(3-\sqrt{2})}}{s + \sqrt{2}} + \frac{-\frac{2}{7}}{s+3} + \frac{-3}{s+1} \quad [1+2=3 \text{ marks}]$$

$$Y = -\frac{s^3 - s^2 - 12s + 2}{(s^2 - 2)(s+3)(s+1)} = \frac{\frac{5}{(1+\sqrt{2})(3+\sqrt{2})}}{s - \sqrt{2}} + \frac{\frac{5}{(1-\sqrt{2})(3-\sqrt{2})}}{s + \sqrt{2}} + \frac{\frac{1}{7}}{s+3} + \frac{6}{s+1}$$

Therefore

$$x(t) = \frac{5\sqrt{8}+20}{\sqrt{8}(1+\sqrt{2})(3+\sqrt{2})}e^{\sqrt{2}t} + \frac{5\sqrt{8}-20}{\sqrt{8}(1-\sqrt{2})(3-\sqrt{2})}e^{-\sqrt{2}t} - \frac{2}{7}e^{-3t} - 3e^{-t}$$

$$y(t) = \frac{5}{(1+\sqrt{2})(3+\sqrt{2})}e^{\sqrt{2}t} + \frac{5}{(1-\sqrt{2})(3-\sqrt{2})}e^{-\sqrt{2}t} + \frac{1}{7}e^{-3t} + 6e^{-t} \quad [1 \text{ mark}]$$

□

Example 32.69. Solve the system of ODEs

$$x' + y' + 5x + 3y = e^{-t}, \quad 2x' + y' + x + y = 3, \quad x(0) = 2, y(0) = 1.$$

Solution. Taking Laplace transform, we obtain

$$sX - 2 + sY - 1 + 5X + 3Y = \frac{1}{s+1} \Rightarrow (s+5)X + (s+3)Y = \frac{1}{s+1} + 3$$

$$2sX - 4 + sY - 1 + X + Y = \frac{3}{s} \Rightarrow (2s+1)X + (s+1)Y = \frac{3}{s} + 5$$

$$\Rightarrow \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} s+5 & s+3 \\ 2s+1 & s+1 \end{bmatrix}^{-1} \begin{bmatrix} \frac{3s+4}{s} \\ \frac{5s+3}{s} \end{bmatrix} = \begin{bmatrix} -\frac{11}{6(s+2)} - \frac{9}{2s} + \frac{25/3}{s-1} \\ \frac{11}{2(s+2)} + \frac{1}{2(s+1)} + \frac{15}{2s} - \frac{25}{2(s-1)} \end{bmatrix}$$

$$\text{Hence, } \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} -\frac{11}{6}e^{-2t} - \frac{9}{2} + \frac{25}{3}e^t \\ \frac{11}{2}e^{-2t} + \frac{1}{2}e^{-t} + \frac{15}{2} - \frac{25}{2}e^t \end{bmatrix} \quad \square$$

Example 32.70. [Johnson, 1997, p. 46] Consider an undamped mass-spring mechanical system with the following equations of motion

$$\begin{aligned} m_1 y_1'' &= k_2(y_2 - y_1) - k_1 y_1, & y_1(0) &= y_2(0) = 0 \\ m_2 y_2'' &= -k_2(y_2 - y_1), & y_1'(0) &= 0, \quad y_2'(0) = 24, \end{aligned}$$

where $k_1/m_1 = 12$, $k_2/m_1 = k_2/m_2 = 8$. Find the solution of the initial value problem.

Solution. Using the given values, the equations of motion are

$$\begin{aligned} y_1'' &= -20y_1 + 8y_2, \\ y_2'' &= 8y_1 - 8y_2. \end{aligned}$$

Taking Laplace transforms we obtain

$$\begin{cases} s^2 Y_1 = -20Y_1 + 8Y_2, \\ s^2 Y_2 - 24 = 8Y_1 - 8Y_2 \end{cases} \Rightarrow \begin{cases} (s^2 + 20)Y_1 - 8Y_2 = 0, \\ -8Y_1 + (s^2 + 8)Y_2 = 24 \end{cases}$$

Solving for

$$Y_1 = \frac{192}{(s^2 + 20)(s^2 + 8) - 64} = \frac{192}{(s^2 + 4)(s^2 + 24)} = 4.8 \frac{2}{s^2 + 2^2} - \frac{9.6}{\sqrt{24}} \cdot \frac{\sqrt{24}}{s^2 + (\sqrt{24})^2}.$$

Therefore $y_1 \approx 4.8 \sin 2t - 1.96 \sin \sqrt{24}t$.

and

$$\begin{aligned} y_2 &= \frac{1}{8}(\ddot{y}_1 + 20y_1) = \frac{1}{8} \left(-19.2 \sin 2t + 47.04 \sin \sqrt{24}t + 96 \sin 2t - 39.2 \sin \sqrt{24}t \right) \\ &= 9.6 \sin 2t + 0.98 \sin \sqrt{24}t. \end{aligned}$$

□

Example 32.71. Solve the system $\begin{cases} x'' - 2x' + 3y' + 2y = 4, & x(0) = x'(0) = 0, \\ 2y' - x' + 3y = 0, & y(0) = 0. \end{cases}$

Solution. Taking Laplace transform:

$$\begin{aligned} s^2 X - 2sX + 3sY + 2Y &= \frac{4}{s} \Rightarrow (s^2 - 2s)X + (3s + 2)Y = \frac{4}{s} \\ 2sY - sX + 3Y &= 0 \Rightarrow -sX + (2s + 3)Y = 0 \end{aligned}$$

Solving the linear algebra problem:

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} s^2 - 2s & 3s + 2 \\ -s & 2s + 3 \end{bmatrix}^{-1} \begin{bmatrix} \frac{4}{s} \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{6(s+2)} - \frac{7}{2s} - \frac{3}{s^2} + \frac{10}{3(s-1)} \\ \frac{1}{3(s+2)} - \frac{1}{s} + \frac{2}{3(s-1)} \end{bmatrix}.$$

$$\text{Hence, the solution is } \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} \frac{1}{6}e^{-2t} - \frac{7}{2} - 3t + \frac{10}{3}e^t \\ \frac{1}{3}e^{-2t} - 1 + \frac{2}{3}e^t \end{bmatrix} \quad \square$$

Example 32.72 (Final Exam Sept 2014, Q4(b)). Use Laplace transform to find the solution $x(t)$ for the system of ODEs:

$$\begin{aligned} x' + 2y' + 4x + 2y &= e^{-t}, & x(0) &= 2, \\ 2x' + y' + x + y &= 1, & y(0) &= 1. \end{aligned} \quad (13 \text{ marks})$$

Solution. Taking Laplace transform:

$$\begin{cases} sX - x(0) + 2(sY - y(0)) + 4X + 2Y = \frac{1}{s+1} \\ 2(sX - x(0)) + sY - y(0) + X + Y = \frac{1}{s} \end{cases} \quad [2 \text{ marks}]$$

After rearranging, we have
$$\begin{cases} (s+4)X + (2s+2)Y = 4 + \frac{1}{s+1} \\ (2s+1)X + (s+1)Y = 5 + \frac{1}{s} \end{cases} \dots\dots\dots [2 \text{ marks}]$$

and solving the algebra equation leads to

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \frac{1}{(s+4)(s+1) - (2s+2)(2s+1)} \begin{bmatrix} s+1 & -2s-2 \\ -2s-1 & s+4 \end{bmatrix} \begin{bmatrix} \frac{4s+5}{s+1} \\ \frac{5s+1}{s} \end{bmatrix}. \quad [2 \text{ marks}]$$

The solution $x(t)$ can be calculated as follows:

$$x(t) = \mathcal{L}^{-1} \left\{ \frac{(4s+5) - \frac{(2s+2)(5s+1)}{s}}{s^2 + 5s + 4 - 4s^2 - 6s - 2} \right\} \quad [1 \text{ mark}]$$

$$= \mathcal{L}^{-1} \left\{ \frac{6s^2 + 7s + 2}{s(3s^2 + s - 2)} \right\} = \mathcal{L}^{-1} \left\{ \frac{6s^2 + 7s + 2}{s(3s-2)(s+1)} \right\} \quad [2 \text{ marks}]$$

$$= \mathcal{L}^{-1} \left\{ \frac{-1}{s} + \frac{42/5}{3s-2} + \frac{1/5}{s+1} \right\} \quad [2 \text{ marks}]$$

$$= -1 + \frac{14}{5}e^{2/3t} + \frac{1}{5}e^{-t}. \quad [2 \text{ marks}]$$

□

Using transfer function, engineers can handle the design of simple closed-loop feedback control system of a system block $G(s)$ and an feedback element $H(s)$ (e.g. for PID control, $H(s) = K + \frac{I}{s} + Ds$) through

$$Y(s) = \frac{G(s)}{1 + G(s)H(s)}U(s) \quad (32.12)$$

e:feedback

using various strategies such as Walter R. Evans' root locus analysis, Bode analysis and Nyquist analysis.

Despite the convenience of transfer function (or Laplace transfor) for simple systems, a large system of ODEs is better modelled through the state-space model, which is discussed in Topic 5, and the numerical computation of ODE will be discussed in Chapter 45.

32.8 Working with Maxima

desolve() uses Laplace transform to solve linear ODEs or systems of linear ODEs.

To solve Example 32.46, Example 32.67 and Example 32.68:

```
desolve(diff(y(t),t)+3*y(t)=(sin(3*t))^2, y(t));

desolve([
  20*i(t) + 0.5*diff(i(t),t) + 8*i1(t) = 200,
  diff(i2(t),t) + 10*i2(t) - 8*i1(t) = 0,
  i(t)-i1(t)-i2(t) = 0
], [i(t), i1(t), i2(t)]);

desolve([
  diff(x(t),t) = x(t) + y(t) + exp(-3*t),
  diff(y(t),t) = x(t) - y(t) + 3*exp(-t)
], [x(t), y(t)]);
```

Chapter 33

Systems of ODEs

ch:ode-sys

Very few real-life engineering system can be modelled successfully by a single ODE. Typically, we need to keep track of several “states” at the same time, and thus need several real-valued functions, which are described by a system of ODEs [Schroers, 2011, 27]. For example, in electrical circuit theory, we want to monitor the voltages or currents of many nodes; in mechanical systems, we want to know the forces of every nodes; in chemical engineering, we want to know the concentrations of various reactors; in control engineering, we want to know the states in a system which can be observed and controlled. In the numerical methods, when applying the finite-difference approximation to the space of a PDE leads to a systems of ODEs. This technique is called the *method of lines*.

Modern engineers find that an appropriate *continuous time state space model*:

$$\begin{aligned}\dot{x}_1 &= f_1(t, x_1, x_2, \dots, x_n; u_1, \dots, u_m), \\ \dot{x}_2 &= f_2(t, x_1, x_2, \dots, x_n; u_1, \dots, u_m), \\ &\dots \\ \dot{x}_n &= f_n(t, x_1, x_2, \dots, x_n; u_1, \dots, u_m)\end{aligned}\tag{33.1}$$

e:state-space

can model an engineering system accurately.

Mathematicians study the (33.1) without the system inputs $u_1, \dots, u_m$, i.e. a *system of ODEs*:

$$\begin{aligned}\dot{x}_1 &= f_1(t, x_1, x_2, \dots, x_n), \\ \dot{x}_2 &= f_2(t, x_1, x_2, \dots, x_n), \\ &\dots \\ \dot{x}_n &= f_n(t, x_1, x_2, \dots, x_n)\end{aligned}\tag{33.2}$$

e:ode:sys

where $x_1, \dots, x_n$ are the n states of a system. In topics 1 to 3, the notations y and x were used to denote the state and “time” respectively. The notations are inherited from Calculus and the theory of two-point boundary value problems. However, in modern ODE and engineering textbooks, the notations x and t are used instead and this topic will follow this convention.

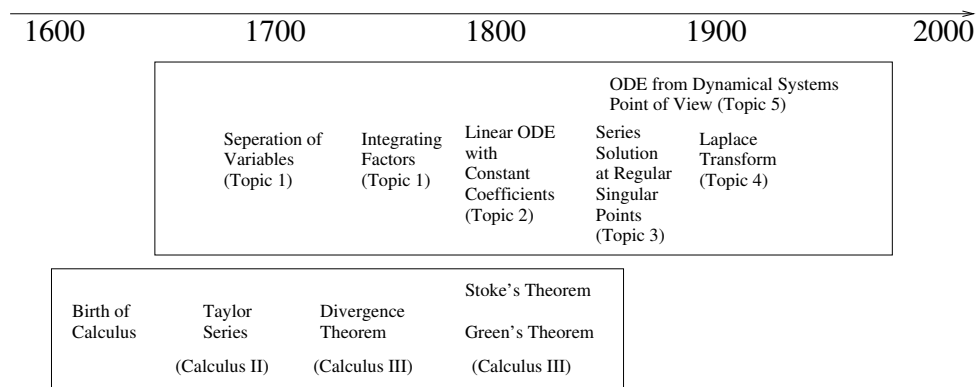
In Section 33.1, we will roughly review the techniques we have learned so far and how could we generalise them to solve (33.2). We will discuss a “qualitative technique” for investigating (33.2) from Section 33.2 to Section 33.12. This technique was first introduced by a French mathematician Henri Poincaré around 1892 to solve the *three-body problem* and was later developed into a subject called *dynamical systems* by many mathematicians.

Due to time constraint, very limited examples are discussed, for much more examples, consult Be-tounes [2010], Logan [2006, Chapters 5 and 6] and Robinson [2004, Part V and VI]. For more theoretical treatment, Arnold [1992] and Chicone [1999] are good choices.

33.1 Development of Solution Methods

sec:review

From Topics 1 to 4, we have learned several solution techniques as well as their limitations — they are too many ODEs which could not be solved analytically, i.e. written as an explicit solution in the form of composites of well-known functions. If we refer to the time-line of Calculus, https://en.wikipedia.org/Timeline_of_calculus_and_mathematics_analysis,



we will see that this course covers the mathematical knowledge of over two-hundred years.

The solution techniques in Topic 1 and 2 belong to a class called the *solution by quadrature*, which is closely related to the notion of *integrability* of ODE in modern mathematics. The series solution in Topic 3 could be generalised to power series solutions for higher dimensional system (which is not very useful due to slow convergence). It can also be generalised to the *theory of complex ODE* and is closely related to *resurgent analysis*.

The Laplace transform technique in Topic 4 could be generalised to *transform methods* and works with hyperfunction theory [Graf, 2010].

Topic 5 is the basic theory in ODE as a development of *three-body problem* (https://en.wikipedia.org/wiki/Three-body_problem). Since the time of Newton, the three-body problem was a major source of development of analysis — Euler, Clairaut, d'Alembert, Laplace, Lagrange, Jacobi, Cauchy, etc. worked on it. At the end of the nineteenth century, Poincaré opened a new era of research by introducing geometry, topological and probabilistic methods to understand qualitatively the incredibly complicated behaviour of most of the solutions of this problem. At the same time, he analysed the methods used by the astronomers in order to understand the short-term motions using divergent series [Diacu, 2000].

According to Chenciner [2012], the landmark in Poincaré's works on the three-body problem is the famous Memoir *Sur le problème des trois corps et les équations de la dynamique*, winner in 1889 of the prize given on the occasion of the 60th birthday of the King of Sweden, and even more *Les méthodes nouvelles de la mécanique céleste* whose three volumes, totalling almost 1300 pages, appear respectively in 1892, 1893 and 1899. Vastly enlarging the scope of the Memoir, this extraordinary work, which encompasses and develops most of Poincaré's previous researches on the three-body problem, is the source of a major part of the modern theory of *Dynamical Systems: normal forms, exponents, invariant manifolds, homoclinic and heteroclinic solutions, integrability* (http://en.wikipedia.org/wiki/Integrable_system), *analytic non-integrability, divergence of the perturbation series and exponentially small splitting of separatrices, variational equations and integral invariants, generating functions, recurrence theorem, surfaces of section and return maps, twisting property*, all of them are part of the present landscape and they paved the way for *bifurcation studies and the theory of singularities, symbolic dynamics, invariant measures and ergodic theory, KAM theory, symplectic geometry*, and also a wealth of computer experiments.

33.2 System of First Order ODEs and Phase Space

sec:phase-space

Since it is impossible to obtain an analytic solution for (33.2) in general, mathematicians study the local structure of a solution and then extend and piece the local solutions together to form a global solution. This idea leads to the development of various “geometrical” concepts and the mathematical subject called *dynamical systems* which will be briefly stated below.

Definition 33.1 (Anosov and Arnold [1988]). A *direction field* in a domain $U \in \mathbb{R}^n$ is a map which associates to each point $\mathbf{x} \in U$ a straight line passing through $\mathbf{x}$.

A *vector field*, defined on a n -dimensional domain $U \subset \mathbb{R}^n$, is a map which associates to each point $\mathbf{x} \in U$ a vector $\mathbf{f}(\mathbf{x})$ in $\mathbb{R}^n$.

def:nonauto

Definition 33.2 (Anosov and Arnold [1988]). Suppose $\mathbf{f}$ is a vector field which depends on time t , i.e. $\mathbf{f} : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$. A *non-autonomous ODE* or *non-autonomous system* is an equation where the right-hand vector field is the time-dependent $\mathbf{f}$:

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}), \quad (t, \mathbf{x}) \in \mathbb{R} \times U, \quad U \subset \mathbb{R}^n. \quad (33.3)$$

e:ode:nonauto

In component form, this is the system of ODEs we have seen in (33.2), where $\mathbf{x} = \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix}$ and

$$\mathbf{f}(t, \mathbf{x}) = \begin{bmatrix} f_1(t, x_1, \dots, x_n) \\ \vdots \\ f_n(t, x_1, \dots, x_n) \end{bmatrix}.$$

Example 33.1. The system of first order ordinary differential equations below

$$\begin{cases} \dot{x}_1 = x_1 - 3x_2 + t^2 \\ \dot{x}_2 = tx_1 \sin x_2 \end{cases}$$

is a non-autonomous system because the time-variable t appears on the right-hand side.

The explicit n th order ODE which we have studied in Topic 2

$$y^{(n)} = f(x, y, y', \dots, y^{(n-1)}) \quad (33.4)$$

can be transformed to a system of first order ODEs (33.3) as follows:

$$\begin{cases} y_1' = y_2 \\ y_2' = y_3 \\ \vdots \\ y_{n-1}' = y_n = f(x, y_1, y_2, \dots, y_{n-1}) \end{cases} \quad (33.5)$$

where $y_1 = y$. We call (33.5) the *state space representation* of the higher order ODE.

Example 33.2. [Schroers, 2011, 29] Let G and M be real constants. Write the following system of ODEs or ODEs as first order systems:

$$\begin{aligned} 1. \quad \begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{bmatrix} &= -\frac{GM}{(x^2 + y^2 + z^2)^{3/2}} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \dots\dots\dots \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \\ \dot{u} \\ \dot{v} \\ \dot{w} \end{bmatrix} = \begin{bmatrix} u \\ v \\ w \\ -GMx/(x^2 + y^2 + z^2)^{3/2} \\ -GM y/(x^2 + y^2 + z^2)^{3/2} \\ -GM z/(x^2 + y^2 + z^2)^{3/2} \end{bmatrix} \\ 2. \quad x'' + x' + 4x &= 0 \dots\dots\dots \begin{aligned} \dot{x} &= y \\ \dot{y} &= -4x - y \end{aligned} \\ 3. \quad x\ddot{x} + (\dot{x})^2 &= 0 \dots\dots\dots \begin{aligned} \dot{x} &= y \\ \dot{y} &= -y^2/x \end{aligned} \end{aligned}$$

Definition 33.3. An *initial value problem* (of a system of ODEs) is a set of inter-related first order ODEs (33.2) together with the initial conditions

$$x_1(t_0) = x_1^0, x_2(t_0) = x_2^0, \dots, x_n(t_0) = x_n^0, \quad (33.6)$$

where time t_0 is in an interval $I = (a, b)$.

The possibility of writing any ODE as a first order system (33.3) is important because it provides a universal framework for proving theorems (the Picard-Lindelöf Theorem in Topic 1 can be generalised for (33.3)) and for numerical work.

Theorem 33.3. Let each of the functions (33.2) and its partial derivatives $\frac{\partial F_i}{\partial x_j}$ of the initial value problems of a system of ODEs be continuous in a domain $\mathbb{R} \times U$ containing an initial point $(t_0, x_1^0, \dots, x_n^0)$. Then there is an interval $|t - t_0| < h$ in which there exists a unique solution $x_1 = \varphi_1(t), \dots, x_n = \varphi_n(t)$ of the system of ODEs (33.2).

According to Henrici [1962], when an appropriate numerical method is applied to solve the first order ODE, the truncation error is not growing and the rounding error is found to decrease most of the time.

Transforming (33.4) to (33.5) is important in the solution of higher-order IVP in *numerical methods (UECM3033)*. This is how one use the commercial software MATLAB to solve higher-order IVP. In the following example, plotutils and Python with SciPy are used to illustrate how the procedure works.

Example 33.4. A hoop of radius a rotates with constant angular velocity ω about a vertical diameter. A bead on the hoop has a position vector which makes an angle θ with the vertical. It is given that the differential equation for the motion of the bead is

$$\ddot{\theta} + \left(\frac{g}{a} - \omega^2 \cos \theta\right) \sin \theta = 0. \quad (33.7)$$

If $g = 10$, $\omega = 10$, $a = 1$, investigate the software which can solve the IVP and plot the numerical solution of θ for the range $[0, 5]$ given the initial condition $\theta(0) = 0.1$ and $\dot{\theta}(0) = 0$.

Solution: To use a numerical software to solve the ODE (33.7), we need to express it in the form (33.5) by letting $\dot{\theta} = v$. Then the IVP becomes

$$\begin{aligned} \dot{\theta} &= v, & \theta(0) &= 0.1 \\ \dot{v} &= -\left(\frac{g}{a} - \omega^2 \cos \theta\right) \sin \theta, & v(0) &= 0 \text{ because } v(t) = \dot{\theta}(t) \end{aligned} \quad (33.8)$$

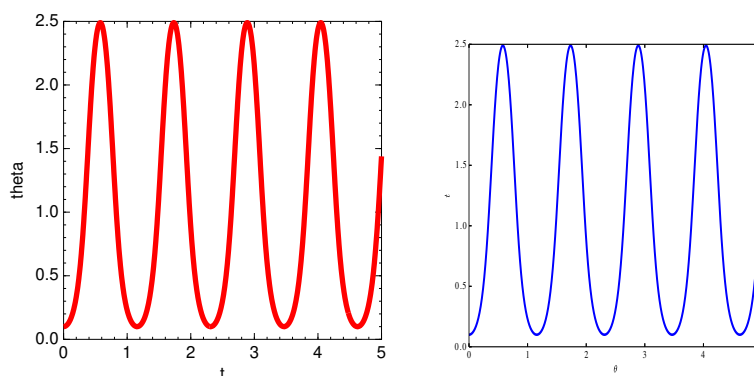
In plotutils, we will first define the constants a , g and ω and then setup (33.8) and then ask the program ode to compute the numerical value θ over the range $[0, 5]$:

```

1 # This example simulates a bead sliding on a smooth circular wire. The
2 # output displays the height of the bead as a function of time.
3 #
4 # You may run this example by doing:
5 #
6 #     ode < bead.ode | graph -T X -C
7 #
8 # or alternatively, to get a real-time plot,
9 #
10 #     ode < bead.ode | graph -T X -C -x 0 5 -y 0 2.5
11 #
12
13 a = 1 # radius of circular wire
14 g = 10 # acceleration due to gravity
15 w = 10 # angular velocity of circular wire
16
17 the' = vthe
18 vthe' = (w^2)*sin(the)*cos(the) - (g/a)*sin(the)
19 the = 0.1
20 vthe = 0
21
22 print t, the
23 step 0,5

```

Plotting the numerical values, we obtain the left graph below:



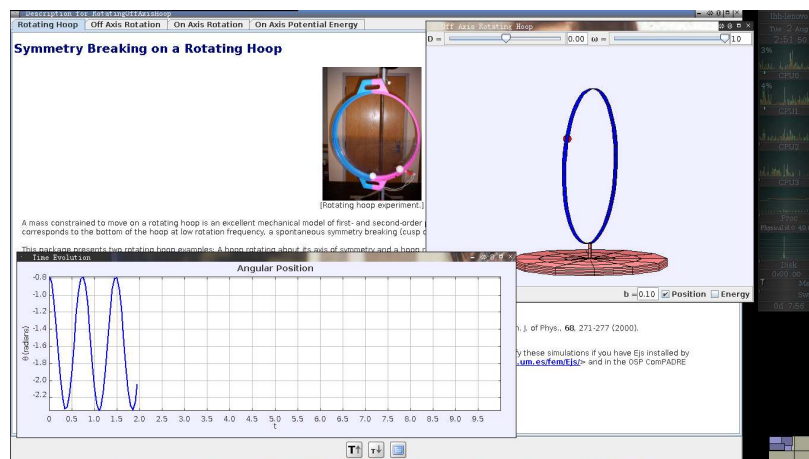
If one is using Python (or MATLAB), then one needs to be familiar with the working environment. With the right modules, it is easy to generate the same graph (the right picture in the previous page) as plotutils as illustrated below.

```

1 from scipy.integrate import odeint
2 from numpy import linspace, zeros, sin, cos
3 from matplotlib.pyplot import plot, show, xlabel, ylabel, savefig
4
5 def bead_motion(state,t):
6     a = 1.0; g = 10.0; w = 10.0
7     dx = zeros(2)
8     theta, v = state[0], state[1]
9     dx[0] = v
10    dx[1] = -(g/a - w**2*cos(theta))*sin(theta)
11    return dx
12
13 tgrid = linspace(0, 5, 501)
14 xgrid = odeint(bead_motion, [0.1,0.0], tgrid)
15 theta = xgrid[:,0]
16 ylabel("$t$"); xlabel(r"$\theta$");
17 plot(tgrid, theta, linewidth=3); savefig("plotbead2.eps"); #show()

```

Remark: The more advanced tool for building interactive application is the EjsS (Java, <http://www.um.es/fem/EjsWiki/pmwiki.php>). However, an advanced Java programming skill is required. An implementation of bead system in EjsS (which is not very flexible) is given in <http://www.opensourcephysics.org/items/detail.cfm?ID=7890> and a screen shot is shown.



The most advanced tools are all programming language based such as ArtiSynth (Java), SimBody (C++), etc.

def:auto

Definition 33.4 (Anosov and Arnold [1988]). An *autonomous (systems of) ODE* associated with a vector field $\mathbf{f}$ which does not depend on time t is

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}), \quad \mathbf{x} \in U. \quad (33.9)$$

e:ode:auto

where $\mathbf{x}$ is a function of “time” $t \in \mathbb{R}$. The domain U is called the *phase space* of the ODE (33.9), and the direct product $\mathbb{R} \times U$ is called the *extended phase space*.

e:ode:auto

Remark 33.5. When U is an open subset of $\mathbb{R}^n$ of dimension n , (33.9) may be expressed as

$$\begin{aligned} \frac{dx_1}{dt} &= f_1(x_1, x_2, \dots, x_n) \\ \frac{dx_2}{dt} &= f_2(x_1, x_2, \dots, x_n) \\ &\vdots \\ \frac{dx_n}{dt} &= f_n(x_1, x_2, \dots, x_n) \end{aligned}$$

in which the functions $f_1, \dots, f_n$ do not depend explicitly on the independent variable t .

Example 33.6. The system

$$\begin{cases} \dot{x} = y \\ \dot{y} = f(x, y) \end{cases}$$

is an autonomous system because the right-hand side does not contain any time-variable t .

Definition 33.5. A *solution* (or *trajectory* or *path*) of the ODE (33.3) (or (33.2), which includes (33.9) as a special case) is a differentiable map φ of an interval $I = (a, b) \subset \mathbb{R}$ to the phase space U such that for all $\tau \in I$, we have

$$\left. \frac{d}{dt} \varphi(t) \right|_{t=\tau} = \mathbf{f}(\tau, \varphi(\tau)).$$

The solution which pass through a point $\mathbf{x}_0$ is denoted as $\varphi(\mathbf{x}_0, t)$. The set of solutions (or trajectories) in the phase space is called the *phase portrait*.

Arnold [1992] gave a very nice “singular pendulum” example:

$$\begin{aligned} \dot{\theta} &= \omega \\ \dot{\omega} &= -A \sin \theta \end{aligned}$$

to illustrate the usefulness of the “phase space” concept. Here θ is the angle of the pendulum with respect to the direction of gravitational force. The phase space of the single pendulum is a cylinder:

$$(\theta, \omega) \in U = S^1 \times \mathbb{R}$$

where S^1 is a circle (0 to 360 degree).

Definition 33.6 (Anosov and Arnold [1988]). An *integral curve* of (33.3) or (33.9) is the graph of a solution, i.e. $\{(t, \varphi(t)) : t \in I\}$. A *phase curve* is the projection of an integral curve on to the phase space along the t -axis.

Example 33.7. Consider the non-autonomous system

$$\dot{x} = t - x + 1. \quad (33.10)$$

1. Write down the solution of (33.10).
2. Write down the integral curve of (33.10).
3. Rewrite (33.10) as a autonomous system and sketch the direction field for the autonomous system derived from (33.10) in the region $(-3, 3) \times (-3, 3)$. (Check out http://en.wikipedia.org/wiki/Slope_field)

Solution: Note that the phase space $U = \mathbb{R}$.

1. A solution of (33.10) has the form

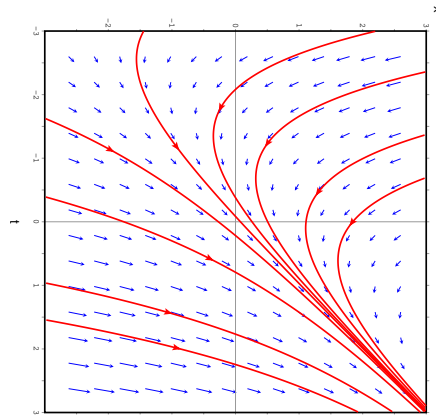
$$\varphi_C : \mathbb{R} \rightarrow \mathbb{R}, \quad t \mapsto \varphi_C(t) = x(t) = t + Ce^{-t}$$

where C is a fixed constant.

2. The integral curves are $\{(t, t + Ce^{-t}) : t \in \mathbb{R}\}$.
3. A standard procedure for transforming a non-autonomous system (33.3) to a autonomous system is

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{f}(t, \mathbf{x}) = t - x + 1, \\ \dot{t} &= 1, \end{aligned}$$

where the phase space $\tilde{U} = \mathbb{R}^2$ and a sketch of the directional field in the region $(-3, 3) \times (-3, 3)$ is



33.2.1 Critical Point

From Example 33.7, we can see that for an autonomous system, all solutions form a bundle of “curves” in the phase space. In general, it is difficult for us to know the solution without solving them numerically. However, there is a kind of “degenerate” curve which is a “point”. With calculus, it is possible for us to know how normal curves behave near a “point”, which we call “critical point”.

Definition 33.7 (Critical Point). A *critical point* (or an *equilibrium point*) of the autonomous system (33.9) are the points $\mathbf{x}_c$ in the phase space U such that

$$\mathbf{f}(\mathbf{x}_c) = \mathbf{0}.$$

Remark 33.8. The trajectory of an equilibrium solution consists of just a single point, i.e. the critical point. Critical points are points where the motion of the state vector is at rest. Hence, an critical point is also called a *steady state point*.

Example 33.9. Find the critical points of the following autonomous system.

1. [E&P, 9.1, Q23] $\dot{x} = 2x - y$, $\dot{y} = x - 3y$;

Solution: Solving

$$2x - y = 0$$

$$x - 3y = 0$$

gives $x = y = 0$, which means that $(0, 0)$ is the (only) critical point.

2. [Nagle, P274;5.4, Q3] $\dot{x} = x - y$, $\dot{y} = x^2 + y^2 - 1$.

Solution.

$$x - y = x^2 + y^2 - 1 = 0 \Rightarrow x = y = \pm 1/\sqrt{2}$$

□

Example 33.10. Sketch the phase portrait of the autonomous system

$$\dot{x} = y,$$

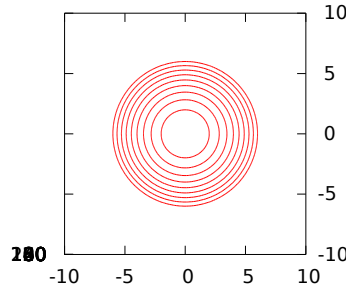
$$\dot{y} = -x.$$

around the critical point.

Solution: It is easy to find that $(0, 0)$ is the critical point.

$$\frac{\dot{y}}{\dot{x}} = \frac{dy}{dx} = -\frac{x}{y} \Rightarrow \frac{y^2}{2} = -\frac{x^2}{2} + \frac{C}{2} \Rightarrow x^2 + y^2 = C.$$

The trajectories $(x(t), y(t))$ consist of the origin $(0, 0)$ and the circles $x^2 + y^2 = C > 0$ and the phase portrait around the critical point is



33.2.2 (Lyapunov) Stability of Critical Point

Engineers are concerned about the stability of a critical point. A critical point $\mathbf{x}_c$ is “stable” if any trajectory that begins near (within δ of) the point $\mathbf{x}_c$ remains near (within ϵ of) this point. When all the trajectories that start close to a stable critical point actually approach it as $t \rightarrow \infty$, then the critical point is asymptotically stable. Formally, we have the following definitions.

Definition 33.8 ((Lyapunov) Stability). A critical point (steady state) $\mathbf{x}_c$ of (33.9) is said to be *(Lyapunov) stable* if for any $\epsilon > 0$, there exists $\delta > 0$ such that for all $\mathbf{x}_0$ with $\|\mathbf{x}_0 - \mathbf{x}_c\| < \delta$, we have $\|\varphi(\mathbf{x}_0, t) - \mathbf{x}_c\| < \epsilon$ for all $t \geq 0$.

Definition 33.9. A critical point $\mathbf{x}_c$ is said to be *unstable* if it is not (Lyapunov) stable.

Definition 33.10 (Asymptotic Stability). A critical point $\mathbf{x}_c$ is said to be *locally asymptotically stable* if it is Lyapunov stable and there is a $\rho > 0$ such that for all $\mathbf{x}_0$ with $\|\mathbf{x}_0 - \mathbf{x}_c\| < \rho$ we have $\|\varphi(\mathbf{x}_0, t) - \mathbf{x}_c\| \rightarrow 0$ as $t \rightarrow \infty$.

Roughly speaking, all solutions starting close enough to a stable critical point $\mathbf{x}_c$ remain close to it for all subsequent time. All solutions starting sufficiently close to an asymptotically stable critical point not only remain close to it for all subsequent time but, in fact, approach $\mathbf{x}_c$ as $t \rightarrow \infty$.

Note that asymptotic stability implies stability. However, the converse is not true as illustrated by the following example.

Example 33.11. Consider the two-dimensional linear system

$$\dot{x} = y, \quad \dot{y} = -x.$$

Show that the critical point $\mathbf{x}_c = \mathbf{0}$ is stable but not asymptotically stable.

Proof: The general solution of the linear system is

$$\mathbf{x}(t) = C_1 \begin{bmatrix} \cos t \\ -\sin t \end{bmatrix} + C_2 \begin{bmatrix} \sin t \\ \cos t \end{bmatrix}$$

For any $\epsilon > 0$, there is a $\delta = \epsilon$ such that when

$$\|\mathbf{x}(0) - \mathbf{x}_c\| = \|\mathbf{x}(0)\| = \left\| \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} \right\| = \sqrt{C_1^2 + C_2^2} < \delta,$$

we have

$$\begin{aligned} \|\mathbf{x}(t) - \mathbf{x}_c\| &= \sqrt{(C_1 \cos t + C_2 \sin t)^2 + (-C_1 \sin t + C_2 \cos t)^2} \\ &= \sqrt{C_1^2 \cos^2 t + C_2^2 \sin^2 t + C_1^2 \sin^2 t + C_2^2 \cos^2 t} = \sqrt{C_1^2 + C_2^2} < \delta = \epsilon. \end{aligned}$$

By definition, $\mathbf{x}_c$ is stable and since $\lim_{t \rightarrow \infty} \mathbf{x}(t) \neq \mathbf{x}_c$, by definition, $\mathbf{x}_c$ is not asymptotically stable.

Example 33.12. The origin of

$$\begin{aligned}\dot{x} &= -x - y + x(x^2 + y^2), \\ \dot{y} &= x - y + y(x^2 + y^2)\end{aligned}$$

is locally asymptotically stable.

33.3 Homogeneous Linear System with Constant Coefficients

A special case of autonomous system (33.9) is the linear autonomous equation (or system) with constant matrix A

$$\dot{\mathbf{x}} = A\mathbf{x}, \quad \mathbf{x} \in \mathbb{R}^n. \quad (33.11)$$

This kind of system constitutes almost the only large class of ODE for which there is a complete mathematical theory. This theory, which is essentially a branch of linear algebra, enables all linear autonomous equation to be solved completely [Anosov and Arnold, 1988, 19].

Similar to the solution of linear ODE with constant coefficients in Topic 2, we assume that the solution has the form $\mathbf{x}(t) = \mathbf{v}e^{\lambda t}$, where $\mathbf{v}$ is a non-zero constant vector and λ is yet to be determined. Substituting it into (33.11) we have

$$\dot{\mathbf{x}} = \lambda \mathbf{v}e^{\lambda t} = A\mathbf{v}e^{\lambda t} \Rightarrow \boxed{A\mathbf{v} = \lambda \mathbf{v}}.$$

Since $\mathbf{v}$ is a non-zero constant vector, the problem of solving a homogeneous system with constant coefficients is reduced to an *eigenvalue problem*.

Before we proceed to the discussion of the theory of homogeneous linear system with constant coefficients (33.11), we will recall a bit of the knowledge from *Fundamentals of Linear Algebra (UECM1313)*.

Definition 33.11 (Eigenvalues and Eigenvectors). Let A be a constant $n \times n$ matrix. The *eigenvalues* of A are the values of λ for which the system of equations

$$(A - \lambda I)\mathbf{v} = \mathbf{0}$$

has at least one nonzero solution $\mathbf{v}$. They can be found by solving the *characteristic equation*

$$|A - \lambda I| = 0.$$

The nonzero solutions $\mathbf{v}$ are called the *eigenvectors* of A corresponding to λ .

Recall from linear algebra that if A is a matrix there are three cases for the eigenvalues:

1. All eigenvalues are real and different.
2. Some of the eigenvalues are complex conjugate pairs.
3. Some eigenvalues are repeated.

33.3.1 Different Real Eigenvalues

If all n real eigenvalues $\lambda_1, \dots, \lambda_n$ for $A\mathbf{v} = \lambda\mathbf{v}$ are different, then the corresponding eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_n$ are linearly independent. By using linear algebra theory, we can have derived the following theorem.

Theorem 33.13 (General Solution of (33.11)). If the $n \times n$ constant matrix A has n linearly independent eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ corresponding to the real eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$, then $\{e^{\lambda_1 t} \mathbf{v}_1, e^{\lambda_2 t} \mathbf{v}_2, \dots, e^{\lambda_n t} \mathbf{v}_n\}$ constitutes a fundamental set of the homogeneous system (33.11). The general solution of (33.11) has the following form

$$\mathbf{x}(t) = C_1 e^{\lambda_1 t} \mathbf{v}_1 + C_2 e^{\lambda_2 t} \mathbf{v}_2 + \dots + C_n e^{\lambda_n t} \mathbf{v}_n \quad (33.12)$$

where $C_1, C_2, \dots, C_n$ are arbitrary constants.

A diagonalizable matrix Λ could be obtained by performing a *similarity transform* to A , i.e. there is a matrix P such that $AP = P\Lambda$, where matrices $P = [\mathbf{v}_1 \cdots \mathbf{v}_n]$ and $\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$. Let

$\mathbf{x} = P\mathbf{y}$, (33.11) can be transformed to a form involving Λ :

$$\dot{\mathbf{x}} = P\dot{\mathbf{y}} = AP\mathbf{y} \Rightarrow \dot{\mathbf{y}} = P^{-1}AP\mathbf{y} = \Lambda\mathbf{y}.$$

Since $\dot{\mathbf{y}} = \Lambda\mathbf{y}$ is a decoupled linear system and the solutions are easy to obtain and the general solution (33.12) for the linear system (33.11) can be obtained.

Let us look at some examples.

Example 33.14. Find the general solution of the system

$$\begin{cases} \dot{x} = -2x + y \\ \dot{y} = -4x + 3y \end{cases} \quad (33.13)$$

by diagonalizing the coefficient matrix.

Solution: First, we identify the matrix $A = \begin{bmatrix} -2 & 1 \\ -4 & 3 \end{bmatrix}$ from the system of linear equations and then find the eigenvalues and eigenvectors of A :

$$\begin{vmatrix} -2 - \lambda & 1 \\ -4 & 3 - \lambda \end{vmatrix} = \lambda^2 - \lambda - 2 = 0 \Rightarrow \lambda_1 = -1, \lambda_2 = 2, \quad \mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$$

Let $P = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$. By introducing the transform $\mathbf{x} = P\mathbf{y}$ such that

$$\dot{\mathbf{x}} = P\dot{\mathbf{y}} = A\mathbf{x} = AP\mathbf{y} \Rightarrow \dot{\mathbf{y}} = P^{-1}AP\mathbf{y} \Rightarrow \begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

It is easy to solve for y_1 and y_2 using the techniques in Topic 1 and obtain

$$y_1(t) = C_1 e^{-t}, \quad y_2(t) = C_2 e^{2t}.$$

The general solution of (33.13) is

$$\mathbf{x}(t) = P\mathbf{y}(t) = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} C_1 e^{-t} \\ C_2 e^{2t} \end{bmatrix} = \begin{bmatrix} C_1 e^{-t} + C_2 e^{2t} \\ C_1 e^{-t} + 4C_2 e^{2t} \end{bmatrix}.$$

Example 33.15. Find the general solution of $\dot{\mathbf{x}} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \mathbf{x}$ where

Solution: After solving the characteristic equation

$$\begin{vmatrix} 2 - \lambda & 3 \\ 2 & 1 - \lambda \end{vmatrix} = \lambda^2 - 3\lambda - 4 = 0$$

we obtain the eigenvalues $\lambda_1 = -1$ and $\lambda_2 = 4$.

The eigenvector corresponding to $\lambda = -1$ is $\mathbf{v}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

The eigenvector corresponding to $\lambda = 4$ is $\mathbf{v}_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$.

Hence, the general solution is

$$\mathbf{x}(t) = C_1 \mathbf{v}_1 e^{-t} + C_2 \mathbf{v}_2 e^{4t} = C_1 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-t} + C_2 \begin{bmatrix} 3 \\ 2 \end{bmatrix} e^{4t}.$$

eg:real-eigenv2

Example 33.16. Solve the ODE $\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$.

Solution. Solving the characteristic equation

$$\begin{vmatrix} 1-\lambda & 1 \\ 4 & 1-\lambda \end{vmatrix} = \lambda^2 - 2\lambda - 3 = 0,$$

we have $\lambda_1 = -1$ and $\lambda_2 = 3$.

For $\lambda = -1$, the corresponding eigenvector is

$$\begin{bmatrix} 2 & 1 \\ 4 & 2 \end{bmatrix} \mathbf{v}_1 = \mathbf{0} \Rightarrow \mathbf{v}_1 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

For $\lambda = 3$, the corresponding eigenvector is

$$\begin{bmatrix} -2 & 1 \\ 4 & -2 \end{bmatrix} \mathbf{v}_2 = \mathbf{0} \Rightarrow \mathbf{v}_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Therefore, the general solution is

$$\mathbf{x} = C_1 \mathbf{x}_1 + C_2 \mathbf{x}_2 = C_1 e^{-t} \begin{bmatrix} 1 \\ -2 \end{bmatrix} + C_2 e^{3t} \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

□

Example 33.17 (Final Exam Sept 2014, Q2(c)). Find the general solution of the following system of ODEs:

$$\dot{\mathbf{x}} = \begin{bmatrix} 3 & -2 \\ 2 & -2 \end{bmatrix} \mathbf{x}$$

where $\mathbf{x} = (x_1, x_2)^T$ and $\dot{\mathbf{x}} = (\frac{dx_1}{dt}, \frac{dx_2}{dt})^T$. (10 marks)

Solution. To find the general solution, we first calculate

$$\begin{vmatrix} 3-\lambda & -2 \\ 2 & -2-\lambda \end{vmatrix} = 0 \Rightarrow -6 - \lambda + \lambda^2 + 4 = \lambda^2 - \lambda - 2 = (\lambda - 2)(\lambda + 1) = 0 \quad [3 \text{ marks}]$$

and obtain the eigenvalues $\lambda = 2, -1$ [1 mark]

When $\lambda = 2$, the eigenvector is

$$\begin{bmatrix} 3-2 & -2 \\ 2 & -2-2 \end{bmatrix} \mathbf{v}_1 = \mathbf{0} \Rightarrow \mathbf{v}_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \quad [2 \text{ marks}]$$

When $\lambda = -1$, the eigenvector is

$$\begin{bmatrix} 3+1 & -2 \\ 2 & -2+1 \end{bmatrix} \mathbf{v}_2 = \mathbf{0} \Rightarrow \mathbf{v}_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad [2 \text{ marks}]$$

The general solution is

$$\mathbf{x} = C_1 \begin{bmatrix} 2 \\ 1 \end{bmatrix} e^{2t} + C_2 \begin{bmatrix} 1 \\ 2 \end{bmatrix} e^{-t}. \quad [2 \text{ marks}]$$

□

Example 33.18 (Final Exam Sept 2016, Q5(a)). Consider the following system of differential equations

$$\begin{aligned}x' &= -2x + 3y \\ y' &= x - 4y\end{aligned}$$

- (i) Find a general solution to the above system using linear algebra method. (10 marks)

Solution. First, we determine the eigenvalues of the system

$$\begin{vmatrix} -2-\lambda & 3 \\ 1 & -4-\lambda \end{vmatrix} = \lambda^2 + 6\lambda + 8 - 3 = (\lambda + 5)(\lambda + 1) = 0 \quad [2 \text{ marks}]$$

Therefore, the eigenvalues are -5 and -1 [1 mark]

For $\lambda = -5$, the eigenvector $\mathbf{v}_1$ can be found by solving

$$\begin{bmatrix} 3 & 3 \\ 1 & 1 \end{bmatrix} \mathbf{v}_1 = \mathbf{0} \Rightarrow \mathbf{v}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad [1.5 \text{ marks}]$$

For $\lambda = -1$, the eigenvector $\mathbf{v}_2$ can be found by solving

$$\begin{bmatrix} -1 & 3 \\ 1 & -3 \end{bmatrix} \mathbf{v}_2 = \mathbf{0} \Rightarrow \mathbf{v}_2 = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \quad [1.5 \text{ marks}]$$

Let $P = \begin{bmatrix} 3 & 1 \\ 1 & -1 \end{bmatrix}$, $A = \begin{bmatrix} -2 & 3 \\ 1 & -4 \end{bmatrix}$ and let $\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix} = P\mathbf{y}$. The system can be transformed to

$$P\mathbf{y}' = AP\mathbf{y} \Rightarrow \mathbf{y}' = P^{-1}AP\mathbf{y} = \begin{bmatrix} -1 & 0 \\ 0 & -5 \end{bmatrix} \mathbf{y} \quad [2 \text{ marks}]$$

The general solutions of this system and the original system are respectively

$$\mathbf{y} = \begin{bmatrix} C_1 e^{-t} \\ C_2 e^{-5t} \end{bmatrix} \Rightarrow \mathbf{x} = P\mathbf{y} \Rightarrow \begin{cases} x(t) = 3C_1 e^{-t} + C_2 e^{-5t} \\ y(t) = C_1 e^{-t} - C_2 e^{-5t} \end{cases} \quad [2 \text{ marks}]$$

□

- (ii) Find the trajectory of the system which starts from the state $(4,6)$ at $t = 0$ and describe the behaviour of the trajectory after a long time. (3 marks)

Solution. From part (i), we have

$$\begin{cases} x(0) = 3C_1 + C_2 = 4 \\ y(0) = C_1 - C_2 = 6 \end{cases} \Rightarrow C_1 = 2.5, C_2 = -3.5$$

The trajectory is $\begin{cases} x(t) = 7.5e^{-t} - 3.5e^{-5t} \\ y(t) = 2.5e^{-t} + 3.5e^{-5t} \end{cases}$ [2 marks]

After a long time, the trajectory approaches the origin $(0,0)^T$ [1 mark]

□

33.3.2 Some Complex Eigenvalues

Since A is a real-valued matrix, if some of the eigenvalues for $A\mathbf{v} = \lambda\mathbf{v}$ are complex, their eigenvalues and eigenvectors are always come in as conjugate pairs. This means that if $\lambda = \alpha + i\beta$ is an eigenvalue with eigenvector $\mathbf{v} = \mathbf{a} + i\mathbf{b}$ for $A\mathbf{v} = \lambda\mathbf{v}$ then $\lambda^* = \alpha - i\beta$ is an eigenvalue with the corresponding eigenvector $\mathbf{v}^* = \mathbf{a} - i\mathbf{b}$. Thus the pair of solutions are

$$\mathbf{x}_1 = \mathbf{v}_1 e^{(\alpha+i\beta)t}, \quad \mathbf{x}_2 = \mathbf{v}_2 e^{(\alpha-i\beta)t}$$

and by substitution and let $\widetilde{C}_1 = C_1 + C_2$ and $\widetilde{C}_2 = C_1 i - C_2 i$, we obtain

$$\begin{aligned}\mathbf{x} &= C_1 \mathbf{x}_1 + C_2 \mathbf{x}_2 = e^{\alpha t} (C_1 \mathbf{v}_1 (\cos \beta t + i \sin \beta t) + C_2 \mathbf{v}_2 (\cos \beta t - i \sin \beta t)) \\ &= e^{\alpha t} [(\widetilde{C}_1 \mathbf{a} + \widetilde{C}_2 \mathbf{b}) \cos \beta t + (\widetilde{C}_2 \mathbf{a} - \widetilde{C}_1 \mathbf{b}) \sin \beta t] \\ &= \widetilde{C}_1 e^{\alpha t} (\mathbf{a} \cos \beta t - \mathbf{b} \sin \beta t) + \widetilde{C}_2 e^{\alpha t} (\mathbf{a} \sin \beta t + \mathbf{b} \cos \beta t)\end{aligned}$$

Summarising our analysis, we obtain the following theorem.

Theorem 33.19 (Real Solution Corresponding to Complex Eigenvalues). *If $\alpha \pm i\beta$ are complex conjugate eigenvalues of the real matrix A with corresponding eigenvectors $\mathbf{a} \pm i\mathbf{b}$, then*

$$\mathbf{x}_1 = e^{\alpha t}(\mathbf{a} \cos \beta t - \mathbf{b} \sin \beta t) \text{ and } \mathbf{x}_2 = e^{\alpha t}(\mathbf{b} \cos \beta t + \mathbf{a} \sin \beta t), \quad -\infty < t < \infty$$

are linearly independent real-valued solutions of (33.11).

Example 33.20. Solve $\dot{\mathbf{x}} = \begin{bmatrix} 1 & 5 \\ -1 & -1 \end{bmatrix} \mathbf{x}$.

Solution: First, we find the eigenvalues:

$$\begin{vmatrix} 1 - \lambda & 5 \\ -1 & -1 - \lambda \end{vmatrix} = \lambda^2 - 1 + 5 = 0 \Rightarrow \lambda = \pm 2i$$

When $\lambda = \pm 2i$, the corresponding eigenvector(s) are

$$\begin{bmatrix} 1 \mp 2i & 5 \\ -1 & -1 \mp 2i \end{bmatrix} \mathbf{v} = \mathbf{0} \Rightarrow \mathbf{v} = \begin{bmatrix} 5 \\ -1 \pm 2i \end{bmatrix} = \begin{bmatrix} 5 \\ -1 \end{bmatrix} \pm i \begin{bmatrix} 0 \\ 2 \end{bmatrix} =: \mathbf{a} \pm i\mathbf{b}$$

respectively. The general solution is

$$\mathbf{x}(t) = c_1 \left(\begin{bmatrix} 5 \\ -1 \end{bmatrix} \cos 2t - \begin{bmatrix} 0 \\ 2 \end{bmatrix} \sin 2t \right) + c_2 \left(\begin{bmatrix} 0 \\ 2 \end{bmatrix} \cos 2t + \begin{bmatrix} 5 \\ -1 \end{bmatrix} \sin 2t \right)$$

Example 33.21. Solve the linear system $\begin{cases} \dot{x}_1 = -x_1 - x_2, \\ \dot{x}_2 = 2x_1 - x_2. \end{cases}$

Solution. Solving

$$\begin{vmatrix} -1 - \lambda & -1 \\ 2 & -1 - \lambda \end{vmatrix} = \lambda^2 + 2\lambda + 3 = 0$$

we have $\lambda = -1 \pm \sqrt{2}i$.

For $\lambda = \alpha + \beta i = -1 + \sqrt{2}i$, we have

$$\begin{bmatrix} -1 + 1 - \sqrt{2}i & -1 \\ 2 & -1 + 1 - \sqrt{2}i \end{bmatrix} \mathbf{v}_1 = \mathbf{0} \Rightarrow \mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + i \begin{bmatrix} 0 \\ -\sqrt{2} \end{bmatrix}$$

For $\lambda = -1 - \sqrt{2}i$, we have $\mathbf{v}_2 = \mathbf{v}_1^*$.

Thus two linearly independent solutions are

$$\mathbf{x}_1 = e^{\alpha t}(\mathbf{a} \cos \beta t - \mathbf{b} \sin \beta t) = e^{-t} \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix} \cos \sqrt{2}t - \begin{bmatrix} 0 \\ -\sqrt{2} \end{bmatrix} \sin \sqrt{2}t \right)$$

$$\mathbf{x}_2 = e^{\alpha t}(\mathbf{b} \cos \beta t + \mathbf{a} \sin \beta t) = e^{-t} \left(\begin{bmatrix} 0 \\ -\sqrt{2} \end{bmatrix} \cos \sqrt{2}t + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \sin \sqrt{2}t \right)$$

Hence the general solution is

$$\mathbf{x} = C_1 \mathbf{x}_1 + C_2 \mathbf{x}_2 = e^{-t} \begin{bmatrix} C_1 \cos \sqrt{2}t + C_2 \sin \sqrt{2}t \\ C_1 \sqrt{2} \sin \sqrt{2}t - C_2 \sqrt{2} \cos \sqrt{2}t \end{bmatrix}.$$

□

33.3.3 Repeated Eigenvalues

If one of the eigenvalue is repeated eigenvalue for $A\mathbf{v} = \lambda\mathbf{v}$ then the system (33.11) is called *degenerate*. To better classify the “level” of degeneration, we will introduce the following notions.

Definition 33.12 (Algebraic Multiplicity). An eigenvalue is of (*algebraic*) *multiplicity* k if it is a k -fold root of the characteristic equation $|A - \lambda I| = 0$.

Definition 33.13. An eigenvalue of multiplicity k is said to be *complete* if it has k linearly independent associated eigenvectors. An eigenvalue of multiplicity k is called *defective* if it is not complete.

If every eigenvalue of an $n \times n$ matrix A is complete, then A will have n linearly independent eigenvectors associated with its eigenvalues. For this type of $n \times n$ matrices A , it may be possible to find m linearly independent eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_m$ corresponding to λ of multiplicity $m \leq n$. The general solution of the system will contain linear combination $C_1 \mathbf{v}_1 e^{\lambda t} + C_2 \mathbf{v}_2 e^{\lambda t} + \dots + C_m \mathbf{v}_m e^{\lambda t}$.

Let us look at an example with an eigenvalue of multiplicity 2 but is complete.

Example 33.22. Solve $\dot{\mathbf{x}} = \begin{bmatrix} 0 & 0 & -2 \\ 1 & 2 & 1 \\ 1 & 0 & 3 \end{bmatrix} \mathbf{x}$.

Solution. Solving the characteristic equation

$$\begin{vmatrix} -\lambda & 0 & -2 \\ 1 & 2-\lambda & 1 \\ 1 & 0 & 3-\lambda \end{vmatrix} = -\lambda(2-\lambda)(3-\lambda) - 2(-(2-\lambda)) = (2-\lambda)(\lambda^2 - 3\lambda + 2) = 0$$

gives the eigenvalues $\lambda_1 = 1, \lambda_2 = 2, \lambda_3 = 2$.

When $\lambda = 1$,

$$\begin{bmatrix} -1 & 0 & -2 \\ 1 & 1 & 1 \\ 1 & 0 & 2 \end{bmatrix} \mathbf{v}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{v}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \mathbf{v}_1 = \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}$$

When $\lambda = 2$,

$$\begin{bmatrix} -2 & 0 & -2 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \mathbf{v} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{v} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \mathbf{v} = \underbrace{\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}}_{\mathbf{v}_2} p_1 + \underbrace{\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}}_{\mathbf{v}_3} p_2$$

where p_1 and p_2 are two parameters.

The general solution is therefore

$$\mathbf{x}(t) = C_1 \mathbf{v}_1 e^{\lambda_1 t} + C_2 \mathbf{v}_2 e^{\lambda_2 t} + C_3 \mathbf{v}_3 e^{\lambda_2 t} = C_1 \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix} e^t + C_2 \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} e^{2t} + C_3 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} e^{2t}.$$

□

If some of the eigenvalues of A are defective, then our usual method will produce fewer than the needed n linearly independent solutions. To construct the general solution of the system, we need to find the “*missing solutions*” using a different approach.

The system is two-fold degenerate if the same eigenvalue appears twice and the eigenvalue is called a *double-eigenvalue*. In this case we might have two linear independent eigenvectors if A is Hermitian or one eigenvector if A is non-Hermitian.

If we have only one eigenvector $\mathbf{v}$ for two repeating eigenvalues λ , we are missing one of the two linearly independent solution vectors in the fundamental matrix. Assume that the second solution to the ODE has the form $\mathbf{x}_2 = (t\mathbf{v}_1 + \mathbf{v}_2)e^{\lambda t}$, the missing eigenvector $\mathbf{v}_2$ could be found by solving the *associated eigenvalue problem*

$$(A - \lambda I)\mathbf{v}_2 = \mathbf{v}_1.$$

The resulting vector $\mathbf{v}_2$ is called a *generalised eigenvector*.

Eigenvalue of Multiplicity Two ($m = 2$)

If A has only one eigenvector associated with an eigenvalue λ of multiplicity two, then we can find two linearly independent solutions of the form

$$\begin{aligned}\mathbf{x}_1 &= \mathbf{v}_1 e^{\lambda t} \\ \mathbf{x}_2 &= (\mathbf{v}_1 t + \mathbf{v}_2) e^{\lambda t}\end{aligned}$$

where the columns vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ must satisfy

$$(A - \lambda \mathbf{I})\mathbf{v}_1 = \mathbf{0}, \quad (A - \lambda \mathbf{I})\mathbf{v}_2 = \mathbf{v}_1.$$

Example 33.23 (Defective Multiplicity 2 Eigenvalues). Solve the ODE $\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$.

Solution: Solving the characteristic equation

$$\begin{vmatrix} 1 - \lambda & -1 \\ 1 & 3 - \lambda \end{vmatrix} = \lambda^2 - 4\lambda + 4 = 0$$

we have $\lambda_1 = \lambda_2 = 2$, a two-fold degenerate eigenvalue.

When $\lambda = 2$, we have

$$\begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix} \mathbf{v}_1 = \mathbf{0} \Rightarrow \mathbf{v}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

The associated eigenvalue problem is

$$\begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix} \mathbf{v}_2 = \mathbf{v}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \Rightarrow \mathbf{v}_2 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

Hence the two independent solutions are

$$\mathbf{x}_1 = e^{2t} \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \mathbf{x}_2 = t e^{2t} \begin{bmatrix} 1 \\ -1 \end{bmatrix} + e^{2t} \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

and the general solution is

$$\mathbf{x} = C_1 e^{2t} \begin{bmatrix} 1 \\ -1 \end{bmatrix} + C_2 e^{2t} \begin{bmatrix} t + 1 \\ -t - 2 \end{bmatrix}.$$

Example 33.24. Solve $\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{x}$.

Solution: The eigenvalues are $\lambda_1 = \lambda_2 = 0$. For $\lambda = 0$,

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{v}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

The associated eigenvalue problem is

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{v}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \Rightarrow \mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Therefore, the general solution is

$$\mathbf{x}(t) = C_1 \mathbf{v}_1 + C_2 (\mathbf{v}_1 t + \mathbf{v}_2) = C_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + C_2 \begin{bmatrix} t \\ 1 \end{bmatrix}.$$

Example 33.25. Solve $\dot{\mathbf{x}} = \begin{bmatrix} -2 & 1 \\ -1 & -4 \end{bmatrix} \mathbf{x}$.

Solution. From the characteristic equation

$$\begin{vmatrix} -2-\lambda & 1 \\ -1 & -4-\lambda \end{vmatrix} = \lambda^2 + 6\lambda + 9 = 0$$

we obtain $\lambda_1 = \lambda_2 = -3$.

For $\lambda = -3$,

$$[A - \lambda I]\mathbf{v}_1 = \mathbf{0} \Rightarrow \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix} \mathbf{v}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \mathbf{v}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

We then solve the associated eigenvalue problem:

$$\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix} \mathbf{v}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \Rightarrow \mathbf{v}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

The general solution is

$$\mathbf{x}(t) = C_1 \mathbf{v}_1 e^{-3t} + C_2 (\mathbf{v}_1 t + \mathbf{v}_2) e^{-3t} = \begin{bmatrix} C_1 + C_2 + C_2 t \\ -C_1 - C_2 t \end{bmatrix} e^{-3t}$$

□

Eigenvalue of multiplicity Three ($m = 3$)

If A has only one eigenvector associated with an eigenvalue λ of multiplicity three, then we can find three linearly independent solutions of the form

$$\mathbf{x}_1 = \mathbf{v}_1 e^{\lambda t}$$

$$\mathbf{x}_2 = (\mathbf{v}_1 t + \mathbf{v}_2) e^{\lambda t}$$

$$\mathbf{x}_3 = \left(\frac{1}{2!} \mathbf{v}_1 t^2 + \mathbf{v}_2 t + \mathbf{v}_3 \right) e^{\lambda t}$$

where the columns vectors $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ must satisfy

$$(A - \lambda I)\mathbf{v}_1 = \mathbf{0}, \quad (A - \lambda I)\mathbf{v}_2 = \mathbf{v}_1, \quad (A - \lambda I)\mathbf{v}_3 = \mathbf{v}_2.$$

Example 33.26 (Defective Multiplicity 3 Eigenvalues, Zill, P466). Solve $\dot{\mathbf{x}} = \begin{bmatrix} 2 & 1 & 6 \\ 0 & 2 & 5 \\ 0 & 0 & 2 \end{bmatrix} \mathbf{x}$.

Solution: Solving the characteristic equation

$$|A - \lambda I| = 0 \Rightarrow \begin{vmatrix} 2-\lambda & 1 & 6 \\ 0 & 2-\lambda & 5 \\ 0 & 0 & 2-\lambda \end{vmatrix} = (2-\lambda)^3 = 0$$

gives the eigenvalue of multiplicity three: $\lambda_1 = \lambda_2 = \lambda_3 = 2$.

For $\lambda = 2$,

$$[A - \lambda I]\mathbf{v}_1 = \mathbf{0} \Rightarrow \begin{bmatrix} 0 & 1 & 6 \\ 0 & 0 & 5 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{v}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

The associated eigenvalue problems are

$$[A - \lambda I]\mathbf{v}_2 = \mathbf{v}_1 \Rightarrow \begin{bmatrix} 0 & 1 & 6 \\ 0 & 0 & 5 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{v}_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

and

$$[A - \lambda \mathbf{I}] \mathbf{v}_3 = \mathbf{v}_2 \Rightarrow \begin{bmatrix} 0 & 1 & 6 \\ 0 & 0 & 5 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{v}_3 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \Rightarrow \mathbf{v}_3 = \begin{bmatrix} 0 \\ -\frac{6}{5} \\ \frac{1}{5} \end{bmatrix}.$$

Therefore the general solution is

$$\begin{aligned} \mathbf{x}(t) &= \left\{ C_1 \mathbf{v}_1 + C_2 (\mathbf{v}_1 t + \mathbf{v}_2) + C_3 (\mathbf{v}_1 (t^2/2) + \mathbf{v}_2 t + \mathbf{v}_3) \right\} e^{\lambda_1 t} \\ &= \left\{ C_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + C_2 \left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} t + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right) + C_3 \left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \frac{t^2}{2!} + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} t + \begin{bmatrix} 0 \\ -6/5 \\ 1/5 \end{bmatrix} \right) \right\} e^{2t} \end{aligned}$$

Example 33.27 (E&P, P456, 7.5 Q11). Solve $\dot{\mathbf{x}} = \begin{bmatrix} -3 & 0 & -4 \\ -1 & -1 & -1 \\ 1 & 0 & 1 \end{bmatrix} \mathbf{x}$.

Solution. The characteristic equation

$$\begin{aligned} \begin{vmatrix} -3-\lambda & 0 & -4 \\ -1 & -1-\lambda & -1 \\ 1 & 0 & 1-\lambda \end{vmatrix} &= (-3-\lambda)(-1-\lambda)(1-\lambda) - 4(1+\lambda) \\ &= -(1+\lambda)((-3-\lambda)(1-\lambda) + 4) = -(1+\lambda)(\lambda^2 + 2\lambda + 1) = 0 \end{aligned}$$

gives an eigenvalue of multiplicity three: $\lambda_1 = \lambda_2 = \lambda_3 = -1$.

For $\lambda = -1$,

$$\begin{bmatrix} -2 & 0 & -4 \\ -1 & 0 & -1 \\ 1 & 0 & 2 \end{bmatrix} \mathbf{v}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 0 & 2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{v}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \mathbf{v}_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

The associated eigenvalue problems are

$$\begin{bmatrix} -2 & 0 & -4 \\ -1 & 0 & -1 \\ 1 & 0 & 2 \end{bmatrix} \mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 0 & 2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \Rightarrow \mathbf{v}_2 = \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}$$

$$\text{and } \begin{bmatrix} -2 & 0 & -4 \\ -1 & 0 & -1 \\ 1 & 0 & 2 \end{bmatrix} \mathbf{v}_3 = \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 0 & 2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{v}_3 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \Rightarrow \mathbf{v}_3 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

The general solution is

$$\begin{aligned} \mathbf{x}(t) &= C_1 \mathbf{v}_1 e^{-t} + C_2 [\mathbf{v}_1 t e^{-t} + \mathbf{v}_2 e^{-t}] + C_3 [\mathbf{v}_1 (t^2/2) e^{-t} + \mathbf{v}_2 t e^{-t} + \mathbf{v}_3 e^{-t}] \\ &= \left\{ C_1 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + C_2 \left(\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} t + \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix} \right) + C_3 \left(\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \frac{t^2}{2!} + \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix} t + \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \right) \right\} e^{-t}. \end{aligned}$$

□

General Case

If A has only one eigenvector associated with an eigenvalue λ of multiplicity m , then we can find m

linearly independent solutions of the form

$$\begin{aligned}\mathbf{x}_1 &= \mathbf{v}_1 e^{\lambda t} \\ \mathbf{x}_2 &= (\mathbf{v}_1 t + \mathbf{v}_2) e^{\lambda t} \\ &\vdots \\ \mathbf{x}_m &= \left(\frac{1}{(m-1)!} \mathbf{v}_1 t^{m-1} + \cdots + \frac{1}{2!} \mathbf{v}_{m-2} t^2 + \mathbf{v}_{m-1} t + \mathbf{v}_m \right) e^{\lambda t}\end{aligned}$$

where the column vectors $\mathbf{v}_i$ must satisfy

$$\begin{aligned}(A - \lambda \mathbf{I})\mathbf{v}_1 &= \mathbf{0} \\ (A - \lambda \mathbf{I})\mathbf{v}_2 &= \mathbf{v}_1 \\ &\vdots \\ (A - \lambda \mathbf{I})\mathbf{v}_{m-1} &= \mathbf{v}_{m-2}\end{aligned}$$

One can refer to Weintraub [2008] for the detail discussion of the theory and more examples.

33.4 Non-Homogeneous Linear System with Constant Coefficients

non-homog:const

In this section, we will investigate the generalisation of the homogeneous linear system with constant coefficients from the previous section to the non-homogeneous case:

$$\dot{\mathbf{x}} = A\mathbf{x} + \mathbf{f}(t) \quad (33.14)$$

e:ode:lin:cons

where A is a constant matrix.

The eigenvalues of A can be very complicated when A has degenerate eigenvalues. We will limit ourselves to the case when A is a diagonalizable $n \times n$ matrix, i.e. we will just generalise Subsection 33.3.1, i.e. we will just investigate the case when A has n linearly independent eigenvectors. These eigenvectors form columns of an invertible matrix P such that

sec:eigv:real:diff

$$P^{-1}AP = \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

with eigenvalues down the main diagonal in the order corresponding to the eigenvector columns of P .

Using the substitution $\mathbf{x} = P\mathbf{y}$, the system (33.14) is transformed into an *uncoupled system*

e:ode:lin:const:nonhomog

$$P\dot{\mathbf{y}} = AP\mathbf{y} + \mathbf{f}(t) \Rightarrow \dot{\mathbf{y}} = P^{-1}AP\mathbf{y} + P^{-1}\mathbf{f}(t) = \Lambda\mathbf{y} + \mathbf{g}(t)$$

which is an array of independent ODEs
$$\begin{cases} \dot{y}_1 = \lambda_1 y_1 + g_1(t), \\ \dot{y}_2 = \lambda_2 y_2 + g_2(t), \\ \vdots \\ \dot{y}_n = \lambda_n y_n + g_n(t), \end{cases} \quad \text{where } \mathbf{g}(t) := P^{-1}\mathbf{f}(t) = \begin{bmatrix} g_1(t) \\ \vdots \\ g_n(t) \end{bmatrix}.$$

By using the “integrating factor” technique from Topic 1, we can obtain the solution

$$y_i = e^{\lambda_i t} \left[\int e^{-\lambda_i t} g_i(t) dt + C_i \right]$$

and the solution to the original problem is $\mathbf{x}(t) = P\mathbf{y}(t)$.

We will illustrate this with the following examples.

Example 33.28. Find the general solution of the linear system
$$\begin{cases} \dot{x}_1 = -2x_1 + x_2 + 1, \\ \dot{x}_2 = x_1 - 2x_2. \end{cases}$$

Solution: First, let us rewrite the ODE in matrix form:

$$\dot{\mathbf{x}} = \begin{bmatrix} -2 & 1 \\ 1 & -2 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Let $A = \begin{bmatrix} -2 & 1 \\ 1 & -2 \end{bmatrix}$ and determine the eigenvalues and eigenvectors of the coefficient matrix A :

$$\begin{vmatrix} -2-\lambda & 1 \\ 1 & -2-\lambda \end{vmatrix} = \lambda^2 + 4\lambda + 3 = 0 \Rightarrow \lambda_1 = -3, \lambda_2 = -1$$

$$\text{When } \lambda = -3, (A - (-3)I_2)\mathbf{v}_1 = \mathbf{0} \Rightarrow \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \mathbf{v}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \mathbf{v}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$\text{When } \lambda = -1, (A - (-1)I_2)\mathbf{v}_1 = \mathbf{0} \Rightarrow \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \mathbf{v}_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Let $P = [\mathbf{v}_2, \mathbf{v}_1] = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ and $\mathbf{x} = P\mathbf{y}$. Then the linear system becomes

$$\underbrace{\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix}}_{\dot{\mathbf{y}}} = \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}}_{P^{-1}} \underbrace{\begin{bmatrix} -2 & 1 \\ 1 & -2 \end{bmatrix}}_A \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}}_P \underbrace{\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}}_{\mathbf{y}} + \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}}_{P^{-1}} \underbrace{\begin{bmatrix} 1 \\ 0 \end{bmatrix}}_{\mathbf{f}(t)} = \underbrace{\begin{bmatrix} -1 & 0 \\ 0 & -3 \end{bmatrix}}_A \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \underbrace{\begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}}_{\mathbf{g}(t)}$$

We need to solve two independent ODEs here:

$$\dot{y}_1 = -y_1 + 1/2 \Rightarrow y_1 = e^{-t} \left[\int \frac{1}{2} e^t dt + C_1 \right] = C_1 e^{-t} + \frac{1}{2}$$

and

$$\dot{y}_2 = -3y_2 + 1/2 \Rightarrow y_2 = e^{-3t} \left[\int \frac{1}{2} e^{3t} dt + C_2 \right] = C_2 e^{-3t} + \frac{1}{6}$$

The general solution of the original ODE is

$$\mathbf{x}(t) = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} C_1 e^{-t} + C_2 e^{-3t} + \frac{2}{3} \\ C_1 e^{-t} - C_2 e^{-3t} + \frac{1}{3} \end{bmatrix}.$$

Example 33.29. Find a general solution of the nonhomogeneous system

$$\dot{\mathbf{x}} = \begin{bmatrix} 1 & 3 \\ 5 & 3 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 12t - 11 \\ -3 \end{bmatrix}, \quad -\infty < t < \infty. \quad (33.15)$$

e:eg:lin-sys2b

Solution. We need to determine the eigenvectors of the coefficient matrix to determine P :

$$\begin{vmatrix} 1-\lambda & 3 \\ 5 & 3-\lambda \end{vmatrix} = \lambda^2 - 4\lambda + 12 = 0 \Rightarrow \lambda_1 = -2, \lambda_2 = 6. \Rightarrow \mathbf{v}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 3 \\ 5 \end{bmatrix}.$$

Let $\mathbf{x} = \begin{bmatrix} 1 & 3 \\ -1 & 5 \end{bmatrix} \mathbf{y}$. Then [le:eg:lin-sys2b](#) (33.15) becomes

$$\dot{\mathbf{y}} = \begin{bmatrix} 1 & 3 \\ -1 & 5 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 3 \\ 5 & 3 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ -1 & 5 \end{bmatrix} \mathbf{y} + \begin{bmatrix} 1 & 3 \\ -1 & 5 \end{bmatrix}^{-1} \begin{bmatrix} 12t - 11 \\ -3 \end{bmatrix} = \begin{bmatrix} -2 & 0 \\ 0 & 6 \end{bmatrix} \mathbf{y} + \begin{bmatrix} \frac{15t}{2} - \frac{23}{4} \\ \frac{3t}{2} - \frac{7}{4} \end{bmatrix}.$$

Hence,

$$y_1(t) = C_1 e^{-2t} + e^{-2t} \int e^{2t} \left[\frac{15t}{2} - \frac{23}{4} \right] dt = C_1 e^{-2t} + \frac{15t}{4} - \frac{19}{4}$$

$$y_2(t) = C_2 e^{6t} + e^{6t} \int e^{-6t} \left[\frac{3t}{2} - \frac{7}{4} \right] dt = C_2 e^{6t} + \frac{1}{4} - \frac{t}{4}$$

Hence, the general solution of (33.15) is le;eq:lin-sys2b

$$\mathbf{x}(t) = P\mathbf{y}(t) = C_1 \begin{bmatrix} e^{-2t} \\ -e^{-2t} \end{bmatrix} + C_2 \begin{bmatrix} 3e^{6t} \\ 5e^{6t} \end{bmatrix} + \begin{bmatrix} 3t - 4 \\ 6 - 5t \end{bmatrix}.$$

□

Example 33.30. Solve the initial value problem $\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ 2 \end{bmatrix}$, $\mathbf{x}(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

Solution. The eigenvalues are found to be

$$\begin{bmatrix} -\lambda & 1 \\ 1 & -\lambda \end{bmatrix} = \lambda^2 - 1 = 0 \Rightarrow \lambda_1 = 1, \lambda_2 = -1.$$

When $\lambda = 1$, the eigenvector, $\mathbf{v}_1$, is

$$\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \mathbf{v}_1 = \mathbf{0} \Rightarrow \mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

When $\lambda = -1$, the eigenvector, $\mathbf{v}_2$, is

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \mathbf{v}_2 = \mathbf{0} \Rightarrow \mathbf{v}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Let $P = [\mathbf{v}_1 \ \mathbf{v}_2] = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ and by applying the transform $\mathbf{x} = P\mathbf{y}$, the initial problem becomes

$$\dot{\mathbf{y}} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \mathbf{y} + \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 2 \end{bmatrix} \Rightarrow \begin{cases} \dot{y}_1 = y_1 + 1 \\ \dot{y}_2 = -y_2 - 1 \end{cases}$$

where the general solution for the decoupled ODE of $\mathbf{y}$ is

$$\begin{cases} y_1 = C_1 e^t - 1 \\ y_2 = C_2 e^{-t} - 1 \end{cases}$$

and the general solution for the ODE of $\mathbf{x}$ is

$$\mathbf{x} = P\mathbf{y} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} C_1 e^t - 1 \\ C_2 e^{-t} - 1 \end{bmatrix} = \begin{bmatrix} C_1 e^t + C_2 e^{-t} - 2 \\ C_1 e^t - C_2 e^{-t} \end{bmatrix}$$

From the initial condition $\mathbf{x}(0) = \begin{bmatrix} C_1 + C_2 - 2 \\ C_1 - C_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \Rightarrow C_1 = C_2 = \frac{3}{2}$

Hence, the solution $\mathbf{x}(t) = \begin{bmatrix} \frac{3}{2}(e^t + e^{-t}) - 2 \\ \frac{3}{2}(e^t - e^{-t}) \end{bmatrix}$. □

g:final200901q5

Example 33.31 (Final Exam April 2009, Q5). Consider the following system of first order differential equations $\frac{dx}{dt} = 5x + 4y - e^{2t}$, $\frac{dy}{dt} = 2x + 3y + 2$. Given that when $t = 0$, $x = 0$, $y = 1$, find $x(t)$ and $y(t)$. (25 marks)

Solution: First, we need to solve the homogeneous problem $\dot{\mathbf{x}} = A\mathbf{x}$ where

$$A = \begin{bmatrix} 5 & 4 \\ 2 & 3 \end{bmatrix} \Rightarrow (\lambda - 5)(\lambda - 3) - 8 = \lambda^2 - 8\lambda + 7 = 0 \Rightarrow \lambda = 7, 1.$$

The corresponding eigenvector to $\lambda = 7, 1$ are $\mathbf{v}_1$ and $\mathbf{v}_2$ which can be found from

$$\begin{bmatrix} -2 & 4 \\ 2 & -4 \end{bmatrix} \mathbf{v}_1 = \mathbf{0} \wedge \begin{bmatrix} 4 & 4 \\ 2 & 2 \end{bmatrix} \mathbf{v}_2 = \mathbf{0} \Rightarrow \mathbf{v}_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Then, we let $P = [\mathbf{v}_1, \mathbf{v}_2]$ and introduce a transform $\mathbf{x} = P\mathbf{u}$ so that

$$\begin{bmatrix} 2 & 1 \\ 1 & -1 \end{bmatrix} \dot{\mathbf{u}} = \begin{bmatrix} 5 & 4 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & -1 \end{bmatrix} \mathbf{u} + \begin{bmatrix} -e^{2t} \\ 2 \end{bmatrix} \Rightarrow \dot{\mathbf{u}} = \begin{bmatrix} 7 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{u} + \frac{1}{-3} \begin{bmatrix} -1 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} -e^{2t} \\ 2 \end{bmatrix}$$

By using what we learn from Topic 1, we have

$$\begin{aligned} \dot{u}_1 - 7u_1 &= \frac{2 - e^{2t}}{3} \Rightarrow u_1 = e^{7t} \int e^{-7t} \times \frac{2 - e^{2t}}{3} dt = \frac{e^{7t}}{3} \left[\frac{e^{-5t}}{5} - \frac{2e^{-7t}}{7} + C_1 \right] \\ \dot{u}_2 - u_2 &= \frac{-4 - e^{2t}}{3} \Rightarrow u_2 = e^t \int e^{-t} \times \frac{-4 - e^{2t}}{3} dt = e^t \left[\frac{4}{3} e^{-t} - \frac{1}{3} e^t + C_2 \right] \end{aligned}$$

Therefore, the general solution is $\mathbf{x} = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \frac{e^{2t}}{15} - \frac{2}{21} + \frac{C_1}{3} e^{7t} \\ \frac{4}{3} - \frac{1}{3} e^{2t} + C_2 e^t \end{bmatrix}$

When $t = 0$, we have $\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \frac{1}{15} - \frac{2}{21} + \frac{C_1}{3} \\ \frac{4}{3} - \frac{1}{3} + C_2 \end{bmatrix} \Rightarrow C_1 = \frac{38}{35}, \quad C_2 = -\frac{5}{3}.$

Alternatively, we can apply the Laplace transform technique from Topic 4 to obtain

$$\begin{cases} sX = 5X + 4Y - \frac{1}{s-2} \\ sY - 1 = 2X + 3Y + \frac{2}{s} \end{cases} \Rightarrow \begin{bmatrix} s-5 & -4 \\ -2 & s-3 \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} -\frac{1}{s-2} \\ \frac{s+2}{s} \end{bmatrix}$$

Therefore, $\begin{bmatrix} X \\ Y \end{bmatrix} = \frac{1}{(s-5)(s-3)-8} \begin{bmatrix} s-3 & 4 \\ 2 & s-5 \end{bmatrix} \begin{bmatrix} -\frac{1}{s-2} \\ \frac{s+2}{s} \end{bmatrix} = \begin{bmatrix} \frac{3s^2 + 3s - 16}{s(s-2)(s^2 - 8 + 7)} \\ \frac{s(s-2)(s^2 - 8 + 7)}{s^3 - 5s^2 - 6s + 20} \end{bmatrix}$

Taking inverse Laplace transform, we have

$$\begin{aligned} x(t) &= \mathcal{L}^{-1} \left\{ \frac{8}{7s} - \frac{5}{3(s-1)} - \frac{1}{5(s-2)} + \frac{76}{105(s-7)} \right\} = \frac{8}{7} - \frac{5}{3}e^t - \frac{1}{5}e^{2t} + \frac{76}{105}e^{7t} \\ y(t) &= \mathcal{L}^{-1} \left\{ -\frac{10}{7s} + \frac{5}{3(s-1)} + \frac{2}{5(s-2)} + \frac{38}{105(s-7)} \right\} = -\frac{10}{7} + \frac{5}{3}e^t + \frac{2}{5}e^{2t} + \frac{38}{105}e^{7t} \end{aligned}$$

33.5 Fundamental Matrix and Transition Matrix

This section studies the generalisation of linear systems with constant coefficients (33.11) to time-varying coefficients:

$$\dot{\mathbf{x}} = A(t)\mathbf{x}, \quad \mathbf{x} \in \mathbb{R}^n \quad (33.16)$$

where $A(t) = \begin{bmatrix} a_{11}(t) & \cdots & a_{1n}(t) \\ \vdots & \ddots & \vdots \\ a_{n1}(t) & \cdots & a_{nn}(t) \end{bmatrix}$ and the coefficients $a_{ij}(t)$ are continuous functions on a given open interval I .

Similar to the general homogeneous linear equation we encountered in Topic 2, Section 2.4, there is no general technique for solving (33.16). Again, similar to Section 2.4, although we cannot find the general solution analytically, we can study the properties of the general solution.

33.5.1 Matrix Functions

Fundamental matrix is a special type of matrix functions associated with (33.16). This section will study the general properties of matrix functions.

Definition 33.14. A matrix (or vector) whose elements are functions of a real variable is called a *matrix function*.

Definition 33.15. A matrix function $A(t)$ in (33.16) is said to be *continuous* at $t = t_0$ (or on an open interval I) if each function $a_{ij}(t)$ is continuous at t_0 (or on I). $A(t)$ is said to be *differentiable* on I if each $a_{ij}(t)$ is differentiable on I , and its *derivative* is defined by

$$A'(t) = \frac{dA(t)}{dt} = \left[\frac{da_{ij}(t)}{dt} \right].$$

The *integral* of $A(t)$ is defined as

$$\int_a^b A(t) dt = \left[\int_a^b a_{ij}(t) dt \right].$$

Example 33.32. Given $A(t) = \begin{bmatrix} \sin t & t \\ 1 & \cos t \end{bmatrix}$, calculate $A'(t)$ and $\int_0^\pi A(t) dt$.

Solution: $A'(t) = \begin{bmatrix} \cos t & 1 \\ 0 & -\sin t \end{bmatrix}$, $\int_0^\pi A(t) dt = \begin{bmatrix} 2 & \pi^2/2 \\ \pi & 0 \end{bmatrix}$.

The following are some of the rules of differentiation for matrix functions:

1. $\frac{d}{dt}(kA(t)) = k \frac{dA(t)}{dt}$, where k is a scalar;
2. $\frac{d}{dt}(A(t) + B(t)) = \frac{dA(t)}{dt} + \frac{dB(t)}{dt}$;
3. $\frac{d}{dt}(CA(t)) = C \frac{dA(t)}{dt}$, $\frac{d}{dt}(A(t)C) = \frac{dA(t)}{dt}C$, where C is a constant matrix;
4. $\frac{d}{dt}(A(t)B(t)) = A(t) \frac{dB(t)}{dt} + \frac{dA(t)}{dt} B(t)$.

33.5.2 Fundamental Matrices

Recall the concept of fundamental set of solutions is related to the general solution of a linear ODE (Definition 2.1.9). The fundamental matrix is a similar concept associated to the general solution of a “system of linear ODEs” (33.16).

The following two theorems characterise the homogeneous linear system (33.16).

Theorem 33.33 (Principles of Superposition). *If the vector function $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ are solutions for the system (33.16), then the linear combination $C_1\mathbf{x}_1(t) + C_2\mathbf{x}_2(t)$ is also a solution for any constants C_1 and C_2 .*

When all the linearly independent solutions of the homogeneous equation (33.16) are known, any solution of (33.16) can be written as a linear combination of the linearly independent solutions.

Definition 33.16 (Linearly Dependent and Independent Vector Functions). The vector functions $\mathbf{x}^{(1)}(t), \mathbf{x}^{(2)}(t), \dots, \mathbf{x}^{(n)}(t)$ are *linearly dependent* on an interval I if there exist constants $C_1, C_2, \dots, C_n$, not all zero, for which

$$C_1\mathbf{x}^{(1)}(t) + C_2\mathbf{x}^{(2)}(t) + \dots + C_n\mathbf{x}^{(n)}(t) = \mathbf{0} \quad \forall t \in I.$$

If the vectors are not linearly dependent on I , they are said to be *linearly independent* on I .

Remark 33.34. The superscripts in $\mathbf{x}^{(1)}(t), \dots, \mathbf{x}^{(n)}(t)$ are just indices and have no other meaning.

Now, we define the “fundamental set of solutions” for the system of linear ODEs (33.16) le:ode:lin:homog

Definition 33.17 (Fundamental Set of Solutions). Any set of vector functions $\mathbf{x}^{(1)}(t), \dots, \mathbf{x}^{(n)}(t)$ which is *linearly independent* and satisfy the system (33.16) le:ode:lin:homog for each point in the interval I is called a *fundamental set of solutions* for the interval I .

eg:HS

Example 33.35. Show that $\mathbf{x}^{(1)}(t) = \begin{bmatrix} e^{-2t} \\ -e^{-2t} \end{bmatrix}$ and $\mathbf{x}^{(2)}(t) = \begin{bmatrix} 3e^{6t} \\ 5e^{6t} \end{bmatrix}$ form a fundamental set of solutions of the linear system $\dot{\mathbf{x}} = \begin{bmatrix} 1 & 3 \\ 5 & 3 \end{bmatrix} \mathbf{x}$ on $(-\infty, \infty)$. Hence, write down the general solution of this system on this interval.

Solution: First, we need to show that $\mathbf{x}^{(1)}(t)$ and $\mathbf{x}^{(2)}(t)$ are the solutions of the linear system:

- $\begin{bmatrix} 1 & 3 \\ 5 & 3 \end{bmatrix} \mathbf{x}^{(1)}(t) = \begin{bmatrix} (1-3)e^{-2t} \\ (5-3)e^{-2t} \end{bmatrix} = \begin{bmatrix} -2e^{-2t} \\ 2e^{-2t} \end{bmatrix} = \dot{\mathbf{x}}^{(1)}(t)$. Hence, $\mathbf{x}^{(1)}(t)$ is a solution
- $\begin{bmatrix} 1 & 3 \\ 5 & 3 \end{bmatrix} \mathbf{x}^{(2)}(t) = \begin{bmatrix} (3+3 \times 5)e^{6t} \\ (5 \times 3 + 3 \times 5)e^{6t} \end{bmatrix} = \begin{bmatrix} 18e^{6t} \\ 30e^{6t} \end{bmatrix} = \dot{\mathbf{x}}^{(2)}(t)$. Hence, $\mathbf{x}^{(2)}(t)$ is a solution

Then, we need to show that $\mathbf{x}^{(1)}(t)$ and $\mathbf{x}^{(2)}(t)$ are linearly independent:

$$C_1 \mathbf{x}^{(1)}(t) + C_2 \mathbf{x}^{(2)}(t) = \begin{bmatrix} C_1 e^{-2t} + 3C_2 e^{6t} \\ -C_1 e^{-2t} + 5C_2 e^{6t} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow C_1 = 0, C_2 = 0.$$

Hence, $\mathbf{x}^{(1)}(t)$ and $\mathbf{x}^{(2)}(t)$ form a fundamental set of solutions and the general solution of the linear system is

$$\mathbf{x}(t) = C_1 \mathbf{x}^{(1)}(t) + C_2 \mathbf{x}^{(2)}(t) = \begin{bmatrix} C_1 e^{-2t} + 3C_2 e^{6t} \\ -C_1 e^{-2t} + 5C_2 e^{6t} \end{bmatrix}.$$

In general, the fundamental set of solutions always exists as stated below.

Theorem 33.36 (Existence of a Fundamental Set). *There exists a fundamental set of solutions for the homogeneous system (33.16) le:ode:lin:homog on an interval I .*

Definition 33.18 (Fundamental Matrix). A matrix which its columns are composed by a fundamental set of solutions:

$$\mathbf{X}(t) = [\mathbf{x}^{(1)}(t), \dots, \mathbf{x}^{(n)}(t)]$$

is called a *fundamental matrix*.

A general solution of the homogeneous system (33.16) le:ode:lin:homog can be written in terms of the fundamental set of solutions or the fundamental matrix.

Theorem 33.37 (General Solution of Homogeneous System). *If $\mathbf{x}^{(1)}(t), \mathbf{x}^{(2)}(t), \dots, \mathbf{x}^{(n)}(t)$ is a fundamental set of solutions of the linear homogeneous system (33.16) le:ode:lin:homog on an interval I where the elements of $A(t)$ are continuous, then the general solution $\mathbf{x}(t)$ of this system can be written in the form of*

$$\mathbf{x}(t) = C_1 \mathbf{x}^{(1)}(t) + C_2 \mathbf{x}^{(2)}(t) + \dots + C_n \mathbf{x}^{(n)}(t) = \mathbf{X}(t) \mathbf{c}$$

where $C_1, C_2, \dots, C_n$ are constants and $\mathbf{c} = [C_1, \dots, C_n]^T$ is an arbitrary constant column vector and the fundamental matrix satisfies the ODE $\dot{\mathbf{X}} = A(t)\mathbf{X}$.

The fundamental matrix of a system (33.16) le:ode:lin:homog is *not unique* because $\tilde{c}\mathbf{X}$ are all fundamental matrices of (33.16) le:ode:lin:homog. Here, $\tilde{c}$ is a non-zero constant.

Definition 33.19 (Wronskian). The *Wronskian* of the n (column) vector functions $\mathbf{x}^{(1)}(t), \dots, \mathbf{x}^{(n)}(t)$ is the determinant

$$W[\mathbf{x}^{(1)}(t), \dots, \mathbf{x}^{(n)}(t)] = \det \mathbf{X}(t) = \begin{vmatrix} \mathbf{x}^{(1)}(t) & \dots & \mathbf{x}^{(n)}(t) \end{vmatrix}.$$

Similar to Topic 2, the Wronskian can be used to test linear independence of vector functions.

Theorem 33.38 (Wronskian Test for Linearly Independence). A set $\{\mathbf{x}^{(1)}(t), \dots, \mathbf{x}^{(n)}(t)\}$ of solutions of the linear homogeneous system (33.16) is linearly independent on an open interval I iff their Wronskian is never zero on I .

Example 33.39. Determine the (a) fundamental set, (b) fundamental matrix and (c) Wronskian of the linear ODE:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

Solution: To find the fundamental matrix, we need to determine the eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ of the ODE as was done in Section 33.3:

$$\begin{vmatrix} -\lambda & 1 & 0 \\ 0 & -\lambda & 1 \\ -6 & -1 & 4-\lambda \end{vmatrix} = -\lambda(\lambda^2 - 4\lambda + 1) - 6 = -(\lambda^3 - 4\lambda^2 + \lambda + 6) = 0$$

Using a modern scientific calculator (or by trying the factors of 6) will give us the eigenvalues

$$\lambda_1 = -1, \lambda_2 = 2, \lambda_3 = 3.$$

When $\lambda = -1$,

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ -6 & -1 & 5 \end{bmatrix} \mathbf{v}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \mathbf{v}_1 = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}.$$

When $\lambda = 2$,

$$\begin{bmatrix} -2 & 1 & 0 \\ 0 & -2 & 1 \\ -6 & -1 & 2 \end{bmatrix} \mathbf{v}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \mathbf{v}_2 = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}.$$

When $\lambda = 3$,

$$\begin{bmatrix} -3 & 1 & 0 \\ 0 & -3 & 1 \\ -6 & -1 & 1 \end{bmatrix} \mathbf{v}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \mathbf{v}_3 = \begin{bmatrix} 1 \\ 3 \\ 9 \end{bmatrix}.$$

(a) The linear ODE has the following fundamental set of solutions:

$$\mathbf{x}_1(t) = \begin{bmatrix} e^{-t} \\ -e^{-t} \\ e^{-t} \end{bmatrix}, \quad \mathbf{x}_2(t) = \begin{bmatrix} e^{2t} \\ 2e^{2t} \\ 4e^{2t} \end{bmatrix}, \quad \mathbf{x}_3(t) = \begin{bmatrix} e^{3t} \\ 3e^{3t} \\ 9e^{3t} \end{bmatrix}.$$

(b) It has the following fundamental matrix

$$\mathbf{X}(t) = \begin{bmatrix} e^{-t} & e^{2t} & e^{3t} \\ -e^{-t} & 2e^{2t} & 3e^{3t} \\ e^{-t} & 4e^{2t} & 9e^{3t} \end{bmatrix}.$$

(c) The Wronskian is

$$\begin{vmatrix} e^{-t} & e^{2t} & e^{3t} \\ -e^{-t} & 2e^{2t} & 3e^{3t} \\ e^{-t} & 4e^{2t} & 9e^{3t} \end{vmatrix} = e^{-t+2t+3t} \begin{vmatrix} 1 & 1 & 1 \\ -1 & 2 & 3 \\ 1 & 4 & 9 \end{vmatrix} = 12e^{4t}.$$

33.5.3 Transition Matrix

If the initial conditions $\mathbf{x}(t_0) = [x_1^0 \cdots x_n^0]^T = \mathbf{x}_0$ are given, then the general solution $\mathbf{x}(t) = \mathbf{X}(t)\mathbf{c}$ of the homogeneous system (33.16) implies

$$\mathbf{x}(t_0) = \mathbf{X}(t_0)\mathbf{c} \Rightarrow \mathbf{c} = \mathbf{X}^{-1}(t_0)\mathbf{x}_0.$$

and the solution becomes

$$\mathbf{x}(t) = \underbrace{\mathbf{X}(t)\mathbf{X}^{-1}(t_0)}_{\Phi(t, t_0)} \mathbf{x}_0.$$

The column vectors in $\mathbf{X}(t)$ are linearly independent, therefore $\mathbf{X}(t)$ is invertible.

Definition 33.20. The matrix $\Phi(t, t_0)$ is called a *transition matrix*.

The effect of the pre-multiplication of transition matrix $\Phi(t, t_0)$ to the initial vector $\mathbf{x}_0$ is to perform a “transition” from the point $\mathbf{x}_0 \in \mathbb{R}^n$ to a new point $\mathbf{x}(t)$. Note that the transition matrix is fundamental matrix with special property that when $t = t_0$, it is an identity matrix, i.e.

$$\Phi(t_0, t_0) = \mathbf{I}.$$

Example 33.40. Use the result from Example 33.16 to find the transitional matrix $\Phi(t, 0)$ of the ODE

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Solution: From Example 33.16 in page 669, the fundamental matrix is

$$\mathbf{X}(t) = \begin{bmatrix} e^{-t} & e^{3t} \\ -2e^{-t} & 2e^{3t} \end{bmatrix}$$

When $t = 0$, $\mathbf{X}(0) = \begin{bmatrix} 1 & 1 \\ -2 & 2 \end{bmatrix}$ and $\mathbf{X}^{-1}(0)$ is the inverse of matrix $\mathbf{X}(0)$, hence the transition matrix is

$$\Phi(t, 0) = \mathbf{X}(t)\mathbf{X}^{-1}(0) = \begin{bmatrix} e^{-t} & e^{3t} \\ -2e^{-t} & 2e^{3t} \end{bmatrix} \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} \\ \frac{1}{2} & \frac{1}{4} \end{bmatrix} = \begin{bmatrix} \frac{1}{2}(e^{-t} + e^{3t}) & \frac{1}{4}(e^{3t} - e^{-t}) \\ e^{3t} - e^{-t} & \frac{1}{2}(e^{-t} + e^{3t}) \end{bmatrix}$$

Example 33.41. If $\begin{bmatrix} 2e^t & e^{6t} \\ -3e^t & e^{6t} \end{bmatrix}$ is a fundamental matrix of $\dot{\mathbf{x}} = \begin{bmatrix} 4 & 2 \\ 3 & 3 \end{bmatrix} \mathbf{x}$. Find the transition matrix $\Phi(t, 0)$.

Solution:

$$\Phi(t, 0) = \begin{bmatrix} 2e^t & e^{6t} \\ -3e^t & e^{6t} \end{bmatrix} \begin{bmatrix} 2 & 1 \\ -3 & 1 \end{bmatrix}^{-1} = \frac{1}{5} \begin{bmatrix} 2e^t + 3e^{6t} & -2e^t + 2e^{6t} \\ -3e^t + 3e^{6t} & 3e^t + 2e^{6t} \end{bmatrix}.$$

Example 33.42. If $\begin{bmatrix} 2e^{-5t} & (1+5t)e^{-5t} \\ e^{-5t} & \frac{5}{2}te^{-5t} \end{bmatrix}$ is a fundamental matrix of $\dot{\mathbf{x}} = \begin{bmatrix} 0 & -10 \\ \frac{5}{2} & -10 \end{bmatrix} \mathbf{x}$. Find the transition matrix $\Phi(t, 0)$.

Solution.

$$\Phi(t, 0) = \begin{bmatrix} 2e^{-5t} & (1+5t)e^{-5t} \\ e^{-5t} & \frac{5}{2}te^{-5t} \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix}^{-1} = \begin{bmatrix} (1+5t)e^{-5t} & -10te^{-5t} \\ \frac{5}{2}te^{-5t} & (1-5t)e^{-5t} \end{bmatrix}$$

□

eg:HS1

Example 33.43. By using Example [33.35](#), write down the transition matrix for $\dot{\mathbf{x}} = \begin{bmatrix} 1 & 3 \\ 5 & 3 \end{bmatrix} \mathbf{x}$ on $(-\infty, \infty)$. If the initial condition is $\mathbf{x}(0) = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, find the solution.

Solution. A transition matrix is

$$\Phi(t, 0) = \mathbf{X}(t)\mathbf{X}(0)^{-1} = \begin{bmatrix} e^{-2t} & 3e^{6t} \\ -e^{-2t} & 5e^{6t} \end{bmatrix} \cdot \frac{1}{8} \begin{bmatrix} 5 & -3 \\ 1 & 1 \end{bmatrix} = \frac{1}{8} \begin{bmatrix} 5e^{-2t} + 3e^{6t} & -3e^{-2t} + 3e^{6t} \\ -5e^{-2t} + 5e^{6t} & 3e^{-2t} + 5e^{6t} \end{bmatrix}.$$

The solution is

$$\mathbf{x}(t) = \Phi(t, 0)\mathbf{x}(0) = \frac{1}{8} \begin{bmatrix} 5e^{-2t} + 3e^{6t} & -3e^{-2t} + 3e^{6t} \\ -5e^{-2t} + 5e^{6t} & 3e^{-2t} + 5e^{6t} \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} e^{-2t} \\ -e^{-2t} \end{bmatrix}$$

□

33.6 Non-Homogeneous Linear Systems

The linear system with constant coefficients ([33.14](#)) and the homogeneous system ([33.16](#)) generalise to a linear system with time-varying coefficients as follows:

$$\dot{\mathbf{x}} = A(t)\mathbf{x} + \mathbf{f}(t), \quad \mathbf{f}(t) = \begin{bmatrix} f_1(t) \\ \vdots \\ f_n(t) \end{bmatrix} \quad (33.17)$$

where the nonhomogeneous terms $f_i(t)$ are continuous functions on a given open interval I .

The above system is a special case of ([33.2](#)) where each of the functions $f_1, \dots, f_n$ in ([33.2](#)) is a linear combination of variables $x_1, \dots, x_n$. Equation ([33.17](#)) has a component form as follows:

$$\begin{aligned} \dot{x}_1 &= a_{11}(t)x_1 + a_{12}(t)x_2 + \dots + a_{1n}(t)x_n + f_1(t), \\ \dot{x}_2 &= a_{21}(t)x_1 + a_{22}(t)x_2 + \dots + a_{2n}(t)x_n + f_2(t), \\ &\dots \\ \dot{x}_n &= a_{n1}(t)x_1 + a_{n2}(t)x_2 + \dots + a_{nn}(t)x_n + f_n(t). \end{aligned} \quad (33.18)$$

We will investigate some theoretical questions about linear system ([33.17](#)) such as the existence of solution, the structure of the solution and the techniques for finding a solution in Section [33.6.1](#).

Techniques for solving ([33.17](#)) include

- Laplace transform: This technique only works for a non-homogeneous linear system with constant coefficients. In this case, the ODE may be transformed to algebraic equation which is easier to solve and then the solution can be obtained by taking inverse Laplace transform. This was already discussed in Topic 4 and was illustrated in Example [33.31](#).
- Undetermined coefficients: The applications of method of undetermined coefficients is similar to the single ODE case in Topic 2. We need to guess a trial function by examining the form of non-homogeneous term $\mathbf{f}(t)$. This will be discussed in Section [33.6.4](#).
- Variation of parameters: Analogous to the single ODE case, we will need to find the homogeneous solution $\mathbf{x}_h = \mathbf{X}\mathbf{c}$ and substitute $\mathbf{x} = \mathbf{X}\mathbf{y}$ in ([33.17](#)) and try to solve for $\mathbf{y}$. This is discussed in Section [33.6.2](#).

33.6.1 Theory of Linear Systems

lin-sys:theory

A special case of Theorem 33.3 states that (33.17) has a unique solution if the matrix $A(t)$ and the vector $\mathbf{f}(t)$ is continuous.

ar-sys:uniq-exist

Theorem 33.44 (Existence and Uniqueness of Solutions to Linear Systems). *If the elements of the matrix $A(t)$ and the column vector $\mathbf{f}(t)$ in (33.17) are continuous on an open interval I that contains the point t_0 , then for any given vector $\mathbf{x}_0$, the initial value problem*

$$\dot{\mathbf{x}}(t) = A(t)\mathbf{x}(t) + \mathbf{f}(t), \quad \mathbf{x}(t_0) = \mathbf{x}_0$$

has a unique solution on I .

This theorem can be applied to the study of the initial value problem of an n th order linear ODE

$$y^{(n)} + a_{n-1}(t)y^{(n-1)} + \cdots + a_1(t)y' + a_0(t)y = f(t) \quad (33.19)$$

e:ode:lin:n

with initial conditions

$$y(t_0) = a_0, \quad y'(t_0) = a_1, \quad \dots, \quad y^{(n-1)}(t_0) = a_{n-1}. \quad (33.20)$$

e:ini:lin:n

By defining

$$x_1 = y, \quad x_2 = y', \quad x_3 = y'', \quad \dots, \quad x_n = y^{(n-1)},$$

the initial value problem (33.19) and (33.20) can be written as a (first order) linear system (33.17) (this technique was mentioned in Section 33.2):

$$\begin{cases} \dot{x}_1(t) = y'(t) = x_2(t), & x_1(0) = a_0, \\ \dot{x}_2(t) = y''(t) = x_3(t), & x_2(0) = a_1, \\ \dots & \dots \\ \dot{x}_{n-1}(t) = y^{(n-1)}(t) = x_n(t), & x_{n-1}(0) = a_{n-2}, \\ \dot{x}_n(t) = y^{(n)}(t) = -a_n(t)x_n(t) - \cdots - a_2(t)x_2(t) - a_1(t)x_1(t) + f(t), & x_n(0) = a_{n-1}. \end{cases}$$

Re-write into matrix form, we obtain (33.17) with

$$A(t) = \begin{bmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ -a_1(t) & -a_2(t) & \cdots & -a_{n-1}(t) & -a_n(t) \end{bmatrix}, \quad \mathbf{f}(t) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} f(t).$$

As mentioned earlier (Section 33.2), one reason for studying systems of ODEs lies in the fact that in numerical methods (*UECM3033*), higher order ODEs are written as an equivalent system of first-order ODEs before they can be numerically integrated.

The following two examples demonstrate how to transform higher order ODEs to systems of first-order ODEs.

eg:5.5.3

Example 33.45. Rewrite the initial value problem

$$y'' + 3y' + 2y = e^t, \quad y(0) = 4, y'(0) = 5$$

as a system of two first-order equations.

Solution: Let $x_1 = y$, $x_2 = y'$ then

$$\begin{aligned} \dot{x}_1 &= x_2, & x_1(0) &= 4, \\ \dot{x}_2 &= e^t - 3x_2 - 2x_1, & x_2(0) &= 5. \end{aligned}$$

Example 33.46. Re-write the third-order ODE $y''' - 4y'' + y' + 6y = f(t)$ into a first order ODE in matrix form.

Solution. Let $x_1 = y$, $x_2 = y'$, $x_3 = y''$. Then

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ f(t) \end{bmatrix}.$$

□

Similar to the constant coefficient system, the solution of general non-homogeneous linear system (33.17) can also be written as the addition of a homogeneous solution and a particular solution.

Definition 33.21. Equation (33.17) (and (33.18)) is called a *(first-order) system of linear ODEs* or a *linear system*. When $\mathbf{f}(t) \neq \mathbf{0}$, the linear system (33.18) is called *non-homogeneous* and when $\mathbf{f}(t) = \mathbf{0}$, it is called *homogeneous*.

Theorem 33.47 (General Solution of a Nonhomogeneous System). Let $\mathbf{x}_p(t)$ be a particular solution of the nonhomogeneous linear system (33.17) on an interval I , and let

$$\mathbf{x}_h(t) = C_1 \mathbf{x}^{(1)}(t) + C_2 \mathbf{x}^{(2)}(t) + \cdots + C_n \mathbf{x}^{(n)}(t)$$

denote the general solution on I of the corresponding homogeneous system

$$\dot{\mathbf{x}} = A(t)\mathbf{x}.$$

Then the general solution of the nonhomogeneous system (33.17) can be written in the form

$$\mathbf{x}(t) = \mathbf{x}_h(t) + \mathbf{x}_p(t).$$

33.6.2 Variation Of Parameters

In Topic 2, we have discussed two techniques for solving non-homogeneous linear ODE — Method of Undetermined Coefficients and Variation of Parameters. The former method is mostly suitable for constant coefficient problems while if the homogeneous solutions is known, the later will work nicely.

In Section 33.4, we learn that to solve a linear system with constant coefficients, we need to introduce the transform $\mathbf{x} = P\mathbf{y}$. The generalisation of this technique is called the *variation of parameters*, where P is replaced by the fundamental matrix $\mathbf{X}$ as follows

$$\mathbf{x} = \mathbf{X}(t)\mathbf{y}(t).$$

By substituting into (33.17), we obtain

$$\dot{\mathbf{x}} = \frac{d}{dt}(\mathbf{X}(t)\mathbf{y}(t)) = A(t)(\mathbf{X}(t)\mathbf{y}(t)) + \mathbf{f}(t) \Rightarrow \dot{\mathbf{X}}(t)\mathbf{y}(t) + \mathbf{X}(t)\dot{\mathbf{y}}(t) = A(t)\mathbf{X}(t)\mathbf{y}(t) + \mathbf{f}(t)$$

by using the properties of matrix functions in Section 33.5.1.

Note that $\dot{\mathbf{X}}(t) = A(t)\mathbf{X}(t)$, by cancelling terms, we obtain

$$\mathbf{X}(t)\dot{\mathbf{y}}(t) = \mathbf{f}(t) \Rightarrow \dot{\mathbf{y}}(t) = \mathbf{X}^{-1}(t)\mathbf{f}(t) \Rightarrow \mathbf{y}(t) = \int \mathbf{X}^{-1}(t)\mathbf{f}(t)dt + \mathbf{c}$$

where $\mathbf{c}$ is a vector of arbitrary constants.

Summarising the above discussion, we have

Theorem 33.48 (Variation Of Parameters). If $\mathbf{X}(t)$ is a fundamental matrix for the corresponding homogeneous equation of (33.17):

$$\dot{\mathbf{x}} = A(t)\mathbf{x}$$

on some interval where $A(t)$ and $\mathbf{f}(t)$ are continuous, then the general solution of the nonhomogeneous system (33.17) is

$$\mathbf{x}(t) = \mathbf{X}(t) \left[\mathbf{c} + \int \mathbf{X}^{-1}(s)\mathbf{f}(s)ds \right] = \underbrace{\mathbf{X}(t)\mathbf{c}}_{\mathbf{x}_h} + \underbrace{\mathbf{X}(t) \int \mathbf{X}^{-1}(s)\mathbf{f}(s)ds}_{\mathbf{x}_p} \quad (33.21)$$

e:nonhomog-gen-s

In particular, the solution of the initial value problem

$$\dot{\mathbf{x}} = A(t)\mathbf{x} + \mathbf{f}(t), \quad \mathbf{x}(t_0) = \mathbf{x}_0$$

is

$$\mathbf{x}(t) = \mathbf{X}(t)\mathbf{X}^{-1}(t_0)\mathbf{x}_0 + \int_{t_0}^t \mathbf{X}(t)\mathbf{X}^{-1}(s)\mathbf{f}(s)ds. \quad (33.22)$$

e:lin-sys:ivp

Remark 33.49. We can compare the above solution of the initial value problem to the result obtained in Topic 4 page 25:

$$\mathbf{x}(t) = \underbrace{\Phi(t)\mathbf{x}(0)}_{\text{natural response}} + \underbrace{\int_0^t \Phi(t-\tau)B\mathbf{u}(\tau)d\tau}_{\text{forced response}}, \quad \Phi(t) := \mathcal{L}^{-1}\{(s\mathbf{I} - A)^{-1}\}$$

in which we find that $\Phi(t) = \mathbf{X}(t)\mathbf{X}^{-1}(0)$.

Let us study an example to learn how the variation of parameters work.

Example 33.50. (a) Use variation of parameters to find the general solution of

$$\dot{\mathbf{x}} = \begin{bmatrix} 1 & 1 \\ 4 & 1 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 2e^{3t} \\ 0 \end{bmatrix}.$$

(b) Find the solution of the initial value problem

$$\dot{\mathbf{x}} = \begin{bmatrix} 1 & 1 \\ 4 & 1 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 2e^{3t} \\ 0 \end{bmatrix}, \quad \mathbf{x}(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Solution: (a) To apply the variation of parameters, we need to find the fundamental matrix of the corresponding homogeneous equation:

$$\dot{\mathbf{x}} = \begin{bmatrix} 1 & 1 \\ 4 & 1 \end{bmatrix} \mathbf{x}$$

in which by using the eigenvector method mentioned in the earlier section, we can obtain

$$\mathbf{X}(t) = \begin{bmatrix} e^{3t} & e^{-t} \\ 2e^{3t} & -2e^{-t} \end{bmatrix}.$$

The general solution is

$$\begin{aligned} \mathbf{x}(t) &= \mathbf{X}(t)\mathbf{c} + \mathbf{X}(t) \int \mathbf{X}^{-1}(s)\mathbf{f}(s)ds \\ &= \begin{bmatrix} e^{3t} & e^{-t} \\ 2e^{3t} & -2e^{-t} \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} + \begin{bmatrix} e^{3t} & e^{-t} \\ 2e^{3t} & -2e^{-t} \end{bmatrix} \int \frac{1}{-4e^{2t}} \begin{bmatrix} -2e^{-t} & -e^{-t} \\ -2e^{3t} & e^{3t} \end{bmatrix} \begin{bmatrix} 2e^{3t} \\ 0 \end{bmatrix} dt \\ &= \begin{bmatrix} C_1e^{3t} + C_2e^{-t} \\ 2C_1e^{3t} - 2C_2e^{-t} \end{bmatrix} + \begin{bmatrix} e^{3t} & e^{-t} \\ 2e^{3t} & -2e^{-t} \end{bmatrix} \int \begin{bmatrix} 1 \\ e^{4t} \end{bmatrix} dt \\ &= \begin{bmatrix} C_1e^{3t} + C_2e^{-t} \\ 2C_1e^{3t} - 2C_2e^{-t} \end{bmatrix} + \begin{bmatrix} e^{3t} & e^{-t} \\ 2e^{3t} & -2e^{-t} \end{bmatrix} \begin{bmatrix} t \\ \frac{1}{4}e^{4t} \end{bmatrix} = \begin{bmatrix} C_1e^{3t} + C_2e^{-t} \\ 2C_1e^{3t} - 2C_2e^{-t} \end{bmatrix} + \begin{bmatrix} t + \frac{1}{4} \\ 2t - \frac{1}{2} \end{bmatrix} e^{3t} \end{aligned}$$

(b) Although we can find the solution of the initial value problem using the formula (33.22). However, it is easier to use the general solution found above:

$$\mathbf{x}(0) = \begin{bmatrix} C_1 + C_2 \\ 2C_1 - 2C_2 \end{bmatrix} + \begin{bmatrix} \frac{1}{4} \\ -\frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \Rightarrow C_1 = \frac{1}{2}, \quad C_2 = \frac{1}{4}.$$

e:lin-sys:ivp

Therefore,

$$\mathbf{x}(t) = \frac{1}{4} \begin{bmatrix} (4t+3)e^{3t} + e^{-t} \\ (8t+2)e^{3t} - 2e^{-t} \end{bmatrix}.$$

Example 33.51 (Final Exam Sept 2016, Q5(b). From Boyce, Di Prima, 9e: p462 Q3). Let $A = \begin{bmatrix} 2 & -5 \\ 1 & -2 \end{bmatrix}$.

Find the general solution of the following non-homogeneous linear system

$$\dot{\mathbf{x}} = A\mathbf{x} + \begin{bmatrix} -\cos t \\ \sin t \end{bmatrix}$$

given that the fundamental matrix of $\dot{\mathbf{x}} = A\mathbf{x}$ is

$$\Phi(t) = \begin{bmatrix} 2\cos t - \sin t & 2\sin t + \cos t \\ \cos t & \sin t \end{bmatrix}.$$

(7 marks)

Solution. One can use method of undetermined coefficients or variation of parameters to find the general solution. For the later, let $\mathbf{x} = \Phi\mathbf{u}$, then

$$\dot{\mathbf{x}} = \dot{\Phi}\mathbf{u} + \Phi\dot{\mathbf{u}} = A\Phi\mathbf{u} + \begin{bmatrix} -\cos t \\ \sin t \end{bmatrix}. \quad [1 \text{ mark}]$$

Note that $\dot{\Phi} = A\Phi$ by definition, therefore, ... [1 mark]

$$\dot{\mathbf{u}} = [\Phi]^{-1} \begin{bmatrix} -\cos t \\ \sin t \end{bmatrix} = \frac{1}{-1} \begin{bmatrix} \sin t & -2\sin t - \cos t \\ -\cos t & 2\cos t - \sin t \end{bmatrix} \begin{bmatrix} -\cos t \\ \sin t \end{bmatrix}. \quad [1 \text{ mark}]$$

Simplifying and integrating gives

$$\dot{\mathbf{u}} = \begin{bmatrix} \sin 2t + (1 - \cos 2t) \\ -\cos 2t - \sin 2t \end{bmatrix} \Rightarrow \mathbf{u} = \begin{bmatrix} -\frac{\cos 2t}{2} + t - \frac{\sin 2t}{2} + C_1 \\ -\frac{\sin 2t}{2} + \frac{\cos 2t}{2} + C_2 \end{bmatrix} \quad [2+1=3 \text{ marks}]$$

The general solution is

$$\begin{aligned} \mathbf{x} &= \begin{bmatrix} 2\cos t - \sin t & 2\sin t + \cos t \\ \cos t & \sin t \end{bmatrix} \begin{bmatrix} t - \frac{\cos 2t}{2} - \frac{\sin 2t}{2} + C_1 \\ -\frac{\sin 2t}{2} + \frac{\cos 2t}{2} + C_2 \end{bmatrix} \\ &= C_1 \begin{bmatrix} 2\cos t - \sin t \\ \cos t \end{bmatrix} + C_2 \begin{bmatrix} 2\sin t + \cos t \\ \sin t \end{bmatrix} + \begin{bmatrix} t(2\cos t - \sin t) \\ t\cos t \end{bmatrix} + \begin{bmatrix} -\frac{3\sin t + \cos t}{2} \\ -\frac{\sin t + \cos t}{2} \end{bmatrix} \end{aligned} \quad [1 \text{ mark}]$$

If we let $C_1 = \widetilde{C}_1 - \frac{1}{2}$ and $C_2 = \widetilde{C}_2 + \frac{1}{2}$, then

$$\mathbf{x} = \widetilde{C}_1 \begin{bmatrix} 2\cos t - \sin t \\ \cos t \end{bmatrix} + \widetilde{C}_2 \begin{bmatrix} 2\sin t + \cos t \\ \sin t \end{bmatrix} + \begin{bmatrix} t(2\cos t - \sin t) \\ t\cos t \end{bmatrix} - \begin{bmatrix} \cos t \\ \cos t \end{bmatrix}$$

□

33.6.3 Matrix Exponential Solutions

sys:var-param2

What happens to $\mathbf{X}(t)$ when $A(t) = A$ is a constant matrix? We will find out in this section that the solution can be expressed beautifully as matrix exponential, which is defined below.

Definition 33.22 (Matrix Exponential). For any $n \times n$ matrix A , the *matrix exponential (of A)* is an $n \times n$ matrix

$$e^{At} = \sum_{n=0}^{\infty} \frac{(At)^n}{n!} = I + tA + \frac{t^2}{2!}A^2 + \frac{t^3}{3!}A^3 + \dots$$

where I is an $n \times n$ identity matrix.

The following are some basic properties of the matrix exponential:

1. $e^0 = I$;
2. $\frac{d}{dt}e^{At} = Ae^{At}$;
3. $e^{At}e^{As} = e^{A(t+s)}$.

Using the notion of matrix exponential, we can express each of the solution of (33.11) and (33.14) elegantly as follows.

Theorem 33.52 (Solution of Homogeneous Systems with Constant Coefficients). *The solution of (33.11) with initial conditions $\mathbf{x}(t_0) = \mathbf{x}_0$ is*

$$\mathbf{x} = e^{At}\mathbf{x}_0$$

and e^{At} is a fundamental matrix of the homogeneous linear system with constant coefficients (33.11).

Theorem 33.53 (Solution of Nonhomogeneous Systems with Constant Coefficients). *The general solution of the nonhomogeneous system (33.14) is*

$$\mathbf{x} = e^{At}\mathbf{c} + e^{At} \int e^{-At}\mathbf{f}(t)dt \quad (33.23)$$

where $\mathbf{c}$ is a vector of arbitrary constants. The initial value problem with $\mathbf{x}(t_0) = \mathbf{x}_0$ has a solution

$$\mathbf{x} = e^{At}\mathbf{x}_0 + e^{At} \int_{t_0}^t e^{-As}\mathbf{f}(s)ds = e^{At}\mathbf{x}_0 + \int_{t_0}^t e^{A(t-s)}\mathbf{f}(s)ds.$$

The equation (33.23) is very elegant theoretically but is very inconvenient to use direct as demonstrated by the following two examples.

Example 33.54. Solve the initial value problem

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \mathbf{x}(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Solution: Let $A = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix}$ and we know that

$$A^k = \begin{cases} A, & k = 1, 3, 5, \dots, \\ -A, & k = 2, 4, 6, \dots \end{cases}$$

Therefore,

$$\begin{aligned} e^{At} &= I + \sum_{k=1,3,5,\dots} \frac{(At)^k}{k!} + \sum_{k=2,4,6,\dots} \frac{(At)^k}{k!} = I + \sum_{k=1,3,5,\dots} \frac{t^k}{k!} A - \sum_{k=2,4,6,\dots} \frac{t^k}{k!} A \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \left(\sum_{k=1,3,5,\dots} \frac{(-t)^k}{k!} + \sum_{k=2,4,6,\dots} \frac{(-t)^k}{k!} \right) \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - (e^{-t} - 1) \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 1 - e^{-t} \\ 0 & e^{-t} \end{bmatrix} \end{aligned}$$

The solution of the initial value problem ($t_0 = 0$) is

$$\begin{aligned} \mathbf{x} &= e^{At}\mathbf{x}_0 + e^{At} \int_0^t e^{-As}\mathbf{f}(s)ds = \begin{bmatrix} 1 & 1 - e^{-t} \\ 0 & e^{-t} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 - e^{-t} \\ 0 & e^{-t} \end{bmatrix} \int_0^t \begin{bmatrix} 1 & 1 - e^{-s} \\ 0 & e^{-s} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} ds. \\ &= \begin{bmatrix} 1 & 1 - e^{-t} \\ 0 & e^{-t} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 - e^{-t} \\ 0 & e^{-t} \end{bmatrix} \int_0^t \begin{bmatrix} 1 - e^{-s} \\ e^{-s} \end{bmatrix} ds \\ &= \begin{bmatrix} 1 & 1 - e^{-t} \\ 0 & e^{-t} \end{bmatrix} \begin{bmatrix} t + e^{-t} \\ 1 - e^{-t} \end{bmatrix} = \begin{bmatrix} t + e^{-t} \\ t + 2e^{-t} - te^{-t} - 2e^{-2t} \end{bmatrix} \end{aligned}$$

Example 33.55. Solve the system $\dot{\mathbf{x}} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \mathbf{x} + \begin{bmatrix} t \\ e^{4t} \end{bmatrix}$.

Solution: Since

$$e^{At} = \sum_{n=0}^{\infty} \frac{t^n}{n!} A^n = \sum_{n=0}^{\infty} \frac{t^n}{n!} \begin{bmatrix} 1^n & 0 \\ 0 & 2^n \end{bmatrix} = \begin{bmatrix} e^t & 0 \\ 0 & e^{2t} \end{bmatrix},$$

the general solution is

$$\begin{aligned} \mathbf{x}(t) &= e^{At} \mathbf{c} + \int_0^t e^{A(t-s)} \begin{bmatrix} s \\ e^{4s} \end{bmatrix} ds = \begin{bmatrix} e^t & 0 \\ 0 & e^{2t} \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} + \int_0^t \begin{bmatrix} e^{t-s} & 0 \\ 0 & e^{2(t-s)} \end{bmatrix} \begin{bmatrix} s \\ e^{4s} \end{bmatrix} ds \\ &= \begin{bmatrix} C_1 e^t \\ C_2 e^{2t} \end{bmatrix} + \begin{bmatrix} e^t - t - 1 \\ \frac{e^{4t}}{2} - \frac{e^{2t}}{2} \end{bmatrix} = \begin{bmatrix} \tilde{C}_1 e^t \\ \tilde{C}_2 e^{2t} \end{bmatrix} + \begin{bmatrix} -t - 1 \\ \frac{e^{4t}}{2} \end{bmatrix}. \end{aligned}$$

33.6.4 Undetermined Coefficients

n-sys:undeterm

When $A(t)$ in the non-homogeneous equation (33.17) is a constant matrix, the method of undetermined coefficients could be applied to determine the particular solution $\mathbf{x}_p$ by guessing a trial function based on the form of non-homogeneous term $\mathbf{f}(t)$. We have applied this method in determining the particular solution of linear ODE of constant coefficients in Topic 2.

Example 33.56. Write down the form of the particular solution $\mathbf{x}_p(t)$ for

$$\begin{aligned} \frac{dx}{dt} &= 5x + 3y - 2e^{-t} + 1, \\ \frac{dy}{dt} &= -x + y + e^{-t} - 5t + 7. \end{aligned}$$

You *do not* need to find the coefficients of $\mathbf{x}_p(t)$ explicitly.

Solution: To determine the form for $\mathbf{x}_p(t)$, we need to check if $\begin{bmatrix} -2 \\ 1 \end{bmatrix} e^{-t}$ appears in the general solution of the associated homogeneous system

$$\dot{\mathbf{x}} = \begin{bmatrix} 5 & 3 \\ -1 & 1 \end{bmatrix} \mathbf{x}$$

By finding the eigenvalues and eigenvectors:

$$\begin{vmatrix} 5 - \lambda & 3 \\ -1 & 1 - \lambda \end{vmatrix} = \lambda^2 - 6\lambda + 5 + 3 = 0 \Rightarrow \lambda_1 = 2, \lambda_2 = 4 \Rightarrow \mathbf{v}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} 3 \\ -1 \end{bmatrix},$$

the homogeneous solution is

$$\mathbf{x}_h(t) = C_1 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{2t} + C_2 \begin{bmatrix} 3 \\ -1 \end{bmatrix} e^{4t}.$$

Hence, the particular solution has the form

$$\mathbf{x}_p(t) = \begin{bmatrix} a_1 \\ b_1 \end{bmatrix} e^{-t} + \begin{bmatrix} a_2 t + a_3 \\ b_2 t + b_3 \end{bmatrix}.$$

Example 33.57. Use the method of undetermined coefficients to solve the system

$$\dot{\mathbf{x}} = \begin{bmatrix} 1 & 3 \\ 5 & 3 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 12 \\ 0 \end{bmatrix} t + \begin{bmatrix} -11 \\ -3 \end{bmatrix}$$

on $(-\infty, \infty)$. Hence, write down the general solution of this system on this interval.

Solution: The nonhomogeneous term is a polynomial of degree 1 and we assume

$$\mathbf{x}_p(t) = \begin{bmatrix} a_1 t + a_2 \\ b_1 t + b_2 \end{bmatrix}$$

By substituting into the ODE, we obtain

$$\dot{\mathbf{x}}_p = \begin{bmatrix} a_1 \\ b_1 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 5 & 3 \end{bmatrix} \begin{bmatrix} a_1 t + a_2 \\ b_1 t + b_2 \end{bmatrix} + \begin{bmatrix} 12t - 11 \\ -3 \end{bmatrix} \Rightarrow \begin{bmatrix} (a_1 + 3b_1 + 12)t + a_2 + 3b_2 - 11 - a_1 \\ (5a_1 + 3b_1)t + 5a_2 + 3b_2 - 3 - b_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

This implies

$$\left. \begin{aligned} a_1 + 3b_1 + 12 &= 0 \\ a_2 + 3b_2 - 11 - a_1 &= 0 \\ 5a_1 + 3b_1 &= 0 \\ 5a_2 + 3b_2 - 3 - b_1 &= 0 \end{aligned} \right\} \Rightarrow a_1 = 3, a_2 = -4, b_1 = -5, b_2 = 6.$$

From Example 33.35, we know that the general solution of the corresponding homogeneous equation

$$\dot{\mathbf{x}} = \begin{bmatrix} 1 & 3 \\ 5 & 3 \end{bmatrix} \mathbf{x}$$

is

$$\mathbf{x}_h(t) = C_1 \begin{bmatrix} e^{-2t} \\ -e^{-2t} \end{bmatrix} + C_2 \begin{bmatrix} 3e^{6t} \\ 5e^{6t} \end{bmatrix}.$$

Therefore, the general solution of the nonhomogeneous equation is

$$\mathbf{x}(t) = \mathbf{x}_h(t) + \mathbf{x}_p(t) = C_1 \begin{bmatrix} e^{-2t} \\ -e^{-2t} \end{bmatrix} + C_2 \begin{bmatrix} 3e^{6t} \\ 5e^{6t} \end{bmatrix} + \begin{bmatrix} 3t - 4 \\ 6 - 5t \end{bmatrix}.$$

33.7 Application-Specific Problems

A system of linear ODEs has many applications such as in control systems, nuclear physics, mechanical system, etc. In this section, we only illustrate how the system of linear ODEs (in state space form) are used in practice with a few examples. More examples are available from Holzbecher [2007], Betounes [2010] and Logan [2006, Chapters 5 and 6] or control system textbooks, Goodwin et al. [2000].

In a control system, the first order approximation of the system is the *LTI state space system* of the representation (32.7) [Åström and Murray, 2008, Section 2.2].

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \quad (33.24)$$

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u} \quad (33.25)$$

where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ are constant matrices. The matrix $\mathbf{A}$ is called the *dynamics matrix*, the matrix $\mathbf{B}$ is called the *control matrix*, the matrix $\mathbf{C}$ is called the *sensor matrix* and the matrix $\mathbf{D}$ is called the *direct term*. The state variables $\mathbf{x}$ are normally not observable while the system output variables $\mathbf{y}$ are any system variable of interest.

The following are a few software implementation of the LTI (32.7).

- The python-control package is a set of python classes and functions that implement common operations for the analysis and design of feedback control systems. The initial goal is to implement all of the functionality required to work through the examples in the textbook Feedback Systems by Åström and Murray. A MATLAB compatibility package (control.matlab) is available that provides many of the common functions corresponding to commands available in the MATLAB Control Systems Toolbox. See <https://github.com/python-control/python-control> or <http://python-control.sourceforge.net/manual/index.html>.
- SciLab
- Octave control toolbox.
- MATLAB Control Systems Toolbox. See <https://www.mathworks.com/products/control.html>

We will illustrate the link between state space system above and transfer function in Topic 4 and how to use the software with a few examples.

Example 33.58. Find the transfer function $\frac{Y(s)}{U(s)}$ of the state space system based on linear system of ODEs:

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -3 & -4 & -2 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u$$

$$y = [5 \quad 1 \quad 0] \mathbf{x} + 0 \cdot u$$

Solution: It is not difficult to obtain the transfer function using Laplace transfer function

$$\frac{Y(s)}{U(s)} = C(sI - A)^{-1}B + D = \frac{s + 5}{s^3 + 2s^2 + 4s + 3}.$$

The following is how to find the transfer function by using Scilab:

```
1 A = [0 1 0; 0 0 1; -3 -4 -2];
2 B = [0; 0; 1];
3 C = [5 1 0];
4 D = [0];
5 sl = syslin('c', A, B, C, D)
6 TF = ss2tf(sl)
```

which returns

```
1 TF =
2
3          2
4  5 + s - 6.661D-15s
5  -----
6          2    3
7  3 + 4s + 2s + s
```

this is due to floating error which will be discussed in UECM3033 Numerical Methods. To have a nice output, one can add the `clean` function to `ss2tf`.

Example 33.59. Find the state space system for the following transfer function

$$\frac{Y(s)}{U(s)} = \frac{s + 5}{s^3 + 2s^2 + 4s + 3}.$$

Solution: There are infinitely many different state-space system giving the same transfer function, so using the `tf2ss` from Scilab:

```
1 num = poly([5 1], 's', 'c');
2 den = poly([3 4 2 1], 's', 'c');
3 TF = num/den
4 Sys = tf2ss(TF)
```

will give a different state-space system:

```
1 Sys(1) (state-space system:)
2
3 !lss A B C D X0 dt !
4
5
6 Sys(2)= A matrix =
7
8 0.1923077 1.0384615 0.
9 -0.7393162 -0.3473146 -1.2869953
10 1.7105685 2.3411532 -1.8449931
11
12
13 Sys(3)= B matrix =
14
15 0.
16 -0.9251849
17 2.2680461
```

18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39

```
Sys(4)= C matrix =
-1.040833    0.    0.

Sys(5)= D matrix =
0.

Sys(6)= X0 (initial state) =
0.
0.
0.

Sys(7)= Time domain =
[]
```

Scilab has the canon function which allows us to extract the canonical controllable form using `canon(Sys[2],Sys[3])`.

In nuclear physics, much more PDE are used instead of ODE, however, for simple decay, it is possible to model with system of linear ODEs.

Example 33.60 (Systems of Equations in Radioactive Decay, Farlow, 6.4, Q31, P353). The radioactive isotope of iodine, ^{135}I , decays into the radioactive isotope of xenon, ^{135}Xe , which in turns decays into something else (not important). The half-lives of iodine and xenon are 6.7 hr and 9.2 hr, respectively.

1. Write the system of ODEs that describes the amount of ^{135}I and ^{135}Xe present.
2. What is the general solution of the system found in part 1.?

Solution: Let $k_1 = \ln \frac{2}{6.7}$ and $k_2 = \ln \frac{2}{9.2}$.

1. Let I and X be the amounts of ^{135}I and ^{135}Xe respectively. The radioactive decay is described by
$$\begin{cases} \dot{I} = -k_1 I, \\ \dot{X} = k_1 I - k_2 X. \end{cases}$$
2. Using the techniques from Topic 1 or previous sections, it is easy to obtain

$$\begin{bmatrix} I(t) \\ X(t) \end{bmatrix} = C_1 e^{-k_1 t} \begin{bmatrix} 1 \\ -1 \end{bmatrix} + C_2 e^{-k_2 t} \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

An important application of systems of linear ODEs is in the modelling of mechanical oscillations. Mechanical vibrations are crucial to the study of vehicle (automobiles, trains, etc.) dynamics [Mitschke and Wallentowitz, 2004, Chapter IX]. For example, a car, is modelled by mass-spring systems [Mitschke and Wallentowitz, 2004]:

$$\begin{aligned} m_3 \ddot{z}_3 + k_3(\dot{z}_3 - \dot{z}_2) + c_3(z_3 - z_2) &= 0 \\ m_2 \ddot{z}_2 - k_3(\dot{z}_3 - \dot{z}_2) - c_3(z_3 - z_2) + k_2(\dot{z}_2 - \dot{z}_1) + c_2(z_2 - z_1) &= 0 \\ m_1 \ddot{z}_1 - k_2(\dot{z}_2 - \dot{z}_1) - c_2(z_2 - z_1) + c_1 z_1 &= c_1 h \end{aligned}$$

where m_i are the mass, k_i are the damping coefficients, h is the external forcing.

The linear algebra techniques (especially the eigenvalue problems) that were discussed in previous sections is important in the solution of periodic solutions of ODE as demonstrated by the following example.

To add in future:

Hurwitz-Lyapunov criteria for the characterisation of the stability of filters and zero dynamics.
Linear fractional transformation or Redheffer star product.

33.8 Introduction to Nonlinear Systems

Many engineering problems can be approximated by the linear system (33.14) for a limited range only. Majority of the systems are nonlinear, such as celestial system [Bate et al., 1971], mechanical system, optimal control [Luenberger, 1979], etc.

Definition 33.23. A system of ODEs (33.2) or (33.3) is called a *nonlinear system* or *nonlinear (systems of) ODE* if it cannot be written in the form (33.18) or (33.17).

As mentioned in Section 33.3, the theory of linear systems is complete with linear algebra as the foundation. However, the theory of nonlinear systems is very incomplete and extremely complex due to its rich behaviour in the dynamics.

Since many nonlinear systems in physics and engineering are autonomous, we will limit ourselves to the discussion of autonomous systems (33.9). The non-autonomous system (33.3) is an important part of modern control theory and will be more appropriate for a graduate level study.

As mentioned in the beginning of Section 33.2, to study the global structure of a nonlinear system (33.9), we will be breaking down the phase space U to multiple pieces.

For a very small piece of open set around a point $\mathbf{x}_0$ where $\mathbf{f}(\mathbf{x}_0) \neq \mathbf{0}$. The *straightening out theorem* (or the *flow box theorem*) characterises the phase portrait around $\mathbf{x}_0$.

Theorem 33.61. (Straightening Out Theorem, [Chicone, 1999, 67]) Suppose that $\mathbf{x}_0 \in U$ and $\mathbf{f}(\mathbf{x}_0)$ in (33.9). Then there are open sets $D_1, D_2 \subset U$ with $\mathbf{x}_0 \in D_1$ and a diffeomorphism $g : D_1 \rightarrow D_2$ such that the ODE (33.9) in the new coordinates is

$$\dot{\mathbf{y}} = Dg(g^{-1}(\mathbf{y}))\mathbf{f}(g^{-1}(\mathbf{y})) = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

At a “nice” critical point $\mathbf{x}_c$, the phase portrait around $\mathbf{x}_c$ could be characterised by a linearised ODE.

Theorem 33.62. (Hartman-Grobman Theorem, [Chicone, 1999, 324]) Let $\mathbf{x}_c$ be a hyperbolic critical point (the real part of all the eigenvalues are non-zero) of the autonomous system (33.9). Suppose $\varphi_t : \mathbb{R}^n \rightarrow \mathbb{R}^n$ and $\psi_t : \mathbb{R}^n \rightarrow \mathbb{R}^n$ are the flows of (33.9) and

$$\dot{\mathbf{x}} = D\mathbf{f}(\mathbf{x}_c)(\mathbf{x} - \mathbf{x}_c), \quad D\mathbf{f}(\mathbf{x}_c) \text{ is a Jacobian of } \mathbf{f} \text{ at } \mathbf{x}_c.$$

Then there is an open set $U \subset \mathbb{R}^n$ containing $\mathbf{x}_c$ and a homeomorphism $\mathbf{g}$ with domain U such that

$$\mathbf{g}(\varphi_t(\mathbf{x})) = \psi_t(\mathbf{g}(\mathbf{x}))$$

whenever $\mathbf{x} \in U$ and both sides of the equation are defined.

33.9 One-Dimensional Phase Space Analysis

Let us study the case when the dimension of the phase space $U \subset \mathbb{R}$ is $n = 1$. In this case, U can be described by an x -axis. Without solving the autonomous differential equation

$$\dot{x} = f(x),$$

we can use its critical points and the algebraic sign of $f(x)$ on the x -axis to determine qualitative behaviour of its solution curves. This is called the *phase space analysis* of a one-dimensional system.

Assume that f and f' are continuous functions on some interval U of the x -axis. Suppose that the ODE has n critical points:

$$f(x) = 0 \Rightarrow x = c_1, \dots, c_n, \quad c_1 < \dots < c_n.$$

Then the critical points divide the phase space, i.e. the x -axis, into $n + 1$ horizontal intervals $U_1, \dots, U_{n+1}$. The non-constant solutions can only lie in one of the interval according to the following analysis.

1. Each non-constant solution is confined to a single interval U_i , because otherwise some solution curve must touch or cross a critical point c_j , violating the *Existence and Uniqueness Theorem* 33.3.
2. The continuity of f dictates that f cannot change signs in each U_i . It follows that a non-constant solution is either increasing or decreasing in U_i .
3. If $x(t)$ is bounded above (or below) by c_i , then either $\lim_{t \rightarrow \infty} x(t) = c_i$ or $\lim_{t \rightarrow -\infty} y(t) = c_i$.
4. If $x(t)$ is bounded above and below by two consecutive critical points, then $x(t)$ must approach the critical points, one as $t \rightarrow \infty$ and the other as $t \rightarrow -\infty$.

Let us apply the above rules to study a logistic equation.

Example 33.63. Consider the logistic equation

$$\frac{dP}{dt} = P(a - bP), \quad P \in \mathbb{R}$$

where a and b are positive constants. Find the equilibrium points and sketch the phase portrait of this equation.

Solution: From the logistic equation, we learn that the phase space $U = \mathbb{R}$.

Step 1: Solving $f(P) = P(a - bP) = 0$ gives the critical points, i.e. the equilibrium points

$$P_1 = 0, \quad P_2 = \frac{a}{b}.$$

Step 2: Study the sign of $f(P)$ over the intervals $U_1 = (-\infty, 0)$, $U_2 = (0, \frac{a}{b})$ and $U_3 = (\frac{a}{b}, \infty)$:

- Over U_1 , $f(P) = -\text{ve} \times +\text{ve} < 0$
- Over U_2 , $f(P) = +\text{ve} \times +\text{ve} > 0$
- Over U_3 , $f(P) = +\text{ve} \times -\text{ve} < 0$

Step 3: Sketch the phase portrait over the phase line $\mathbb{R}$:

$$\longleftarrow 0 \longrightarrow \frac{a}{b} \longleftarrow$$

Example 33.64 (Boyce, 2.6, P63, Q9). Sketch the phase portrait of the equation

$$\dot{x} = x^2(x^2 - 1), \quad x \in \mathbb{R}.$$

Solution: Step 1: Solve $x^2(x^2 - 1) = 0$ gives three equilibrium points

$$x_1 = -1, \quad x_2 = 0, \quad x_3 = 1.$$

Step 2: Determine the sign of $f(x) = x^2(x^2 - 1)$ for the following intervals:

x	$(-\infty, -1)$	$(-1, 0)$	$(0, 1)$	$(1, \infty)$
$f(x)$	> 0	< 0	< 0	> 0

Step 3: Sketch the phase portrait over the phase line $\mathbb{R}$:

$$\longrightarrow -1 \longleftarrow 0 \longleftarrow 1 \longrightarrow$$

Based on the concept of “attraction”, the critical points can be classified into sinks, sources and nodes.

Definition 33.24. A *sink* (or *attractor*) is a critical point $\mathbf{x}_c$ where the solutions with initial conditions sufficiently close to $\mathbf{x}_c$ are asymptotic to $\mathbf{x}_c$ as $t \rightarrow \infty$ (arrows on the phase line point toward the point).

A *source* (or *repeller*) is a critical point $\mathbf{x}_c$ where the solutions with initial conditions sufficiently close to $\mathbf{x}_c$ move away from $\mathbf{x}_c$ as $t \rightarrow \infty$.

By definition, -1 is a sink, 1 is a source and 0 is neither a sink nor a source in Example 33.64. leg:boyce:p63q9

Example 33.65 (Final Exam Sept 2016, Q6(a)). Consider the nonlinear differential equation on $\mathbb{R}$: $\dot{x} = (x-1)^2(x^2-4)$.

- Find all of the critical points. (2 marks)

Solution. $\dot{x} = 0 \Rightarrow x = 1, 2, -2$[0.5+0.7+0.8=2 marks] □

- Classify the stability of each critical point. Justify your answers. (6 marks)

Solution. Let ϵ be a very small value.

For $x \in (1-\epsilon, 1+\epsilon)$, when $x < 1$, $\dot{x} < 0$ and when $x > 1$, $\dot{x} < 0$. Therefore, $x = 1$ is a **semi-stable** critical point.[2 marks]

For $x \in (2-\epsilon, 2+\epsilon)$, when $x < 2$, $\dot{x} < 0$ and when $x > 2$, $\dot{x} > 0$. Therefore, $x = 2$ is an **unstable** critical point.[2 marks]

For $x \in (-2-\epsilon, -2+\epsilon)$, when $x < -2$, $\dot{x} > 0$ and when $x > -2$, $\dot{x} < 0$. Therefore, $x = -2$ is a **stable** critical point.[2 marks] □

33.10 Two-Dimensional Phase Space Analysis

In this section, we will investigate the phase space analysis of a two-dimensional autonomous system.

Definition 33.25. When $n = 2$, the system (33.9) is called a *plane* (or *two-dimensional*) *autonomous system* and is written as

$$\begin{aligned}\dot{x} &= f(x, y), \\ \dot{y} &= g(x, y), \quad (x, y) \in U \subset \mathbb{R}^2.\end{aligned}\tag{33.26}$$

The phase space U is also called a *phase plane*.

33.10.1 Classification of Integral Curve Portrait of Autonomous Constant Linear System

When (33.26) is an autonomous constant linear system, it has the form

$$\dot{\mathbf{x}} = A\mathbf{x} + \mathbf{b} \tag{*}$$

where A is a 2×2 constant matrix and $\mathbf{b}$ is a constant vector. The critical point could be found by solving

$$A\mathbf{x}_c + \mathbf{b} = \mathbf{0} \Rightarrow \mathbf{x}_c = -A^{-1}\mathbf{b}.$$

when the determinant of A is not zero.

Mathematicians discovered that by introducing the transform $\bar{\mathbf{x}} = \mathbf{x} - \mathbf{x}_c$, we can transform (*) to a homogeneous linear system (33.11):

$$\dot{\bar{\mathbf{x}}} + \dot{\mathbf{x}}_c = A(\bar{\mathbf{x}} + \mathbf{x}_c) + \mathbf{b} = A\bar{\mathbf{x}} + A(-A^{-1}\mathbf{b}) + \mathbf{b} \Rightarrow \dot{\bar{\mathbf{x}}} = A\bar{\mathbf{x}}. \tag{33.27}$$

When the determinant of A is not zero, $\bar{\mathbf{x}} = \mathbf{0}$ is the only critical point of the system (33.27) and all the eigenvalues λ_1 and λ_2 of A are non-zero. leg:ode:auto2d:lin

By applying the technique in Section 33.3, the trajectories of the system (33.27) can be obtained. leg:ode:auto2d:lin

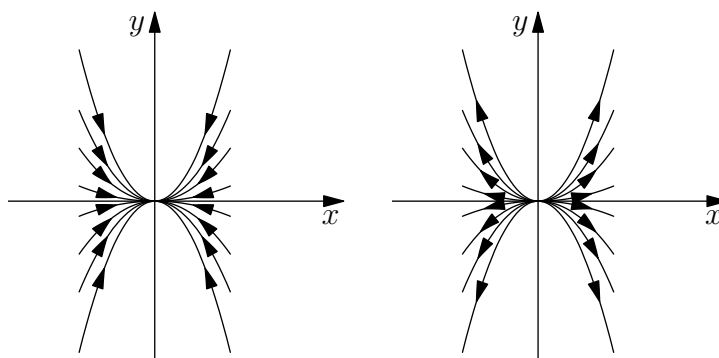
Real unequal eigenvalues of the same sign $\Rightarrow$ Node

In this case, either $\lambda_1 < \lambda_2 < 0$ or $0 < \lambda_1 < \lambda_2$. The general solution of the system has the form

$$\mathbf{x}(t) = C_1 e^{\lambda_1 t} \mathbf{v}_1 + C_2 e^{\lambda_2 t} \mathbf{v}_2$$

where $\mathbf{v}_1$ and $\mathbf{v}_2$ are linearly independent eigenvectors corresponding to λ_1 and λ_2 , respectively.

The critical point $\mathbf{0}$ of this type is called a *node*. The type of trajectories looks the same except that all the arrow heads on the trajectories as demonstrated below.



Real unequal eigenvalues of opposite sign $\Rightarrow$ Saddle Point

In this case, either $\lambda_2 < 0 < \lambda_1$ or $\lambda_1 < 0 < \lambda_2$. The general solution of the system has

$$\mathbf{x}(t) = C_1 e^{\lambda_1 t} \mathbf{v}_1 + C_2 e^{\lambda_2 t} \mathbf{v}_2.$$

If a solution starts on the line through $\mathbf{v}_1$ (hence $c_2 = 0$), it will remain on that line.

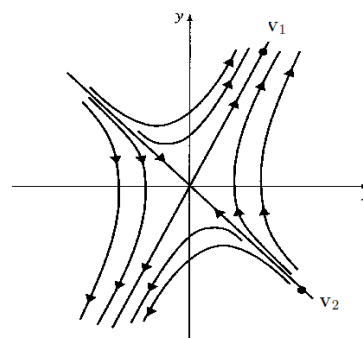
Also, since $\lambda_1 > 0$, the solution $\mathbf{x}(t)$ approaches infinity (i.e., the distance from $\mathbf{x}(t)$ to the origin approaches infinity) as $t \rightarrow \infty$.

Similarly, if a solution starts on the line through $\mathbf{v}_2$, it will remain on that line except that now the solution $\mathbf{x}(t)$ approaches the origin as $t \rightarrow \infty$.

When t is a large positive number, the term involving the positive exponential λ_1 is the dominant term, the solution $\mathbf{x}(t)$ will approach infinity in a direction tangent to the line determined by $\mathbf{v}_1$.

When t is a large negative number, the term involving the negative eigenvalue λ_2 is the dominant term, so the solution asymptotically “comes from” a direction determined by $\mathbf{v}_2$.

The critical point of this type is called a *saddle point*.



Saddle point real eigenvalues of opposite signs

Equal eigenvalues $\Rightarrow$ Proper and Improper Nodes

In this case, $\lambda_2 = \lambda_1 = \lambda$. There are two possible outcomes for the eigenvectors of A :

1. There are two independent eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$: In this case, the critical point is a *proper node*, where the general solution is

$$\mathbf{x}(t) = e^{\lambda t} (C_1 \mathbf{v}_1 + C_2 \mathbf{v}_2)$$

which is a product of a constant vector and a scalar function of t . Hence, the trajectories are lines through the origin.

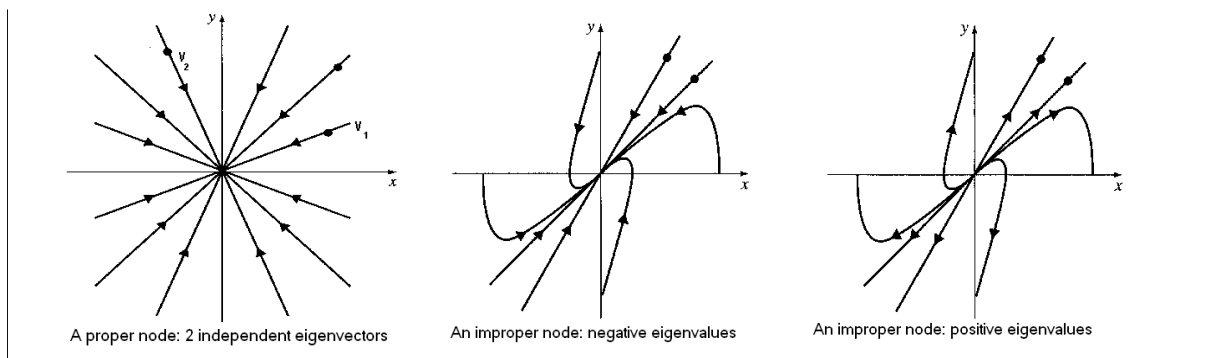
2. One linearly independent eigenvector $\mathbf{v}_1$: In this case, the critical point is an *improper node*, where the general solution is

$$\mathbf{x}(t) = e^{\lambda t} [C_1 \mathbf{v}_1 + C_2 (t\mathbf{v}_1 + \mathbf{v}_2)]$$

where $\mathbf{v}_2$ is found using the technique mentioned in Subsection [33.3.3](#). [sec:eigv:repeat]

If $\lambda < 0$, for large t , the dominant term is $C_2 t e^{\lambda t} \mathbf{v}_1$. Hence, as $t \rightarrow \infty$, all solutions approach $(0, 0)$ tangent to the line through $\mathbf{v}_1$.

When $\lambda > 0$, all solutions would move away from the improper node.



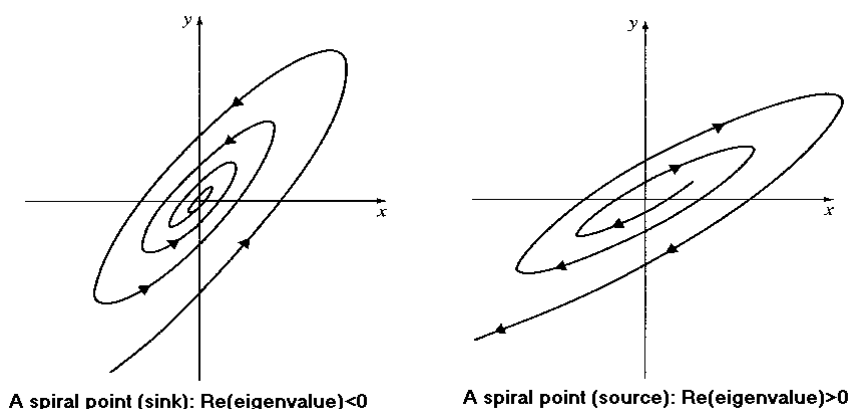
Complex conjugate but not purely imaginary eigenvalues $\Rightarrow$ Spiral (or Focus) Point

In this case, $\lambda_1 = \alpha + i\beta$, $\lambda_2 = \alpha - i\beta$, $\beta \neq 0$. If the corresponding eigenvectors are $\mathbf{v} = \mathbf{a} + i\mathbf{b}$ and $\mathbf{v}^* = \mathbf{a} - i\mathbf{b}$, the general solution can be written in the form

$$\mathbf{x}(t) = C_1 e^{\alpha t} (\mathbf{a} \cos \beta t - \mathbf{b} \sin \beta t) + C_2 e^{\alpha t} (\mathbf{b} \cos \beta t + \mathbf{a} \sin \beta t).$$

Since $\mathbf{a} \cos \beta t - \mathbf{b} \sin \beta t$ and $\mathbf{b} \cos \beta t + \mathbf{a} \sin \beta t$ oscillates, the trajectories will spiral around the critical point $(0,0)$. This type of critical point is called a *spiral (point)* or a *focus*.

- When $\alpha < 0$, the trajectories spiral in toward $(0,0)$.
- When $\alpha > 0$, the trajectories spiral out away from $(0,0)$.



Purely imaginary eigenvalues $\Rightarrow$ Centre

In this case, $\lambda_1 = \beta i$, $\lambda_2 = -\beta i$, $\beta \neq 0$. If the corresponding eigenvectors are $\mathbf{v} = \mathbf{a} + i\mathbf{b}$ and $\mathbf{v}^* = \mathbf{a} - i\mathbf{b}$, the general solution can be written in the form

$$\mathbf{x}(t) = C_1 (\mathbf{a} \cos \beta t - \mathbf{b} \sin \beta t) + C_2 (\mathbf{b} \cos \beta t + \mathbf{a} \sin \beta t) = \begin{bmatrix} R_1 \cos(\beta t - \delta_1) \\ R_2 \cos(\beta t - \delta_2) \end{bmatrix}$$

for some $R_1, R_2, \delta_1, \delta_2$. Hence, the trajectories are ellipses centred at $(0,0)$. This type of critical point is called a *centre point*.

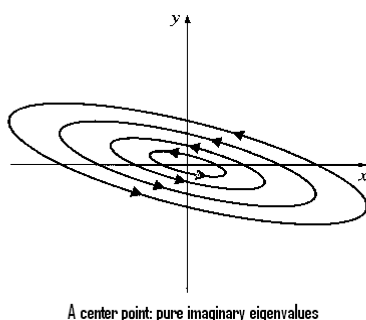


Table 33.1 summarises the classification of the types of critical point $(0,0)$ of (33.27) where $|A| \neq 0$.

Table 33.1: Classification of critical points

tbl:ode5-1

Eigenvalues of A	Type of critical point
Real, unequal, same sign	Node
Real, unequal, opposite sign	Saddle point
Real and equal	Proper or improper node
Complex conjugates but not pure imaginary	Spiral point
Pure imaginary	Centre point

Example 33.66 (Farlow, Ex8.2, Q1-8). Determine whether the critical point $(0, 0)$ of the system

$$\dot{x} = -2x + y, \quad \dot{y} = -x - 2y$$

is a node, saddle point, spiral point, or centre point.

Solution: From the characteristic equation

$$\begin{vmatrix} -2 - \lambda & 1 \\ -1 & -2 - \lambda \end{vmatrix} = \lambda^2 + 4\lambda + 4 + 1 = 0$$

we obtain the eigenvalues $\lambda = \pm 2i$, therefore $(0, 0)$ is a spiral point.

33.10.2 Classification of Phase Portrait of Autonomous Constant Linear System

In the previous subsection, the classification of integral curves concerns only with the shape of the solutions and the direction is unimportant. In this subsection, we classify the phase portrait (the set of solutions), where the direction is related to the (Lyapunov) stability (refer to Subsection 33.2.2) of the steady solution (i.e. the critical point).

First, we will state the stability theorem.

thm:STBLS

Theorem 33.67 (Stability Theorem). *Let λ_1 and λ_2 be the eigenvalues of the non-singular matrix A in the two-dimensional linear system (33.27). Then the critical point $(0, 0)$ is*

1. asymptotically stable if the real parts of λ_1 and λ_2 are both negative;
2. stable but not asymptotically stable if the real parts of λ_1 and λ_2 are both zero, i.e., $\lambda_1, \lambda_2 = \pm i\beta$;
3. unstable if either λ_1 or λ_2 has a positive real part.

The classification of the origin as a critical point with the introduction of stability is summarised in Table 33.2.

Table 33.2: Classification of critical points using stability

tbl:ode5-2

Eigenvalues	Type of critical point	Stability
$\lambda_1 > \lambda_2 > 0$	Node	Unstable
$\lambda_1 < \lambda_2 < 0$	Node	Asymptotically stable
$\lambda_2 < 0 < \lambda_1$	Saddle point	Unstable
$\lambda_1 = \lambda_2 > 0$	Proper or improper node	Unstable
$\lambda_1 = \lambda_2 < 0$	Proper or improper node	Asymptotically stable
$\lambda_1, \lambda_2 = \alpha \pm \beta i$		
$\alpha > 0$	Spiral point	Unstable
$\alpha < 0$	Spiral point	Asymptotically stable
$\lambda_1, \lambda_2 = \pm \beta i$	Centre point	Stable

Example 33.68. Classify the critical point at the origin for the system

$$\dot{\mathbf{x}} = \begin{bmatrix} 1 & -3 \\ -3 & 1 \end{bmatrix} \mathbf{x}.$$

Solution: The eigenvalues are found to be $\lambda = -2, 4$. Therefore $(0, 0)$ is an unstable saddle point.

Example 33.69 (Stability of the Mass-Spring). Find the critical point and determine the stability of the frictionless mass-spring equation

$$\ddot{x} + \omega_0^2 x = 0.$$

Sketch the phase portrait at the critical point.

Solution: Rewriting the ODE as a system of first order ODE:

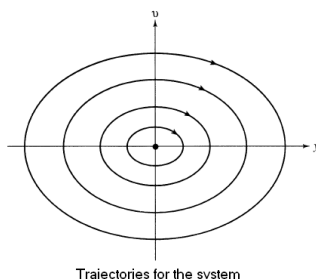
$$\begin{cases} \dot{x} = y \\ \dot{y} = -\omega_0^2 x \end{cases} \Rightarrow \begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega_0^2 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

The critical point is $(0, 0)$ and the eigenvalues at this critical point is

$$\lambda_1, \lambda_2 = \pm i\omega_0.$$

This is a centre. It is stable but not asymptotically stable. This reflects the fact that if the mass in a frictionless mass-spring system is moved slightly from the equilibrium position, it will not move far away from this position, but it will also not return to equilibrium.

The sketch of the phase portrait at critical point $(0, 0)$ is shown below



Example 33.70 (Phase Plane Analysis of Vibrating Spring). The equation of motion of the vibrating mass-spring system is

$$m\ddot{u} + d\dot{u} + ku = 0$$

where m , d , and k are positive constants.

Write the second-order equation as a system of two first-order equations by letting $x = u$, $y = \dot{u}$, and show that $x = y = 0$ is a critical point of this system. Classify this type of critical point as a function of m , d , and k . Determine whether the critical point is stable, asymptotically stable, or unstable.

Solution: By letting $x = u$, $y = \dot{u} = \dot{x}$, the equation of motion becomes

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= -\left(\frac{d}{m}\right)y - \left(\frac{k}{m}\right)x \end{aligned}$$

$(0, 0)$ is the critical point and the characteristic equation at $(0, 0)$ is

$$\begin{vmatrix} -\lambda & 1 \\ -\frac{k}{m} & -\frac{d}{m} - \lambda \end{vmatrix} = \lambda^2 + \frac{d}{m}\lambda + \frac{k}{m} = 0 \Rightarrow \lambda = \frac{-\frac{d}{m} \pm \sqrt{\left(\frac{d}{m}\right)^2 - 4\left(\frac{k}{m}\right)}}{2}$$

- When $d^2 - 4mk < 0$, $(0, 0)$ is a spiral point (asymptotically stable);
- When $d^2 - 4mk = 0$, $(0, 0)$ is an improper node (asymptotically stable) because the eigenvalue is repeated and degenerate.

- When $d^2 - 4mk > 0$, $(0, 0)$ is a proper node (asymptotically stable).

By a simple change of variables, Theorem 33.67 can be used to classify critical points that are not located at the origin. However, (33.27) implies that the nonhomogeneous equation and its corresponding homogeneous equation has the same eigenvalues.

Example 33.71. Given the linear system $\frac{dx}{dt} = 2x + y + 3$, $\frac{dy}{dt} = -3x - 2y - 4$. Determine (a) the critical point (x_c, y_c) ; (b) the type and stability of (x_c, y_c) .

Solution: (a) The critical point $(x_c, y_c) = (-2, 1)$ is calculated by definition:

$$2x + y + 3 = -3x - 2y - 4 = 0 \Rightarrow x = -2, y = 1$$

By change of coordinate $x = u + 2$, $y = v - 1$, we know that

$$\begin{aligned} \dot{u} &= 2u + v, \\ \dot{v} &= -3x - 2y. \end{aligned} \Rightarrow (2 - \lambda)(-2 - \lambda) + 3 = \lambda^2 - 1 = 0 \Rightarrow \lambda = \pm 1$$

The critical point is an unstable saddle point.

Example 33.72 (Final Exam Sept 2014, Q4(c)). By using linear algebra technique but without solving the equations, determine whether the solutions $x_1(t)$, $x_2(t)$ and $x_3(t)$ of the system of equations below is stable, asymptotically stable, or unstable:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} -3 & 0 & 0 \\ -2 & -1 & 2 \\ -1 & -2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

(6 marks)

Solution. The characteristic equation is

$$\begin{vmatrix} -3 - \lambda & 0 & 0 \\ -2 & -1 - \lambda & 2 \\ -1 & -2 & -1 - \lambda \end{vmatrix} = (-3 - \lambda) \begin{vmatrix} -1 - \lambda & 2 \\ -2 & -1 - \lambda \end{vmatrix} = (-3 - \lambda)(\lambda^2 + 2\lambda + 1 + 4) = 0 \quad [2 \text{ marks}]$$

The characteristic values are $\lambda = -3, -1 + 2i, -1 - 2i$ [3 marks]

Since the real parts are all negative, therefore x_1, x_2 and x_3 are all asymptotically stable.

..... [1 mark]

□

33.10.3 Classification of Phase Portrait of Almost Linear Systems

In this section, we generalise the classification of phase portrait to almost linear system, a nonlinear system which resembles a linear system around the critical points.

Definition 33.26 (Almost Linear System). A two-dimensional autonomous system (33.9) is called an *almost linear system at an critical point* $\mathbf{x}_c$ if

1. $\mathbf{f}(\mathbf{x})$ is a continuous vector-valued function whose component functions possess continuous partial derivatives in an open region containing the critical point $\mathbf{x}_c$.
2. The Jacobian matrix $A = D\mathbf{f}(\mathbf{x}_c)$ is invertible.
3. $\mathbf{f}$ satisfies $\lim_{\mathbf{x} \rightarrow \mathbf{x}_c} \frac{\|\mathbf{f}(\mathbf{x}) - A(\mathbf{x} - \mathbf{x}_c)\|}{\|\mathbf{x} - \mathbf{x}_c\|} = 0$.

Remark 33.73. $\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2}$ is the *length* of a vector $\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$.

For an almost linear system (33.9) (or (33.26)), the trajectories of the linearization approximate those of the nonlinear system in a neighbourhood of $\mathbf{x}_c$. Since the nonlinear system behaves essentially like its linearization in the neighbourhood of $\mathbf{x}_c$, its phase portrait should be more or less the same as its linearization defined below (we can treat this as a special case of Hartman-Grobman Theorem).

Definition 33.27 (Linearization). Let (x_c, y_c) be an critical point of the nonlinear system (33.26), where f and g are twice-differentiable functions. The *linearization* of (33.26) at the critical point (x_c, y_c) is the linear system

$$\begin{bmatrix} \dot{v} \\ \dot{w} \end{bmatrix} = \underbrace{\begin{bmatrix} f_x(x_c, y_c) & f_y(x_c, y_c) \\ g_x(x_c, y_c) & g_y(x_c, y_c) \end{bmatrix}}_{\text{Jacobian matrix, } D\mathbf{f}(\mathbf{x}_c)} \begin{bmatrix} v \\ w \end{bmatrix}$$

Example 33.74. Determine all critical points for the system

$$\dot{x} = x(1 - y), \quad \dot{y} = y(2 - x),$$

and determine the linearization to the system at each critical point.

Solution: Solving $x(1 - y) = y(2 - x)$ gives the critical points

$$(x_c, y_c) = (0, 0), \quad (2, 1)$$

The Jacobian matrix is

$$D\mathbf{f}(x, y) = \begin{bmatrix} 1 - y & -x \\ -y & 2 - x \end{bmatrix}$$

The linearization at $(0, 0)$ is

$$\begin{bmatrix} \dot{v} \\ \dot{w} \end{bmatrix} = D\mathbf{f}(0, 0) \begin{bmatrix} v \\ w \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} \Rightarrow \begin{cases} \dot{v} = v \\ \dot{w} = 2w \end{cases}$$

The linearization at $(2, 1)$ is

$$\begin{bmatrix} \dot{v} \\ \dot{w} \end{bmatrix} = D\mathbf{f}(2, 1) \begin{bmatrix} v \\ w \end{bmatrix} = \begin{bmatrix} 0 & -2 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} \Rightarrow \begin{cases} \dot{v} = -2w \\ \dot{w} = -v \end{cases}$$

It seems reasonable to expect that we can determine the stability and type of a given critical point of a nonlinear system from a study of its linearization. This is indeed true in most cases according to the following theorem by Poincaré.

Theorem 33.75 (Poincaré's Stability Theorem). If (33.26) is almost linear at an critical point, $\mathbf{x}_c$. Let $\dot{\mathbf{z}} = A\mathbf{z}$ be the associated linearized system (ref. Definition 33.27) where

$$A = D\mathbf{f}(x_c, y_c) = \begin{bmatrix} f_x(x_c, y_c) & f_y(x_c, y_c) \\ g_x(x_c, y_c) & g_y(x_c, y_c) \end{bmatrix}.$$

Then the type and stability of the critical point $\mathbf{x}_c = (x_c, y_c)$ of the almost linear system and are the same as those of the critical point $\mathbf{0} = (0, 0)$ for the associated linear system as follows:

1. If $\mathbf{z} = \mathbf{0}$ is an asymptotically stable critical point of $\dot{\mathbf{z}} = A\mathbf{z}$, then $\mathbf{x}_c$ is an asymptotically stable critical point of (33.26).
2. If $\mathbf{z} = \mathbf{0}$ is an unstable critical point of $\dot{\mathbf{z}} = A\mathbf{z}$, then $\mathbf{x}_c$ is an unstable critical point of (33.26).
3. If $\mathbf{z} = \mathbf{0}$ is a stable (but not asymptotically stable) critical point of $\dot{\mathbf{z}} = A\mathbf{z}$, then no conclusions can be drawn about stability properties of critical point $\mathbf{x}_c$.

The classification of the type of critical points for almost linear system and its associated linear system according to Poincaré Stability Theorem is given in Table 33.3.

Remark 33.76. When the associated linearized system is asymptotically stable, both eigenvalues of the coefficient matrix A have negative real parts. Thus, $\|\mathbf{z}\|$ is exponentially decreasing to 0. If the nonlinear perturbation is sufficiently weak, we might expect this behaviour to persist in the nonlinear system.

Similarly, if the linearized system is unstable, at least one of the eigenvalues of A has a positive real part. The linearized system therefore has some solutions that grow exponentially in norm. In this case, we might expect instability to persist when the nonlinear perturbation is sufficiently weak.

Table 33.3: Classification of critical points of an almost linear system.

tbl:ode5-3

Eigenvalues	Associated linear system		Almost linear system	
	Type	Stability	Type	Stability
$\lambda_1 > \lambda_2 > 0$	Node	Unstable	Node	Unstable
$\lambda_1 < \lambda_2 < 0$	Node	Asymptotically stable	Node	Asymptotically stable
$\lambda_2 < 0 < \lambda_1$	Saddle point	Unstable	Saddle point	Unstable
$\lambda_1 = \alpha + \beta i$ $\lambda_2 = \alpha - \beta i$ $\alpha > 0$ $\alpha < 0$	Spiral point	unstable	Spiral point	unstable
	Spiral point	Asymptotically stable	Spiral point	Asymptotically stable
$\lambda_1 = \lambda_2 > 0$	Proper or improper node	Unstable	Proper, improper node or Spiral point	Unstable
$\lambda_1 = \lambda_2 < 0$	Proper or improper node	Asymptotically stable	Proper, improper node or Spiral point	Asymptotically stable
$\lambda_1, \lambda_2 = \pm \beta i$	Centre point	stable	Centre point or Spiral point	Indeterminate

Finally, if the linearized system is stable but not asymptotically stable, then the eigenvalues of A are purely imaginary. In this case, the linearized system is sitting on the fence between stability and instability. It is possible for the nonlinear perturbation (however small) to tip the balance either way—causing the nonlinear system to be stable or unstable.

Let us look at a few examples of the Poincaré's Stability Theorem.

Example 33.77. Show that the system is almost linear at the critical point $(0, 0)$, and discuss the type and stability of the critical point $(0, 0)$ of the almost linear system

$$\dot{x} = x + y - 2xy, \quad \dot{y} = -2x + y + 3y^2.$$

Solution: (a) The system is almost linear at $(0, 0)$ because

$$\begin{aligned}
 \lim_{(x,y) \rightarrow (0,0)} \frac{\left\| \begin{bmatrix} x + y - 2xy \\ -2x + y + 3y^2 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \right\|}{\left\| \begin{bmatrix} x \\ y \end{bmatrix} \right\|} &= \lim_{(x,y) \rightarrow (0,0)} \frac{\sqrt{(-2xy)^2 + (3y^2)^2}}{\sqrt{x^2 + y^2}} \\
 &= \lim_{r \rightarrow 0^+} \frac{\sqrt{4r^4 \cos^2 \theta \sin^2 \theta + 9r^4 \sin^4 \theta}}{\sqrt{r^2 \cos^2 \theta + r^2 \sin^2 \theta}} \\
 &= \lim_{r \rightarrow 0^+} r \sin \theta \sqrt{4 \cos^2 \theta + 9 \sin^2 \theta} = 0
 \end{aligned}$$

The linearization has the following characteristic equation at $(0, 0)$:

$$\begin{vmatrix} 1 - \lambda & 1 \\ -2 & 1 - \lambda \end{vmatrix} = (1 - \lambda)^2 + 2 = 0 \Rightarrow \lambda = 1 \pm \sqrt{2}i.$$

Therefore $(0, 0)$ is an unstable spiral point.

Example 33.78. Verify that $(1, 0)$ is an critical point of the system

$$\dot{x} = x(x - 1), \quad \dot{y} = y(2 + xy^2)$$

and show that the system is almost linear, and discuss the type and stability of the critical point $(1, 0)$.

Solution: It is easy to verify $\mathbf{x}_c = (1, 0)$ is a critical point since

$$x(x-1) = 1(1-1) = 0, \quad y(2+xy^2) = 0(2+1 \cdot 0^2) = 0.$$

$$A = D\mathbf{f}(1, 0) = \begin{bmatrix} 2x-1 & 0 \\ y^3 & 2+3xy^2 \end{bmatrix} \Big|_{(x,y)=(1,0)} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

To show the system is almost linear by definition, we need to prove

$$\begin{aligned} \lim_{(x,y) \rightarrow (1,0)} \frac{\left\| \begin{bmatrix} x(x-1) \\ y(2+xy^2) \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} x-1 \\ y-0 \end{bmatrix} \right\|}{\left\| \begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\|} &= \lim_{(x,y) \rightarrow (1,0)} \frac{\sqrt{(x-1)^4 + (xy^2)^2}}{\sqrt{(x-1)^2 + y^2}} \\ &= \lim_{r \rightarrow 0^+} \frac{\sqrt{r^4 \cos^4 \theta + r^4 \sin^4 \theta (1+r \cos \theta)^2}}{r} \\ &= \lim_{r \rightarrow 0^+} r \sqrt{\cos^4 \theta + \sin^4 \theta (1+r \cos \theta)^2} = 0 \end{aligned}$$

where $x-1 = r \cos \theta$ and $y = r \sin \theta$.

Therefore, the system is almost linear at $(1, 0)$. The critical point is a node which is asymptotically stable by the Poincaré Stability Theorem.

As can be observed from the above examples, showing that a system is almost linear is complicated. In general, for an analytic system (which is generally the case in this subject), the Hartman-Grobman Theorem (which includes the Poincaré Stability Theorem) says that if all the real part of the eigenvalues of the Jacobian matrix $D\mathbf{f}(\mathbf{x}_c)$ are non-zero. If there is a single zero real part, further analysis is required to determine the exact type.

Example 33.79. Determine the type and stability of the critical point $(0, 0)$ of the system $\dot{\mathbf{x}} = \begin{bmatrix} 1 & 1 \\ -2 & 1 \end{bmatrix} \mathbf{x} + \begin{bmatrix} -2xy \\ 3y^3 \end{bmatrix}$

Solution: The system is analytic and the linearization at $(0, 0)$

$$\dot{\mathbf{x}} = \begin{bmatrix} 1 & 1 \\ -2 & 1 \end{bmatrix} \mathbf{x}$$

has real eigenvalues $\lambda = -1, 3$ which are non-zero. According to Poincaré's Stability Theorem, $(0, 0)$ is a saddle point.

Example 33.80 (Nagle, 12.3, Q9 P764). Find all the critical points for the given system

$$\dot{x} = 16 - xy, \quad \dot{y} = x - y^3$$

and discuss the type and stability of the critical points.

Solution: (a) Determine all critical points:

$$\left. \begin{aligned} 16 - xy &= 0 \\ x - y^3 &= 0 \end{aligned} \right\} \Rightarrow (x = y^3) \wedge (16 = xy) \Rightarrow (x = y^3) \wedge (16 = y^4) \Rightarrow y = \pm 2$$

Therefore, the critical points are $(8, 2)$ and $(-8, -2)$.

(b) The Jacobian of the system is

$$D\mathbf{f}(\mathbf{x}) = \begin{bmatrix} -y & -x \\ 1 & -3y^2 \end{bmatrix}$$

At $\mathbf{x}_c = (8, 2)$,

$$D\mathbf{f}(8, 2) = \begin{bmatrix} -2 & -8 \\ 1 & -12 \end{bmatrix} \Rightarrow (-2 - \lambda)(-12 - \lambda) + 8 = 0 \Rightarrow \lambda^2 + 14\lambda + 32 = 0$$

By using calculator, we can find the answers to be approximately $\lambda \approx -11.1231, -2.8769$. Hence, $(8, 2)$ is asymptotically stable proper node.

At $\mathbf{x}_c = (-8, -2)$,

$$Df(8, 2) = \begin{bmatrix} 2 & 8 \\ 1 & -12 \end{bmatrix} \Rightarrow (2 - \lambda)(-12 - \lambda) - 8 = 0 \Rightarrow \lambda^2 + 10\lambda - 32 = 0.$$

By using calculator, we can find the answers to be approximately $\lambda \approx -12.5498, 2.5498$. Hence, $(-8, -2)$ is an unstable saddle point.

Example 33.81. Determine and classify all critical points for the system

$$\dot{x} = x - y, \quad \dot{y} = y(2x + y - 3).$$

Solution. (a) Determine all critical points:

$$\left. \begin{array}{l} x - y = 0 \\ y(2x + y - 3) = 0 \end{array} \right\} \Rightarrow (x = y) \wedge (y = 0 \vee 3y - 3 = 0) \Rightarrow x = y = 0 \vee x = y = 1.$$

Therefore, we have two critical points $(0, 0)$ and $(1, 1)$.

(b) The Jacobian of the system is

$$Df(\mathbf{x}) = \begin{bmatrix} 1 & -1 \\ 2y & 2x + 2y - 3 \end{bmatrix}$$

At $\mathbf{x}_c = (0, 0)$,

$$Df(0, 0) = \begin{bmatrix} 1 & -1 \\ 0 & -3 \end{bmatrix} \Rightarrow \lambda = 1, -3$$

Therefore, $(0, 0)$ is an unstable saddle point.

At $\mathbf{x}_c = (1, 1)$,

$$Df(1, 1) = \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix} \Rightarrow \lambda = 1 \pm i\sqrt{2}.$$

Therefore, $(1, 1)$ is an unstable spiral point. □

Example 33.82 (Nagle, 12.3, Q1,3,5,7 P764). For each of the following system, (a) Find the Jacobian; (b) determine the type and stability of the critical point $(0, 0)$.

1. $\begin{cases} \dot{x} = 3x + 2y - y^2, \\ \dot{y} = -2x - 2y + xy \end{cases}$ (a) $\begin{bmatrix} 3 & 2 \\ -2 & -2 \end{bmatrix}$ (b) $\lambda = 2, -1$;
Unstable saddle point
2. $\begin{cases} \dot{x} = x - y, \\ \dot{y} = 3x + 5y - x^3 \end{cases}$ (a) $\begin{bmatrix} 1 & -1 \\ 3 & 5 \end{bmatrix}$ (b) $\lambda = 2, 4$;
Unstable proper node
3. $\begin{cases} \dot{x} = x + 5y - y^2, \\ \dot{y} = -x - y - y^2 \end{cases}$ (a) $\begin{bmatrix} 1 & 5 \\ -1 & -1 \end{bmatrix}$ (b) $\lambda = \pm 2i$;
Undetermined spiral?
centre?
4. $\begin{cases} \dot{x} = e^{x+y} \cos x, \\ \dot{y} = \cos y + x - 1 \end{cases}$ (a) $\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$ (b) $\lambda \approx -0.618, 1.618$;
Unstable saddle point

Example 33.83. Discuss the type and stability of the critical points of the first order system corresponding to the equation

$$\ddot{x} + x - 0.25x^2 = 0.$$

Solution: Let $y = \dot{x}$, then the equation becomes

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= \ddot{x} = -x + 0.25x^2 \end{aligned}$$

From

$$\left. \begin{array}{l} y = 0 \\ -x + 0.25x^2 = 0 \end{array} \right\} \Rightarrow (y = 0) \wedge (x = 0 \vee x = 4)$$

So we obtain the critical points $(0, 0)$ and $(4, 0)$.

The Jacobian of the system is

$$D\mathbf{f}(\mathbf{x}) = \begin{bmatrix} 0 & 1 \\ -1 + 0.5x & 0 \end{bmatrix}$$

At $(0, 0)$, $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ has eigenvalues $\lambda = \pm i$. By Poincaré's Stability Theorem, $(0, 0)$ is either a centre point or a spiral point of the nonlinear system.

Remark: This theorem tells us nothing exactly type it is. In the mathematical theory of Hamiltonian mechanics, $(0, 0)$ is found to be a centre point.

At $(4, 0)$, $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ has eigenvalues $\lambda = \pm 1$. By Poincaré's Stability Theorem, $(4, 0)$ is an unstable saddle point.

Example 33.84. Determine the type and stability of each critical point of the autonomous system

$$\dot{x} = 2y, \quad \dot{y} = x(x - 6).$$

Solution:

Solving $2y = 0$ and $x(x - 6) = 0$, we obtain all the critical points $(0, 0)$ and $(6, 0)$.

The Jacobian of the system is

$$D\mathbf{f}(x, y) = \begin{bmatrix} 0 & 2 \\ 2x - 6 & 0 \end{bmatrix}$$

At $(0, 0)$,

$$D\mathbf{f}(0, 0) = \begin{bmatrix} 0 & 2 \\ -6 & 0 \end{bmatrix} \Rightarrow \lambda = \pm\sqrt{12}$$

This implies that $(0, 0)$ unstable saddle point.

At $(6, 0)$,

$$D\mathbf{f}(6, 0) = \begin{bmatrix} 0 & 2 \\ 6 & 0 \end{bmatrix} \Rightarrow \lambda = \pm\sqrt{12}i$$

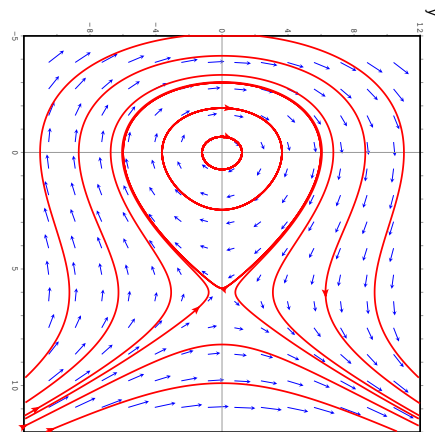
According to Poincaré's Stability Theorem, $(6, 0)$ is either a centre or a spiral point, but its stability cannot be determined.

Again, using the mathematical theory of Hamiltonian system, $(6, 0)$ is found to be a stable center point.

We can use an open source computer algebra software Maxima using the following commands

```
load("plotdf");
plotdf([2*y, x*(x-6)], [x, -5, 12], [y, -12, 12]);
```

to generate the phase portrait of this system. It is shown on the right and we can better understand the meaning of Straightening Out Theorem (when the vector at a point is non-zero, in a small neighbourhood containing the point, the vectors are nearly parallel due to continuing of vector field) and Hartman-Grobman Theorem (which says that the phase portrait at the critical point is very similar to the associated linear system).



Example 33.85 (Final Exam Sept 2014, Q3(b)). Given the following nonlinear ODEs:

$$\begin{aligned}\dot{x} &= x - x^2 - xy, \\ \dot{y} &= 4y - y^2 - 3xy.\end{aligned}$$

- (i) Determine the type of the critical point $(0, 0)$. Explain your answer in detail. (6 marks)

Solution. To study the dynamic behaviour at $\mathbf{0} = (0, 0)$, we determine its Jacobian:

$$D_{\mathbf{x}}\mathbf{f}(\mathbf{0}) = \begin{bmatrix} 1 - 2x - y & -x \\ -3y & 4 - 2y - 3x \end{bmatrix} \bigg|_{(0,0)} = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}. \quad [2+1=3 \text{ marks}]$$

The eigenvalues are

$$\begin{vmatrix} 1 - \lambda & 0 \\ 0 & 4 - \lambda \end{vmatrix} = 0 \Rightarrow \lambda = 1, 4 \quad [2 \text{ marks}]$$

The eigenvalues are real and $(0, 0)$ is a proper node.[1 mark] □

- (ii) Use the result in (i) to determine whether the critical point $(0, 0)$ is asymptotically stable, stable or unstable. Justify your answer. (2 marks)

Solution. Since one of the real part of the eigenvalues are greater than zero, ... [1 mark]

The critical point $(0, 0)$ is unstable.[1 mark] □

- (iii) Find all the other critical points. (6 marks)

Solution. By definition, the critical points satisfy

$$\begin{aligned}x - x^2 - xy &= 0 \Rightarrow x(1 - x - y) = 0 \\ 4y - y^2 - 3xy &= 0 \Rightarrow y(4 - y - 3x) = 0\end{aligned} \quad [1+1=2 \text{ marks}]$$

We can obtain the solutions:

$$\begin{aligned}x &= 0, \quad y = 0; \\ x &= 0, \quad 4 - y - 3x = 0 \Rightarrow y = 4; \\ 1 - x - y &= 0, \quad y = 0 \Rightarrow x = 1; \\ 1 - x - y &= 0, \quad 4 - y - 3x = 0 \Rightarrow 4 + (x - 1) - 3x = 0 \Rightarrow x = \frac{3}{2}, \quad y = -\frac{1}{2}\end{aligned}$$

..... [1+1+2=4 marks] □

Example 33.86 (Final Exam Sept 2014, Q5(b)). Given the system of ODEs

$$\begin{aligned}\dot{x} &= y(y^2 + 1), \\ \dot{y} &= -y - x(x^2 - 1).\end{aligned}$$

- (i) Determine all the critical points. (4 marks)

Solution. $y = 0 \Rightarrow x(x^2 - 1) = 0 \Rightarrow x = 0, 1, -1$[1 mark]

There are three critical points: $(0, 0)$, $(1, 0)$, $(-1, 0)$ [3 marks] □

- (ii) For each of the critical points in (i), determine its type (stable node, unstable node, stable focus (or spiral), unstable focus (or spiral), saddle). (7 marks)

Solution. The Jacobian is

$$D_{\mathbf{x}}\mathbf{f} = \begin{bmatrix} 0 & 3y^2 + 1 \\ -3x^2 + 1 & -1 \end{bmatrix} \quad [1 \text{ mark}]$$

At $(0, 0)$, $D_{\mathbf{x}}\mathbf{f} = \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix}$ has eigenvalues $\lambda = \frac{-1 \pm \sqrt{5}}{2}$[1 mark]

Hence, $(0, 0)$ is a *saddle* [1.5 marks]

At $(1, 0)$ and $(-1, 0)$, $D_{\mathbf{x}}\mathbf{f} = \begin{bmatrix} 0 & 1 \\ -2 & -1 \end{bmatrix}$ has eigenvalues $\lambda = \frac{-1 \pm \sqrt{7}i}{2}$ [2 marks]

Hence, $(1, 0)$ and $(-1, 0)$ are *stable spiral* [1.5 marks]

□

Example 33.87 (Final Exam Sept 2016, Q6(b)). Let a be a *positive constant*. Consider the system of nonlinear differential equations below

$$\begin{aligned}\dot{x} &= -x^3 + y, \\ \dot{y} &= -a(x + y).\end{aligned}$$

- (i) Find all the critical points of the system. (3 marks)

Solution. From

$$\begin{cases} -x^3 + y = 0 \\ -a(x + y) = 0 \end{cases} \Rightarrow y = -x \Rightarrow -x^3 - x = -x(x^2 + 1) = 0 \Rightarrow x = 0 \quad [2 \text{ marks}]$$

There is only one critical point $(0, 0)$ [1 mark]

□

- (ii) Find the Jacobian and calculate the eigenvalues of the Jacobian at the critical point(s) found in part (i). Hence, determine the stability and the type (spiral, node, saddle or centre) of the critical point(s). (5 marks)

Solution.

$$\left. \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{0}} = \begin{bmatrix} -3x^2 & 1 \\ -a & -a \end{bmatrix}_{(x,y)=(0,0)} = \begin{bmatrix} 0 & 1 \\ -a & -a \end{bmatrix} \quad [1 \text{ mark}]$$

The eigenvalues of the Jacobian at $(0, 0)$ can be calculated as follows:

$$\begin{vmatrix} -\lambda & 1 \\ -a & -a - \lambda \end{vmatrix} = \lambda^2 + a\lambda + a = 0 \Rightarrow \lambda = \frac{-a \pm \sqrt{a^2 - 4a}}{2} \quad [2 \text{ marks}]$$

When $a < 4$, the critical point is a stable spiral. [1 mark]

When $a \geq 4$, the critical point is a stable node. [1 mark]

□

- (iii) Sketch the phase portrait of the trajectories near the critical point(s) when $a = 4$. (4 marks)

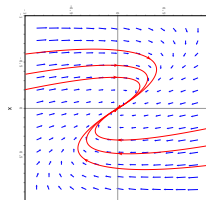
Solution. We have repetitive eigenvalues

$$\begin{vmatrix} -\lambda & 1 \\ -4 & -4 - \lambda \end{vmatrix} = 0 \Rightarrow \lambda^2 + 4\lambda + 4 = 0 \Rightarrow \lambda = -2, -2$$

The corresponding eigenvector is $\begin{bmatrix} 2 & 1 \\ -4 & -2 \end{bmatrix} \mathbf{v} = \mathbf{0} \Rightarrow \mathbf{v} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$ [1 mark]

The phase portrait at $(0, 0)$ is [3 marks]

□



33.11 Three- and Higher-Dimensional Phase Space Analysis

C++ and Boost library

```
1 // http://headmyshoulder.github.io/odeint-v2/examples.html
2
3 #include <iostream>
4 #include <boost/array.hpp>
5
6 #include <boost/numeric/odeint.hpp>
7
```

```

8 using namespace std;
9 using namespace boost::numeric::odeint;
10
11 const double sigma = 10.0;
12 const double R = 28.0;
13 const double b = 8.0 / 3.0;
14
15 typedef boost::array< double , 3 > state_type;
16
17 void lorenz( const state_type &x , state_type &dxdt , double t )
18 {
19     dxdt[0] = sigma * ( x[1] - x[0] );
20     dxdt[1] = R * x[0] - x[1] - x[0] * x[2];
21     dxdt[2] = -b * x[2] + x[0] * x[1];
22 }
23
24 void write_lorenz( const state_type &x , const double t )
25 {
26     cout << t << '\t' << x[0] << '\t' << x[1] << '\t' << x[2] << endl;
27 }
28
29 int main(int argc, char **argv)
30 {
31     state_type x = { 10.0 , 1.0 , 1.0 }; // initial conditions
32     integrate( lorenz , x , 0.0 , 25.0 , 0.1 , write_lorenz );
33 }

```

33.12 Competing Species and Predator-Prey Equation

sec:pred-prey

ODE has been successfully applied to the solution of many engineering problems related to chemical, mechanical and electrical systems. Scientists have attempted to model the dynamics of a group of species using ODE. Both Lotka [1925] and Volterra have written books to summarise these attempts in the early 20th century.

The application of ODE to biology have met with limited success. It is able to characterise the invasion of the host species by parasite [Lotka, 1925, p. 113]. However, due to the complexity of various factors and the difficulty to determine various coefficients in the ODE, the ODE cannot provide very precise prediction for the long term.

The ODE that describes the relation between n species is called the predator-prey equation.

Definition 33.28. A *predator-prey equation (or model)* of n species is an autonomous equation (33.9) such the $f(\mathbf{x})$ is given by

$$f_i(x_1, \dots, x_n) = x_i(b_i + a_{i1}x_1 + \dots + a_{in}x_n)$$

where a_{ij} and b_i are constants which need to be determined using experiments.

When $n = 2$, the predator-prey model is written as

$$\begin{aligned}\dot{x} &= x(a_1 - b_1x - c_1y), \\ \dot{y} &= y(a_2 - b_2x - c_2y).\end{aligned}\tag{33.28}$$

e:prey:2

Here, the variables $x(t)$ and $y(t)$ characterise the populations at time t of two species. The positive constants $a_1, a_2, b_1, b_2, c_1, c_2$ depend on the species and environment under investigation.

Despite its limited success, the analysis of the predator-prey model and other bio-chemical models allows scientists to come with suggestions for governments to consider. A very nice example was given in Zill [2013], which we listed it below.

Example 33.88 (C. J. Knickerbocker (St. Lawrence Univ)). Early in 1995, after much controversy, public debate, and a 70-year absence, grey wolves were reintroduced into the US' Yellowstone National Park and central Idaho. Since the extermination of the wolf in the 1920s, significant changes had been recorded in the populations of other animals residing in the part. For instance, the coyote population (as well as other predator populations) had grown in the absence of competition from the large grey wolf.

Therefore, with the reintroduction of the grey wolf, we anticipate changes in both the predator and the prey animal populations in the Yellowstone Park ecosystem, and the success of the wolf will depend on how it influences and is influenced by the other species in the ecosystem.

Consider the following simplified model for the interaction among the elk, coyotes, and wolves within the ecosystem

$$\begin{aligned}\frac{dE}{dt} &= 0.04E - 0.003EC - 0.85EW \\ \frac{dC}{dt} &= -0.06C + 0.001EC \\ \frac{dW}{dt} &= -0.12W + 0.005EW\end{aligned}$$

where E is the elk population, C is the coyote population and W is the wolf population. All populations are measured in thousands. The variable t represents time measure in years from 1995 and the initial conditions are

$$E(0) = 60.0, \quad C(0) = 2.0, \quad W(0) = 0.015,$$

which means that in year 1995, there are 60,000 elk, 2,000 coyotes, and 15 wolves.

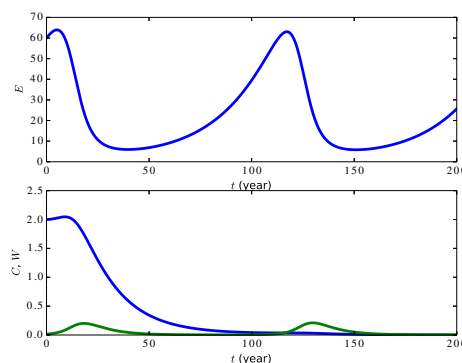
We can use the mathematical tools in this topic to analyse the changes in the populations of elk, coyotes and wolves.

```

1 from scipy.integrate import odeint
2 from numpy import linspace, zeros, sin, cos
3 from matplotlib.pyplot import plot, show, xlabel, ylabel, savefig, subplot
4
5 def predator_preymodel(population,t):
6     E, C, W = population
7     delta_population = zeros(3)
8     delta_population[0] = 0.04*E - 0.003*E*C - 0.85*E*W
9     delta_population[1] = -0.06*C + 0.001*E*C
10    delta_population[2] = -0.12*W + 0.005*E*W
11    return delta_population
12
13 tgrid = linspace(0, 200, 10001)
14 xgrid = odeint(predator_preymodel, [60.0,2.0,0.015], tgrid)
15 subplot(2,1,1)
16 xlabel("$t$ (year)"); ylabel("$E$");
17 plot(tgrid, xgrid[:,0], linewidth=3);
18 subplot(2,1,2)
19 xlabel("$t$ (year)"); ylabel("$C$, $W$");
20 plot(tgrid, xgrid[:,1:3], linewidth=3);
21 savefig("predator_preymodel.eps") #show();

```

According to this model, the changes in the populations of elk, coyotes and wolves for 200 years are shown below.



For documentaries regarding this incident, one can watch <https://www.youtube.com/watch?v=ysa50BhXz-Q>, <https://www.youtube.com/watch?v=YbGFYq0HcRI> and read <https://www.nps.gov/yell/learn/nature/wolves.htm>

Chapter 34

Introduction to PDE and Fourier Series

ch:pde:intro

Partial differential equation is a broad subject that there are more than hundreds of books written on it. The less abstract and classical examples are found in many mathematical physics books.

A large proportion of this subject come from the classical “mathematical physics” which is found in Byerly [1893], a book written more than 100 years ago.

Undergraduate texts are suitable for this subject are Asmar [2005], McOwen [2003], Brown and Churchill [2008] and Duffy [2004]. Some extra and more advanced texts for willing students are Strauss [1992], John [1982], Farlow [1993], Evans [1998], Boggess and Narcowich [2001], Weinberger [1995], etc.

34.1 Introduction to PDE

Many scientific laws are characterised by partial differential equations. The great problems for mathematicians, are to understand the solutions of these equations, using theoretical analysis, numerical simulations, perturbation theory, etc. [Gowers et al., 2008, 455]

34.1.1 Definitions

Definition 34.1. A *partial differential equation* (henceforth abbreviated as PDE) is an equation involving a function u of n variables ($n > 1$) and certain of its partial derivatives of the form

$$F\left(\frac{\partial^k u}{\partial x_1^k}, \dots, \frac{\partial^k u}{\partial x_n^k}, \dots, \frac{\partial^2 u}{\partial x_1^2}, \frac{\partial^2 u}{\partial x_1 \partial x_2}, \dots, \frac{\partial^2 u}{\partial x_n^2}, \frac{\partial u}{\partial x_1}, \dots, \frac{\partial u}{\partial x_n}, u, x_1, \dots, x_n\right) = 0 \quad (34.1)$$

where F is a given function of the independent variables $x_1, \dots, x_n, u$, and a finite number of its partial derivatives:

$$F: \mathbb{R}^{n^k} \times \mathbb{R}^{n^{k-1}} \times \dots \times \mathbb{R}^{n^2} \times \mathbb{R}^n \times \mathbb{R} \times U \rightarrow \mathbb{R}.$$

If the *multi-index notation* (http://en.wikipedia.org/wiki/Multi-index_notation) below is introduced:

$$D^\alpha u = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}}, \quad \alpha = (\alpha_1, \dots, \alpha_n), \quad |\alpha| = \alpha_1 + \dots + \alpha_n,$$

then the PDE (34.1) can be written compactly as [Evans, 1998, 14]

$$F(D^k u(x), \dots, D^2 u(x), Du(x), u(x), x) = 0, \quad x = (x_1, \dots, x_n) \in U \subset \mathbb{R}^n. \quad (34.2)$$

Definition 34.2. The function u is called a (*classical*) *solution* of a PDE (34.2) if after the substitution of u and its partial derivatives into the left-hand-side of (34.2), $F = 0$ in some region of $U \subset \mathbb{R}^n$.

Definition 34.3. The *order* of a PDE is the order of the highest derivative that occurs. Equation (34.2) is a k -th order PDE.

def:linear

Definition 34.4. The PDE (34.2) is called *linear* if it has the form

$$\sum_{|\alpha| \leq k} a_\alpha(x) D^\alpha u = f(x).$$

for given functions a_α ($|\alpha| \leq k$) and f . The linear PDE is said to be *homogeneous* (or *non-homogeneous*) if $f = 0$ (or $f \neq 0$).

Definition 34.5. The PDE (34.2) is called *semi-linear* if it is linear in the derivative of order k with coefficients that depends only on $x_1, \dots, x_n$:

$$\sum_{|\alpha|=k} a_\alpha(x) D^\alpha u + a_0(D^{k-1}u, \dots, Du, u, x) = 0.$$

Definition 34.6. The PDE (34.2) is called *quasi-linear* if it is linear in the derivative of order k with coefficients that depends on $x_1, \dots, x_n$ and the derivatives of order less than k :

$$\sum_{|\alpha|=k} a_\alpha(D^{k-1}u, \dots, Du, u, x) D^\alpha u + a_0(D^{k-1}u, \dots, Du, u, x) = 0.$$

Remark 34.1. The following relation explains the link between the three types of PDEs above:

$$\text{Quasi-linear} \supset \text{Semi-linear} \supset \text{Linear}.$$

Definition 34.7. The PDE (34.2) is called *fully (or complete) nonlinear* if it depends nonlinearly upon the highest order derivatives.

Definition 34.8. Several PDEs involving one or more unknown functions and their derivatives constitute a *system (of PDEs)*. A k^{th} -order system can be written as

$$\mathbf{F}(D^k \mathbf{u}(x), D^{k-1} \mathbf{u}(x), \dots, D^2 \mathbf{u}(x), D\mathbf{u}(x), \mathbf{u}(x), x) = \mathbf{0}, \quad x \in U, \quad (34.3)$$

where

$$\mathbf{F} : \mathbb{R}^{n^k} \times \mathbb{R}^{n^{k-1}} \times \dots \times \mathbb{R}^{n^2} \times \mathbb{R}^n \times \mathbb{R} \times U \rightarrow \mathbb{R}^m$$

and $\mathbf{u} : U \rightarrow \mathbb{R}^m$, $\mathbf{u} = (u_1, \dots, u_m)$ is the unknown.

Similar to PDE, systems are classified into linear, semi-linear, quasi-linear and fully nonlinear.

34.1.2 Notations and Examples

Since useful PDEs arise in physics and engineering problems, we will need to introduce a few notations which are used by mathematicians and scientists to denote PDEs. The notations of partial derivatives of u are

$$\frac{\partial u}{\partial x_i}, u_{x_i}, \frac{\partial u}{\partial x^i}, u_{x^i}, \partial_i u.$$

depending on whether the coordinate is written as $(x_1, \dots, x_n)$ or $(x^1, \dots, x^n)$. The *gradient* of u has notations

$$\text{grad } u = \nabla u := \left(\frac{\partial u}{\partial x_1}, \dots, \frac{\partial u}{\partial x_n} \right).$$

The *divergence* of a vector function $\mathbf{F}$ has notations

$$\text{div } \mathbf{F} = \nabla \cdot \mathbf{F} := \sum_{i=1}^{\infty} \frac{\partial F_i}{\partial x_i}.$$

The *Laplacian* of u has notations

$$\Delta u = \nabla^2 u = \nabla \cdot \nabla u := \sum_{n=1}^{\infty} \frac{\partial^2 u}{\partial x_i^2}.$$

With the above notations, we can write down a few important PDEs from science and mathematics below. First, let us write down a few simple linear PDEs (recall Definition 34.4). get linear

Definition 34.9. Let U be an open subset of $\mathbb{R}^n$ and u be a real-valued function with the domain U and $\mathbf{x} \in U$.

1. <i>Laplace equation</i> :	$\Delta u = 0.$
2. <i>Poisson equation</i> :	$\Delta u = f(\mathbf{x}), \quad f(\mathbf{x}) \neq 0.$
3. <i>Helmholtz (or eigenvalue) equation</i> :	$\Delta u = -\lambda u, \quad \lambda > 0.$
4. <i>Heat (or diffusion) equation</i> :	$u_t - k\Delta u = 0, \quad k > 0.$
5. <i>Wave equation</i> :	$u_{tt} - k\Delta u = 0, \quad k > 0.$
6. <i>Telegraph equation</i> :	$u_{tt} + du_t - u_{xx} = 0.$
7. <i>Beam equation</i> :	$u_t + u_{xxxx} = 0.$
8. <i>Schrödinger equation</i> [Abers, 2004]:	$i\hbar\psi_t(\mathbf{x}, t) = -\frac{\hbar^2}{2m}\Delta\psi(\mathbf{x}, t) + V(\mathbf{x})\psi(\mathbf{x}, t).$
9. Fokker-Planck equation:	$u_t - \sum_{i,j=1}^n (a^{ij}u)_{x_i x_j} - \sum_{i=1}^n (b^i u)_{x_i} = 0.$

In this course, we will only be solving some of the simple linear PDEs stated in the example above.

Linear PDEs are relatively simpler to study. However, in practice, we will encounter many nonlinear PDEs which are extremely difficult to solve.

Example 34.2. The following nonlinear PDEs are listed in Evans [1998]:

1. Eikonal equation:	$\ \nabla u\ _2 = f(\mathbf{x}).$
2. Nonlinear Poisson equation:	$\Delta u = -f(u), \quad f \text{ is nonlinear.}$
3. p -Laplacian equation:	$\nabla \cdot (\ \nabla u\ _2^{p-2} \nabla u) = 0.$
4. Minimal surface equation:	$\nabla \cdot \left(\frac{\nabla u}{\sqrt{1 + \nabla u ^2}} \right) = 0.$
5. Hamilton-Jacobi equation:	$u_t + H(Du, x) = 0.$
6. Scalar conservation law:	$u_t - \nabla \cdot \mathbf{F}(u) = 0.$
7. Scalar reaction-diffusion equation:	$u_t - \Delta u = f(u).$
8. Porous medium equation:	$u_t - \Delta(u^\gamma) = 0.$
9. Nonlinear wave equations:	$u_{tt} - \nabla \cdot \mathbf{a}(\nabla u) = f(u).$
10. Korteweg-de Vries (KdV) equation:	$u_t + uu_x + u_{xxx} = 0.$

More examples can be found in Zwillinger [1997] and Markowich [2007]. In practice, we will also encounter systems of PDEs which arises from physical modelling.

Example 34.3. Time-dependent Maxwell equations in free space (i.e. with no charges $\frac{\rho}{\epsilon_0}$ and no currents $\mu_0 \mathbf{j}$) is a linear system [Fitzpatrick, 2008]:

$$\begin{cases} \nabla \times \mathbf{B} = \epsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t}, \\ \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \\ \nabla \cdot \mathbf{B} = 0, \\ \nabla \cdot \mathbf{E} = 0. \end{cases}$$

Note that $\epsilon_0 \mu_0 = \frac{1}{c^2}$ in vacuum.

Example 34.4. The famous Navier-Stokes equations for incompressible, viscous flow is a nonlinear system:

$$\begin{cases} \mathbf{u}_t + \mathbf{u} \cdot \nabla \mathbf{u} - \Delta \mathbf{u} = -\nabla p, \\ \nabla \cdot \mathbf{u} = 0. \end{cases}$$

34.1.3 How to Solve PDEs?

PDEs rarely appear alone. Most problems require us to solve the PDEs subject to appropriate “data” such as boundary conditions and initial conditions.

Definition 34.10. With appropriate data, a given PDE problem is said to be *well-posed* if

- (a) the problem in fact **has a solution**;
- (b) the solution is **unique**;
- (c) the solution depends continuously on the data given in the problem.

The last condition is particularly important for problems arising from physical applications: we would prefer that our (unique) solution changes only a little when the conditions specifying the problem change a little.

It is a fundamental truth that we cannot solve most PDEs, if “solve” means coming up with a more-or-less explicit formula for the answer. However, we can often deduce properties of solutions without actually solving the PDE, either explicitly or numerically and this is the modern research in PDE [Gowers et al., 2008]. For a general or well-posed PDE, the following are some techniques to show that a solution exists and some could possibly provide an explicit form for the solution:

- (1) *Coordinates transformation* and *symmetrical analysis* of the PDE;
- (2) *Separation of variables* and *eigenfunction expansion* (or *Fourier method*) for linear PDE with bounded domain;
- (3) *Integral transforms* (e.g. Fourier transform, Laplace transform) for linear PDE with unbounded domain;
- (4) Transform to *Integral equations* and apply *Sobolev space methods* [Showalter, 1994];
- (5) *Calculus of variation methods*;
- (6) *Operator semigroup*;
- (7) *Generalised function theory* [Wikipedia, a];
- (8) *Differential operator theory* and *Microlocal analysis* [Wikipedia, b];
- (9) *Probabilistic method*;
- (10) *Perturbation analysis*;
- (11) *Numerical methods*.

Methods (2) and (3) are the focus in this course. They constitute parts of the classical theory of PDE which is based on the smoothness or real analytic property of u .

Many basic PDEs can be solved in the classical sense. But many more complicated PDEs cannot be solved using methods (2) and (3). In such cases, advanced methods related to functional analysis have to be used — the idea is to study the PDE in a large function space that has many “non-smooth” functions and then try to check the regularity of the solution of the PDE by embedding it into a function space which has less “non-smooth” functions.

34.2 Fourier Series

sec:fs

In Calculus, we learn that if a function f is sufficiently smooth at the neighbourhood at point $x = x_0$, it is possible to use a Taylor series to approximate the function. This approximation is good for a small neighbourhood of x_0 and has been applied in numerical analysis (UECM3033). However, it is not suitable for partial differential equations, where we need to “approximate” for the entire interval where f is defined.

The French mathematician Joseph Fourier proposed the idea of using the sum of sine and cosine series of different frequencies in Fourier [1878] to approximate **any** function f on an entire interval. This idea was inspired by earlier works on the analysis of wave equations. A brief history of Fourier series is given in Boggess and Narcowich [2001].

34.2.1 (Trigonometric) Fourier Series

Consider a function f which is defined on an interval $[-L, L]$.

Definition 34.11. A (trigonometric) Fourier series of f is

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} \left[a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right] \quad (34.4)$$

where the sum could be finite or infinite.

In order to compute the Fourier coefficients, a_n and b_n in (34.4), we need the following result on the orthogonality of the trigonometric functions.

Definition 34.12 (Orthogonality for Real Functions). Two functions ϕ and ψ over an interval $[-L, L]$ are called *orthogonal* to each other if

$$\int_{-L}^L \phi(x)\psi(x)dx = 0.$$

Definition 34.13. The *Kronecker's Delta* is defined as follows:

$$\delta_{mn} = \begin{cases} 1, & m = n \\ 0, & m \neq n. \end{cases} \quad (34.5)$$

The following theorem tells us that sine and cosine functions of different frequencies are orthogonal.

Theorem 34.5 (Orthogonality of Sine and Cosine Functions). The following integral relations hold:

$$\begin{aligned} \text{(a)} \quad & \frac{1}{L} \int_{-L}^L \cos\left(\frac{k\pi x}{L}\right) \cos\left(\frac{\ell\pi x}{L}\right) dx = \delta_{k\ell} + \delta_{k0}\delta_{\ell 0} = \begin{cases} 2 & \text{if } k = \ell = 0, \\ 1 & \text{if } k = \ell \geq 1, \\ 0 & \text{otherwise.} \end{cases} \\ \text{(b)} \quad & \frac{1}{L} \int_{-L}^L \sin\left(\frac{k\pi x}{L}\right) \sin\left(\frac{\ell\pi x}{L}\right) dx = \delta_{k\ell} = \begin{cases} 1 & \text{if } k = \ell \geq 1, \\ 0 & \text{otherwise.} \end{cases} \\ \text{(c)} \quad & \frac{1}{L} \int_{-L}^L \cos\left(\frac{k\pi x}{L}\right) \sin\left(\frac{\ell\pi x}{L}\right) dx = 0 \quad \text{for all integers } k, \ell. \end{aligned}$$

Proof: We will only prove part (a). Part (b) and (c) are left as exercise.

Using the trigonometric identity $\cos A \cos B = \frac{\cos(A+B) + \cos(A-B)}{2}$, we have

$$\int_{-L}^L \cos\left(\frac{k\pi x}{L}\right) \cos\left(\frac{\ell\pi x}{L}\right) dx = \frac{1}{2} \int_{-L}^L \left\{ \cos \frac{(k+\ell)\pi x}{L} + \cos \frac{(k-\ell)\pi x}{L} \right\} dx.$$

If $k \neq \ell$, we have

$$\int_{-L}^L \cos\left(\frac{k\pi x}{L}\right) \cos\left(\frac{\ell\pi x}{L}\right) dx = \frac{1}{2} \left[\frac{L \sin \frac{(k+\ell)\pi x}{L}}{(k+\ell)\pi} + \frac{L \sin \frac{(k-\ell)\pi x}{L}}{(k-\ell)\pi} \right]_{-L}^L = 0.$$

If $k = \ell \geq 1$, then

$$\int_{-L}^L \cos^2 \frac{k\pi x}{L} dx = \int_{-L}^L \frac{1 + \cos \frac{2k\pi x}{L}}{2} dx = L.$$

If $k = \ell = 0$, then

$$\int_{-L}^L 1 dx = 2L.$$

Using the orthogonality result above, we can derive a formula for computing the coefficients in the Fourier series (34.4).

Definition 34.14. The values a_0 , a_n and b_n in (34.4) are called the (trigonometric) Fourier coefficients of the function f and are given by the following formulas:

$$a_0 = \frac{1}{L} \int_{-L}^L f(x) dx, \quad (34.6)$$

$$a_n = \frac{1}{L} \int_{-L}^L f(x) \cos \frac{n\pi x}{L} dx, \quad n = 1, 2, \dots, \quad (34.7)$$

$$b_n = \frac{1}{L} \int_{-L}^L f(x) \sin \frac{n\pi x}{L} dx, \quad n = 1, 2, \dots. \quad (34.8)$$

How can we actually derive the above formula? Well, we have to “assume” that

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left[a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right], \quad -L \leq x \leq L$$

Let’s say that we want to find the k th value “ a_k ”, then we will multiply the above equation by $\cos \frac{k\pi x}{L}$ to get

$$f(x) \cos \frac{k\pi x}{L} = \frac{a_0}{2} \cos \frac{k\pi x}{L} + \sum_{n=1}^{\infty} \left[a_n \cos \frac{n\pi x}{L} \cos \frac{k\pi x}{L} + b_n \sin \frac{n\pi x}{L} \cos \frac{k\pi x}{L} \right],$$

and we can integrate the two sides over the interval $[-L, L]$:

$$\int_{-L}^L f(x) \cos \frac{k\pi x}{L} dx = \frac{a_0}{2} \int_{-L}^L \cos \frac{k\pi x}{L} dx + \sum_{n=1}^{\infty} \left[a_n \int_{-L}^L \cos \frac{n\pi x}{L} \cos \frac{k\pi x}{L} dx + b_n \int_{-L}^L \sin \frac{n\pi x}{L} \cos \frac{k\pi x}{L} dx \right]$$

According to Theorem 34.5, only when $n = k$, the integral is not zero, others are all zero! So, the right hand side of the above equation will become

$$\int_{-L}^L f(x) \cos \frac{k\pi x}{L} dx = \frac{a_0}{2} \times 0 + \sum_{n=1}^{k-1} a_n \times 0 + a_k \times L + \sum_{n=k+1}^{\infty} a_n \times 0 + \sum_{n=1}^{\infty} b_n \times 0 = a_k L$$

From here, we can obtain (34.7). Other formulas can be derived similarly.

Example 34.6. Find the (trigonometric) Fourier coefficients for the function

$$f(x) = e^{ax}, \quad a \neq 0, \quad -\pi < x < \pi.$$

Solution: Note that $L = \pi$ and the Fourier coefficients are

$$a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} e^{ax} dx = \frac{2 \sinh(a\pi)}{a\pi} \quad (\because \sinh x = \frac{e^x - e^{-x}}{2})$$

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} e^{ax} \cos(nx) dx = \frac{1}{\pi} \left[\frac{e^{ax}}{a^2 + n^2} (a \cos nx + n \sin nx) \right]_{-\pi}^{\pi} = \frac{2a(-1)^n \sinh(a\pi)}{(a^2 + n^2)\pi}$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} e^{ax} \sin(nx) dx = \frac{1}{\pi} \left[\frac{e^{ax}}{a^2 + n^2} (a \sin nx - n \cos nx) \right]_{-\pi}^{\pi} = \frac{-2n(-1)^n \sinh(a\pi)}{(a^2 + n^2)\pi}$$

Hence the (trigonometric) Fourier series is

$$f(x) \sim \frac{\sinh(a\pi)}{a\pi} + \frac{\sinh(a\pi)}{\pi} \sum_{n=1}^{\infty} \frac{2(-1)^n}{a^2 + n^2} (a \cos(nx) - n \sin(nx)).$$

Note that we are defining the Fourier series for a given function defined on the interval $[-L, L]$. Is it possible to define Fourier series for a function f on $\mathbb{R}$? It is impossible except when f is periodic since (34.4) consists of sine and cosine functions which are periodic.

Definition 34.15. A *periodic function* is a function $f : \mathbb{R} \rightarrow \mathbb{R}$ such that there is a constant $T > 0$ such that

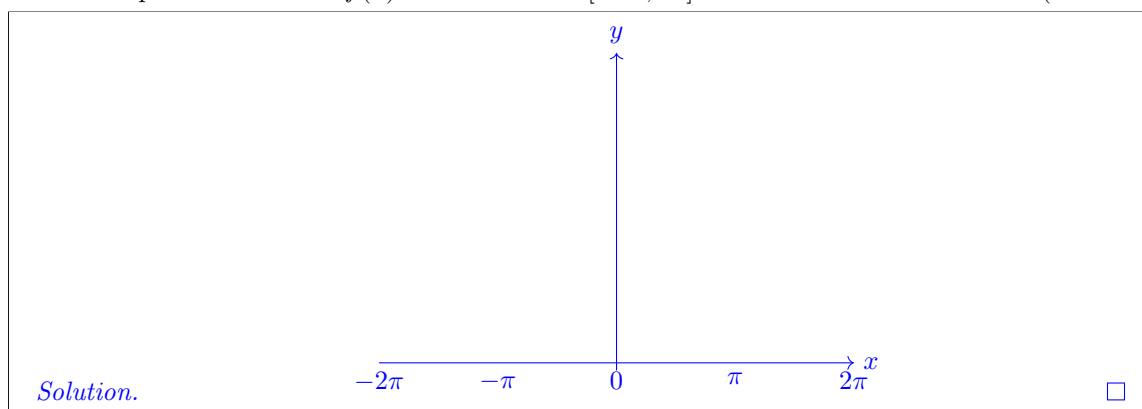
$$f(x + T) = f(x), \quad \forall x \in \mathbb{R}.$$

The minimum positive constant T is called the *period* of the periodic function f . The *frequency* of f is defined as $\frac{1}{T}$ and the *angular frequency* of f is defined as $\frac{2\pi}{T}$, usually denoted as ω .

Example 34.7 (UCCM3003 Final Exam, Sep 2007, Q1(a)). Consider the following periodic function with a period of 2π :

$$f(x) = \begin{cases} x^2, & 0 \leq x < \pi, \\ (x + 2\pi)^2, & -\pi < x < 0. \end{cases}$$

- (i) Sketch the periodic function $f(x)$ for the interval $[-2\pi, 2\pi]$. (2 marks)



- (ii) Find the trigonometric Fourier series of the periodic function $f(x)$. (14 marks)

Solution.

$$\begin{aligned} a_0 &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) dx = \frac{1}{\pi} \left[\int_{-\pi}^0 (x + 2\pi)^2 dx + \int_0^{\pi} x^2 dx \right] \\ &= \frac{1}{\pi} \left[\frac{(2\pi)^3}{3} - \frac{\pi^3}{3} + \frac{\pi^3}{3} - \frac{0^3}{3} \right] = \frac{8\pi^2}{3}. \end{aligned} \quad [2 \text{ marks}]$$

For $n = 1, 2, \dots$,

$$\begin{aligned} a_n &= \frac{1}{\pi} \left[\int_{-\pi}^0 (x + 2\pi)^2 \cos nx dx + \int_0^{\pi} x^2 \cos nx dx \right] \\ &= \frac{1}{\pi} \left[\frac{2(x + 2\pi) \cos nx}{n^2} \Big|_{-\pi}^0 + \frac{2x \cos nx}{n^2} \Big|_{-\pi}^0 \right] \\ &= \frac{4}{n^2} \end{aligned} \quad [1+2+2=5 \text{ marks}]$$

$$\begin{aligned} b_n &= \frac{1}{\pi} \left[\int_{-\pi}^0 (x + 2\pi)^2 \sin nx dx + \int_0^{\pi} x^2 \sin nx dx \right] \\ &= \frac{1}{\pi} \left[\left(\frac{-(x + 2\pi)^2 \cos nx}{n} + \frac{2 \cos nx}{n^3} \right) \Big|_{-\pi}^0 + \left(\frac{-x^2 \cos nx}{n} + \frac{2 \cos nx}{n^3} \right) \Big|_0^{\pi} \right] = \frac{-4\pi}{n} \end{aligned} \quad [1+2+2=5 \text{ marks}]$$

Therefore,

$$f(x) \sim \frac{4\pi^2}{3} + \sum_{n=1}^{\infty} \frac{4}{n^2} \cos nx + \sum_{n=1}^{\infty} \left(-\frac{4\pi}{n} \right) \sin nx \quad [2 \text{ marks}]$$

When a periodic function f is even, its Fourier series will only involve cosines because the integrand

in (34.8) will be odd and the integral will be zero. Therefore,

$$f(x) \sim \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos \frac{n\pi x}{L}$$

with a_0 and a_n given by (34.6) and (34.7) respectively.

When a periodic function f is odd, its Fourier series will only involve sines because the integrals (34.6) and (34.7) will be zero and therefore

$$f(x) \sim \sum_{n=1}^{\infty} b_n \sin \frac{n\pi x}{L}$$

with b_n given by (34.8).

Example 34.8 (UECM3003 Final Exam, Apr 2013, Q1(a)). The Fourier series representation of a function f defined on the interval $(-L, L)$ is given by

$$f^*(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right)$$

where

$$\begin{cases} a_0 = \frac{1}{2L} \int_{-L}^L f(x) dx, \\ a_n = \frac{1}{L} \int_{-L}^L f(x) \cos \frac{n\pi x}{L} dx, \quad n = 1, 2, 3, \dots \\ b_n = \frac{1}{L} \int_{-L}^L f(x) \sin \frac{n\pi x}{L} dx, \quad n = 1, 2, 3, \dots \end{cases}$$

- (i) If $f(x) = g(x) + h(x)$ where $g(x)$ is an even function of x and $h(x)$ is an odd function of x , show that

$$\begin{aligned} a_0 &= \frac{1}{L} \int_0^L g(x) dx, \\ b_n &= \frac{2}{L} \int_0^L h(x) \sin \frac{n\pi x}{L} dx, \quad n = 1, 2, 3, \dots \end{aligned}$$

[You may assume any results regarding odd and even functions given in lectures.] (6 marks)

Solution. We are going to use odd $\times$ odd=even and

$$\begin{aligned} \int_{-L}^L \text{even}(x) dx &= 2 \int_0^L \text{even}(x) dx, \quad \int_{-L}^L \text{odd}(x) dx = 0 \\ a_0 &= \frac{1}{2L} \int_{-L}^L (g(x) + h(x)) dx = \frac{1}{2L} \int_{-L}^L g(x) dx = \frac{1}{L} \int_0^L g(x) dx \\ b_n &= \frac{1}{L} \int_{-L}^L (g(x) + h(x)) \sin \frac{n\pi x}{L} dx = \frac{1}{L} \int_{-L}^L h(x) \sin \frac{n\pi x}{L} dx = \frac{2}{L} \int_0^L h(x) \sin \frac{n\pi x}{L} dx \end{aligned}$$

□

- (ii) If $f(x) = x + x^{100}$ for $-\pi < x < \pi$, find the coefficients b_1 in the Fourier series for f .

(6 marks)

Solution.

$$b_1 = \frac{1}{\pi} \int_{-\pi}^{\pi} (x + x^{100}) \sin x dx = \frac{2}{\pi} \int_0^{\pi} x \sin x dx = \frac{2}{\pi} \times \pi = 2.$$

□

Example 34.9. Show that

$$\frac{x}{2} \sim \sin x - \frac{1}{2} \sin 2x + \frac{1}{3} \sin 3x - \frac{1}{4} \sin 4x + \frac{1}{5} \sin 5x - \dots, \quad -\pi < x < \pi.$$

eg:fourier:x

Proof: Since $f(x) = \frac{x}{2}$ is an odd function on $(-\pi, \pi)$, $a_0 = a_n = 0$ and

$$\begin{aligned} b_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} \frac{x}{2} \sin nx dx = \frac{1}{2\pi} \left[\int_{-\pi}^{\pi} \underbrace{x \sin nx}_{\text{even function}} dx \right] \\ &= \frac{1}{\pi} \int_0^{\pi} x \sin nx dx = \frac{1}{\pi} \left[\frac{-x \cos nx}{n} + \frac{\sin nx}{n^2} \right]_0^{\pi} = \frac{1}{\pi} \left[\frac{-\pi(-1)^n}{n} \right] = \frac{(-1)^{n+1}}{n} \end{aligned}$$

where $n = 1, 2, \dots$.

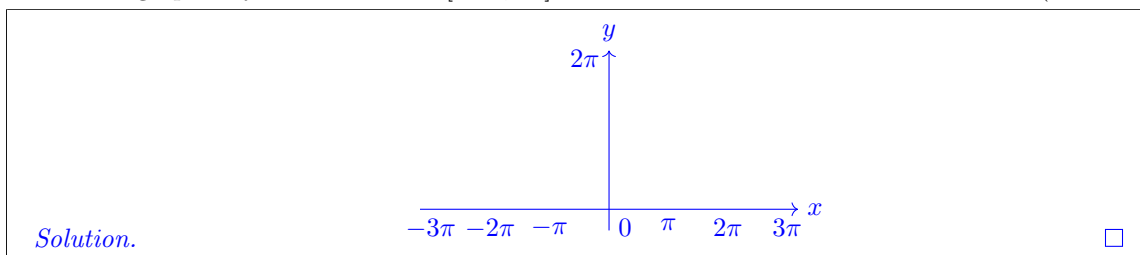
Remark: Fourier derived this formula in Fourier [1878, p149]. This is amazing but his derivation method is incomprehensible to me.

exam:201005:q1a

Example 34.10 (UCCM3003 Final Exam, May 2010, Q1(a)). A periodic function f with period 2π is defined within the period $-\pi < x < \pi$ by

$$f(x) = x + \pi, \quad -\pi < x < \pi.$$

- (i) Sketch the graph of f on the interval $[-3\pi, 3\pi]$. (2 marks)



- (ii) State whether f is even, odd or neither. Give a reason. (2 marks)

Solution. Since $f(-1) = -1 + \pi \neq 1 + \pi = f(1)$ and $f(-1) \neq -f(1)$, f is neither even nor odd. □

- (iii) Find the Fourier series of f .

Solution. For $n = 1, 2, \dots$,

$$\begin{aligned} a_0 &= \frac{1}{\pi} \int_{-\pi}^{\pi} (x + \pi) dx = \frac{1}{\pi} \left[\frac{(x + \pi)^2}{2} \right]_{-\pi}^{\pi} = \frac{4\pi^2 - 0}{2\pi} = 2\pi \\ a_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} (x + \pi) \cos nx dx = \frac{1}{\pi} \left[\int_{-\pi}^{\pi} \underbrace{x \cos nx}_{\text{odd function}} dx + \pi \int_{-\pi}^{\pi} \cos nx dx \right] = 0 \\ b_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} (x + \pi) \sin nx dx = \frac{1}{\pi} \left[\int_{-\pi}^{\pi} \underbrace{x \sin nx}_{\text{even function}} dx + \pi \underbrace{\int_{-\pi}^{\pi} \sin nx dx}_{=0} \right] \\ &= \frac{2}{\pi} \int_0^{\pi} x \sin nx dx = \frac{2}{\pi} \left[\frac{-x \cos nx}{n} + \frac{\sin nx}{n^2} \right]_0^{\pi} = \frac{2}{\pi} \left[\frac{-\pi(-1)^n}{n} \right] = \frac{2(-1)^{n+1}}{n} \end{aligned}$$

Therefore, $f(x) \sim \pi + \sum_{n=1}^{\infty} \frac{2(-1)^{n+1}}{n} \sin nx$. □

In general for a T -periodic function f characterised by g over the interval $[A, A + T]$ we have the following theorem.

thm:fs:gen

Theorem 34.11. If f is a T -periodic function given by a piecewise analytic function $g : [A, A + T] \rightarrow \mathbb{R}$ then f can be expressed as a Fourier series below:

$$f(x) \sim \frac{a_0}{2} + \sum_{n=1}^{\infty} \left[a_n \cos \frac{2n\pi x}{T} + b_n \sin \frac{2n\pi x}{T} \right]$$

where

$$a_0 = \frac{2}{T} \int_A^{A+T} g(x) dx, \quad a_n = \frac{2}{T} \int_A^{A+T} g(x) \cos \frac{2n\pi x}{T} dx, \quad b_n = \frac{2}{T} \int_A^{A+T} g(x) \sin \frac{2n\pi x}{T} dx.$$

eg:sqr-wave

Example 34.12 (Square Waves). Find the Fourier series representation of the periodic function f with period $2L$, where

$$f(x) = \begin{cases} 1, & 0 \leq x < L, \\ -1, & L \leq x < 2L. \end{cases}$$

Solution: By Theorem 34.11, $T = 2L$.

$$\begin{aligned} a_0 &= \frac{2}{2L} \int_0^{2L} f(x) dx = \frac{1}{L} \left[\int_0^L 1 dx + \int_L^{2L} (-1) dx \right] = \frac{L - L}{L} = 0 \\ a_n &= \frac{2}{2L} \int_0^{2L} f(x) \cos \frac{2n\pi x}{2L} dx = \frac{1}{L} \left[\int_0^L \cos \frac{n\pi x}{L} dx + \int_L^{2L} (-1) \cos \frac{n\pi x}{L} dx \right] \\ &= \frac{1}{L} \left[\left[\frac{\sin \frac{n\pi x}{L}}{\frac{n\pi}{L}} \right]_0^L + \left[\frac{\sin \frac{n\pi x}{L}}{\frac{n\pi}{L}} \right]_L^{2L} \right] = 0 \\ b_n &= \frac{1}{L} \left[\int_0^L \sin \frac{n\pi x}{L} dx + \int_L^{2L} (-1) \sin \frac{n\pi x}{L} dx \right] = \frac{1}{L} \left[\left[-\frac{\cos \frac{n\pi x}{L}}{\frac{n\pi}{L}} \right]_0^L + \left[\frac{\cos \frac{n\pi x}{L}}{\frac{n\pi}{L}} \right]_L^{2L} \right] \\ &= \frac{1}{n\pi} [-(-1)^n + 1 + (1 - (-1)^n)] = \frac{2(1 - (-1)^n)}{n\pi} \end{aligned}$$

Therefore, the Fourier series representation of f is

$$f(x) \sim \sum_{n=1}^{\infty} \frac{2(1 - (-1)^n)}{n\pi} \sin \frac{n\pi x}{L} = \sum_{k=1}^{\infty} \frac{4}{(2k-1)\pi} \sin \frac{(2k-1)\pi x}{L}.$$

eg:tri-wave

Example 34.13 (Triangle Waves). Find the Fourier series for the $2L$ -periodic function which is defined as follows for the interval $[0, 2L)$:

$$f(x) = \begin{cases} 2x/L - 1, & 0 \leq x < L; \\ 3 - 2x/L, & L \leq x < 2L. \end{cases}$$

Solution. By Theorem 34.11, $T = 2L$.

$$\begin{aligned} a_0 &= \frac{2}{2L} \int_0^{2L} f(x) dx = \frac{1}{L} \left[\int_0^L \left(\frac{2x}{L} - 1 \right) dx + \int_L^{2L} \left(3 - \frac{2x}{L} \right) dx \right] = \frac{0 - 0}{L} = 0 \\ a_n &= \frac{2}{2L} \int_0^{2L} f(x) \cos \frac{2n\pi x}{2L} dx = \frac{1}{L} \left[\int_0^L \left(\frac{2x}{L} - 1 \right) \cos \frac{n\pi x}{L} dx + \int_L^{2L} \left(3 - \frac{2x}{L} \right) \cos \frac{n\pi x}{L} dx \right] \\ &= \frac{1}{L} \left\{ \left[\left(\frac{2x}{L} - 1 \right) \frac{\sin \frac{n\pi x}{L}}{\frac{n\pi}{L}} + \frac{2}{L} \times \frac{\cos \frac{n\pi x}{L}}{(\frac{n\pi}{L})^2} \right]_0^L + \left[\left(3 - \frac{2x}{L} \right) \frac{\sin \frac{n\pi x}{L}}{\frac{n\pi}{L}} - \frac{2}{L} \times \frac{\cos \frac{n\pi x}{L}}{(\frac{n\pi}{L})^2} \right]_L^{2L} \right\} \\ &= \frac{1}{L} \left\{ \frac{2((-1)^n - 1)L}{(n\pi)^2} - \frac{2(1 - (-1)^n)L}{(n\pi)^2} \right\} = \frac{4((-1)^n - 1)}{(n\pi)^2} \\ b_n &= \frac{1}{L} \left[\int_0^L \left(\frac{2x}{L} - 1 \right) \sin \frac{n\pi x}{L} dx + \int_L^{2L} \left(3 - \frac{2x}{L} \right) \sin \frac{n\pi x}{L} dx \right] \\ &= \frac{1}{L} \left\{ \left[-\left(\frac{2x}{L} - 1 \right) \frac{\cos \frac{n\pi x}{L}}{\frac{n\pi}{L}} + \frac{2}{L} \times \frac{\sin \frac{n\pi x}{L}}{(\frac{n\pi}{L})^2} \right]_0^L + \left[-\left(3 - \frac{2x}{L} \right) \frac{\cos \frac{n\pi x}{L}}{\frac{n\pi}{L}} + \frac{2}{L} \times \frac{\sin \frac{n\pi x}{L}}{(\frac{n\pi}{L})^2} \right]_L^{2L} \right\} \\ &= \frac{1}{L} \left\{ \frac{L(-(-1)^n - 1)}{n\pi} + \frac{L(1 + (-1)^n)}{n\pi} \right\} = 0 \end{aligned}$$

Therefore, the Fourier series representation of f is

$$f(x) \sim \sum_{n=1}^{\infty} \frac{4((-1)^n - 1)}{(n\pi)^2} \cos \frac{n\pi x}{L}.$$

□

From the above examples, we find that computing an integral over the interval $[-L, L]$ is generally simpler than computing an integral over the interval $[A, A + T]$.

eg:saw-wave

Example 34.14 (Sawtooth Waves). Find the Fourier series for $f(x) = \frac{x}{L}$ if $-L < x < L$ and $f(x+2L) = f(x)$.

Solution. Note that f is odd, so $a_0 = a_n = 0$ and the calculation of b_n is similar to Example 34.9. eg:fourier:x

$$b_n = \frac{2}{L} \int_0^L \frac{x}{L} \sin \frac{n\pi x}{L} dx = \frac{2}{L^2} \left[-\frac{(-1)^n L^2}{n\pi} \right] = \frac{2(-1)^{n+1}}{n\pi}.$$

Therefore, the Fourier series representation of f is

$$f(x) \sim \sum_{n=1}^{\infty} \frac{2(-1)^{n+1}}{n\pi} \sin \frac{n\pi x}{L}.$$

□

The three waveform in Examples 34.12, 34.13 and 34.14 are basic waveform generator in digital music. First, let us state some physics before we illustrate how “Fourier series” can be applied in digital music generation. eg:sqrwave,tri-wave,saw-wave

According to Ryan [2014], air pressure is measured by the SI-unit Pa (Pascal) which is equivalent to N/m^2 (force/area). The normal air pressure at sea level is 101,325 Pa. Everyday sounds typically correspond to variations in air pressure of about $0.00002 \sim 2$ Pa, while a jet engine may cause variations as large as 200 Pa. Short exposure to variations of about 20 Pa may lead to hearing damage.

For a human with good hearing to perceive variations in air pressure as sound, the number of variations per second must be in the range $20 \sim 20000$. This means humans can only perceive variations in air pressure as “sound” if the Fourier series of the sound signal contains at least one sufficiently large term with frequency in the range $20 \sim 20000$ Hz.

Definition 34.16. The function $\sin \frac{2\pi t}{T}$ represents a *pure tone* with frequency $f = \frac{1}{T}$. In physics, f has a unit Hz (Hetz) which is the same as s^{-1} .

Example 34.15. Use a computer software (e.g. SoX or ChuckK) to listen to the sound generated from sine wave, triangle wave, sawtooth wave and square wave of 440Hz.

Solution: Using ChuckK (<http://chuck.cs.princeton.edu/>), we can write a program to listen to sound of all waves mentioned.

```
1 SinOsc s1 => dac;
2 1 => s1.gain; 440 => s1.freq;
3 2.0::second => now;
4 TriOsc s2 => dac;
5 1 => s2.gain; 440 => s2.freq;
6 2.0::second => now;
7 SawOsc s3 => dac;
8 1 => s3.gain; 440 => s3.freq;
9 2.0::second => now;
10 SqrOsc s4 => dac;
11 1 => s4.gain; 440 => s4.freq;
12 2.0::second => now;
```

Amazingly, it is even simpler to play them using SoX under bash:

```
1 for i in sine triangle sawtooth square; do
2     play -n synth 2 $i 440;
3 done
```

34.2.2 Fourier Series On the Half Interval

In Topic 2, we will encounter functions which are defined on the interval $[0, L]$ and they are required to be either series of sines or series of cosines. How can we determine the coefficients in this case.

To express f as a series of cosine in the interval $[0, L]$, we consider the *even extension* of f :

$$f_e(x) = \begin{cases} f(x) & \text{if } 0 \leq x \leq L, \\ f(-x) & \text{if } -L \leq x < 0. \end{cases}$$

The function f_e is an even function defined on $[-L, L]$. Therefore, only cosine terms appear in its Fourier expansion $f_e(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos \frac{n\pi x}{L}$, $x \in [-L, L]$. The Fourier series of f_e involves only cosine series and we introduce the following definition.

Definition 34.17. The *Fourier cosine series* for f over the interval $[0, L]$ is

$$f(x) \sim \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos \frac{n\pi x}{L}, \quad x \in [0, L]$$

where

$$a_0 = \frac{2}{L} \int_0^L f(x) dx, \quad a_n = \frac{2}{L} \int_0^L f(x) \cos \frac{n\pi x}{L} dx.$$

Similarly, to express f as a sine series, we consider the odd extension of f :

$$f_o(x) = \begin{cases} f(x) & \text{if } 0 \leq x \leq L, \\ -f(-x) & \text{if } -L \leq x < 0. \end{cases}$$

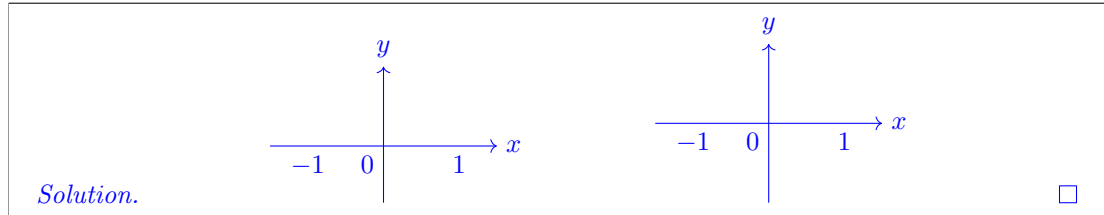
The Fourier series of f_o over the interval $[-L, L]$ will be regarded as the sine expansion for f .

Definition 34.18. The *Fourier-sine series* of a function f over the interval $[0, L]$ is

$$f(x) \sim \sum_{n=1}^{\infty} b_n \sin \frac{n\pi x}{L}, \quad x \in (0, L], \quad \text{where } b_n = \frac{2}{L} \int_0^L f(x) \sin \frac{n\pi x}{L} dx.$$

Example 34.16 (UCCM3003 Final Exam, Sep 2009, Q2). Consider the function $f(x) = x - x^2$ defined over the interval $0 \leq x < 1$.

- (a) (i) Sketch the full-range periodic extension of f over the interval of $-1 < x < 1$. (3 marks)
(ii) Sketch the half-range periodic extension of f with odd symmetry over the interval of $-1 < x < 1$. (4 marks)



- (b) Show that Fourier series for the full range periodic extension of f is

$$\frac{1}{6} - \sum_{n=1}^{\infty} \frac{\cos 2n\pi x}{n^2 \pi^2}. \quad (8 \text{ marks})$$

Solution. Let $n = 1, 2, \dots$. Then

$$\begin{aligned} a_0 &= 2 \int_0^1 (x - x^2) dx = 2 \left[\frac{x^2}{2} - \frac{x^3}{3} \right]_0^1 = \left[\frac{1}{2} - \frac{1}{3} \right] = \frac{1}{3} \\ a_n &= 2 \int_0^1 (x - x^2) \cos 2n\pi x dx = 2 \left[\frac{(x - x^2) \sin 2n\pi x}{2n\pi} - \frac{2x \cos 2n\pi x}{(2n\pi)^2} + \frac{2 \sin 2n\pi x}{(2n\pi)^3} \right]_0^1 \\ &= 2 \left[-\frac{2 \cos 2n\pi}{(2n\pi)^2} - 0 \right] = \frac{-1}{n^2 \pi^2} \end{aligned}$$

Since f is even, $b_n = 0$ and therefore the full-range Fourier series for f is

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos 2n\pi x = \frac{1}{6} + \sum_{n=1}^{\infty} \frac{-1}{n^2 \pi^2} \cos 2n\pi x.$$

- (c) Find the Fourier sine series of f . (7 marks)

Solution.

$$\begin{aligned} b_n &= 2 \int_0^1 (x - x^2) \sin n\pi x dx = 2 \left[-\frac{(x - x^2) \cos n\pi x}{n\pi} - \frac{2x \sin n\pi x}{(n\pi)^2} - \frac{2 \cos n\pi x}{(n\pi)^3} \right]_0^1 \\ &= \frac{4(1 - (-1)^n)}{(n\pi)^3} \end{aligned}$$

The Fourier sine series of f is $f(x) = \sum_{n=1}^{\infty} \frac{4(1 - (-1)^n)}{(n\pi)^3} \sin n\pi x$. □

- (d) Which of the series in Part (b) and Part (c) will converge faster to f in $0 < x < 1$? Explain your answer. (3 marks)

Solution. The series in part (c) will converge faster since the ratio $|b_n/a_n| \leq \frac{8}{n\pi} < 1$ for large n . □

34.2.3 Pointwise Convergence of Fourier Series

So far, I have been using $\sim$ to relate a function f to its Fourier series (34.4). Aren't they equal? Well, it is easy to calculate $f(x_0)$ given x_0 but the Fourier series consists of infinite terms, how do we know that summing them actually gives us something meaningful?

If we put in x_0 into the Fourier series and the sum is meaningful, then we say that the Fourier series converges at the point x_0 . This section explores the “pointwise” convergence of Fourier series by stating some useful theorems. The proofs can be found in textbooks such as Al-Gwaiz [2008], Stein and Shakarchi [2003], Boggess and Narcowich [2001], etc.

To determine the pointwise convergence of Fourier series, we first determine the values of a_n and b_n as $n \rightarrow \infty$.

Lemma 34.1 (The Riemann-Lebesgue Lemma). *Suppose f is a piecewise continuous function on the interval $a \leq x \leq b$. Then*

$$\lim_{n \rightarrow \infty} \int_a^b f(x) \cos \frac{n\pi x}{L} dx = \lim_{n \rightarrow \infty} \int_a^b f(x) \sin \frac{n\pi x}{L} dx = 0.$$

An important consequence of this lemma is that only a finite number of Fourier coefficients are large (in absolute value) than any given positive number. This fact is the basic idea underlying the process of data compression. One way to compress a signal is first to express it as a Fourier series; then discard all the small Fourier coefficients and retain (or transmit) only the finite number of Fourier coefficients that are larger than some given threshold.

The sum appearing in a Fourier expansion generally contains an infinite number of terms. An infinite sum is, by definition, a limit of *partial sums*; i.e.

$$\sum_{n=1}^{\infty} \left\{ a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right\} := \lim_{N \rightarrow \infty} \sum_{n=1}^N \left\{ a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right\}$$

provided that the limit exists. Therefore,

Definition 34.19 (Pointwise Convergence). The Fourier series of f is said to be *converging to $f(x)$ at the point x* if

$$\frac{a_0}{2} + \lim_{N \rightarrow \infty} \sum_{n=1}^N \left\{ a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right\} = f(x).$$

The following theorem on describes the pointwise convergence of Fourier series at a point x .

Theorem 34.17 (Pointwise Convergence Theorem). *Suppose f is a continuous and periodic function on $\mathbb{R}$. Then for each point x , where the derivative of f is defined, the Fourier series of f at x converges to the value $f(x)$.*

Example 34.18. (Fourier [1878, p145]) Show that $\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \dots$. [Hint: Use Example 34.9]

Solution. Take $x = \frac{\pi}{2}$, according to Example 34.9, [eg:fourier:x](#)

$$x \mapsto \frac{x}{2} \in C^1(-\pi, \pi) \bigwedge f'(x) = \frac{1}{2} \xrightarrow[\text{Theorem 34.17}]{\text{Theorem 34.17}} 1 - \frac{1}{2} \cdot 0 + \frac{1}{3}(-1) - \frac{1}{4} \cdot 0 + \frac{1}{5}(1) - \cdots = \frac{\pi}{2}.$$

□

In theorem [thm:fourier:pt-conv](#) 34.17, the function f is required to be continuous and differentiable. The following theorem characterises the “discontinuity” for the case of piece-wise continuous functions.

Theorem 34.19 (Pointwise Convergence Theorem). Suppose a function f is periodic and piecewise continuous. Suppose x is a point where f is left and right differentiable (but not necessarily continuous). Then the Fourier series of f at x converges to $\frac{f(x+0)+f(x-0)}{2}$, i.e.

$$\frac{a_0}{2} + \lim_{N \rightarrow \infty} \sum_{n=1}^N \left\{ a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right\} = \frac{f(x+0) + f(x-0)}{2},$$

where $f(x-0) = \lim_{h \rightarrow 0^+} f(x-h)$, $f(x+0) = \lim_{h \rightarrow 0^+} f(x+h)$.

Remark 34.20. A more general version of this theorem is given in Körner [1988, Chapter 16] which involves “Riemann integrable”, a concept that should be discussed in UECM2033 Elementary Real Analysis.

Example 34.21. Study the pointwise convergence of square wave in Example [eg:sqr-wave](#) 34.12.

Solution: Since f is periodic and continuous on for the range $D = \mathbb{R} \setminus \{0, \pm L, \pm 2L, \dots\}$ and for every $x \in D$, $f'(x) = 0$. The Fourier series of f converges to the value $f(x)$. For $x = 2nL$, where $n \in \mathbb{Z}$, we can derive the convergence without applying Theorem [thm:fourier:pt-conv2](#) 34.19:

$$\sum_{k=1}^{\infty} \frac{4}{(2k-1)\pi} \sin \frac{(2k-1)\pi(2nL)}{L} = 0 = \frac{1+(-1)}{2} = \frac{f(x+0) + f(x-0)}{2}.$$

Similarly for $x = (2n-1)L$, $n \in \mathbb{Z}$, the series converges to 0.

Example 34.22. Study the pointwise convergence of triangular wave in Example [eg:tri-wave](#) 34.13.

Solution: Since f is periodic and continuous on $x \in \mathbb{R}$ and $f'(x)$ exists except for $x \in B = \{0, \pm L, \pm 2L, \dots\}$. However, the left and right derivatives exist at $x \in B$. The Fourier series of f has pointwise convergence at every $x \in \mathbb{R}$ according to Theorem [thm:fourier:pt-conv2](#) 34.19.

Example 34.23. Study the pointwise convergence of sawtooth wave in Example [eg:saw-wave](#) 34.14.

Solution: Let $B = \{\pm L, \pm 3L, \dots\}$. Since f is continuous and has derivatives at every $x \in D = \mathbb{R} \setminus B$ and f has left and right derivatives $\frac{1}{L}$ at $x \in B$. The Fourier series of f has pointwise convergence according to Theorem [thm:fourier:pt-conv2](#) 34.19.

34.2.4 Uniform Convergence and Gibbs Phenomena

The previous section discuss the convergence at a point of the function. This section discusses the (uniform) convergence of the partial sums of Fourier series of function f and what happens if non-uniform convergence arises.

Definition 34.20 (Uniform Convergence). A sequence of functions $\{f_n\}$ converges to f uniformly for an interval $[a, b]$ if the rate of convergence is independent of the point x . In other words, given any small tolerance $\epsilon > 0$, there is a number N which is independent of x such that $|f_n(x) - f(x)| < \epsilon$ for all x and all $n \geq N$.

Definition 34.21. The Fourier series of f is said to *converge uniformly* to f if the sequence of partial sums

$$S_N(x) = \frac{a_0}{2} + \sum_{n=1}^N \left\{ a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right\}$$

converges to f uniformly for $[-L, L]$, where a_0 , a_n and b_n are the Fourier coefficients of f .

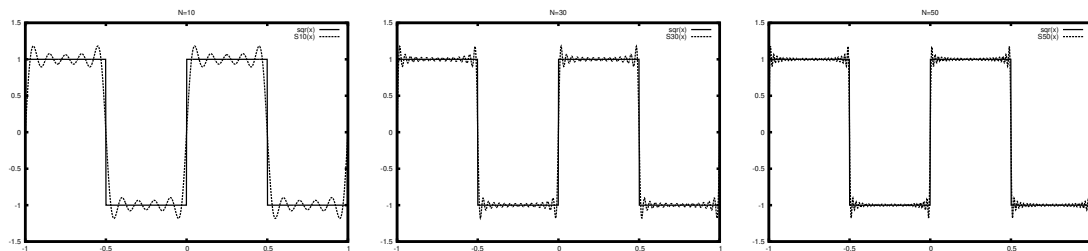
Theorem 34.24. The Fourier series of a $2L$ -periodic piecewise smooth continuous function f converges uniformly to f on $[-L, L]$.

What happens to the Fourier series when f has discontinuous points? A physicist observed something funny with the partial sums of Fourier when f has discontinuity and the strange behaviour was clarified by Gibbs, Körner [1988, Chapter 17] for details.

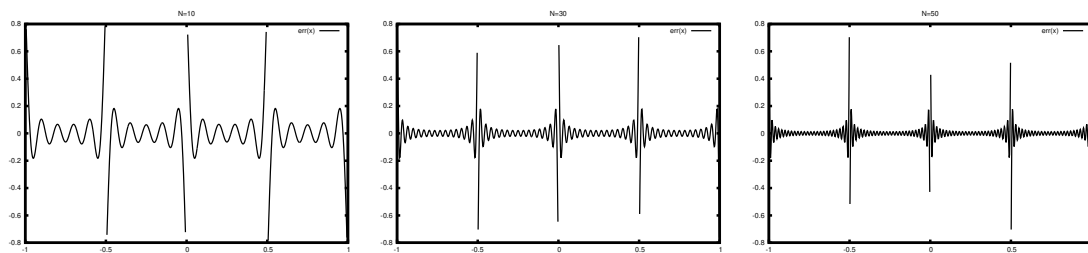
To investigate the convergence property of the partial sums of a Fourier series, we study the square wave function in Example 34.12, which has a partial sum

$$S_N(x) = \sum_{n=1}^N \frac{2(1 - (-1)^n)}{n\pi} \sin \frac{n\pi x}{L}.$$

Consider when $L = \frac{1}{2}$. We want to look at what happens at a discontinuous point $x = 0$. We learned from the previous section that $S_N(0) \rightarrow 0$ as $N \rightarrow \infty$. However, if we plot out the functions $S_{10}(x)$, $S_{30}(x)$ and $S_{50}(x)$, we find that S_N has “spikes” near the discontinuous point $x = 0$. Furthermore, the height of the spikes stays the same for $N = 10, 30, 50$. The width of the peaks shrink as N increase and the spikes move toward the origin. This is called the *Gibbs phenomenon*.



To see how the peaks move, we draw the difference $f(x) - S_N(x)$.



As we can see, the peak near the discontinuous points persist for different N .

In general, Gibbs phenomenon implies that no matter how large N is, there is always a small region around the discontinuity point where the magnitude of the error is independent of N . This behaviour is called “nonuniform convergence”.

34.2.5 Parseval’s Theorem and Mean Square Error

sec:1.5

Uniform convergence is nice because a sequence of continuous functions will still be continuous by taking limit but this is not true for pointwise convergence. However, there are many Fourier series which does not converge uniformly. Is it possible to have a kind of convergence which is nicer than pointwise convergence but is not as restrictive as uniform convergence? Yes, there is and it is based on the idea of “square error” and we will investigate it in this section.

Definition 34.22. A function sequence $\{f_N\}_{N=1}^{\infty}$ over the interval $[a, b]$ is said to *converge in L^2 -norm to a function f* or *converge to a function f in $L^2([a, b])$* if

$$\|f_N - f\|_2 \rightarrow 0 \quad \text{as } N \rightarrow \infty$$

where $\|g\|_2 := \int_a^b |g(x)|^2 dx$ (is a norm and $L^2([a, b])$ is a normed vector space).

Theorem 34.25. Suppose f is piecewise continuous function on $[-L, L]$. The sequence of the partial sums S_N of the Fourier series (34.4) of f converges in L^2 -norm to the function f , i.e.

$$\|S_N - f\|_2 \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

Definition 34.23. The value

$$\frac{1}{b-a} \int_a^b (f(x) - S_N(x))^2 dx$$

is called the *mean-squared error* of the reference function f and its approximation S_N .

By using Theorem [34.5](#), it is possible to obtain the *Parseval's theorem* and *Bessel's inequality* stated below.

Theorem 34.26 (Parseval's Theorem). If f is a piecewise continuous function on $[-L, L]$ and a_0 , a_n and b_n are the Fourier coefficients of f , then

$$\frac{1}{L} \int_{-L}^L [f(x)]^2 dx = \frac{a_0^2}{2} + \sum_{n=1}^{\infty} \{a_n^2 + b_n^2\}.$$

Remark 34.27. The “ L^2 -norm” of a signal is often interpreted as its “energy”. With this physical interpretation, the squares of the Fourier coefficients of a signal measure the energy of the corresponding frequency components. Therefore a physical interpretation of Parseval's theorem is that the *energy of a signal f* is simply the *sum of the energies from each of its frequency components*.

Theorem 34.28 (Bessel's Inequality). If f is a piecewise continuous function on $[-L, L]$ and a_0 , a_n and b_n are the Fourier coefficients of f , then

$$\frac{a_0^2}{2} + \sum_{n=1}^N \{a_n^2 + b_n^2\} \leq \frac{1}{L} \int_{-L}^L (f(x))^2 dx,$$

where N is any integer ≥ 1 . In particular, the series $\sum_{n=1}^{\infty} \{a_n^2 + b_n^2\}$ is convergent.

Example 34.29. Use the Parseval's Theorem to evaluate $\sum_{n=1}^{\infty} \frac{1}{a^2 + n^2}$ [Hint: Use Example [34.6](#)]

Solution.

$$\therefore \frac{1}{\pi} \int_{-\pi}^{\pi} (e^{ax})^2 dx = \frac{\sinh(2a\pi)}{a\pi} = \frac{2 \sinh^2(a\pi)}{a^2 \pi^2} + \frac{\sinh^2(a\pi)}{\pi^2} \sum_{n=1}^{\infty} \left\{ \frac{4a^2}{(a^2 + n^2)^2} + \frac{4(-n)^2}{(a^2 + n^2)^2} \right\}.$$

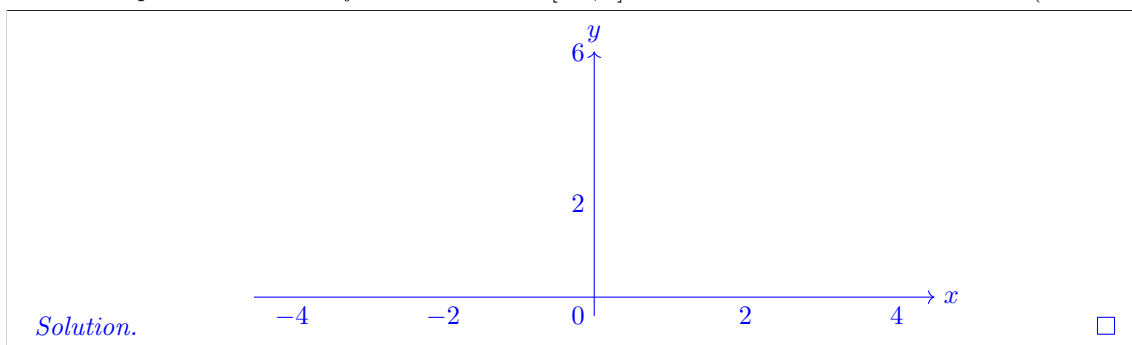
$$\therefore \sum_{n=1}^{\infty} \frac{1}{a^2 + n^2} = \frac{\pi^2}{4 \sinh^2(a\pi)} \times \frac{a\pi \sinh(2a\pi) - 2 \sinh^2(a\pi)}{a^2 \pi^2} = \frac{a\pi \coth(a\pi) - 1}{2a^2}$$

□

Example 34.30 (UCCM3003 Final Exam, Sep 2008, Q1). Consider the following periodic function with a period of 4:

$$f(x) = 2|x| - x, \quad -2 < x \leq 2.$$

1. Sketch the periodic function f for the interval $[-4, 4]$. (4 marks)



2. Find the trigonometric Fourier series of the periodic function f . (14 marks)

Solution.

$$a_0 = \frac{1}{2} \left[\int_{-2}^0 (2(-x) - x) dx + \int_0^2 (2x - x) dx \right] = \frac{1}{2} \left[\frac{3(-2)^2}{2} + \frac{2^2}{2} \right] = 4$$

$$a_n = \frac{1}{2} \left[\int_{-2}^0 (2(-x) - x) \cos \frac{n\pi x}{2} dx + \int_0^2 (2x - x) \cos \frac{n\pi x}{2} dx \right]$$

$$= \frac{1}{2} \left\{ \frac{12((-1)^n - 1)}{n^2\pi^2} + \frac{4((-1)^n - 1)}{n^2\pi^2} \right\} = \frac{8((-1)^n - 1)}{n^2\pi^2}$$

$$b_n = \frac{1}{2} \left[\int_{-2}^0 (2(-x) - x) \sin \frac{n\pi x}{2} dx + \int_0^2 (2x - x) \sin \frac{n\pi x}{2} dx \right]$$

$$= \frac{1}{2} \left\{ \frac{12(-1)^n}{n\pi} + \frac{-4(-1)^n}{n\pi} \right\} = \frac{4(-1)^n}{n\pi}$$

The Fourier series of f is $2 + \sum_{n=1}^{\infty} \left\{ \frac{8((-1)^n - 1)}{n^2\pi^2} \cos \frac{n\pi x}{2} + \frac{4(-1)^n}{n\pi} \sin \frac{n\pi x}{2} \right\}$. □

3. By applying the Parseval's Theorem to (b), show that

$$\sum_{n=1}^{\infty} \frac{16}{(2n-1)^4\pi^4} + \sum_{n=1}^{\infty} \frac{1}{n^2\pi^2} = \frac{1}{3}. \quad (5 \text{ marks})$$

Solution.

$$\frac{4^2}{2} + \sum_{n=1}^{\infty} \frac{16^2}{(2n-1)^4\pi^4} + \sum_{n=1}^{\infty} \frac{16}{n^2\pi^2} = \frac{1}{2} \left[\int_{-2}^0 (2(-x) - x)^2 dx + \int_0^2 (2x - x)^2 dx \right]$$

$$= \frac{1}{2} \left[24 + \frac{8}{3} \right] = \frac{40}{3}$$

This implies

$$16 \left[\sum_{n=1}^{\infty} \frac{16}{(2n-1)^4\pi^4} + \sum_{n=1}^{\infty} \frac{1}{n^2\pi^2} \right] = \frac{16}{3} \Rightarrow \text{above equation}$$
□

Example 34.31 (UCCM3003 Final Exam, Apr 2012, Q1). Suppose that f is a function of period 2π with $f(x) = \cos(\frac{x}{2})$, $-\pi < x < \pi$.

- (a) State whether f is even, odd or neither. Give a reason. (2 marks)

Solution. f is even since $\cos(-\frac{x}{2}) = \cos(\frac{x}{2})$ and it is periodic. □

- (b) Show that the Fourier series of f is

$$f^*(x) = \frac{2}{\pi} - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^n}{4n^2 - 1} \cos nx. \quad (14 \text{ marks})$$

[Hint: $\cos A \cos B = \frac{1}{2}[\cos(A+B) + \cos(A-B)]$, $\sin(A \pm B) = \sin A \cos B \pm \cos A \sin B$.]

Solution. Since f is even, $b_n = 0$.

$$a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} \cos\left(\frac{x}{2}\right) dx = \frac{2}{\pi} \left[2 \sin \frac{x}{2} \right]_0^{\pi} = \frac{4}{\pi}$$

When $n = 1, 2, \dots$,

$$\begin{aligned} a_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} \cos\left(\frac{x}{2}\right) \cos nx dx = \frac{1}{\pi} \int_0^{\pi} \left[\cos\left(\frac{x}{2} + nx\right) + \cos\left(\frac{x}{2} - nx\right) \right] dx \\ &= \frac{1}{\pi} \left[\frac{\sin(n + \frac{1}{2})x}{n + \frac{1}{2}} + \frac{\sin(n - \frac{1}{2})x}{n - \frac{1}{2}} \right]_0^{\pi} = \frac{1}{\pi} \left[\frac{\sin(n + \frac{1}{2})\pi}{n + \frac{1}{2}} + \frac{\sin(n - \frac{1}{2})\pi}{n - \frac{1}{2}} \right] \\ &= \frac{1}{\pi} \left[\frac{\cos(n\pi) \sin \frac{\pi}{2}}{n + \frac{1}{2}} + \frac{-\cos(n\pi) \sin \frac{\pi}{2}}{n - \frac{1}{2}} \right] = \frac{(-1)^n}{\pi} \left[\frac{n - \frac{1}{2} - n - \frac{1}{2}}{n^2 - \frac{1}{4}} \right] = -\frac{4(-1)^n}{\pi(4n^2 - 1)} \end{aligned}$$

Therefore, we obtain the Fourier series above. $\square$

(c) Use the series of part (b) and Parseval's Theorem to deduce the identity $\sum_{n=1}^{\infty} \frac{1}{(4n^2 - 1)^2} = \frac{\pi^2 - 8}{16}$.

[Hint: $\cos^2 A = \frac{1 + \cos 2A}{2}$] (9 marks)

Solution. From $\frac{a_0^2}{2} + \sum_{n=1}^{\infty} (a_n^2 + b_n^2) = \frac{16/\pi^2}{2} + \sum_{n=1}^{\infty} \frac{16}{\pi^2(4n^2 - 1)^2} = \frac{1}{\pi} \int_{-\pi}^{\pi} \cos^2\left(\frac{x}{2}\right) dx$,

we have

$$\frac{16}{\pi^2} \left[\frac{1}{2} + \sum_{n=1}^{\infty} \frac{1}{(4n^2 - 1)^2} \right] = \frac{2}{\pi} \int_0^{\pi} \frac{1 + \cos(x)}{2} dx = \frac{2}{\pi} \times \frac{\pi}{2} \Rightarrow \sum_{n=1}^{\infty} \frac{1}{(4n^2 - 1)^2} = \frac{\pi^2}{16} - \frac{1}{2}$$

$\square$

Example 34.32 (UECM3003 Final Exam, Apr 2013, Q1(c)). Use Parseval's identity to compute $\int_{-\pi}^{\pi} [g(x)]^2 dx$

if $g(x) = 1 + \sum_{n=1}^{\infty} \frac{\cos nx}{2^n}$, $-\pi < x < \pi$. Simplify your answer. (6 marks)

Solution.

$$\frac{1}{\pi} \int_{-\pi}^{\pi} [g(x)]^2 dx = \frac{2^2}{2} + \sum_{n=1}^{\infty} \left(\frac{1}{2^n}\right)^2 = 2 + \sum_{n=1}^{\infty} 2^{-2n} = 2 + \frac{1}{3} \Rightarrow \int_{-\pi}^{\pi} [g(x)]^2 dx = \frac{7}{3}\pi.$$

$\square$

34.2.6 Complex Form of Fourier Series

It is probably more convenient to express Fourier series in complex form using the *complex exponentials*

$$e^{inx} = \exp(inx) = \cos(nx) + i \sin(nx), \quad n \in \mathbb{Z}, \quad i = \sqrt{-1},$$

due to the simple computational properties of these functions:

1. $e^{i(t+2\pi)} = e^{it}$ (periodic)
2. $|e^{it}| = 1$ (modulus of 1)
3. $\overline{e^{it}} = e^{-it}$ (conjugate)
4. $e^{it}e^{is} = e^{i(t+s)}$ (product)
5. $e^{it}/e^{is} = e^{i(t-s)}$ (quotient)
6. $\frac{d}{dt}\{e^{it}\} = ie^{it}$ (differentiation)

for all $t, s \in \mathbb{R}$.

Similar to sine and cosine functions, it is possible to define the “orthogonality” for complex exponential functions.

Definition 34.24 (Orthogonality for Complex Functions). Two periodic complex functions ϕ and ψ over $\mathbb{C}$ with a period $T > 0$ are called *orthogonal with respect to the weight $w(x)$* if

$$\int_a^{a+T} \phi(x) \overline{\psi(x)} w(x) dx = 0$$

where $a \in \mathbb{R}$. When $w(x) = 1$, we say that ϕ and ψ are orthogonal.

Complex exponentials of different integers are orthogonal.

Theorem 34.33. For any integers k and ℓ ,

$$\frac{1}{2L} \int_{-L}^L \exp\left(\frac{ik\pi x}{L}\right) \overline{\exp\left(\frac{i\ell\pi x}{L}\right)} dx = \delta_{k\ell}$$

where the overline denotes the complex conjugate.

Definition 34.25. The *complex Fourier series* associated with f , a function defined over $[-L, L]$ (more generally, $[A, A + T]$) or a $2L$ -periodic (more generally, T -periodic) function over $\mathbb{R}$, is

$$f(x) \sim \sum_{n=-\infty}^{\infty} c_n e^{in\pi x/L}, \quad (\text{more generally } f(x) \sim \sum_{n=-\infty}^{\infty} c_n e^{i2n\pi x/T})$$

where the *complex Fourier coefficients* are

$$c_n = \frac{1}{2L} \int_{-L}^L f(x) e^{-in\pi x/L} dx, \quad (\text{more generally } c_n = \frac{1}{b-a} \int_A^{A+T} f(x) e^{-i2n\pi x/T} dx) \quad n \in \mathbb{Z}.$$

Remark 34.34. The complex Fourier coefficients are related to the trigonometric Fourier coefficients by the relation $c_n = \frac{a_n - ib_n}{2}$. *Discrete Fourier transform* (and *fast Fourier transform*) are defined using the discrete version of complex Fourier series.

Example 34.35 (UCCM3003 Final Exam, Apr 2008, Q1(d)). Find the complex Fourier series of the periodic function $\sin(2x)$ over $\mathbb{R}$. (2 marks)

Solution.

$$\sin(2x) = \frac{e^{i2x} - e^{-i2x}}{2i} = \frac{i}{2} e^{-i2x} + \frac{-i}{2} e^{i2x}. \quad [2 \text{ marks}]$$

□

Example 34.36 (UCCM3003 Final Exam, Apr 2011, Q1(b)). Let f be a 4-periodic function. f is defined within the period $-2 < x < 2$ by

$$f(x) = |x| - 1, \quad -2 < x < 2.$$

- (i) State whether f is even, odd or neither. Justify your answer. (3 marks)

Solution. f is even since $f(-x) = |-x| - 1 = |x| - 1 = f(x)$. □

- (ii) Show that the complex Fourier series of f is
$$\sum_{n=-\infty, n \neq 0}^{\infty} \frac{2((-1)^n - 1)}{n^2 \pi^2} e^{in\pi x/2}. \quad (9 \text{ marks})$$

Solution. When $n \neq 0$,

$$\begin{aligned} c_n &= \frac{1}{2 \times 2} \int_{-2}^2 (|x| - 1) e^{-in\pi x/2} dx = \frac{1}{2} \int_0^2 (|x| - 1) \cos \frac{n\pi x}{2} dx \\ &= \frac{1}{2} \left[\frac{2(x-1) \sin \frac{n\pi x}{2}}{n\pi} + \frac{4 \cos \frac{n\pi x}{2}}{(n\pi)^2} \right]_0^2 = \frac{2((-1)^n - 1)}{n^2 \pi^2} \\ c_0 &= \frac{1}{2 \times 2} \int_{-2}^2 (|x| - 1) dx = \frac{1}{2} \int_0^2 (x - 1) dx = \frac{1}{2} \left[\frac{2^2 - 2(2)}{2} \right] = 0 \end{aligned}$$

Hence, we obtain the complex Fourier series of f . □

Example 34.37 (UECM3003 Final Exam, Apr 2013, Q1(b)). Find the complex Fourier series of the periodic function f if $f(x) = e^{-x}$ for $0 < x < 5$ and $f(x+5) = f(x)$ otherwise. Simplify your answer. (7 marks)

Solution. For $n \in \mathbb{Z}$,

$$c_n = \frac{1}{5} \int_0^5 e^{-x} e^{-i2n\pi x/5} dx = \frac{1}{5} \left[\frac{e^{-x-i2n\pi x/5}}{-1-i2n\pi/5} \right]_0^5 = \frac{1-e^{-5-i2n\pi}}{5+2n\pi i} = \frac{1-e^{-5}}{5+2n\pi i}$$

The complex Fourier series of f is $\sum_{n=-\infty}^{\infty} \frac{1-e^{-5}}{5+2n\pi i} e^{i2n\pi x/5}$. □

thm:parseval:c

Theorem 34.38 (Parseval's Theorem for Complex Fourier Series). *If f is a piecewise continuous function and c_n are the complex Fourier coefficients of f , then*

$$\frac{1}{2L} \int_{-L}^L |f(x)|^2 dx = \sum_{n=-\infty}^{\infty} |c_n|^2.$$

Example 34.39. (a) Write down the complex Fourier series for the 2π -periodic function

$$f(x) = e^{ax}, \quad a \in \mathbb{R} \quad -\pi < x < \pi.$$

(b) Show that $\frac{\pi}{a} \coth(a\pi) = \sum_{n=-\infty}^{\infty} \frac{1}{a^2 + n^2}$.

Solution: (a) The complex Fourier coefficients ($L = \pi$) are

$$\begin{aligned} c_n &= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ax} e^{-inx} dx = \frac{1}{2\pi} \left[\frac{e^{(a-in)x}}{a-in} \right]_{-\pi}^{\pi} \\ &= \frac{1}{2\pi} \left[\frac{e^{-in\pi} e^{a\pi} - e^{in\pi} e^{-a\pi}}{a-in} \right] = \frac{1}{2\pi} \frac{2(-1)^n \sinh(a\pi)}{a-in}. \quad (\because e^{\pm in\pi} = (-1)^n) \end{aligned}$$

Hence the complex Fourier series is

$$f(x) = \frac{\sinh(a\pi)}{\pi} \sum_{n=-\infty}^{\infty} \frac{(-1)^n}{a-in} e^{inx}.$$

(b) By applying the Parseval's theorem, we obtain

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{2ax} dx = \frac{\sinh(2a\pi)}{2a\pi} = \frac{\sinh^2(a\pi)}{\pi^2} \sum_{n=-\infty}^{\infty} \frac{1}{a^2 + n^2}.$$

After simplifying the above equation, we obtain the equality.

Example 34.40 (UCCM3003 Final Exam, Sep 2007, Q1(b)). (i) Find the complex Fourier series of the following periodic function with a period of 4.

$$g(x) = 3x, \quad -2 < x < 2 \quad (6 \text{ marks})$$

Solution.

$$c_0 = \frac{1}{4} \int_{-2}^2 3x dx = \frac{3}{4} \left[\frac{x^2}{2} \right]_{-2}^2 = 0 \quad [1 \text{ mark}]$$

For $n \neq 0$,

$$\begin{aligned} c_n &= \frac{1}{4} \int_{-2}^2 3x e^{-\frac{2\pi i x}{4}} dx = \frac{3}{4} \left[\frac{2ix e^{-in\pi x/2}}{n\pi} + \frac{4e^{-in\pi x/2}}{(n\pi)^2} \right]_{-2}^2 \\ &= \frac{3}{4} \left[\frac{2i[2(-1)^n - (-2)(-1)^n]}{n\pi} + \frac{4[(-1)^n - (-1)^n]}{n^2\pi^2} \right] = \frac{6i(-1)^n}{n\pi} \end{aligned} \quad [4 \text{ marks}]$$

$$f(x) \sim \sum_{\substack{n=-\infty \\ n \neq 0}}^{\infty} \frac{6i(-1)^n}{n\pi} e^{\frac{in\pi x}{2}} \dots\dots\dots [1 \text{ mark}] \quad \square$$

(ii) By applying the Parseval's Theorem

$$\frac{1}{2L} \int_{-L}^L |f(x)|^2 dx = \sum_{n=-\infty}^{\infty} |c_n|^2$$

to part (b)(i) above, show that $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$. (3 marks)

Solution.

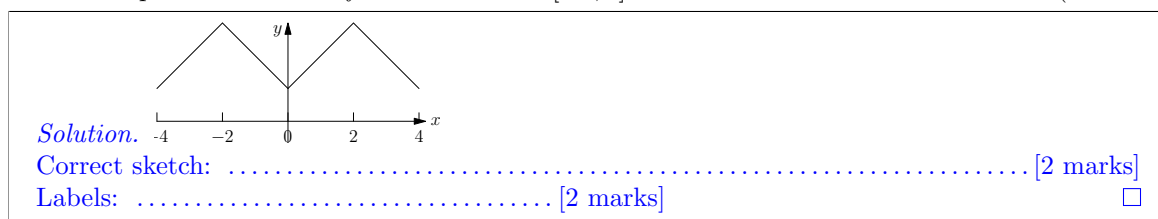
$$\frac{1}{4} \int_{-2}^2 (3x)^2 dx = \frac{9}{4} \left[\frac{x^3}{3} \right]_{-2}^2 = \frac{9 \times 16}{12} = 12 \quad [1 \text{ mark}]$$

$$\sum_{n=-\infty}^{\infty} |c_n|^2 = 2 \sum_{n=1}^{\infty} \left| \frac{6}{n\pi} \right|^2 = \frac{72}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n^2} = 12 \Rightarrow \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} \dots\dots\dots [2 \text{ marks}] \quad \square$$

Example 34.41 (UCCM3003 Final Exam, Apr 2008, Q1). Consider the following periodic function with a period of 4:

$$f(x) = |x| + 1, \quad -2 < x < 2.$$

(a) Sketch the periodic function f for the interval $[-4, 4]$. (4 marks)



(b) Find the trigonometric Fourier series of the periodic function f . (10 marks)

Solution. Since $f(-x) = |-x| + 1 = f(x)$. Hence $f(x)$ is even and

$$b_n = 0 \dots\dots\dots (1 \text{ mark})$$

$$a_0 = \frac{1}{2} \int_{-2}^2 (|x| + 1) dx = \int_0^2 (x + 1) dx \quad (1 \text{ mark})$$

$$= \left[\frac{x^2}{2} + x \right]_0^2 = 2 + 2 = 4 \quad (1 \text{ mark})$$

$$a_n = \int_0^2 (x + 1) \cos \frac{n\pi x}{2} dx \quad (1 \text{ mark})$$

$$= \left[\frac{2(x+1)}{n\pi} \sin \frac{n\pi x}{2} + \frac{4}{n^2\pi^2} \cos \frac{n\pi x}{2} \right]_0^2 \quad (3 \text{ marks})$$

$$= \frac{4}{n^2\pi^2} [(-1)^n - 1] \quad (2 \text{ marks})$$

So $f(x) = 2 + \sum_{n=1}^{\infty} \frac{4}{n^2\pi^2} [(-1)^n - 1] \cos \frac{n\pi x}{2} \dots\dots\dots (1 \text{ mark}) \quad \square$

(c) Find the complex Fourier series of the periodic function f . (6 marks)

Solution. $c_n = \frac{1}{4} \int_{-2}^2 (|x| + 1) e^{-in\frac{\pi}{2}x} dx \quad (1 \text{ mark})$

$$= \frac{1}{4} \int_{-2}^2 (|x| + 1) (\cos \frac{n\pi x}{2} - i \sin \frac{n\pi x}{2}) dx \quad (1 \text{ mark})$$

$$= \frac{1}{4} \int_{-2}^2 (|x| + 1) \cos \frac{n\pi x}{2} dx \quad (1 \text{ mark})$$

$$= \frac{1}{2} a_n = \frac{2}{n^2\pi^2} [(-1)^n - 1], \quad n \neq 0 \quad (1 \text{ mark})$$

$$c_0 = \frac{1}{4} \int_{-2}^2 (|x| + 1) dx = \frac{1}{2} a_0 = 2 \quad (1 \text{ mark})$$

So $f(x) = 2 + \sum_{\substack{n=-\infty \\ n \neq 0}}^{\infty} \frac{2}{n^2\pi^2} [(-1)^n - 1] e^{in\pi x/2} \dots\dots\dots (1 \text{ mark}) \quad \square$

(d) By applying the Parseval's Theorem, show that $\sum_{k=1}^{\infty} \frac{1}{(2k-1)^4} = \frac{\pi^4}{96}$.

Solution.

$$\frac{1}{2} \int_{-2}^2 (|x| + 1)^2 dx = \frac{4^2}{2} + \sum_{n=1}^{\infty} \left(\frac{4}{n^2 \pi^2} [(-1)^n - 1] \right)^2 \quad (1 \text{ mark})$$

$$\Rightarrow \int_0^2 (x+1)^2 dx = 8 + \sum_{k=1}^{\infty} \left(\frac{-8}{(2k-1)^2 \pi^2} \right)^2 \quad (1 \text{ mark})$$

$$\Rightarrow \left[\frac{(x+1)^3}{3} \right]_0^2 = 8 + \sum_{k=1}^{\infty} \frac{64}{(2k-1)^4 \pi^4} \quad (1 \text{ mark})$$

$$\Rightarrow \frac{27}{3} - \frac{1}{3} - 8 = \frac{2}{3} = \sum_{k=1}^{\infty} \frac{64}{(2k-1)^4 \pi^4} \quad (1 \text{ mark})$$

$$\Rightarrow \sum_{k=1}^{\infty} \frac{1}{(2k-1)^4} = \frac{\pi^4}{32 \times 3} = \frac{\pi^4}{96} \quad (1 \text{ mark})$$

□

g:final200901q2

Example 34.42 (UCCM3003 Final Exam, Apr 2009, Q2). Let $f(t) = 3t^2$ in $-1 \leq t < 1$ be a periodic function of period 2.

- (a) Find the trigonometric Fourier series for f . (12 marks)

Solution. Since $f(-t) = f(t)$, f is even, $b_n = 0$ and

$$a_0 = \int_{-1}^1 3t^2 dt = 2$$

$$a_n = 2 \int_0^1 3t^2 \cos n\pi t dt = 2 \left[\frac{3t^2 \sin n\pi t}{n\pi} + \frac{6t \cos n\pi t}{(n\pi)^2} - \frac{6 \sin n\pi t}{(n\pi)^3} \right]_0^1 = \frac{12(-1)^n}{n^2 \pi^2}$$

$$\text{for } n = 1, 2, \dots, f(t) = 1 + \sum_{n=1}^{\infty} \frac{12(-1)^n}{n^2 \pi^2} \cos n\pi t.$$

□

- (b) Find the corresponding complex Fourier series of f . (5 marks)

Solution. When $n \neq 0$,

$$c_n = \frac{1}{2} \int_{-1}^1 3t^2 e^{-in\pi t} dt = \int_0^1 3t^2 \cos n\pi t dt = \frac{6(-1)^n}{n^2 \pi^2}; \quad c_0 = \frac{1}{2} \int_{-1}^1 3t^2 dt = 1.$$

$$\text{The corresponding complex Fourier series of } f \text{ is } 1 + \sum_{n=-\infty, n \neq 0}^{\infty} \frac{6(-1)^n}{n^2 \pi^2} e^{in\pi x}.$$

□

- (c) From the result in part (a), show that $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$. (3 marks)

Solution. The function f is continuous and by the Pointwise Convergence Theorem 34.19, when $x = 1$, we have

$$\frac{f(1+0) + f(1-0)}{2} = 3 = 1 + \sum_{n=1}^{\infty} \frac{12(-1)^n}{n^2 \pi^2} (-1)^n \Rightarrow \frac{12}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n^2} = 2 \Rightarrow \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$

□

- (d) (Also, in UCCM3003 Final Exam, Apr 2011, Q1(b)) By using the Parseval's theorem and the result in part (a), show that $\sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{90}$. (5 marks)

$$\text{Solution. } \frac{2^2}{2} + \sum_{n=1}^{\infty} \frac{12^2}{n^4 \pi^4} = \int_{-1}^1 (3t^2)^2 dt = \frac{18}{5} \Rightarrow \sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{12^2} \left[\frac{18}{5} - 2 \right] = \frac{\pi^4}{90}.$$

□

exam:201005:q1b

Example 34.43 (UCCM3003 Final Exam, May 2010, Q1(b)).

- (i) Show that the complex Fourier series of the 4-periodic function

$$g(x) = \begin{cases} 0, & -2 < x < 0, \\ 1, & 0 < x < 2, \end{cases}$$

is $g^*(x) = \frac{1}{2} + \frac{i}{2\pi} \sum_{n=-\infty, n \neq 0}^{\infty} \frac{(-1)^n - 1}{n} \exp\left(\frac{in\pi x}{2}\right)$. (6 marks)

Solution. When $n \neq 0$,

$$c_n = \frac{1}{2 \times 2} \int_0^2 e^{-in\pi x/2} dx = \frac{1}{4} \left[\frac{2e^{-in\pi x/2}}{-in\pi} \right]_0^2 = \frac{1 - e^{-in\pi}}{2in\pi} = \frac{1 - (-1)^n}{2in\pi}$$

$$c_0 = \frac{1}{2 \times 2} \int_0^2 1 dx = \frac{2}{4} = \frac{1}{2}$$

When $x \neq 0, \pm 4, \pm 8, \dots$, $g(x) = \frac{1}{2} + \sum_{\substack{n=-\infty \\ n \neq 0}}^{\infty} \frac{i((-1)^n - 1)}{2n\pi} e^{in\pi x/2}$. □

(ii) Using the Fourier coefficients of g^* , together with the Parseval's Theorem, show that

$$\sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} = \frac{\pi^2}{8}. \quad (6 \text{ marks})$$

Solution. By Theorem ~~34.38~~^{thm:parseval:c}, we have

$$\left(\frac{1}{2}\right)^2 + \sum_{\substack{n=-\infty \\ n \neq 0}}^{\infty} \left(\frac{1 - (-1)^n}{2n\pi}\right)^2 = \frac{1}{2 \times 2} \int_{-2}^2 f(x)^2 dx = \frac{1}{4} \int_0^2 1 dx = \frac{1}{2}$$

$$\therefore \sum_{\substack{n=-\infty \\ n \neq 0}}^{\infty} \left(\frac{2}{2(2n-1)\pi}\right)^2 = \frac{1}{4} \Rightarrow \sum_{\substack{n=-\infty \\ n \neq 0}}^{\infty} \frac{1}{(2n-1)^2} = 2 \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} = \frac{\pi^2}{4} \Rightarrow \text{above.}$$

□

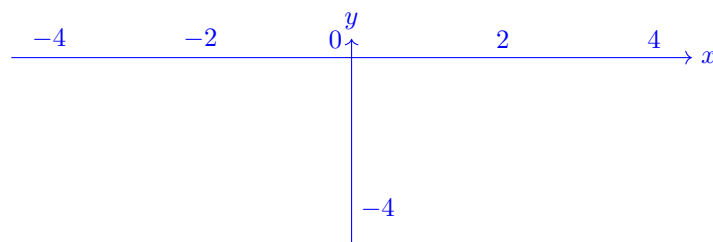
Example 34.44 (UECM3003 Final Exam, Apr 2014, Q1(a)). Let f be a 4-periodic function. Given that in the interval of $0 \leq x \leq 4$, f is defined as follows

$$f(x) = x(x-4), \quad 0 \leq x \leq 4.$$

(i) Sketch the graph of f for the interval $[-4, 4]$. State whether f is even, odd or neither.

(4 marks)

Solution. From the graph below, we observe that the function f is even.



□

(ii) Find the complex Fourier series of f .

(10 marks)

Solution. When $n \neq 0$

$$c_n = \frac{1}{4} \int_0^4 x(x-4) e^{-i2n\pi x/4} dx = \frac{2}{4} \int_0^2 x(x-4) \cos \frac{n\pi x}{2} dx$$

$$= \frac{1}{2} \left[\frac{(x^2 - 4x) \sin \frac{n\pi x}{2}}{\frac{n\pi}{2}} + \frac{(2x-4) \cos \frac{n\pi x}{2}}{(\frac{n\pi}{2})^2} - \frac{2 \sin \frac{n\pi x}{2}}{(\frac{n\pi}{2})^3} \right]_0^2$$

$$= \frac{1}{2} \left[0 + \frac{16}{n^2 \pi^2} + 0 \right] = \frac{8}{n^2 \pi^2}$$

$$c_0 = \frac{1}{4} \int_0^4 x(x-4) dx = \frac{1}{4} \left[\frac{x^3}{3} - 2x^2 \right]_0^4 = -\frac{8}{3}$$

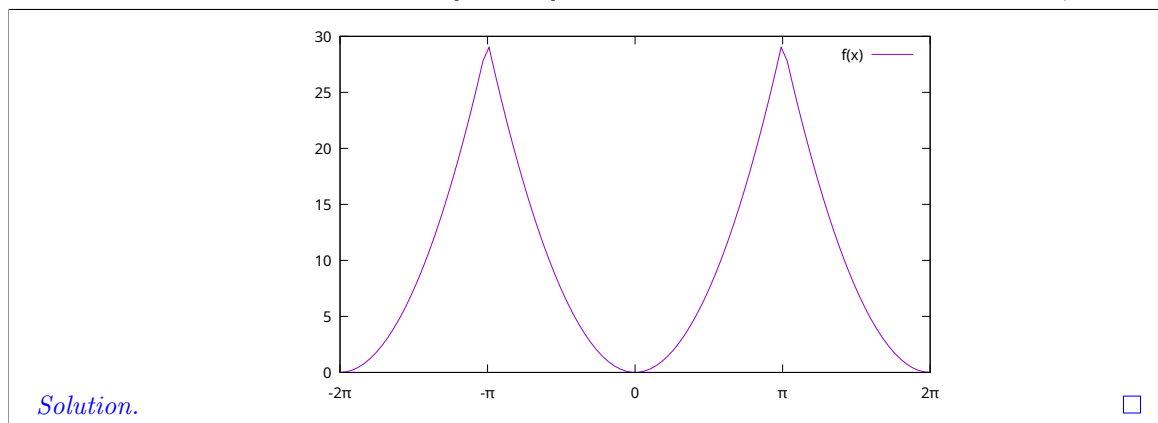
$$\text{Therefore, } f(x) \sim -\frac{8}{3} + \sum_{n=-\infty, n \neq 0}^{\infty} \frac{8}{n^2 \pi^2} e^{in\pi x/2}. \quad \square$$

(iii) Hence evaluate $\sum_{n=1}^{\infty} \frac{1}{n^4}$. (5 marks)

$$\begin{aligned} \text{Solution. } \left(-\frac{8}{3}\right)^2 + 2 \sum_{n=1}^{\infty} \frac{8^2}{n^4 \pi^4} &= \frac{1}{4} \int_0^4 x^2(x-4)^4 dx = \frac{1}{4} \times \frac{512}{15} = \frac{128}{15} \\ \therefore \frac{128}{\pi^4} \sum_{n=1}^{\infty} \frac{1}{n^4} &= \frac{128}{15} - \frac{64}{9} = \frac{64}{45} \Rightarrow \sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{90}. \quad \square \end{aligned}$$

Example 34.45 (UECM3003 Final Exam, Apr 2015, Q1). Let f be a 2π periodic function and $f(x) = 3x^2$ in $-\pi \leq x \leq \pi$.

(a) Sketch the graph of f for the interval $[-2\pi, 2\pi]$. (2 marks)



(b) Show that the complex Fourier series of f is

$$f(x) \sim \pi^2 + \sum_{n=-\infty, n \neq 0}^{\infty} \frac{6(-1)^n}{n^2} e^{inx}. \quad (8 \text{ marks})$$

Solution. This is very similar to Example 34.42 except in this case, $L = \pi$. When $n \neq 0$,

$$\begin{aligned} c_n &= \frac{1}{2\pi} \int_{-\pi}^{\pi} 3x^2 e^{-inx} dx = \frac{3}{\pi} \int_0^{\pi} x^2 \cos nx dx \\ &= \frac{3}{\pi} \left[\frac{x^2 \sin nx}{n} + \frac{2x \cos nx}{n^2} - \frac{2 \sin nx}{n^3} \right]_0^{\pi} = \frac{3}{\pi} \left[0 + \frac{2\pi(-1)^n}{n^2} - 0 \right] = \frac{6(-1)^n}{n^2} \\ c_0 &= \frac{1}{2\pi} \int_{-\pi}^{\pi} 3x^2 dx = \frac{2\pi^3}{2\pi} = \pi^2 \end{aligned}$$

$$\text{Hence, the complex Fourier series of } f \text{ is } \pi^2 + \sum_{\substack{n=-\infty \\ n \neq 0}}^{\infty} \frac{6(-1)^n}{n^2} e^{inx}. \quad \square$$

(c) Find the trigonometric Fourier series of f . (6 marks)

Solution. Since f is even, $b_n = 0$ and

$$a_0 = 2c_0 = 2\pi^2 \quad a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} 3x^2 \cos nx dx = \frac{6}{\pi} \int_0^{\pi} x^2 \cos nx dx = \frac{12(-1)^n}{n^2}.$$

$$\text{The trigonometric Fourier series of } f \text{ is } \pi^2 + \sum_{n=1}^{\infty} \frac{12(-1)^n}{n^2} \cos nx. \quad \square$$

(d) Hence, show that $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$ and $\sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{90}$. (9 marks)

Solution: The derivation is very similar to Example 34.42 part (c) and (d) and is left as an exercise. eg: final200901q2

34.2.7 Application: Isoperimetric Theorem

The Parseval's theorem is useful in proving an interesting theorem for smooth closed curves in a plane. The "isoperimetric problem" is the problem of finding the simple closed plane curve of given perimeter with maximum area [Courant and Hilbert, 1953, 118].

Theorem 34.46. Consider a smooth closed curve in the plane xy . Denote by P its perimeter (total arclength) and by A the area of the region enclosed by the curve. Then we have

$$4\pi A \leq P^2.$$

The equality holds if and only if the curve is a circle.

Proof: A parametric representation of the closed curve is given by

$$(x(t), y(t)), \quad t \in [-\pi, \pi], \quad x(-\pi) = x(\pi), \quad y(-\pi) = y(\pi).$$

The formulae giving length P and area A of the closed curve are

$$P = \int_{-\pi}^{\pi} \sqrt{(\dot{x})^2 + (\dot{y})^2} dt \quad \text{and} \quad A = \int_{-\pi}^{\pi} x \dot{y} dt.$$

where $\dot{x} = \frac{dx}{dt}$ and $\dot{y} = \frac{dy}{dt}$.

Let

$$s(t) = \int_{-\pi}^t \sqrt{(\dot{x}(u))^2 + (\dot{y}(u))^2} du.$$

Then $P = s(\pi)$. Consider the new variable $\tilde{t} = -\pi + \frac{2\pi s(t)}{P}$. If we rewrite the parametric representation in terms of $\tilde{t}$, we get

$$(\tilde{x}(\tilde{t}), \tilde{y}(\tilde{t})) = (x(t), y(t)).$$

Since $\frac{d\tilde{x}}{d\tilde{t}} = \frac{dx}{dt} \frac{dt}{d\tilde{t}} = \frac{\dot{x}}{\frac{d\tilde{t}}{dt}} = \frac{\dot{x}}{2\pi(\dot{x}^2 + \dot{y}^2)/P}$ and $\frac{d\tilde{y}}{d\tilde{t}} = \frac{\dot{y}}{2\pi(\dot{x}^2 + \dot{y}^2)/P}$ we have

$$\left(\frac{d\tilde{x}}{d\tilde{t}}\right)^2 + \left(\frac{d\tilde{y}}{d\tilde{t}}\right)^2 = \frac{P^2}{4\pi^2},$$

i.e. the new variable enables us to reparametrise the curve while assuming the quantity $(\dot{x})^2 + (\dot{y})^2$ to be constant, i.e. $(\dot{x})^2 + (\dot{y})^2 = \frac{P^2}{4\pi^2}$.

Since the curve is smooth, we can expand x and y into Fourier series

$$\begin{cases} x(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \{a_n \cos nt + b_n \sin nt\}, \\ y(t) = \frac{c_0}{2} + \sum_{n=1}^{\infty} \{c_n \cos nt + d_n \sin nt\}. \end{cases}$$

Taking derivative for the above equations gives

$$\begin{cases} \dot{x}(t) = \sum_{n=1}^{\infty} n \{-a_n \sin nt + b_n \cos nt\}, \\ \dot{y}(t) = \sum_{n=1}^{\infty} n \{-c_n \sin nt + d_n \cos nt\}. \end{cases}$$

Parseval's theorem implies

$$\frac{P^2}{2\pi} = \int_{-\pi}^{\pi} ((\dot{x})^2 + (\dot{y})^2) dt = \pi \sum_{n=1}^{\infty} n^2 \{a_n^2 + b_n^2 + c_n^2 + d_n^2\}.$$

On the other hand, using the orthogonality property of sine and cosine functions, we have

$$A = \int_{-\pi}^{\pi} x(t)\dot{y}(t)dt = \pi \sum_{n=1}^{\infty} n\{a_nd_n - b_nc_n\}.$$

Algebraic manipulations imply

$$\frac{P^2}{2\pi} - 2A = \pi \sum_{n=1}^{\infty} \left[n(a_n - d_n)^2 + n(b_n + c_n)^2 + n(n-1)(a_n^2 + b_n^2 + c_n^2 + d_n^2) \right].$$

Since the left hand side of this equality is positive, we deduce

$$P^2 \geq 4\pi A.$$

On the other hand, we will have $\frac{P^2}{2\pi} - 2A = 0$ iff

$$\begin{cases} a_n^2 + b_n^2 + c_n^2 + d_n^2 = 0, & \text{for any } n > 1 \\ a_1 - d_1 & = 0, \\ b_1 + c_1 & = 0. \end{cases}$$

This implies

$$\begin{cases} x(t) = a_0 + a_1 \cos(t) - c_1 \sin(t), \\ y(t) = c_0 + c_1 \cos(t) + a_1 \sin(t) \end{cases}$$

for $t \in [-\pi, \pi]$. Therefore the curve is a circle centred at (a_0, c_0) with radius $\sqrt{a_1^2 + c_1^2}$, which completes the proof of the theorem.

34.3 Multiple Fourier Series

According to Section [34.2](#), a periodic function of a single variable $f(x)$ has a corresponding Fourier series? What about a periodic function of multi-variable, for example, $g(x, y)$, $h(x, y, z)$? Yes, we can and this is discussed in the textbook Grafakos [2004]! However, it is very abstract and we will only discuss the “double Fourier series” which is introduced in Weinberger [1995] and Tolstov [1976, Chapter 7].

Suppose we have a continuously differentiable two-variable function $f(x, y)$ which is periodic:

$$f(x + L, y) = f(x, y + M) = f(x, y), \quad \forall x, y \in \mathbb{R}.$$

For each fixed value of y we can expand $f(x, y)$ in a uniformly convergent Fourier series

$$f(x, y) = \frac{a_0(y)}{2} + \sum_{n=1}^{\infty} \left[a_n(y) \cos \frac{n\pi x}{L} + b_n(y) \sin \frac{n\pi x}{L} \right].$$

The coefficients

$$a_n(y) = \frac{1}{L} \int_{-L}^L f(x, y) \cos \frac{n\pi x}{L} dx, \quad b_n(y) = \frac{1}{L} \int_{-L}^L f(x, y) \sin \frac{n\pi x}{L} dx$$

are continuously differentiable in y . Therefore, they can be expanded in uniformly convergent Fourier series

$$\begin{aligned} a_n(y) &= \frac{a_{n0}}{2} + \sum_{m=1}^{\infty} \left(a_{nm} \cos \frac{m\pi y}{M} + b_{nm} \sin \frac{m\pi y}{M} \right), \quad n = 0, 1, 2, \dots, \\ b_n(y) &= \frac{c_{n0}}{2} + \sum_{m=1}^{\infty} \left(c_{nm} \cos \frac{m\pi y}{M} + d_{nm} \sin \frac{m\pi y}{M} \right), \quad n = 1, 2, \dots \end{aligned}$$

where

$$\begin{aligned} a_{nm} &= \frac{1}{ML} \int_{-L}^L \int_{-M}^M f(x, y) \cos \frac{n\pi x}{L} \cos \frac{m\pi y}{M} dx dy, \quad m = 0, 1, \dots, \quad n = 0, 1, \dots, \\ b_{nm} &= \frac{1}{ML} \int_{-L}^L \int_{-M}^M f(x, y) \cos \frac{n\pi x}{L} \sin \frac{m\pi y}{M} dx dy, \quad m = 1, 2, \dots, \quad n = 0, 1, \dots, \\ c_{nm} &= \frac{1}{ML} \int_{-L}^L \int_{-M}^M f(x, y) \sin \frac{n\pi x}{L} \cos \frac{m\pi y}{M} dx dy, \quad m = 0, 1, \dots, \quad n = 1, 2, \dots, \\ d_{nm} &= \frac{1}{ML} \int_{-L}^L \int_{-M}^M f(x, y) \sin \frac{n\pi x}{L} \sin \frac{m\pi y}{M} dx dy, \quad m = 1, 2, \dots, \quad n = 1, 2, \dots. \end{aligned}$$

Putting the series for the coefficients into the series of $f(x, y)$, we have

$$\begin{aligned} f(x, y) &= \frac{a_{00}}{4} + \frac{1}{2} \sum_{m=1}^{\infty} \left[a_{0m} \cos \frac{m\pi y}{M} + b_{0m} \sin \frac{m\pi y}{M} \right] + \frac{1}{2} \sum_{n=1}^{\infty} \left[a_{n0} \cos \frac{n\pi x}{L} + c_{n0} \sin \frac{n\pi x}{L} \right] \\ &\quad + \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \left[a_{nm} \cos \frac{n\pi x}{L} \cos \frac{m\pi y}{M} + b_{nm} \cos \frac{n\pi x}{L} \sin \frac{m\pi y}{M} \right. \\ &\quad \left. + c_{nm} \sin \frac{n\pi x}{L} \cos \frac{m\pi y}{M} + d_{nm} \sin \frac{n\pi x}{L} \sin \frac{m\pi y}{M} \right]. \end{aligned}$$

The Parseval's equation now looks like

$$\frac{1}{ML} \int_{-L}^L \int_{-M}^M f(x, y)^2 dx dy = \frac{a_{00}^2}{4} + \frac{1}{2} \sum_{n=1}^{\infty} (a_{0n}^2 + b_{0n}^2 + a_{n0}^2 + c_{n0}^2) + \sum_{n,m=1}^{\infty} (a_{nm}^2 + b_{nm}^2 + c_{nm}^2 + d_{nm}^2).$$

This is the Parseval equation for double Fourier series. It has been derived under the hypothesis that $f(x, y)$ is continuously differentiable. However, if we simply assume that f is such that $\int_{-\pi}^{\pi} \int_{-\pi}^{\pi} f^2 dx dy$ is finite, f can be approximated in the mean by continuously differentiable functions. It follows that the Parseval equation remains valid for such functions.

If $f(x, y)$ is continuous and continuously differentiable and if squares of its second partial derivatives have finite integrals, then the double Fourier series converges absolutely and uniformly to $f(x, y)$ as a double series.

If $f(x, y)$ is an odd function of x and y , all the Fourier coefficients except the d_{nm} vanish. The resulting series is a double sine series. Similarly if f is even in x and odd in y , only the b_{nm} are nonzero. If f is prescribed only for $0 < x < \pi$, $0 < y < \pi$, it may be continued as either an odd or an even function of each of the variables. Thus it may be expressed as a double sine series, a double cosine series, a sine-cosine series, or a cosine-sine series [Asmar, 2005, 158].

Example 34.47. Find the double Fourier series for $f(x, y) = x^2 y^2$, $(x, y) \in [-\pi, \pi]^2$.

Solution: Since f is even, we would expect $b_{nm} = c_{nm} = d_{nm} = 0$.

$$a_{00} = \frac{1}{\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} x^2 y^2 dx dy = \frac{4}{\pi^2} \int_0^{\pi} x^2 dx \int_0^{\pi} y^2 dy = \frac{4 \cdot \pi^3 \cdot \pi^3}{3 \cdot 3\pi^2} = \frac{4\pi^4}{9}$$

$$a_{n0} = \frac{1}{\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} x^2 y^2 \cos nx dx dy = \frac{4\pi}{3} \int_0^{\pi} x^2 \cos nx dx = \frac{8\pi^2(-1)^n}{3n^2}$$

$$a_{0m} = \frac{1}{\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} x^2 y^2 \cos my dx dy = \frac{4\pi}{3} \int_0^{\pi} y^2 \cos my dy = \frac{8\pi^2(-1)^m}{3m^2}$$

$$\begin{aligned} a_{nm} &= \frac{1}{\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} x^2 y^2 \cos nx \cos my dx dy \\ &= \frac{4}{\pi^2} \left(\int_0^{\pi} x^2 \cos nx dx \right) \left(\int_0^{\pi} y^2 \cos my dy \right) = \frac{16(-1)^n(-1)^m}{n^2 m^2} \end{aligned}$$

$$x^2 y^2 = \frac{\pi^4}{9} + \frac{4\pi^2}{3} \sum_{n=1}^{\infty} \frac{(-1)^n}{n^2} \cos nx + \frac{4\pi^2}{3} \sum_{m=1}^{\infty} \frac{(-1)^m}{m^2} \cos my + 16 \sum_{n,m=1}^{\infty} \frac{(-1)^{n+m}}{n^2 m^2} \cos nx \cos my.$$

It is possible to develop a general theory for multiple Fourier series. This was done in Grafakos [2004]. However, to avoid the cumbersome “periods”, all of them are normalised to 1, and mathematicians introduced the torus where multiple Fourier series lived on:

$$\mathbb{T}^n = [0, 1]^n / \sim$$

where $\sim$ indicates that the values 0 and 1 are regarded as the same.

Definition 34.26. A function on $\mathbb{T}^n$ is a function on $\mathbb{R}^n$ that satisfies $f(\mathbf{x} + \mathbf{m}) = f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{m} \in \mathbb{Z}^n$.

Definition 34.27. For a complex valued function $f \in L^1(\mathbb{T}^n)$ and $\mathbf{m} \in \mathbb{Z}^n$,

$$\widehat{f}(\mathbf{m}) := \int_{\mathbb{T}^n} f(\mathbf{x}) e^{-2\pi i \mathbf{m} \cdot \mathbf{x}} d\mathbf{x}$$

is called the $\mathbf{m}$ th Fourier coefficients of f .

Note that $L^1(\mathbb{T}^n)$ means all the Lebesgue integrable functions (it should have been mentioned in UECM2033 Elementary Real Analysis). In previous sections, we have limit ourselves to only piecewise continuous functions, which are of interest to physicists. However, the theory of Fourier analysis can be studied on the more general but abstract functions in $L^1(\mathbb{T}^n)$.

Definition 34.28. The (multiple) Fourier series of f at $\mathbf{x} \in \mathbb{T}^n$ is the series

$$\sum_{\mathbf{m} \in \mathbb{Z}^n} \widehat{f}(\mathbf{m}) e^{2\pi i \mathbf{m} \cdot \mathbf{x}}.$$

34.4 Sturm-Liouville (Eigenvalue) Problems

By applying the separation of variables to solve second order partial differential equation problems in Topics 2, 3 and 4 will lead to *Sturm-Liouville (eigenvalue) problems (SL problems for short)*, described by a second order ODE in *self-adjoint form* and additional boundary conditions:

$$\frac{d}{dx} \left(p(x) \frac{du}{dx} \right) - q(x)u + \lambda r(x)u = 0 \quad \text{for } \alpha < x < \beta. \quad (34.9)$$

Definition 34.29. When

- the functions r and q in (34.9) are continuous, and the function p is continuously differentiable;
 - the values $p(x)$ and $r(x)$ are positive and the value $q(x)$ is nonnegative for $\alpha \leq x \leq \beta$,
- the ODE (34.9) is said to be *regular*.

Definition 34.30. If $c_1^2 + c_2^2 \neq 0$, $c_3^2 + c_4^2 \neq 0$, the regular ODE (34.9) and the boundary conditions

$$c_1 u(\alpha) + c_2 u'(\alpha) = 0, \quad c_3 u(\beta) + c_4 u'(\beta) = 0 \quad (34.10)$$

together form a *regular Sturm-Liouville (eigenvalue) problem (or regular SL problem)*.

Definition 34.31. If the ODE (34.9) is not regular or the boundary condition is infinite, then (34.9) and additional boundary conditions form a *singular Sturm-Liouville problem*.

Definition 34.32. A problem involving a regular ODE (34.9) and the following constraints

$$p(\alpha) = p(\beta), \quad u(\alpha) = u(\beta), \quad u'(\alpha) = u'(\beta)$$

is called a *periodic SL problem*.

This section will just introduce the basic results of regular Sturm-Liouville problems and we will investigate relevant singular Sturm-Liouville. The proofs can be found in Körner [1988, Chapter 39], Al-Gwaiz [2008], Haberman [1987, Chapter 5], Bell [1968b], Asmar [2005], Weinberger [1995], Courant and Hilbert [1953], etc.

Definition 34.33. It is easy to check that $u = 0$ is a solution of SL problem (34.9) and (34.10) and it is called a *trivial solution*.

Definition 34.34. The values of λ for which the problem (34.9), (34.10) has a nontrivial solution, i.e. a solution other than $u \equiv 0$, are called the *eigenvalues*. The corresponding solutions are called the *eigenfunctions*.

If $\bar{u}$ is an eigenfunction corresponding to an eigenvalue λ , then any constant multiple of $\bar{u}$ is also an eigenfunction.

Remark 34.48. The SL problem can be seen as an extension of the generalised eigenvalue problem in linear algebra (we only learned how to solve a special case of this problem $A\mathbf{x} = \lambda\mathbf{x}$ in UECM1313 Fundamental of Linear Algebra)

$$(A - \lambda B)\mathbf{x} = 0, \quad A \text{ is symmetric, } B \text{ is symmetric positive definite} \quad (34.11)$$

to the framework of second order ODEs, where the “vector” is now a “function”.

Some well known results for the problem (34.11) are:

- the eigenvalues λ form a finite ordered sequence of real numbers.
- the eigenvectors form a *B-orthogonal basis* for the vector space $\mathbb{R}^n$, i.e. $v_i^T B v_j = \delta_{ij}$.

thm:sl

Theorem 34.49. The eigenvalues and eigenfunctions of a regular Sturm-Liouville problem (34.9) and (34.10) have the following properties

(a) **All eigenvalues are real** and compose a countably infinite collections satisfying

$$\lambda_1 < \lambda_2 < \cdots < \lambda_n \rightarrow \infty \quad \text{as } n \rightarrow \infty.$$

When $c_2 = c_4 = 0$, $\lambda_1 > 0$.

- (b) To each eigenvalue λ_n there corresponds only **one** independent eigenfunction $u_n(x)$.
- (c) The eigenfunctions $u_n(x)$, $n = 1, 2, \dots$, form a “complete” set, that is, any piecewise smooth function f can be represented by a series of the eigenfunctions:

$$f(x) \sim \sum_{n=1}^{\infty} a_n u_n(x)$$

- (d) Eigenfunctions u_k and u_ℓ corresponding to different eigenvalues λ_k and λ_ℓ are **orthogonal relative to the weight function r** , i.e.

$$\int_{\alpha}^{\beta} u_k(x) u_\ell(x) r(x) dx = 0.$$

- (e) Any eigenvalue can be related to its eigenfunction by the Raleigh quotient:

$$\lambda = \frac{-puu'|_{\alpha}^{\beta} + \int_{\alpha}^{\beta} [p(u')^2 + qu^2] dx}{\int_{\alpha}^{\beta} u^2 r dx}.$$

The above theorem may be true for certain singular SL problems (like Example 34.54) as we will see in later topics [Haberman, 1987, Sections 6.7 & 6.8].

The following are some examples of regular Sturm-Liouville problems.

Example 34.50 (UCCM3003 Final Exam, May 2010, Q2(b)). Show that for the case when k is positive, the eigenvalues and the corresponding eigenfunctions of the Sturm-Liouville problem

$$X'' + kX = 0, \quad X(0) = X(\pi) = 0$$

are respectively $k = n^2$ and $X_n(x) = \sin(nx)$ for $n = 1, 2, 3, \dots$. You do not have to consider the cases $k = 0$ and $k < 0$, which do not have any eigenvalues. (6 marks)

Remark: Since $p(x) = 1 > 0$, $q(x) = 0 \geq 0$ and $r(x) = 1 > 0$ for $\frac{1}{2} \leq x \leq 1$, the Sturm-Liouville problem is regular by definition.

Solution: See Topic 2.

Example 34.51 (UCCM3003 Final Exam, May 2010, Q3(b)). The Sturm-Liouville problem

$$x^2 y'' + 2xy' + \frac{\lambda}{x^2} y = 0, \quad y\left(\frac{1}{2}\right) = y(1) = 0$$

has eigenfunctions $y_n(x) = \sin\left(\frac{n\pi}{x}\right)$ for $n = 1, 2, 3, \dots$. Put the equation in self-adjoint form and give an orthogonality relation. (6 marks)

Solution. The Sturm-Liouville problem can be written in the following self-adjoint form:

$$\frac{d}{dx}\left(x^2 \frac{dy}{dx}\right) + \lambda\left(\frac{1}{x^2}\right)y = 0$$

Therefore, $p(x) = x^2 > 0$, $q(x) = 0 \geq 0$ and $r(x) = \frac{1}{x^2} > 0$ for $\frac{1}{2} \leq x \leq 1$, the Sturm-Liouville problem is regular by definition.

The Theorem 34.49 (d) characterises the “orthogonality relation” below:

$$\int_{1/2}^1 \frac{\sin\left(\frac{m\pi}{x}\right) \sin\left(\frac{n\pi}{x}\right)}{x^2} dx = \begin{cases} 0 & \text{when } m \neq n, \\ \frac{1}{2} & \text{when } m = n. \end{cases}$$

Note that when $m = n$, we have

$$\int_{1/2}^1 \frac{\sin^2\left(\frac{n\pi}{x}\right)}{x^2} dx = - \int_2^1 \sin^2(n\pi u) du = \int_1^2 \frac{1 - \cos 2n\pi u}{2} du = \frac{1}{2}.$$

□

Example 34.52 (UECM3003 Final Exam, Apr 2013, Q2(b)). Find the positive eigenvalues and the corresponding eigenfunctions of the Sturm-Liouville problem

$$x^2 y'' + xy' + \lambda y = 0, \quad y(1) = y(5) = 0.$$

Solution. This SL problem is regular in the range $1 < x < 5$. This is Euler type equation and has a solution of the form $y(x) = u(\ln x)$. So $y' = \frac{u'}{x}$, $y'' = \frac{u'' - u'}{x^2}$ and the ODE becomes

$$u'' - u' + u' + \lambda u = u'' + \lambda u = 0 \Rightarrow u(x) = A \cos(\sqrt{\lambda} x) + B \sin(\sqrt{\lambda} x)$$

Therefore,

$$y(x) = A \cos(\sqrt{\lambda} \ln x) + B \sin(\sqrt{\lambda} \ln x)$$

From the BCs, we obtain the eigenvalues $\begin{cases} y(1) = A \cos 0 + B \sin 0 = 0 \Rightarrow A = 0 \\ y(5) = B \sin(\sqrt{\lambda} \ln 5) = 0 \Rightarrow \lambda = \left(\frac{n\pi}{\ln 5}\right)^2 \end{cases}$

where $n = 1, 2, \dots$ and the corresponding eigenfunctions $y_n(x) = B_n \sin \frac{n\pi \ln x}{\ln 5}$. □

The following is a periodic SL problem example.

Example 34.53. Write down the eigenvalues and eigenfunctions for the periodic SL problem below

$$y'' + \lambda y = 0, \quad 0 < x < 2\pi, \quad y(0) = y(2\pi).$$

Solution: The general solution is $y(x) = Ae^{i\sqrt{\lambda}x} + Be^{-i\sqrt{\lambda}x}$ and the eigenvalues and eigenfunctions are $\lambda_n = n^2$, $y_n(x) = e^{inx}$, $n = 0, \pm 1, \pm 2, \dots$.

The following are some examples of singular Sturm-Liouville problems.

eg:bessel

Example 34.54. Consider the SL problem of a (parametric) Bessel equation:

$$x^2 y'' + xy' + (\lambda x^2 - \nu^2)y = 0 \quad \text{or} \quad [xy']' + \left(\lambda^2 x - \frac{\nu^2}{x}\right)y = 0, \quad 0 < x < L. \quad (34.12)$$

e:bessel

- (a) Explain why is the above SL problem (36.23) is singular.

Solution: It is singular at $x = 0$ since $p(x) = x^2$ is zero at $x = 0$.

- (b) If y in (36.23) is required to be bounded and $y(L) = 0$, write down the eigenvalues, eigenfunctions and orthogonal relations.

Solution: The general solution is

$$y(x) = aJ_\nu(\lambda x) + bY_\nu(\lambda x), \quad J_\nu, Y_\nu \text{ are defined in Topic 3}$$

and the eigenvalues and eigenfunctions are

$$\lambda_n = \sqrt{\frac{\nu_n}{L}}, \quad y_n(x) = J_\nu(\lambda_n x), \quad n = 0, \pm 1, \pm 2, \dots$$

where ν_n are the n th-root of the Bessel function J_ν , i.e. $J_\nu(\nu_n) = 0$.

The orthogonal relations are $\int_0^L J_\nu(\lambda_k x) J_\nu(\lambda_\ell x) x dx = C_k \delta_{k\ell}$ where $C_k \neq 0$ are some constants.

Example 34.55. Use the program SLEIGN2 to find the eigenvalues $\lambda_1, \dots, \lambda_5$ of the SL problem of a (parametric) Bessel equation:

$$y'' + \frac{1}{x}y' + \left(\lambda - \frac{9}{16x^2}\right)y = 0, \quad 0 < x < 1, \quad y(1) = 0. \quad (34.13)$$

e:bessel2

Solution: By default, SLEIGN2 can only solve the normalised Bessel equation ($\nu = \frac{3}{4}$) below:

$$u'' + \left(\lambda - \frac{\nu^2 - \frac{1}{4}}{x^2}\right)u = 0, \quad 0 < x < 1, \quad u(1) = 0.$$

By using the relation $y(x) = x^{-1/2}u(x)$ (see Titchmarsh [1962, 24]), we can normalise (34.13). According to the classification of boundary conditions in Titchmarsh [1962], when $0 \leq \nu < 1$ and $\nu \neq 0.5$, the boundary condition at $x = 0$ is of limit cycle non-oscillatory. Therefore, we need to pass the following parameters to xample.x in the file auto.in:

```
1 output: bessel.rep
2 np: 2
3 param: 0.75
4 a: 0.0
5 b: 1.0
6 classa: lcno
7 classb: r
8 bca: 1.0,0.0
9 bcb: d
10 numeig: 1,5
11 end:
```

which will generate the following eigenvalues $\lambda_1, \dots, \lambda_5$ (with the error estimate):

```
1 NUMEIG,EIG,TOL = 1 44.2575531 2.05342971E-07
2 IFLAG = 1
3 NUMEIG,EIG,TOL = 2 96.0716095 1.95415080E-08
4 IFLAG = 2
5 NUMEIG,EIG,TOL = 3 167.625641 4.61892768E-07
6 IFLAG = 2
```

7	NUMEIG,EIG,TOL =	4	258.919250	4.25446160E-07
8	IFLAG =	1		
9	NUMEIG,EIG,TOL =	5	369.952820	1.09298428E-08

Note: There is a Matlab package MATSLISE for solving Sturm-Liouville and Schrödinger equations.

Example 34.56. Consider the singular SL problem of a spherical Bessel equation:

$$x^2 y'' + 2xy' + (\lambda x^2 - n(n+1))y = 0, \quad 0 < x < \infty.$$

Find the eigen-solutions of this problem if y is bounded and $y(L) = 0$.

Solution: Note that it is singular at $x = 0$ since $p(x) = x^2$ is zero at $x = 0$. The general solution of the SL problem is

$$y(x) = A j_n(\sqrt{\lambda}x) + B y_n(\sqrt{\lambda}x), \quad j_n, y_n \text{ are defined in Topic 4}$$

and the eigenvalues and eigenfunctions are

$$\lambda_n = \sqrt{\frac{\alpha_n}{L}}, \quad y_n(x) = j_\mu(\lambda_n x), \quad n = 1, 2, \dots$$

where α_n is the n th-root of j_μ , i.e. $j_\mu(\alpha_n) = 0$.

Example 34.57. Consider the singular SL problem of the Legendre equation

$$(1 - x^2)y'' - 2xy' + \lambda y = 0 \quad \text{or} \quad [(1 - x^2)y']' + \lambda y = 0, \quad -1 < x < 1.$$

Find the eigen-solutions of this problem if $y(\pm 1)$ are bounded.

Solution: Note that $p(x) = 1 - x^2$ and since $p(\pm 1) = 0$, hence the ODE is not regular. The general solution of the ODE is obtained using the Frobenius method (UECM2023 Topic 4 Series Solution):

$$y(x) = A \left(1 - \frac{\lambda}{2!}x^2 - \frac{\lambda(6-\lambda)}{4!}x^4 - \frac{\lambda(6-\lambda)(20-\lambda)}{6!}x^6 + \dots \right) \\ + B \left(x + \frac{2-\lambda}{3!}x^3 + \frac{(2-\lambda)(12-\lambda)}{5!}x^5 - \frac{(2-\lambda)(12-\lambda)(30-\lambda)}{7!}x^7 + \dots \right)$$

and the eigenvalues and eigenfunctions are

$$\lambda_n = n(n+1), \quad y_n(x) = P_n(x), \quad n = 0, 1, 2, \dots$$

where P_n is defined in Topic 4.

34.5 Hilbert Space and Generalised Fourier Series

With the above knowledge, it is time to think how to summarise them. Fourier originally thought that all functions can be expressed as Fourier series. However, a function f that can be expressed as Fourier series (34.4) must at least be “integrable” because the Fourier coefficients (34.6), (34.7) and (34.8) are found using integration!

So what happens if we put all integrable functions together to form a set? Does a set of functions have anything nice?

Well, we can introduce addition and scalar multiplication to the set of functions to turn it into a vector space (there is a group of mathematicians who study function spaces, and they like to use the term “linear space” instead of vector space, for example Zeidler [1995] and Atkinson and Han [2009]).

Definition 34.35. A *vector space* or *linear space* over $\mathbb{F}$ (which is either $\mathbb{R}$ or $\mathbb{C}$) is a set X together with an addition

$$\forall \mathbf{u}, \mathbf{v} \in X, \quad \mathbf{u} + \mathbf{v} \in X$$

and a *scalar multiplication*

$$\forall k \in \mathbb{F}, \forall \mathbf{u} \in X, \quad k\mathbf{u} \in X$$

where all the vector space rules are satisfied (refer to UECM1313 Fundamental of Linear Algebra).

Let $\Omega \subset \mathbb{R}^n$ be an **open set**. In the later topics, we will sometimes use function spaces of continuous functions.

Definition 34.36. $C(\Omega)$ is the set of real-valued functions which are continuous on Ω . For any non-negative integer k , $C^k(\Omega)$ is the set of all functions whose derivatives of order up to k are continuous on Ω . In particular, $C^0(\Omega) = C(\Omega)$.

Note that $C^k(\Omega)$ are vector spaces. Next, we can introduce the notion of “length” to functions.

Definition 34.37. A *norm* of a vector space X is a mapping

$$\|\cdot\| : X \rightarrow \mathbb{R}, \quad \mathbf{x} \mapsto \|\mathbf{x}\|$$

such that for all $\mathbf{u}, \mathbf{v} \in X$ and $k \in \mathbb{F}$, the following statements are true:

$$\|\mathbf{u}\| \geq 0; \quad \|\mathbf{u}\| = 0 \Leftrightarrow \mathbf{u} = \mathbf{0}; \quad \|k\mathbf{u}\| = |k|\|\mathbf{u}\|; \quad \|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|.$$

A *normed space* is a vector space with a given norm. A normed space which is “complete” is called a *Banach space*. The notion of completeness (which says that all the Cauchy sequence in the normed space has a limit) should have been introduced in UECM2033 Elementary Real Analysis.

Now we can define the concept of “distance” using the concept of norm (in lay man term, length).

Definition 34.38. The *distance* of two vectors $\mathbf{u}, \mathbf{v}$ of a normed space is defined as $\|\mathbf{u} - \mathbf{v}\|$.

Let $1 \leq p < \infty$. The most important Banach spaces are the L^p spaces. Its theory is founded in Lebesgue integral and real analysis the subject that discusses Lebesgue integral.

Definition 34.39.

$$L^p(\Omega, \mu) := \{f : \Omega \rightarrow \mathbb{F} \mid \left[\int_{\Omega} |f(\mathbf{x})|^p d\mu \right]^{1/p} < \infty\}.$$

When $d\mu = d\mathbf{x}$, we denote the L^p space as $L^p(\Omega)$. When $d\mu = w(\mathbf{x})d\mathbf{x}$, where $w(\mathbf{x})$ is a weight function, we denote the L^p space $L_w^p(\Omega)$.

Theorem 34.58. $\|f\|_{p,\mu} := \left[\int_{\Omega} |f(\mathbf{x})|^p d\mu \right]^{1/p}$ is a norm of the Banach space $L^p(\Omega, \mu)$.

Not all normed spaces can have “angles”. Those that allow the notion of angles belong to the following class.

Definition 34.40. Let X be a vector space over $\mathbb{F}$. An *inner product* on X assigns to each pair $\mathbf{u}, \mathbf{v} \in X$ a number

$$\langle \mathbf{u}, \mathbf{v} \rangle \quad \text{or} \quad \langle \mathbf{u} | \mathbf{v} \rangle \in \mathbb{F}$$

such that the following hold for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in X$ and $k_1, k_2 \in \mathbb{F}$:

1. $\langle \mathbf{u} | \mathbf{u} \rangle \geq 0$ and $\langle \mathbf{u} | \mathbf{u} \rangle = 0$ iff $\mathbf{u} = \mathbf{0}$;
2. $\langle \mathbf{u} | k_1 \mathbf{v} + k_2 \mathbf{w} \rangle = k_1 \langle \mathbf{u} | \mathbf{v} \rangle + k_2 \langle \mathbf{u} | \mathbf{w} \rangle$;
3. $\langle \mathbf{u} | \mathbf{v} \rangle = \overline{\langle \mathbf{v} | \mathbf{u} \rangle}$.

A vector space with an inner product is called an *inner product space* and a complete inner product space is called a *Hilbert space*. The norm of the inner product space is given by $\|\mathbf{x}\| := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$.

Remark 34.59. Physicists prefer to use $\langle \mathbf{u} | \mathbf{v} \rangle$ while mathematicians normally use $\langle \mathbf{u}, \mathbf{v} \rangle$. Some mathematicians also use $(\mathbf{u}, \mathbf{v})$ for inner product, but it is sometimes confusing.

Theorem 34.60. $\mathbb{R}^n$ together with the dot product forms a Hilbert space.

Theorem 34.61. $L_w^2(\Omega)$ together with $(f, g) \mapsto \int_{\Omega} f(\mathbf{x})g(\mathbf{x})w(\mathbf{x})d\mathbf{x}$ forms a Hilbert space.

Definition 34.41. Let X be an inner product space. For $\mathbf{u}, \mathbf{v} \in X$, $\mathbf{u}$ and $\mathbf{v}$ are said to be *orthogonal* if $\langle \mathbf{u} | \mathbf{v} \rangle = 0$. A set $S \subset X$ is called an *orthogonal set* if for any two different elements $\mathbf{u}, \mathbf{v} \in S$ are orthogonal. An orthogonal set S is said to be *complete* if there is no non-zero element in X which is orthogonal to every $\mathbf{v} \in S$. A complete orthogonal set S is said to be a *complete orthonormal system* if $\langle \mathbf{u} | \mathbf{u} \rangle = 1$ for every $\mathbf{u} \in S$.

By using Zorn lemma (from set theory), mathematicians show that for any inner product space of more than one element has a complete orthogonal set.

Now can define the generalised Fourier series for Hilbert space.

Definition 34.42. Let X be a Hilbert space and $\mathbf{x} \in X$. For a given complete orthonormal system $S := \{\mathbf{e}_n \mid n \in \mathbb{Z}\}$ of X , the *generalised Fourier series* of X is

$$\mathbf{x} = \sum_{n \in \mathbb{Z}} \langle \mathbf{x} | \mathbf{e}_n \rangle \mathbf{e}_n$$

and the numbers $\langle \mathbf{x} | \mathbf{e}_n \rangle$ are called the *Fourier coefficients* of $\mathbf{x}$ with respect to S .

In the earlier sections, we have come across various complete orthonormal system for various Hilbert spaces, such as

Hilbert space	Inner Product, $\langle f g \rangle$	Complete Orthonormal system
$L^2([-L, L])$	$\int_{-L}^L f(x)g(x)dx$ $\int_{-L}^L f(x)\overline{g(x)}dx$	$\left\{ \frac{1}{\sqrt{2L}}, \frac{1}{\sqrt{L}} \cos \frac{n\pi x}{L}, \frac{1}{\sqrt{L}} \sin \frac{n\pi x}{L} \mid n = 1, 2, \dots \right\}$ $\left\{ \frac{1}{\sqrt{2L}} \exp \frac{n\pi x}{L} \mid n = 0, \pm 1, \pm 2, \dots \right\}$
$L^2([\alpha, \beta], r(x)dx)$	$\int_{\alpha}^{\beta} f(x)g(x)r(x)dx$	$\left\{ \frac{1}{\sqrt{\int_{\alpha}^{\beta} u_n(x)r(x)dx}} u_n(x) \mid n = 0, 1, 2, \dots \right\}$

In Topics 3 and 4, we will be applying some generalised Fourier series to approximate the solutions of partial differential equations in cylindrical and spherical coordinates.

Both the Bessel's inequality and Parseval's Theorem in the previous section are valid for generalised Fourier series.

Theorem 34.62. If X is a Hilbert space with a given complete orthonormal system $\{\mathbf{e}_n \mid n \in \mathbb{Z}\}$, then for any $\mathbf{x} \in X$, $\sum_{n=1}^N \langle \mathbf{x} | \mathbf{e}_n \rangle^2 \leq \|\mathbf{x}\|^2$ and $\sum_{n=-\infty}^{\infty} \langle \mathbf{x} | \mathbf{e}_n \rangle^2 = \|\mathbf{x}\|^2$. In particular, for the eigenfunctions $\{u_n\}$ of regular Sturm-Liouville problem (34.9) and (34.10), we have

$$\sum_{n=1}^N \int_{\alpha}^{\beta} f(x)u_n(x)r(x)dx \Big/ \int_{\alpha}^{\beta} (u_n(x))^2 r(x)dx \leq \int_{\alpha}^{\beta} (f(x))^2 r(x)dx \quad \text{Bessel's Inequality}$$

$$\sum_{k=n}^{\infty} \int_{\alpha}^{\beta} f(x)u_n(x)r(x)dx \Big/ \int_{\alpha}^{\beta} (u_n(x))^2 r(x)dx = \int_{\alpha}^{\beta} (f(x))^2 r(x)dx. \quad \text{Parseval's Theorem}$$

Useful Programs

fft.py

```

1 def plotfn2fft(fn, a, b, nterms, N=2**11):
2     #from scipy.fftpack import fft, ifft
3     #http://stackoverflow.com/questions/32161734/slightly-different-fft-results-from-matlab-fft-and-
4     from pyfftw.interfaces.scipy_fftpack import fft, ifft

```

```

5     from numpy import arange,abs,pi,floor,where,concatenate
6     import matplotlib.pyplot as plt
7     from matplotlib.pyplot import xlabel, ylabel
8
9     if b <= a: return None
10    if N > 2**16:
11        print("N = {N} is too large:".format(N=N))
12        return None
13
14    # N is sample
15    T = float(b-a)
16    t = arange(a,b,T/N)
17    x = fn(t)
18    ### Plot time signal
19    plt.subplot(2,2,1)
20    plt.plot(t,x); xlabel('$t$'); ylabel('$x$')
21
22    # To use FFT, function must be defined in the range [0,T]
23    # Using pivot to shift!
24    pivot = (floor(a/T)+1)*T
25    x = concatenate((x[where(t>=pivot)],x[where(t<pivot)]))
26    xhat = fft(x,N)/N
27    #y = stop*ifft(xhat) # ifft = Inverse FFT
28    #plt.plot(t,x,t,y); xlabel('$t$'); ylabel('$x$')
29    half = N/2
30    freqs = arange(0,half)/T
31    mag = 2.*abs(xhat[:half])
32    mag[0] /= 2.
33    plt.subplot(2,2,2)
34    plt.stem(freqs[:half], mag); xlabel('$f$'); ylabel('$2|c_n|$')
35    # http://www.chem.mtu.edu/~tbco/cm416/fft1.pdf
36    an = 2*(xhat[:half].real)
37    an[0] /= 2.0
38    bn = -2*(xhat[:half].imag)
39    plt.subplot(2,2,3)
40    xl = '$f=%f$ Hz' % (1/T)
41    plt.stem(arange(nterms), an[:nterms]); xlabel(xl); ylabel('$a_n$')
42    plt.axis([-1,nterms,min(an[:nterms]),max(an[:nterms])])
43    plt.subplot(2,2,4)
44    plt.stem(arange(nterms), bn[:nterms]); xlabel(xl); ylabel('$b_n$')
45    #plt.savefig("fft.eps")
46    plt.show()
47
48    if __name__=='__main__':
49        from numpy import sin, cos, exp, pi
50
51        ### FFT is doing fine with the following functions with n=2**11
52        #def f(t): return 5 - 2*cos(t) + 8*sin(3*t)
53        #def f(x): return exp(x)
54        #def f(x): return x + pi
55
56        ### FFT is not accurate enough with the following function
57        def f(x):
58            from numpy import cos, sin
59            return cos(x)**5*sin(x)
60        plotfn2fft(f, -pi, pi, nterms=10,N=2**14)

```

fig/pde/gibbs.plt

```

1  sqr(x) = (x<-0.5)?1.0:(x<0.0)?-1.0:(x<0.5)?1.0:-1.0
2  fc(x,n)= sin(2*pi*n*x)/n
3  S10(x) = 4.0/pi*(fc(x,1.0)+fc(x,3.0)+fc(x,5.0)+fc(x,7.0)+fc(x,9.0))
4  C20(x) = 4.0/pi*(fc(x,11.0)+fc(x,13.0)+fc(x,15.0)+fc(x,17.0)+fc(x,19.0))
5  C30(x) = 4.0/pi*(fc(x,21.0)+fc(x,23.0)+fc(x,25.0)+fc(x,27.0)+fc(x,29.0))
6  S30(x) = S10(x) + C20(x) + C30(x)
7  C40(x) = 4.0/pi*(fc(x,31.0)+fc(x,33.0)+fc(x,35.0)+fc(x,37.0)+fc(x,39.0))
8  C50(x) = 4.0/pi*(fc(x,41.0)+fc(x,43.0)+fc(x,45.0)+fc(x,47.0)+fc(x,49.0))

```

```

9  S50(x) = S30(x) + C40(x) + C50(x)
10
11  set samples 1000
12  set border linewidth 0.2
13  set terminal postscript eps linewidth 5 enhanced
14
15  set output "gibbs1.eps"
16  set title "N=10"
17  plot [-1:1] sqr(x), S10(x)
18  ##pause -1
19  set output "gibbs2.eps"
20  set title "N=30"
21  plot [-1:1] sqr(x), S30(x)
22  set output "gibbs3.eps"
23  set title "N=50"
24  plot [-1:1] sqr(x), S50(x)
25
26  set output "gibbs1e.eps"
27  set title "N=10"
28  err(x) = abs(sqr(x)-S10(x))>0.8?NaN:sqr(x)-S10(x)
29  plot [-1:1] err(x)
30  set output "gibbs2e.eps"

```

Chapter 35

Boundary-Value Problems in Rectangular Coordinates

We start with first order quasi-linear PDE follow by a classification of second order linear PDEs. The second order linear PDEs that arise in the modelling of physical problems involving functions of more than one variables can be solved using the *method of separation of variables* (or *Fourier method*). This method enables us to reduce PDE of several variables to familiar ODEs. The solution of PDE can then be expanded in a series in terms of certain special functions derived from the ODEs.

35.1 First Order Quasi-Linear PDEs

First order PDEs appear in dynamics, continuum mechanics and optics. For example, in the *scalar conservation law*

$$u_t + \nabla \cdot \mathbf{F}(u) = 0 \quad (35.1)$$

e:conservation

the unknown u is the density of some physical quantity and the nonlinear vector field $\mathbf{F}$ is the flux.

Definition 35.1. A *first order quasi-linear PDE* has the form

$$\sum_{k=1}^n A_k(\mathbf{x}, u) \frac{\partial u}{\partial x_k} = f(\mathbf{x}, u), \quad (35.2)$$

e:1stOrderPDE

where $\mathbf{x} = (x_1, \dots, x_n)$ and $A_i(\mathbf{x}, u)$ and $f(\mathbf{x}, u)$ are smooth functions.

A simple example of (35.1) and (35.2) is the *transport problem* described below.

Example 35.1 (Dr Goh Yong Kheng, 2014). Imagine a long pipe containing water moving with velocity c along the length of the pipe, say let call this direction the x -direction. Let A be the cross sectional area of the pipe. A drop of dye (or any substance) is injected into the water and let $u(x, t)$ be the concentration profile of the dye spread across in the water. Assuming no diffusion occurs and water only flow in the x -direction.

Consider the amount of dye molecules that pass through a cross sectional slice on the pipe at $[x, x + \Delta x]$ per unit time. The law of mass conservation requires that the net amount of dye entering the slice volume is the same as the change of mass in the slice volume. The change of mass in the slice volume is

$$(\text{change of dye mass per unit time}) \times (\text{volume}) = \frac{\partial u}{\partial t}(x, t) \times A \Delta x,$$

while the net amount of dye entering the slice is

$$(\text{flow of dye entering the slice}) - (\text{flow of dye leaving the slice}) = cu(x, t) \times A - cu(x + \Delta x, t) \times A.$$

Putting all terms together, we have

$$\frac{\partial u}{\partial t}(x, t) A \Delta x = cA [u(x, t) - u(x + \Delta x, t)]. \quad (35.3)$$

eq:dye

Taking Δx to the right hand side and let $\Delta x \rightarrow 0$, then we have derived what is called the *transport equation* or the *advection equation*:

$$\frac{\partial u}{\partial t} = - \lim_{\Delta x \rightarrow 0} c \frac{u(x + \Delta x, t) - u(x, t)}{\Delta x} = -c \frac{\partial u}{\partial x},$$

or

$$u_t(x, t) + c u_x(x, t) = 0. \quad (35.4)$$

eq:transport

35.1.1 Change of Variables and Method of Characteristic

To tackle the transport equation (35.4), we can try to perform a change of variables and hopefully it will be converted into a familiar equation which is easy to solve. Let us introduce a new coordinate system (ξ, η) where

$$\xi = x - ct, \quad \eta = t.$$

Here, η is essentially the same time variable as the old time coordinate t . While ξ now is a new “moving” coordinate that moves with a constant speed of c .

One can imagine that the original coordinate (x, t) is the coordinate system use by a passenger on the platform of a train station, while (ξ, η) is the coordinate system by a passenger on the train that moves to the right with speed c .

Let $v(\xi, \eta)$ be the value of u observed by the passenger on the train. Note that essentially $u(x, t) = v(\xi, \eta)$, they are the same physical quantity but observed by two different observers, one stationary and one moves with speed c . In fact, we could use $u(\xi, \eta)$ to emphasize that, they are the same quantity. However, for the sake of clarity, we will use u and v in this example.

Since $u(x, t) = v(\xi, \eta)$, we have

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial v}{\partial \xi} \frac{\partial \xi}{\partial t} + \frac{\partial v}{\partial \eta} \frac{\partial \eta}{\partial t} = -c \frac{\partial v}{\partial \xi} + \frac{\partial v}{\partial \eta}, \\ \frac{\partial u}{\partial x} &= \frac{\partial v}{\partial \xi} \frac{\partial \xi}{\partial x} + \frac{\partial v}{\partial \eta} \frac{\partial \eta}{\partial x} = \frac{\partial v}{\partial \xi}. \end{aligned}$$

Substitute into the transport equation (35.4) and we will obtain

$$\boxed{\frac{\partial v}{\partial \eta} = 0} \Rightarrow v(\xi, \eta) = f(\xi),$$

where $f \in C^1$ is arbitrary. Thus, for the passenger on the train, she will observe a stationary wave that the wave profile will remains and do not change in time. To get what will be a stationary observer on the platform sees, re-write the solution back into the (x, t) coordinate, we have

$$u(x, t) = f(x - ct).$$

Theorem 35.2. Let $\phi \in C^1$. The transport equation (35.4) and the data $u(x, 0) = \phi(x)$ is well-formed with a unique solution $u(x, t) = \phi(x - ct)$.

The *method of characteristic* tries to transform the first order PDE (35.2) using the “characteristic curves” into a system of ODEs, which may sometimes be easier to solve.

Definition 35.2. A *characteristic curve* of the PDE (35.2) is a curve on the graph of u :

$$\ell : s \mapsto (\mathbf{x}(s), u(\mathbf{x}(s))) \in \text{graph}(u)$$

such that the **normal vector** $(\nabla u, -1)$ along the curve ℓ is orthogonal to the **vector field** on the **surface** $\text{graph}(u) := \{(x_1, \dots, x_n, u(x_1, \dots, x_n)) : x_i \in \mathbb{R}\}$:

$$\underbrace{(A_1(\mathbf{x}, u), \dots, A_n(\mathbf{x}, u), f(\mathbf{x}, u))}_{\text{vector field on } S} \cdot \underbrace{(\nabla u, -1)}_{\text{normal vector}} = 0. \quad (35.5)$$

e:characteristic

By using the knowledge of Calculus, we know that the derivative of u along the characteristic curve ℓ is

$$\frac{d}{ds} u(\mathbf{x}(s)) = \sum_{k=1}^n \frac{\partial u}{\partial x_k} \frac{dx_k}{ds} = \left(\frac{dx_1}{ds}, \dots, \frac{dx_n}{ds} \right) \cdot \nabla u \Rightarrow \left(\frac{dx_1}{ds}, \dots, \frac{dx_n}{ds}, \frac{d}{ds} u(\mathbf{x}(s)) \right) \cdot (\nabla u, -1) = 0$$

By comparing with (35.5), it is possible for us to related the quasi-linear PDE (35.2) with a system of ODEs below

$$\frac{dx_i}{ds} = A_i(\mathbf{x}, u), \quad \frac{du}{ds} = f(\mathbf{x}, u). \quad (35.6)$$

When $f(\mathbf{x}, u) = 0$ in (35.5), we also regard the projection of the characteristic curves onto $\mathbb{R}^n$ as the characteristic curves.

35.1.2 Constant Coefficient and Homogeneous Examples

Let us look at some constant coefficient examples to understand what we mean.

Example 35.3. Find the general solution of $4u_x - 3u_t = 0$.

Solution: To find the characteristic curves, we write down the equivalent system of ODEs:

$$\begin{cases} \frac{dx}{ds} = 4 \\ \frac{dt}{ds} = -3 \\ \frac{du}{ds} = 0 \end{cases} \Rightarrow \begin{cases} x = 4s + x_0 \\ t = -3s + t_0 \\ u = u_0 \end{cases}$$

where (t_0, x_0, u_0) is a point on the surface $S = \{(t, x, u) : u = u(x, t)\}$.

So what is the problem? We have not find the solution $u(x, t)$ and we don't really know what is S exactly. But note that $f(\mathbf{x}, u) = 0$ in this case, so we can just consider the projection of the characteristic curve, i.e.

$$(t, x) = (-3s + t_0, 4s + x_0), \quad s \in \mathbb{R}.$$

By recalling what we learned from UECM1313 Fundamental of Linear Algebra, we know that the above equation describes a line which can be re-written as

$$4t + 3x = 4t_0 + 3x_0. \quad (35.7)$$

Along this characteristic line, u as a function of s is constant, i.e.

$$u(x(s), t(s)) = u_0. \quad (35.8)$$

Since $4t_0 + 3x_0$ is also a constant, so it is relate to the constant u_0 by some unknown (but general) function f :

$$u_0 = f(4t_0 + 3x_0) \Rightarrow u(x, t) = f(4t + 3x)$$

and we obtain the general solution above.

We can single out an unique solution for (35.4) if appropriate data is given as follows.

Example 35.4. Find the solution to $4u_x - 3u_t = 0$ if $u(x, 0) = x^2$.

Solution. From Example 35.3, the general solution is $u(x, t) = f(3x + 4t)$.

Since $u(x, 0) = x^2 = f(3x)$, we have $f(x) = x^2/9$. So the solution is

$$u(x, t) = f(3x + 4t) = \frac{(3x + 4t)^2}{9}.$$

□

Example 35.5. Find the solution to $4u_x - 3u_t = 0$ if $u(x, t) = \sin(t)$ when $x = t$.

Solution: Along $x = t$, we have

$$u(x, t) = u(t, t) = f(3t + 4t) = f(7t) = \sin(t)$$

thus $f(t) = \sin(t/7)$ and the solution

$$u(x, t) = f(3x + 4t) = \sin\left(\frac{3}{7}x + \frac{4}{7}t\right).$$

However, not all curves are suitable as “initial condition”. For example, assuming that u is $\sin(t)$ along the curve $x = 1 - 4/3t$. We have $u(1 - 4/3t, t) = f(1) = \sin(t)$ which is obviously contradicting because $f(1)$ is constant while $\sin$ is not!

The difference between the first two curves and the last is that the last curve $x = 1 - 4/3t$ is a characteristic curve, but the others are not.

Since u is known to be constant on the characteristic curves, this means that every point on the characteristics is providing the same information. Thus, for single out a unique solution from the collection of possible solutions, *we need to provide information on u other than the characteristics!*

Definition 35.3. The problem of finding a solution of a linear PDE (35.2) taking on prescribed values on a given curve is called the *Cauchy problem*, and the information given on the curve is called *Cauchy data*.

The examples above seem to suggest that we can expect a unique solution of a Cauchy problem, as long as the curve which data is specified is not a characteristic curve.

35.1.3 Variable Coefficient or Non-Homogeneous Examples

Now, we will see that the method of characteristic also works with variable coefficient first-order PDE if we can solve the corresponding ODE.

Example 35.6. Solve the following Cauchy problem:

$$\begin{aligned} u_t + x u_x &= 0, \\ u(x, 0) &= x^2. \end{aligned}$$

Solution: The corresponding system of ODE is (note that $'$ means the derivative with respect to variable s):

$$\begin{cases} t' = 1, \\ x' = x, \\ u' = 0 \end{cases} \Rightarrow \begin{cases} t = s + t_0, \\ x = x_0 e^s, \\ u = u_0. \end{cases}$$

Without the loss of generalisation, we will let the “time variable” $t_0 = 0$. Hence,

$$x_0 = x e^{-s} = x e^{-t}.$$

Because the constant u_0 is x_0 is related by some function f , we obtain the general solution

$$u_0 = f(x_0) \Rightarrow \boxed{u(x, t) = f(x e^{-t})}.$$

Given the along $t = 0$, $u(x, 0) = f(x e^0) = x^2$, thus $f(x) = x^2$. The specific solution is

$$u(x, t) = x^2 e^{-2t}.$$

Example 35.7 (UECM3003 Final Exam, Apr 2014, Q2(a)). Consider the following first order linear PDE:

$$2 \frac{\partial u}{\partial x} + y \frac{\partial u}{\partial y} = 0.$$

- (i) Write down the characteristic curve for the PDE. (4 marks)

Solution. The corresponding system of ODEs is

$$\begin{cases} x' = 2 \\ y' = y \\ u' = 0 \end{cases} \Rightarrow \begin{cases} x = 2s + x_0 \\ y = y_0 e^s \\ u = u_0 \end{cases} \Rightarrow y_0 = y e^{-s} \Rightarrow \boxed{y e^{-x/2} = \text{constant}, y_0 e^{-x_0/2}}.$$

□

- (ii) Given the value of u is $\frac{y}{x^2 + y^2}$ along the curve $y = 5e^{x/2}$. Is the problem of finding a solution to the PDE well-posed? Explain. (3 marks)

Solution. No! It is not well-posed because $y = 5e^{x/2} \Rightarrow ye^{-x/2} = 5$ is a characteristic curve. Along this curve, u is a constant but

$$\frac{y}{x^2 + y^2} = \frac{5e^{x/2}}{x^2 + 25e^x}$$

is not a constant! □

- (iii) Given that the value of u is $\ln y$ along $x = 0$, find a specific solution to the PDE. (5 marks)

Solution. The general solution is $u(x, y) = f(ye^{-x/2})$ and so the solution can be obtained:

$$f(y) = u(0, y) = \ln y \Rightarrow u(x, y) = \ln(ye^{-x/2}) = \ln y - \frac{x}{2}.$$

□

Example 35.8 (UECM3003 Final Exam, Apr 2015, Q3(a)). By using an appropriate change of variable, find a general solution to the following first order PDE

$$u_x - xu_y = 4. \quad (6 \text{ marks})$$

Solution. The corresponding system of ODEs is

$$\begin{cases} x' = 1 \\ y' = -x \\ u' = 4 \end{cases} \Rightarrow \begin{cases} \frac{dy}{dx} = \frac{y'}{x'} = -x \\ \frac{du}{dx} = \frac{u'}{x'} = 4 \end{cases} \Rightarrow \begin{cases} y = -\frac{x^2}{2} + C_1 \\ u = 4x + C_2 \end{cases}$$

where C_1 and C_2 are some constants related by some arbitrary function f . Hence, we can obtain the general solution

$$C_2 = f(C_1) \Rightarrow u(x, y) = 4x + f\left(y + \frac{x^2}{2}\right).$$

□

Finally, we will look at a more complicated example, which hopes to remind us that transforming to ODE may not be helpful if the ODE is impossible to solve.

Example 35.9. Solve the PDE

$$y(u^2 - x^2 - y^2) \frac{\partial u}{\partial x} + xu \frac{\partial u}{\partial y} = xy$$

subject to BC $u|_{x=0} = 2y$.

Solution: The PDE is equivalent to the system of ODE:

$$x' = y(u^2 - x^2 - y^2), \quad (35.9)$$

$$y' = xu, \quad (35.10)$$

$$u' = xy. \quad (35.11)$$

From (35.10) and (35.11) we have the following relationship which can be solved using the *separation of variables* (UECM2023 ODE):

$$\frac{du}{dy} = \frac{u'}{y'} = \frac{xy}{xu} = \frac{y}{u} \Rightarrow u^2 - y^2 = C_1. \quad (\dagger)$$

Hence, from (35.9) and (35.10) we have

$$\frac{dy}{dx} = \frac{y'}{x'} = \frac{xu}{y(u^2 - x^2 - y^2)} = \frac{\pm x \sqrt{C_1 + y^2}}{y(C_1 - x^2)} \Rightarrow \pm \frac{y dy}{\sqrt{C_1 + y^2}} = \frac{x dx}{C_1 - x^2}$$

or

$$\pm \sqrt{C_1 + y^2} = -\frac{1}{2} \ln |C_1 - x^2| + C_2 \xRightarrow{(\dagger)} \mp |u| + C_2 = \ln \sqrt{|u^2 - x^2 - y^2|}$$

e:2a

e:2b

e:2c

Since both C_1 and C_2 are constants and they are related by a general function f :

$$\ln \sqrt{|u^2 - x^2 - y^2|} \pm |u| = C_2 = f(C_1) = f(u^2 - y^2).$$

From the initial condition, we have

$$\ln \sqrt{|(2y)^2 - y^2|} \pm |2y| = \ln \sqrt{|3y^2|} \pm 2|y| = f((2y)^2 - y^2) = f(3y^2).$$

Hence

$$f(z) = \ln \sqrt{|z|} \pm \frac{2}{\sqrt{3}} \sqrt{|z|}$$

and the solution is

$$\ln \sqrt{|u^2 - x^2 - y^2|} \pm |u| = \ln \sqrt{|u^2 - y^2|} \pm \frac{2}{\sqrt{3}} \sqrt{|u^2 - y^2|}.$$

35.2 Second Order Linear PDEs

Second-order PDEs model a significantly wider variety of physical phenomena than do first-order equations. This is why this subject focuses on the solution of second order linear PDEs. PDEs of order greater than two are much less common. Generally speaking, higher order PDEs do not represent fundamental physical laws, but are rather derived from such. Beam equation and KdV equation in Topic 1 are such examples.

In this section, we will first classify the “constant coefficient” second order PDEs follow by the formulation of the heat equations, Laplace equations and wave equations in rectangular coordinates.

35.2.1 Hadamard’s Classification

An n -variable second order constant coefficient linear PDE has the form:

$$\sum_{i,j=1}^n a_{ij} u_{x_i x_j} + \sum_{i=1}^n b_i u_{x_i} + cu = d. \quad (35.12)$$

e:pde:lin:o2

Theoretical physicists will normally express the second order linear PDE in the Einstein summation form

$$a^{\alpha\beta} \partial_\alpha \partial_\beta u + b^\alpha \partial_\alpha u + cu = d.$$

The *principal part* of (35.12) can be written as

$$P(\alpha) = \alpha^T A \alpha, \quad A = (a_{ij}), \quad a_{ij} = a_{ji} \quad (\text{physicists will express it as } a^{\mu\nu} = a^{\nu\mu})$$

In linear algebra (UECM1303), we learn that real, symmetric $n \times n$ matrix has n real eigenvalues. Therefore, a real symmetric matrix A has P positive eigenvalues, Z zero eigenvalues and $n - P - Z$ negative eigenvalues.

Theorem 35.10 (Classification of Second Order Constant-Coefficient PDEs). *There is a linear change of variables such that when the coefficients are constant and $d = 0$, PDE (35.12) can be transformed into the following canonical forms [Chateau and Zachmann, 1986]:*

- $u_{\eta_1 \eta_1} - \sum_{i=2}^n u_{\eta_i \eta_i} = 0$ if $Z = 0, P = 1$ or $Z = 0, P = n - 1$ (hyperbolic).
- $\sum_{i=1}^{n-Z} \pm u_{\eta_i \eta_i} = 0$ if $Z > 0$ (equivalently, $\det A = 0$) (parabolic).
- $\sum_{i=1}^n u_{\eta_i \eta_i} = 0$ if $Z = 0, P = n$ or $Z = 0, P = 0$ (elliptic).
- $\sum_{i=1}^P u_{\eta_i \eta_i} - \sum_{i=P+1}^n u_{\eta_i \eta_i} = 0$ if $Z = 0, 1 < P < n - 1$ (ultrahyperbolic).

Example 35.11. Laplace equation is elliptic, heat equation is parabolic and wave equation is hyperbolic.

Corollary 35.12. A two-variable second order linear PDE has the form:

$$au_{xx} + bu_{xy} + cu_{yy} + du_x + eu_y + fu = g \quad (35.13) \quad \text{e:2.1}$$

where a to g are functions of variables x and y .

Associated with (35.13) is the polynomial

$$P(\alpha, \beta) := a\alpha^2 + b\alpha\beta + c\beta^2 + d\alpha + e\beta + f.$$

$P(\alpha, \beta)$ and (hence) PDE (35.13) are classified into the following types using the discriminant:

- hyperbolic: $b^2 - 4ac > 0$;
- parabolic: $b^2 - 4ac = 0$;
- elliptic: $b^2 - 4ac < 0$.

By a suitable transformation of the independent variables, the linear PDE (35.13) can be reduced to one of the following three canonical forms:

- Hyperbolic: $u_{\xi\eta} + \Phi(\xi, \eta, u, u_\xi, u_\eta) = 0$;
- Parabolic: $u_{\xi\xi} + \Phi(\xi, \eta, u, u_\xi, u_\eta) = 0$;
- Elliptic: $u_{\xi\xi} + u_{\eta\eta} + \Phi(\xi, \eta, u, u_\xi, u_\eta) = 0$.

The method of characteristic can be extended to second order PDE to find a suitable transformation when the roots of the auxilliary equation below are all real:

$$ar^2 + br + c = 0$$

where the characteristic lines are given by $\frac{dy}{dx} = -r_i$, $i = 1, 2$.

Example 35.13 (Dr Goh Yong Kheng, 2014). Solve the Cauchy problem for $6u_{xx} + 5u_{xy} + u_{yy} = 0$, with the Cauchy data $u = \cos x + \sin x$ and $u_y = \cos x + 2\sin x$ along the line $y = 0$.

Solution: First, we need to find the roots r_1 and r_2 of the auxilliary equation

$$6r^2 + 5r + 1 = (3r + 1)(2r + 1) = 0 \Rightarrow r_1 = -\frac{1}{3}, r_2 = -\frac{1}{2}.$$

The first characteristic equation governs the following characteristic curves:

$$\frac{dy}{dx} = -r_1 = \frac{1}{3} \Rightarrow 3y - x = k_1.$$

The second characteristic equation governs the following characteristic curves:

$$\frac{dy}{dx} = -r_2 = \frac{1}{2} \Rightarrow 2y - x = k_2.$$

The new coordinate is $\xi = 3y - x$, $\eta = 2y - x$. Apply the chain rule and express the PDE $6u_{xx} + 5u_{xy} + u_{yy} = 0$ in terms of $u(\xi, \eta)$, we have the canonical form

$$u_{\xi\eta} = 0.$$

It is easy to solve it:

$$u(\xi, \eta) = f(\xi) + g(\eta) \Rightarrow u(x, y) = f(3y - x) + g(2y - x), \quad f, g \in C^2.$$

From the initial conditions

$$\begin{aligned} u(x, 0) &= f(-x) + g(-x) = \cos x + \sin x \Rightarrow -f'(-x) - g'(-x) = -\sin x + \cos x, \\ u_y(x, 0) &= 3f'(-x) + 2g'(-x) = \cos x + 2\sin x, \end{aligned}$$

we can obtain

$$\begin{cases} f'(-x) = 3\cos x, \\ g'(-x) = \sin x - 4\cos x \end{cases} \Rightarrow \begin{cases} f'(x) = 3\cos x, \\ g'(x) = -\sin x - 4\cos x \end{cases} \Rightarrow \begin{cases} f(x) = 3\sin x, \\ g(x) = \cos x - 4\sin x \end{cases}$$

and

$$u(x, y) = 3 \sin(3y - x) + \cos(2y - x) - 4 \sin(2y - x).$$

Example 35.14. Find the general solution of $u_{xx} + 4u_{xy} + 4u_{yy} = 0$.

Solution: First, we need to determine the root of the auxilliary equation

$$r^2 + 4r + 4 = 0 \Rightarrow r = -2.$$

The characteristic line is given by

$$\frac{dy}{dx} = -r = 2 \Rightarrow y - 2x = k.$$

The new coordinate is $\boxed{\eta = y - 2x}$, $\boxed{\xi = x}$. With this transform, we obtain the canonical form:

$$u_{\xi\xi} = 0.$$

Therefore, the general solution is

$$u(\xi, \eta) = \xi f(\eta) + g(\eta) \Rightarrow \boxed{u(x, y) = xf(y - 2x) + g(y - 2x)}, \quad f, g \in C^2.$$

Example 35.15 (UCCM3003 Final Exam, Apr 2009, Q1). Consider the following partial differential equation

$$3 \frac{\partial^2 u}{\partial x^2} + 2 \frac{\partial^2 u}{\partial x \partial y} - \frac{\partial^2 u}{\partial y^2} = 0.$$

- (a) Show whether the equation is elliptic, parabolic or hyperbolic? (4 marks)

Solution. Note that $a = 3$, $b = 2$, $c = -1$ and by Corollary 35.12, cor:classification

$$b^2 - 4ac = 4 - 4(3)(-1) = 16 > 0,$$

the equation is hyperbolic. □

- (b) By changing the variables $s = x - y$ and $t = x + 3y$, show that the above equation can be reduced to $\frac{\partial^2 u}{\partial s \partial t} = 0$. (11 marks)

Solution. Since $(x, y) \leftrightarrow (s, t)$ are one-to-one continuous differentiable on $\mathbb{R}^2$, we have the following chain rules:

$$\begin{aligned} \frac{\partial u}{\partial x} &= \frac{\partial u}{\partial s} \frac{\partial s}{\partial x} + \frac{\partial u}{\partial t} \frac{\partial t}{\partial x} = \underbrace{\frac{\partial u}{\partial s} + \frac{\partial u}{\partial t}} \\ \frac{\partial u}{\partial y} &= \frac{\partial u}{\partial s} \frac{\partial s}{\partial y} + \frac{\partial u}{\partial t} \frac{\partial t}{\partial y} = \underbrace{-\frac{\partial u}{\partial s} + 3 \frac{\partial u}{\partial t}} \end{aligned}$$

which implies

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2} &= \frac{\partial^2 u}{\partial s^2} + \frac{\partial^2 u}{\partial s \partial t} + \frac{\partial^2 u}{\partial t \partial s} + \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial s^2} + 2 \frac{\partial^2 u}{\partial s \partial t} + \frac{\partial^2 u}{\partial t^2} \\ \frac{\partial^2 u}{\partial x \partial y} &= -\frac{\partial^2 u}{\partial s^2} + 3 \frac{\partial^2 u}{\partial s \partial t} - \frac{\partial^2 u}{\partial t \partial s} + 3 \frac{\partial^2 u}{\partial t^2} = -\frac{\partial^2 u}{\partial s^2} + 2 \frac{\partial^2 u}{\partial s \partial t} + 3 \frac{\partial^2 u}{\partial t^2} \\ \frac{\partial^2 u}{\partial y^2} &= -(-\frac{\partial^2 u}{\partial s^2}) - 3 \frac{\partial^2 u}{\partial s \partial t} + 3 \left[-\frac{\partial^2 u}{\partial t \partial s} + 3 \frac{\partial^2 u}{\partial t^2} \right] = \frac{\partial^2 u}{\partial s^2} - 6 \frac{\partial^2 u}{\partial s \partial t} + 9 \frac{\partial^2 u}{\partial t^2} \end{aligned}$$

Therefore,

$$0 = 3 \frac{\partial^2 u}{\partial x^2} + 2 \frac{\partial^2 u}{\partial x \partial y} - \frac{\partial^2 u}{\partial y^2} = \underbrace{\left[3 - 2 - 1 \right] \frac{\partial^2 u}{\partial s^2}}_{=0} + \underbrace{\left[6 + 4 + 6 \right] \frac{\partial^2 u}{\partial s \partial t}}_{\neq 0} + \underbrace{\left[3 + 6 - 9 \right] \frac{\partial^2 u}{\partial t^2}}_{=0}.$$

□

- (c) Solve the above partial differential equation for $x > 0$, given that when $y = 0$, $u(x, 0) = 2 \ln(x)$ and $\frac{\partial u}{\partial y}(x, 0) = \frac{2}{x}$. (10 marks)

Solution. The general solution of $\frac{\partial^2 u}{\partial s \partial t} = 0$ is

$$u(s, t) = f(s) + g(t) \Rightarrow u(x, y) = f(x - y) + g(x + 3y).$$

From

$$u(x, 0) = f(x) + g(x) = 2 \ln(x) \Rightarrow f'(x) + g'(x) = \frac{2}{x}$$

$$\frac{\partial u}{\partial y}(x, 0) = -f'(x) + 3g'(x) = \frac{2}{x}$$

we have $4g'(x) = \frac{4}{x} \Rightarrow g(x) = \ln(x)$ and $f(x) = \ln(x)$. Hence,

$$u(x, y) = \ln(x - y) + \ln(x + 3y). \quad \square$$

Example 35.16 (UCCM3003 Final Exam, Sep 2009, Q1(a)). Consider the following partial differential equation

$$25u_{xx} - 30u_{xy} + 9u_{yy} = 0.$$

- (i) Determine if the partial differential equation is elliptic, parabolic or hyperbolic. (3 marks)

Solution. Note that $a = 25$, $b = -30$, $c = 9$ and

$$b^2 - 4ac = 900 - 4(25)(9) = 0,$$

the partial differential equation is parabolic. $\square$

- (ii) Given $s = \frac{3}{5}x + y$ and $t = x$, show that the above equation can be reduced to $\frac{\partial^2 u}{\partial t^2} = 0$. Hence, find the general solution to the above equation. (10 marks)

Solution.

$$\begin{aligned} u_x &= u_s \frac{ds}{dx} + u_t \frac{dt}{dx} = \frac{3}{5}u_s + u_t \\ \Rightarrow u_{xx} &= \frac{3}{5}(u_x)_s + (u_x)_t = \frac{9}{25}u_{ss} + \frac{3}{5}u_{ts} + \frac{3}{5}u_{st} + u_{tt} \end{aligned}$$

$$u_y = u_s \frac{ds}{dy} + u_t \frac{dt}{dy} = u_s \Rightarrow \begin{cases} u_{yy} = u_{ss} \\ u_{xy} = (\frac{3}{5}u_s + u_t)_s = \frac{3}{5}u_{ss} + u_{st} \end{cases}$$

Hence,

$$25u_{xx} - 30u_{xy} + 9u_{yy} = \left(9u_{ss} + 30u_{st} + 25u_{tt}\right) - 30\left(\frac{3}{5}u_{ss} + u_{st}\right) + 9u_{ss} = 25u_{tt} = 0$$

i.e. $u_{tt} = 0$. $\square$

- (iii) Find the specific solution, given that when $x = 0$, $u(0, y) = \sin y$, and $\frac{\partial u}{\partial x}(0, y) = 2y$.

(5 marks)

Solution. The general solution is

$$u_{tt} = 0 \Rightarrow u_t = f(s) \Rightarrow u(s, t) = f(s)t + g(s) \Rightarrow u(x, y) = f\left(\frac{3}{5}x + y\right)x + g\left(\frac{3}{5}x + y\right)$$

When $x = 0$, we have

$$u(0, y) = g(y) = \sin y$$

and

$$\frac{\partial u}{\partial x}(0, y) = \frac{3}{5}f'(y) \times 0 + f(y) + \frac{3}{5}g'(y) = 2y \Rightarrow f(y) = 2y - \frac{3}{5}\cos(y).$$

The specific solution is

$$u(x, y) = \left[2\left(\frac{3}{5}x + y\right) - \frac{3}{5}\cos\left(\frac{3}{5}x + y\right)\right]x + \sin\left(\frac{3}{5}x + y\right) \quad \square$$

Example 35.17 (UCCM3003 Final Exam, Apr 2011, Q4(a)). (i) Show that the partial differential equation

tion

$$\frac{\partial^2 u}{\partial x^2} - 3 \frac{\partial^2 u}{\partial x \partial y} - 10 \frac{\partial^2 u}{\partial y^2} = 0$$

could be put into a canonical form of $u_{st} = 0$ where $s = y - 2x$ and $t = y + 5x$. (9 marks)

Solution. $u_x = u_s \frac{ds}{dx} + u_t \frac{dt}{dx} = -2u_s + 5u_t$; $u_y = u_s \frac{ds}{dy} + u_t \frac{dt}{dy} = u_s + u_t$

$$u_{xx} = (-2u_s + 5u_t)_x = -2(-2u_{ss} + 5u_{st}) + 5(-2u_{st} + 5u_{tt}) = 4u_{ss} - 20u_{st} + 25u_{tt}$$

$$u_{xy} = (-2u_s + 5u_t)_y = (-2u_{sy} + 5u_{ty}) = (-2u_{ss} + 3u_{st} + 5u_{tt})$$

$$u_{yy} = (u_s + u_t)_y = u_{sy} + u_{ty} = u_{ss} + u_{st} + u_{st} + u_{tt} = u_{ss} + 2u_{st} + u_{tt}$$

Hence,

$$u_{xx} - 3u_{xy} - 10u_{yy} = (4 + 6 - 10)u_{ss} + (-20 - 9 - 20)u_{st} + (25 - 15 - 10)u_{tt} = -49u_{st} = 0.$$

i.e. $u_{st} = 0$. □

(ii) Hence, find the particular solution that satisfies the following boundary-value conditions:

$$u(0, y) = y, \quad \frac{\partial u}{\partial x}(0, y) = \sin y. \quad (10 \text{ marks})$$

Solution. $u_{st} = 0 \Rightarrow u = f(s) + g(t) = f(y - 2x) + g(y + 5x)$.

From $u(0, y) = y$ and $u_x(0, y) = \sin y$, we have

$$\begin{cases} f(y) + g(y) = y & \Rightarrow f'(y) + g'(y) = 1 \\ -2f'(y) + 5g'(y) = \sin y \end{cases}$$

$$\begin{bmatrix} f'(y) \\ g'(y) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -2 & 5 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ \sin y \end{bmatrix} = \frac{1}{7} \begin{bmatrix} 5 & -1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \sin y \end{bmatrix} = \begin{bmatrix} (5 - \sin y)/7 \\ (2 + \sin y)/7 \end{bmatrix}.$$

Therefore $f(y) = \frac{5}{7}y + \frac{1}{7}\cos y$ and $g(y) = \frac{2}{7}y - \frac{1}{7}\cos y$.

The particular solution is

$$u(x, y) = \frac{5}{7}(y - 2x) + \frac{1}{7}\cos(y - 2x) + \frac{2}{7}(y + 5x) - \frac{1}{7}\cos(y + 5x).$$

□

Note that the type changes with position in the xy -plane unless a , b and c are constants. For example, the Tricomi equation $yu_{xx} + u_{yy} = 0$ is elliptic when $y > 0$ and is hyperbolic when $y < 0$. The following example is another example which behave differently in different region.

Example 35.18 (UCCM3003 Final Exam, Apr 2011, Q1(a)). Sketch the region in the xy -plane where the following second order partial differential equation is hyperbolic. Explain your answer.

$$y \frac{\partial^2 u}{\partial x^2} - 2 \frac{\partial^2 u}{\partial x \partial y} + x \frac{\partial^2 u}{\partial y^2} = 0. \quad (6 \text{ marks})$$

Solution. The PDE is hyperbolic if

$$(-2)^2 - 4yx > 0 \Rightarrow xy > 1 \Rightarrow \text{in first and third quadrants.}$$

□

35.2.2 Heat or Diffusion Equation

The three most fundamental second order PDEs come from relatively simple physics phenomena. We will first “formulate” the heat equation, then Laplace equation and finally, the wave equation.

The *heat equation* (or *diffusion equation*) is an important PDE which describes the variation of temperature (or concentration, etc.) in a given domain over time. In the special case of heat propagation

in an *isotropic and homogeneous medium* in the n -dimensional space, a (*homogeneous*) *heat equation* has the form (Topic 1 Definition 1.1.10):

$$u_t - k\Delta u = 0, \quad u : \mathbb{R}^n \times \mathbb{R}^+ \rightarrow \mathbb{R}. \quad (35.14) \quad \text{e:heat}$$

A general heat equation not only characterises the diffusion of heat, it also governs particle diffusion as well as the propagation of action potential in nerve cells. Although they are not diffusive in nature, some quantum mechanics problems are also governed by a mathematical analog of the heat equation. It also can be used to model some phenomena (like options) arising in finance, like the Black-Scholes option pricing model or the Ornstein-Uhlenbeck processes. The equation, and various non-linear analogues, has also been used in image analysis (e.g. level set methods). An abstract form of heat equation on manifolds provides a major approach to the *Atiyah-Singer index theorem*, and has led to much further work on heat equations in Riemannian geometry.

The solution of the heat equation (38.7) for a rectangular domain can be approximated numerically using the *Crank-Nicolson method* (UECM3033 Numerical Methods, Topic 8) and this method can be extended to many of the models with no closed form solution [Wilmott et al., 1995, §8.7].

The heat equation (38.7) can be formulated for n -dimensional domain. However, in this subject, we will only focus on the formulation which has physically meaning in a 3-D world, i.e. $n \leq 3$. Note that the Laplacian Δ defined in Topic 1 (and used in (38.7)) has the following forms depending on the “space dimension”, or the number of “space variables” to be more precise:

- 1-dimensional (1-D): $\Delta u = u_{xx}$.
- 2-dimensional (2-D): $\Delta u = u_{xx} + u_{yy}$.
- 3-dimensional (3-D): $\Delta u = u_{xx} + u_{yy} + u_{zz}$.

We will learn that the Laplacian Δ takes different forms in other coordinate systems.

When $n = 1$, the “temperature” $u = u(x, t)$ at position x and time t is characterised by the “one-dimensional (space)” heat equation:

$$u_t - ku_{xx} = 0, \quad u : \mathbb{R}^1 \times \mathbb{R}^+ \rightarrow \mathbb{R}. \quad (35.15) \quad \text{e:heat1d}$$

When $n = 2$, we have a “two-dimensional (space)” heat equation

$$u_t - k(u_{xx} + u_{yy}) = 0, \quad u : \mathbb{R}^2 \times \mathbb{R}^+ \rightarrow \mathbb{R}. \quad (35.16) \quad \text{e:heat2d}$$

When $n = 3$, we have a “three-dimensional (space)” heat equation

$$u_t - k(u_{xx} + u_{yy} + u_{zz}) = 0, \quad u : \mathbb{R}^3 \times \mathbb{R}^+ \rightarrow \mathbb{R}. \quad (35.17) \quad \text{e:heat3d}$$

Physicists normally requires the above equations to be derived from physics principles. Although it is not compulsory for mathematicians to derive the above equations, it is usefully to build some intuition about the equation based on the derivations.

Example 35.19 (Dr Goh Yong Kheng, 2014). Consider a dye dropped into the liquid in a tiny pipe, what is the diffusion of the dye to the surrounding? According to Fick’s law of diffusion, the dye particles will move from regions of higher concentration to regions of lower concentration, i.e. the particle flux is proportional to the concentration gradient u_x at that point.

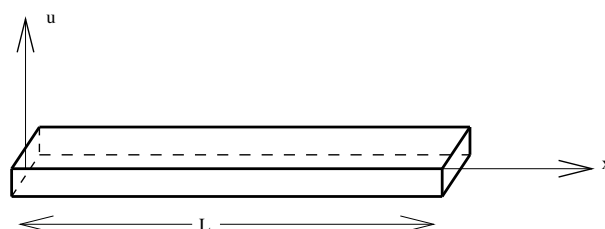
Let the proportional constant be k , i.e. flux, $\phi(x, t) = -ku_x$ then we have

$$\frac{\partial u}{\partial t}(x, t) A\Delta x = A[-k u_x(x, t) + k u_x(x + \Delta x, t)].$$

Divide Δx and let it tends to zero, we will get the diffusion equation (35.15). e:heat1d

Example 35.20 (Dr Goh Yong Kheng, 2014). A heat conduction of a rod or a bar under the following conditions is modelled by (35.15): e:heat1d

1. The rod/bar is made of a single homogeneous conducting material;
2. The rod/bar is laterally insulated (heat flows only in the x -direction);
3. The rod/bar is thin (the temperature at all points of a cross section is constant).



The derivation of the heat equation is essentially the same as the diffusion process in Example 35.19. The total heat in the volume is

$$dQ = c\rho A \Delta x u(x, t),$$

where c is the specific heat capacity and ρ is the mass density. According to the Fourier's law of heat conduction, the rate of heat flow is proportional to the negative temperature gradient,

$$q(x, t) = -K_0 \frac{\partial u}{\partial x}.$$

Together with the law of energy conservation, we have

$$\frac{\partial}{\partial t} [s(x)u(x, t)\rho(x)A\Delta x] = q(x, t) - q(x + \Delta x, t).$$

Simplify the expression and we have the heat equation

$$\frac{\partial u}{\partial t} = \frac{K_0}{s\rho} \frac{\partial^2 u}{\partial x^2}. \quad \text{eq:heat}$$

Example 35.21 (Dr Goh Yong Kheng, 2014). Consider a dye release in a large tank of liquid, what is the diffusion of the dye to the surrounding? The Fick's law of diffusion says

$$\phi(x, t) = -k\nabla u$$

and the diffusion flow is the surface integration of the dye particle flux through the normal of the volume boundary

$$\iiint_D u_t dx dy dz = \oiint_{\partial D} k\nabla u \cdot \hat{\mathbf{n}} dS.$$

Apply the divergence theorem on the surface integration

$$\oiint_{\partial D} k\nabla u \cdot \hat{\mathbf{n}} dS = \iiint_D k\nabla \cdot (\nabla u) dx dy dz \quad \text{e:thm:div} \quad (35.19)$$

and we will have the 3D diffusion equation (35.17).

Example 35.22. Let k be a material-specific quantity depending on the thermal conductivity, the density and the heat capacity. From Fourier's law of cooling (which states that the time rate of heat transfer through a material is proportional to the negative gradient in the temperature and to the area at right angles, to that gradient), i.e. the heat flux $\mathbf{q}$ is related to the temperature u by

$$\mathbf{q} = -k\nabla u$$

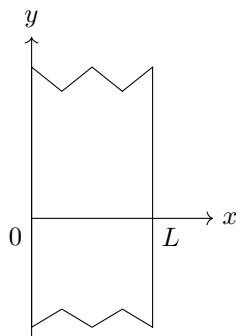
and the diffusion of heat is the surface integration of the flux through the normal of the volume boundary

$$\iiint_D u_t dx dy dz = \oiint_{\partial D} k\nabla u \cdot \hat{\mathbf{n}} dS.$$

Apply the divergence theorem on the surface integration (35.19) and we will obtain (35.17)

It is necessary to emphasise that the “dimensional of space” in the equation has nothing to do with real-world problem. For example, it is possible to approximate a 2-D problem with a “1-D (space)” heat equation.

Example 35.23.



The heat conduction on a slab under the following conditions is modelled by (35.15):

1. the slab is made of a single **homogeneous** conducting material;
2. the slab is very long in the y -direction that heat flows only in the x -direction.

35.2.3 Laplace Equation and Poisson Equation

In physics, Laplace equation arises in electrostatics modelling. In electrostatics, one of the Maxwell's equation

$$\nabla \times \mathbf{E} + \mu_0 \frac{\partial \mathbf{H}}{\partial t} = \mathbf{0}.$$

which governs the variations in “electricity and magnetism” becomes

$$\nabla \times \mathbf{E} = \mathbf{0}.$$

Since this holds *everywhere* in space it is possible to derive the electrostatic “potential” u such that

$$\mathbf{E} = -\nabla u$$

Combining with another Maxwell's equation for vacuum

$$\nabla \cdot \mathbf{E} = \frac{1}{\epsilon_0} \rho(\mathbf{x}) =: -f(\mathbf{x}),$$

we obtain the *Poisson equation* (Topic 1 Definition 1.1.10)

$$-\nabla \cdot \nabla u = -f(\mathbf{x}), \quad \text{i.e.} \quad \Delta u = f(\mathbf{x}), \quad u : \mathbb{R}^n \rightarrow \mathbb{R} \quad (35.20) \quad \text{e:poisson}$$

where ρ is the charge density and ϵ_0 is a constant, the electric permittivity of free space [Butkov, 1968, 295].

When $f \equiv 0$ in (47.4), we obtain the Laplace equation

$$\Delta u = 0, \quad u : \mathbb{R}^n \rightarrow \mathbb{R}. \quad (35.21) \quad \text{e:lapl}$$

Laplace equation (35.21) is also related to the heat equation (38.7) in a “limiting” case (related to “thermal equilibrium”) where $u_t \rightarrow 0$ as $t \rightarrow \infty$.

As mentioned above, we will limit ourselves to $n \leq 3$. In one-dimensional space, (35.21) becomes $\frac{\partial^2 u}{\partial x^2} = 0$ which has a solution $u = ax + b$ which is not very interesting. So we will only consider the “two-dimensional (space)” Laplace equation:

$$u_{xx} + u_{yy} = 0, \quad u : \mathbb{R}^2 \rightarrow \mathbb{R} \quad (35.22) \quad \text{e:rect:laplace}$$

and the “three-dimensional (space)” Laplace equation:

$$u_{xx} + u_{yy} + u_{zz} = 0, \quad u : \mathbb{R}^3 \rightarrow \mathbb{R}. \quad (35.23) \quad \text{e:rect:laplace}$$

35.2.4 Wave Equation

An n -dimensional (*homogeneous*) wave equation has the following form

$$\square u = u_{tt} - c^2 \Delta u = 0, \quad u : \mathbb{R}^n \times \mathbb{R}^+ \rightarrow \mathbb{R}. \quad (35.24) \quad \text{e:wave}$$

Note that we replace k with c^2 in Topic 1, Definition 1.1.10. When $n = 1$, we have a “1-D (space)” wave equation

$$u_{tt} - c^2 u_{xx} = 0, \quad u : \mathbb{R}^1 \times \mathbb{R}^+ \rightarrow \mathbb{R} \quad (35.25) \quad \text{e:rect:wave1d}$$

When $n = 2$, we have a “2-D (space)” wave equation

$$u_{tt} - c^2(u_{xx} + u_{yy}) = 0, \quad u : \mathbb{R}^2 \times \mathbb{R}^+ \rightarrow \mathbb{R}. \quad (35.26) \quad \text{e:rect:wave2d}$$

When $n = 3$, we have a “3-D (space)” wave equation

$$u_{tt} - c^2(u_{xx} + u_{yy} + u_{zz}) = 0, \quad u : \mathbb{R}^3 \times \mathbb{R}^+ \rightarrow \mathbb{R}. \quad (35.27) \quad \text{e:rect:wave3d}$$

Physically, the derivation of the motions of springs and strings can lead to (35.25). Consider a stretched string (e.g. a guitar string or a taut violin string) of length L with ends fastened on the x -axis at $x = 0$ and $x = L$. Suppose that the string is set to vibrate by displacing it from its equilibrium position and then releasing it. Assuming that the string vibrates only in a fixed plane, we let $u(x, t)$ denote the transverse displacement at time $t \geq 0$ of the point on the string at position x . In particular, $u(x, 0)$ denotes the initial shape of the string. We wish to determine the subsequent motion of the string by finding $u(x, t)$ for $t > 0$ and $0 < x < L$.

To derive a mathematical equation, we need to make the following simplifying assumptions that are not far from being realistic:

1. The string has a constant mass density ρ (a homogeneous string).
2. The string is perfectly elastic and offers no resistance to bending.
3. The motion of the string is transverse only (this implies the horizontal component of the tension is the same at all points). The transverse vibrations of the string are small and take place in a plane containing the x -axis, the xu -plane. Also, the slope at any point of the string, $\frac{\partial u}{\partial x}$, is small.

Under these assumptions, we will be able to obtain a 1-D wave equation (35.25) for the free vibrations of the string, where $c^2 = \frac{\tau}{\rho}$, τ has units of mass-length/time² and ρ has units of mass/length, so c^2 has units of length²/time² and c represents the *velocity* (of the wave).

The method of characteristics allows us to obtain the solution of wave equation (35.25).

Example 35.24 (UCCM3003 Final Exam, Apr 2011, Q3(a)). Let g be a continuous function of a single variable. Show that $u(x, t) = \frac{1}{2c} \int_{\alpha}^{x+ct} g(s) ds$ is a solution to the wave equation (35.25), where α and c are constants.

Solution.

$$\begin{aligned}\frac{\partial u}{\partial t} &= \frac{1}{2c} [g(x+ct) \cdot c - g(x-ct) \cdot (-c)] = \frac{1}{2} [g(x+ct) + g(x-ct)] \\ \frac{\partial^2 u}{\partial t^2} &= \frac{1}{2} [g'(x+ct) \cdot c + g'(x-ct) \cdot (-c)] = \frac{c}{2} [g'(x+ct) - g'(x-ct)] \\ \frac{\partial u}{\partial x} &= \frac{1}{2c} [g(x+ct) \cdot 1 - g(x-ct) \cdot 1] \\ \frac{\partial^2 u}{\partial x^2} &= \frac{1}{2c} [g'(x+ct) - g'(x-ct)].\end{aligned}$$

Hence, $u_{tt} = c^2 u_{xx}$. □

Example 35.25 (UECM3003 Final Exam, Apr 2014, Q3(a)). Let g be a continuous function of one variable. Show that

$$u(x, t) = \frac{1}{2c} \int_{\alpha}^{x+ct} g(s) ds$$

is a solution to the initial value problem of

$$\begin{aligned}\frac{\partial^2 u}{\partial t^2} &= c^2 \frac{\partial^2 u}{\partial x^2}, \\ u(x, 0) &= 0, \quad \text{and} \quad \frac{\partial u}{\partial t}(x, 0) = g(x).\end{aligned}$$

Proof: The proof of the PDE is the same as the April 2011 final exam problem (Example 35.24). Here, in addition, we need to show:

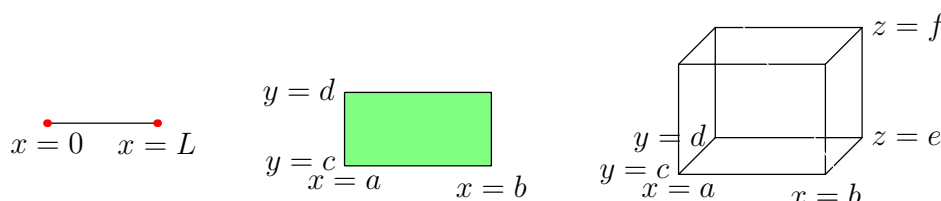
$$u(x, 0) = \frac{1}{2c} \int_{\alpha}^x g(s) ds = 0$$

and

$$\frac{\partial u}{\partial t}(x, 0) = \frac{1}{2} [g(x+c \times 0) + g(x-c \times 0)] = \frac{1}{2} (2g(x)) = g(x).$$

35.2.5 Domains, Boundary Conditions and Initial Conditions

In the following sections, we will learn that homogeneous PDEs in rectangular coordinates mentioned above can be only be solved “analytically” for “rectangular domains” of different dimensions below:



PDEs very rarely appear alone. Most scientific problems require us to solve the PDEs subject to extra “data”, i.e. appropriate *boundary conditions* as well as *initial conditions*. These data together with the PDE are normally “well-posed” (Topic 1, Definition 1.1.14), therefore, we can obtain a unique analytic solution.

We know that for $n = 1, 2$ and 3 , the rectangular domains will have 2, 4 and 6 “faces” respectively which forms the boundary. Each “face” of the domain can be imposed a *linear boundary condition* (BCs for short) and linear BCs are normally categorised as follows.

def:linear-bc

Definition 35.4. Let B_1 be a portion of the boundary of the domain D where the PDE is defined. Then

- $u|_{B_1} = g$ is called a *type 1 BC* or *Dirichlet condition*;
- $\frac{\partial u}{\partial n}|_{B_1} = g$ is called a *type 2 BC* or *Newmann condition* or *flux condition*;
- $\left(\alpha u + \beta \frac{\partial u}{\partial n}\right)|_{B_1} = g$ is called a *type 3 BC* or *Robin condition* or *radiation condition*.

def:linear-bc

Definition 35.5. The linear boundary condition in Definition 35.4 is called a *homogeneous boundary condition* if $g = 0$ on the boundary B_1 .

Definition 35.6. If **both** the linear PDE and the boundary conditions are homogeneous, the problem is said to be *homogeneous*. Otherwise, the problem is said to be *non-homogeneous*.

In the context of heat equation, each of the three BCs has the following physical meaning respectively:

- Temperature specified on the boundary;
- Heat flow across the boundary specified, the normal n is the *outward normal* direction to the boundary;
- A mix of temperature and heat flow of boundary and *surrounding medium* are specified.

What is so special about homogeneous problem? The theory of homogeneous problem in linear algebra (UECM1313) says that if $\mathbf{u}$ and $\mathbf{v}$ are the solution of a homogeneous problem $\mathbf{Ax} = \mathbf{0}$, then $a\mathbf{u} + b\mathbf{v}$ is also a solution, where a, b are arbitrary constants. Generalising this to homogeneous PDE problem, we obtain the following theorem [Weinberger, 1995, pp29–33]:

thm:superposition

Theorem 35.26 (Principle of Superposition). *If ϕ and ψ are solutions to a linear partial differential equation and satisfy a linear homogeneous boundary condition, then the linear combination $u = c_1\phi + c_2\psi$ is also a solution and satisfies the same boundary condition. Here c_1 and c_2 are arbitrary constants.*

When $n = 1$ (1-D space problem), the boundary conditions are related to physics in the following manner [Haberman, 1987, 145].

	Heat flow	Vibrating string	Terminology
$u = 0$	Fixed (zero) temperature	Fixed (zero) displacement	Type 1 BC
$u_x = 0$	Insulated	Free	Type 2 BC
$u_x = hu$	(Homogeneous) Newton's law of cooling 0° outside temperature	(Homogeneous) Elastic boundary condition	Type 3 BC
$u(-L) = u(L)$	Perfect thermal contact	–	Periodic condition
$u_x(-L) = u_x(L)$	Perfect thermal contact	–	Periodic condition
$ u(0) < \infty$	Bounded temperature	–	Singularity condition

For PDE problems with first and second order derivative with respect to time t , we need to further impose initial conditions for $u|_{t=0}$ or $\frac{\partial u}{\partial t}|_{t=0}$.

For 2-D Laplace problem, if the boundary is not rectangular but polygonal then we may need to apply *conformal mapping* (Ref: UECM3103 Complex Analysis Topic 8 and 9) to transform the original polygon to rectangles. For 3-D Laplace problem, we need to apply the variational method which is beyond the scope of this subject. In *potential theory*, the *maximum principle* implies that when n -dimensional Laplace equations (including (35.22) and (35.23), $n \geq 2$) are imposed with appropriate boundary condition (e.g. Dirichlet BC), there is at most one solution and this solution varies continuously with the boundary values.

35.3 1-D Heat Equation

We will start our study of second order PDE problems from the analysis of 1-D heat equation (35.15). Since there is a first-order differentiation with respect to t in (35.15), we need to impose an initial “temperature distribution” which is given by a function f so that we have an *initial condition*:

$$u(x, 0) = f(x), \quad x \in [0, L]. \quad (35.28)$$

A rectangular domain is just an interval $[0, L]$. The boundary of the interval consist of two points, $\{0, L\}$, called the *end points*. According to Section 35.2.5, either type 1, 2 or 3 homogeneous boundary condition can be imposed on the end points, forming $3^2 = 9$ possibilities. It is possible to solve these homogeneous boundary value problems using the *separation of variables* method, but we will restrict ourselves to only some of the possibilities. We will then explain how to generalise the separation of variables method to handle nonhomogeneous problems.

35.3.1 (Homogeneous) Type 1 BCs

If the end points are imposed type 1 BC:

$$u(0, t) = u(L, t) = 0, \quad t > 0, \quad (35.29)$$

to find a solution of (35.15) which is not identically zero and satisfies the BCs (35.29), we let

$$u(x, t) = T(t)X(x) = TX \neq 0 \quad (\text{using the abbreviations } T = T(t), X = X(x)). \quad (35.30)$$

Substituting (35.30) back into (35.15), one obtains

$$T'X = kTX'' \Rightarrow \frac{T'(t)}{kT(t)} = \frac{X''(x)}{X(x)} \quad (35.31)$$

where $T' = \frac{dT}{dt}$ and $X' = \frac{dX}{dx}$. Here $'$ is used to denote the differentiation with respect to the corresponding function's variable.

The equation (35.31) is “separated” in the sense that the left side of the equation is a function of t alone, while the right side is a function of x alone. Since variables t and x are independent of each other, the only way to get equality is to have the functions on both sides constant and equal. Let the constant be $-\lambda$ (the reason for taking a negative sign will be known below). In this case, (35.31) becomes

$$\frac{T'(t)}{kT(t)} = \frac{X''(x)}{X(x)} = -\lambda. \quad (35.32)$$

Thus we have two sets of ODE:

$$T'(t) = -\lambda kT(t) \quad (35.33)$$

and

$$X''(x) = -\lambda X(x). \quad (35.34)$$

Do not forget the type 1 boundary condition (35.29) plays an important role for us to find the solution to the original problem which composed of (35.15), (35.28) and (35.29). If we substitute (35.30) into (35.29), we will obtain

$$X(0)T(t) = X(L)T(t) = 0.$$

Since t is arbitrary and $T \neq 0$ (according to (35.30)), we must have

$$X(0) = X(L) = 0. \quad (35.35)$$

Let us now analyse the solution to the Sturm-Liouville problem (35.34) and (35.35) below.

Example 35.27. What is/are the solution(s) for (35.34) and (35.35) when $\lambda < 0$?

Solution. Since λ is a constant, we can use characteristic equation (UECM2023 ODE Topic 2) to find the general solution for (35.34):

$$\omega^2 + \lambda = 0 \Rightarrow \omega = \pm\sqrt{-\lambda} \Rightarrow X(x) = Be^{\sqrt{-\lambda}x} + Ce^{-\sqrt{-\lambda}x} \quad (35.36)$$

From (35.35), we obtain

$$\begin{aligned} X(0) &= B + C = 0 \\ X(L) &= Be^{\sqrt{-\lambda}L} + Ce^{-\sqrt{-\lambda}L} = 0 \end{aligned}$$

Since $\sqrt{-\lambda}L \neq 0$,

$$\begin{vmatrix} 1 & 1 \\ e^{\sqrt{-\lambda}L} & e^{-\sqrt{-\lambda}L} \end{vmatrix} = e^{-\sqrt{-\lambda}L} - e^{\sqrt{-\lambda}L} = -2\sinh(\sqrt{-\lambda}L) \neq 0$$

The simultaneous equations above lead to $B = C = 0$ which implies $X = 0$. Therefore $u = XT = 0$ is trivial. $\square$

Example 35.28. What is/are the solution(s) for (35.34) and (35.35) when $\lambda = 0$?

Solution. When $\lambda = 0$, the solution of (35.34) is

$$X(x) = Bx + C.$$

for some constants B and C . From the boundary condition (35.29) we have

$$\begin{aligned} X(0) &= C = 0 \\ X(L) &= BL + C = 0 \Rightarrow B = 0. \end{aligned}$$

Hence, we obtain a trivial solution $X(x) = 0$. Therefore $u = 0$ is trivial. $\square$

Example 35.29. What is/are the solutions for (35.34) and (35.35) when $\lambda > 0$?

Solution. When $\lambda > 0$, from the characteristic equation, we have

$$\omega^2 + \lambda = 0 \Rightarrow \omega = \pm\sqrt{\lambda}i$$

and the general solution of (35.34) is

$$X(x) = B_\lambda \sin(\sqrt{\lambda}x) + C_\lambda \cos(\sqrt{\lambda}x)$$

where B_λ and C_λ are real constants.

From the boundary conditions (35.29), we obtain

$$\begin{aligned} X(0) &= C_\lambda = 0, \\ X(L) &= B_\lambda \sin(\sqrt{\lambda}L) + C_\lambda \cos(\sqrt{\lambda}L) = 0 \end{aligned}$$

From this simultaneous equations, we obtain $C_\lambda = 0$ and

$$B_\lambda \sin(\sqrt{\lambda}L) = 0 \Rightarrow \begin{cases} B_\lambda = 0 \Rightarrow X = 0 \Rightarrow u = 0 \\ B_\lambda \neq 0 \Rightarrow \sin(\sqrt{\lambda}L) = 0 \Rightarrow \sqrt{\lambda}L = n\pi \end{cases}$$

where n are positive integers because $\sqrt{\lambda}L$ needs to be positive.

Therefore, when

$$\lambda = \left(\frac{n\pi}{L}\right)^2, \quad n = 1, 2, \dots$$

we have non-trivial solutions, i.e. the eigenfunctions

$$X_n(x) = B_n \sin \frac{n\pi x}{L}.$$

Here we let $B_\lambda = B_n$ since we found that λ is given by n . $\square$

Once we have determined the value for λ , we can substitute into (36.43) and obtain the solution

$T_n(t) = A_n e^{-\frac{n^2 \pi^2 k t}{L^2}}$. Since $u = TX$, we have a sequence of special solutions

$$u_n(x, t) = A_n e^{-\frac{n^2 \pi^2 k t}{L^2}} B_n \sin \frac{n \pi x}{L}, \quad n = 1, 2, \dots$$

Since the heat equation (35.15) and the boundary conditions are **linear homogeneous**, the sum of all solutions u_n is still a solution according to the **principle of superposition** (Theorem 35.26). This gives the *general solution*

$$u(x, t) = \sum_{n=1}^{\infty} b_n e^{-\frac{n^2 \pi^2 k t}{L^2}} \sin \frac{n \pi x}{L} \quad (35.37)$$

where $b_n = A_n B_n$. The constants b_n can be determined from the initial condition (35.28):

$$b_n = \frac{2}{L} \int_0^L f(x) \sin \frac{n \pi x}{L} dx.$$

Note that b_n is either Fourier sine series coefficients or Fourier coefficients of the “odd-function” f (Topic 1).

35.3.2 (Homogeneous) Type 2 BCs

When a bar has insulated ends, the end points are characterised by type 2 BCs:

$$\frac{\partial u}{\partial x}(0, t) = \frac{\partial u}{\partial x}(L, t) = 0, \quad t > 0, \quad (35.38)$$

Note that (35.15) and (35.38) form a homogeneous linear system. Therefore, the separation of variables method can be applied to find a solution. Assume $u = TX$, we will two ODEs (36.43) and (35.34) similar to the previous section. However, the implication of type 2 BCs are different:

$$T(t)X'(0) = T(t)X'(L) = 0 \Rightarrow X'(0) = X'(L) = 0. \quad (35.39)$$

Let us analyse the Sturm-Liouville problem (35.34) and (35.39) below.

Example 35.30. What is/are the solution(s) for (35.34) and (35.39) when $\lambda < 0$?

Solution. When $\lambda < 0$, there exist real numbers B, C such that

$$X(x) = Ae^{\sqrt{-\lambda}x} + Be^{-\sqrt{-\lambda}x}.$$

The boundary conditions (35.39) imply the following simultaneous equations:

$$\begin{aligned} X'(0) &= \sqrt{-\lambda}A - \sqrt{-\lambda}B = 0, \\ X'(L) &= \sqrt{-\lambda}Ae^{\sqrt{-\lambda}L} - \sqrt{-\lambda}Be^{-\sqrt{-\lambda}L} = 0. \end{aligned}$$

The determinant

$$\begin{vmatrix} \sqrt{-\lambda} & -\sqrt{-\lambda} \\ \sqrt{-\lambda}e^{\sqrt{-\lambda}L} & -\sqrt{-\lambda}e^{-\sqrt{-\lambda}L} \end{vmatrix} = \lambda e^{-\sqrt{-\lambda}L} - \lambda e^{\sqrt{-\lambda}L} = -2\lambda \sinh(\sqrt{-\lambda}L) \neq 0.$$

This implies $A = B = 0$ and so $X = 0$ is the only solution, which is trivial. $\square$

Example 35.31. What is/are the solution(s) for (35.34) and (35.39) when $\lambda = 0$?

Solution. When $\lambda = 0$, there exist real numbers A_0, B_0 such that the general solution of (35.34) is

$$X(x) = A_0 + B_0 x.$$

Substituting it into the boundary conditions (35.39), we obtain

$$B_0 = 0.$$

This implies $X(x) = A_0$ and we have a non-trivial solution. $\square$

Example 35.32. What is/are the solution(s) for (35.34) and (35.39) when $\lambda > 0$?

Solution. When $\lambda > 0$, the general solution of (35.34) is

$$X(x) = A_\lambda \cos(\sqrt{\lambda}x) + B_\lambda \sin(\sqrt{\lambda}x).$$

Substituting it into the boundary conditions (35.39), we obtain the simultaneous equations:

$$\begin{aligned} X'(0) &= -\sqrt{\lambda}A_\lambda \times 0 + \sqrt{\lambda}B_\lambda \times 1 = 0 \\ X'(L) &= -\sqrt{\lambda}A_\lambda \sin(\sqrt{\lambda}L) + \sqrt{\lambda}B_\lambda \cos(\sqrt{\lambda}L) = 0 \end{aligned}$$

This implies that

$$B_\lambda = 0 \wedge \sqrt{\lambda}A_\lambda \sin(\sqrt{\lambda}L) = 0.$$

If we do not want a trivial solution, we must have $A_\lambda \neq 0$. Note that $\lambda > 0$ and so

$$\sin(\sqrt{\lambda}L) = 0 \Rightarrow \sqrt{\lambda}L = n\pi, \quad n = 1, 2, \dots$$

Hence,

$$\lambda = \left(\frac{n\pi}{L}\right)^2, \quad n = 1, 2, \dots$$

Hence, we obtain the eigenfunctions (with $A_n = A_\lambda$) and the solution to T :

$$X_n(x) = A_n \cos \frac{n\pi x}{L}$$

which are non-trivial solutions. □

Example 35.33 (UECM3003 Final Exam, Apr 2014, Q5(a)). Find all possible eigenfunctions that solve the following Sturm-Liouville problem

$$\frac{d^2 y}{dt^2} + ky = 0,$$

subject to the boundary conditions $y'(0) = 0$ and $y'(1) = 0$.

Solution: We have our answer

$$k = 0, (n\pi)^2 : \quad y_0(x) = A_0, \quad y_n(x) = A_n \cos n\pi x, \quad n = 1, 2, \dots$$

by combining all the answers from Examples 35.30, 35.31 and 35.32 with λ to k and let $L = 1$.

Now, let us write down all the non-trivial solutions for the heat equation (35.15) and the type 2 boundary conditions (35.29).

- When $\lambda = 0$, the equation (36.43) becomes

$$T' = 0 \Rightarrow T(t) = C_0$$

and the non-trivial solution is

$$u_0(x, t) = T(t)X(x) = C_0 A_0 =: a_0.$$

- When $\lambda = \left(\frac{n\pi}{L}\right)^2$, $n = 1, 2, \dots$, we have non-trivial solution X and the solution to (36.43) is

$$T(t) = C_n e^{-(n\pi/L)^2 kt}$$

and the non-trivial solution to the PDE problem is

$$u_n(x, t) = T(t)X(x) = C_n A_n e^{-(n\pi/L)^2 kt} \cos \frac{n\pi x}{L}.$$

Since the heat equation (35.15) and the boundary conditions (35.38) are linear homogeneous, we can apply the principle of superposition to obtain a general solution

$$u(x, t) = a_0 + \sum_{n=1}^{\infty} a_n e^{-(n\pi/L)^2 kt} \cos \frac{n\pi x}{L} \quad (35.40)$$

where $a_n = C_n A_n$.

Example 35.34. Determine the values of a_0 and a_n from the initial condition (35.28).

Solution. The initial condition is

$$u(x, 0) = a_0 + \sum_{n=1}^{\infty} a_n \cos \frac{n\pi x}{L} = f(x), \quad x \in [0, L]$$

The left-hand-side is the cosine Fourier series of the function f , therefore, using the knowledge from Topic 1, we have

$$a_0 = \frac{1}{L} \int_0^L f(x) dx, \quad a_n = \frac{2}{L} \int_0^L f(x) \cos \frac{n\pi x}{L} dx.$$

□

Example 35.35 (UCCM3003 Final Exam, Apr 2008, Q2). Consider the following heat equation in rectangular coordinate system:

$$\begin{cases} u_t = 4u_{xx}, & 0 < x < 5, \\ \frac{\partial u}{\partial x} \Big|_{x=0} = \frac{\partial u}{\partial x} \Big|_{x=5} = 0, & t > 0, \\ u(x, 0) = 1 + 2 \cos \frac{\pi x}{5} + 3 \cos \pi x, & 0 \leq x \leq 5. \end{cases}$$

(a) Find the solution $u(x, t)$ of the heat equation using the method of separation of variables.

(23 marks)

Solution. Let $u = XT$.

PDE implies $\frac{T'}{4T} = \frac{X''}{X} = -\lambda$, where λ is a constant (2 marks)

From the BC, we have $X'(0) = X'(5) = 0$ (2 marks)

The S-L problem $X'' + \lambda X = 0$, $X'(0) = X'(5) = 0$ is a type 2 homogeneous problem. We will analyse the value of λ as follows:

(a) $\lambda < 0$:

$$X = Ae^{\sqrt{-\lambda}x} + Be^{-\sqrt{-\lambda}x}, \quad X' = A\sqrt{-\lambda}e^{\sqrt{-\lambda}x} - B\sqrt{-\lambda}e^{-\sqrt{-\lambda}x} \quad \dots (1 \text{ mark})$$

$$X'(0) = A\sqrt{-\lambda} - B\sqrt{-\lambda} = 0$$

$$X'(5) = A\sqrt{-\lambda}e^{\sqrt{-\lambda}5} - B\sqrt{-\lambda}e^{-\sqrt{-\lambda}5} = 0 \quad \dots (1 \text{ mark})$$

Solving it gives $A = B = 0$. X only have trivial solutions. (1 mark)

(b) $\lambda = 0$:

$$X = A + Bx, \quad X'(0) = X'(5) = 0 \text{ implies } B = 0. \quad \dots (1 \text{ mark})$$

Hence A can be any value. This implies that we have an eigenfunction $X_0(x) = A_0$ and $\lambda_0 = 0$ is an eigenvalue. (1 mark)

(c) $\lambda > 0$:

$$X = A \cos \sqrt{\lambda}x + B \sin \sqrt{\lambda}x, \quad X' = -A\sqrt{\lambda} \sin \sqrt{\lambda}x + B\sqrt{\lambda} \cos \sqrt{\lambda}x \quad \dots (1 \text{ mark})$$

$$X'(0) = 0 \text{ implies } B = 0 \quad \dots (1 \text{ mark})$$

$$X'(5) = 0 \text{ implies } \sin \sqrt{\lambda}5 = 0. \text{ Solving it gives the eigenvalue } \lambda_n = \left(\frac{n\pi}{5}\right)^2, \quad n = 1, 2, \dots$$

$$\text{and the eigenfunctions } X_n(x) = A_n \cos \frac{n\pi x}{5}. \quad \dots (1 \text{ mark})$$

When $\lambda = 0$ and $\left(\frac{n\pi}{5}\right)^2$, $n = 1, 2, \dots$, we have

$$T_0(t) = C_0, \quad T_n(t) = C_n e^{-4\left(\frac{n\pi}{5}\right)^2 t}. \quad (2 \text{ marks})$$

Hence the general solution is

$$u(x, t) = \sum_{k=0}^{\infty} T_k(t) X_k(x) = a_0 + \sum_{n=1}^{\infty} a_n e^{-4\left(\frac{n\pi}{5}\right)^2 t} \cos \frac{n\pi x}{5} \quad (2 \text{ marks})$$

where $a_0 = A_0 C_0$ and $a_n = A_n C_n$ are constants.

From the initial condition,

$$u(x, 0) = a_0 + \sum_{n=1}^{\infty} a_n \cos \frac{n\pi x}{5} = 1 + 2 \cos \frac{\pi x}{5} + 3 \cos \pi x \quad (2 \text{ marks})$$

Since $\{\cos \frac{n\pi x}{5}\}_{n=0}^{\infty}$ are linearly independent, by comparing coefficients, we have

$$a_0 = 1, a_1 = 2, a_5 = 3, a_n = 0 \text{ for } n \neq 0, 1, 5. \quad (1+1+1+1=4 \text{ marks})$$

Hence the solution of the heat equation is

$$u(x, t) = 1 + 2e^{-\frac{4\pi^2}{25}t} \cos \frac{\pi x}{5} + 3e^{-4\pi^2 t} \cos \pi x. \quad (1 \text{ mark})$$

□

- (b) What is steady-state solution of the heat problem in part (a)? (2 marks)

Solution. Steady-state solution = $\lim_{t \rightarrow \infty} u(x, t) = 1$ (2 marks) □

Remark 35.36. Steady-state solution refers to a “solution” which is independent of time, i.e. $u_t = 0$.

35.3.3 (Homogeneous) Type 3 BCs

A type 3 BC can be seen as an intermediate state between type 1 BC and type 2 BC. When the two end points are restricted to homogeneous type 3 BCs, we have

$$\frac{\partial u}{\partial x}(0, t) - h_1 u(0, t) = \frac{\partial u}{\partial x}(L, t) + h_2 u(L, t) = 0, \quad t > 0 \quad (35.41)$$

where $h_1, h_2 > 0$ are constants.

Since the problem is homogeneous, by applying the separation of variable $u = TX$, the PDE (35.15) and the BCs (35.41) lead to an SL problem

$$X'' + \lambda X = 0, \quad X'(0) - h_1 X(0) = X'(L) + h_2 X(L) = 0. \quad (35.42)$$

We need to find the eigenvalues and eigenfunctions for (35.42) by the following analysis.

- Suppose that $\lambda < 0$. Then there exist real numbers A, B such that $X(x) = Ae^{\sqrt{-\lambda}x} + Be^{-\sqrt{-\lambda}x}$. The BCs in (35.42) implies that

$$\begin{bmatrix} \sqrt{-\lambda} - h_1 & -\sqrt{-\lambda} - h_1 \\ e^{\sqrt{-\lambda}L}(\sqrt{-\lambda} + h_2) & e^{-\sqrt{-\lambda}L}(-\sqrt{-\lambda} + h_2) \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Since the determinant of the coefficient matrix

$$\begin{aligned} & \lambda(e^{-\sqrt{-\lambda}L} - e^{\sqrt{-\lambda}L}) + h_1 h_2 (e^{\sqrt{-\lambda}L} - e^{-\sqrt{-\lambda}L}) + (h_1 + h_2)\sqrt{-\lambda}(e^{-\sqrt{-\lambda}L} + e^{\sqrt{-\lambda}L}) \\ &= -2\lambda \sinh(\sqrt{-\lambda}L) + 2h_1 h_2 \sinh(\sqrt{-\lambda}L) + 2(h_1 + h_2)\sqrt{-\lambda} \cosh(\sqrt{-\lambda}L) \\ &= 2(h_1 h_2 - \lambda) \sinh(\sqrt{-\lambda}L) + 2(h_1 + h_2)\sqrt{-\lambda} \cosh(\sqrt{-\lambda}L) > 0, \end{aligned}$$

we have $A = B = 0$ which implies X is trivial.

- Suppose that $\lambda = 0$. Then there exist real numbers A_0, B_0 such that $X(x) = A_0 + B_0 x$. The BCs in (35.42) implies that

$$\begin{bmatrix} -h_1 & 1 \\ h_2 & 1 + Lh_2 \end{bmatrix} \begin{bmatrix} A_0 \\ B_0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The determinant

$$-h_1(1 + Lh_2) - h_2 < 0$$

so $B_0 = A_0 = 0$ and X is trivial.

- When $\lambda > 0$. There exist real numbers A_λ and B_λ such that $X(x) = A_\lambda \cos(\sqrt{\lambda}x) + B_\lambda \sin(\sqrt{\lambda}x)$. From BCs (35.42) we have

$$\begin{bmatrix} -h_1 & \sqrt{\lambda} \\ -\sqrt{\lambda} \sin \sqrt{\lambda}L + h_2 \cos \sqrt{\lambda}L & \sqrt{\lambda} \cos \sqrt{\lambda}L + h_2 \sin \sqrt{\lambda}L \end{bmatrix} \begin{bmatrix} A_\lambda \\ B_\lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (35.43)$$

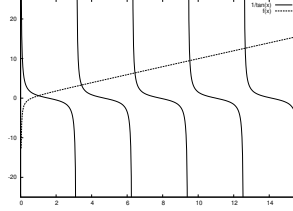
The determinant has to be zero as follows in order to have eigenfunctions:

$$-h_1 \sqrt{\lambda} \cos \sqrt{\lambda}L - h_1 h_2 \sin \sqrt{\lambda}L + \lambda \sin \sqrt{\lambda}L - h_2 \sqrt{\lambda} \cos \sqrt{\lambda}L = 0$$

Simplifying the above equation, we have

$$\begin{aligned}\sqrt{\lambda}(h_1 + h_2) \cos \sqrt{\lambda}L &= (\lambda - h_1 h_2) \sin \sqrt{\lambda}L \\ \Rightarrow \cot \sqrt{\lambda}L &= \frac{\lambda - h_1 h_2}{\sqrt{\lambda}(h_1 + h_2)} = \frac{\sqrt{\lambda}}{h_1 + h_2} - \frac{h_1 h_2}{(h_1 + h_2)\sqrt{\lambda}}\end{aligned}$$

Let $s = \sqrt{\lambda}L$, we can sketch the function $y = \cot s$ and $y = \frac{s}{(h_1 + h_2)L} - \frac{h_1 h_2 L}{(h_1 + h_2)s}$ as follows



we obtain eigenvalues $\lambda = \beta_n^2$, $n = 1, 2, \dots$ which satisfy the equation $\cot \beta_n L = \frac{\beta_n}{h_1 + h_2} - \frac{h_1 h_2}{(h_1 + h_2)\beta_n}$. The corresponding eigenfunctions are

$$X_n(x) = A_n \cos \beta_n x + B_n \sin \beta_n x = A_n \left(\cos \beta_n x + \frac{h_1}{\beta_n} \sin \beta_n x \right)$$

since $h_1 > 0$, $\sqrt{\lambda} > 0$ and according to (35.43), we have

$$-h_1 A_\lambda + \sqrt{\lambda} B_\lambda = 0 \Rightarrow B_\lambda = \frac{h_1}{\beta_n} A_\lambda.$$

For $\lambda = \beta_n^2$, $n = 1, 2, \dots$, solving the ODE (36.43) of T leads to

$$T_n(t) = C_n e^{-\beta_n^2 kt}.$$

and the general solution for u is

$$u(x, t) = \sum_{n=1}^{\infty} a_n e^{-\beta_n^2 kt} \left(\cos \beta_n x + \frac{h_1}{\beta_n} \sin \beta_n x \right)$$

where $a_n = C_n A_n$ are constants. They can be calculated from the initial condition (35.28) as follows

$$\sum_{n=1}^{\infty} a_n \left(\cos \beta_n x + \frac{h_1}{\beta_n} \sin \beta_n x \right) = f(x) \Rightarrow a_n = \frac{\int_0^L f(x) \left[\cos \beta_n x + \frac{h_1}{\beta_n} \sin \beta_n x \right] dx}{\int_0^L \left[\cos \beta_n x + \frac{h_1}{\beta_n} \sin \beta_n x \right]^2 dx}$$

where $n = 1, 2, \dots$. Note that we apply the orthogonal property of eigenfunctions of Sturm-Liouville problem from Topic 1.

35.3.4 (Homogeneous) Mixed BCs

It is possible to have any combination of type 1, 2 and 3 BCs. However, as we have seen above, they can be very complicated, a few examples are given in Asmar [2005, Chapter 3].

Here, we will consider when one end of the bar is insulated while the other end is kept at zero degree, then one end is characterised by a type 2 boundary condition and the other end is characterised by type 1 boundary condition:

$$\frac{\partial u}{\partial x}(0, t) = u(L, t) = 0, \quad t > 0, \quad (35.44)$$

The boundary conditions are equal to 0 and hence “homogeneous”. Again, we can apply the separation of variables by assuming $u = TX$ to obtain two ODEs (35.34) and (36.43) and the following boundary conditions:

$$T(t)X'(0) = T(t)X(L) = 0 \Rightarrow X'(0) = X(L) = 0. \quad (35.45)$$

Example 35.37. Find the non-trivial solutions (eigenfunctions) for the Sturm-Liouville problem (35.34) and (35.45).

Solution. • When $\lambda < 0$, the general solution of (35.34) is

$$X(x) = Ae^{\sqrt{-\lambda}x} + Be^{-\sqrt{-\lambda}x}.$$

From the boundary conditions (35.45), we have

$$\begin{aligned} X'(0) = 0 &\Rightarrow \sqrt{-\lambda}A - \sqrt{-\lambda}B = 0, \\ X(L) = 0 &\Rightarrow Ae^{\sqrt{-\lambda}L} + Be^{-\sqrt{-\lambda}L} = 0. \end{aligned}$$

Since the determinant

$$\begin{vmatrix} \sqrt{-\lambda} & -\sqrt{-\lambda} \\ e^{\sqrt{-\lambda}L} & e^{-\sqrt{-\lambda}L} \end{vmatrix} = \sqrt{-\lambda}(e^{\sqrt{-\lambda}L} + e^{-\sqrt{-\lambda}L}) \neq 0,$$

we have $A = B = 0$ which implies X is trivial.

- When $\lambda = 0$, the general solution of (35.34) is

$$X(x) = A_0 + B_0x.$$

The boundary conditions (35.45) imply

$$X'(0) = B_0 = 0 \wedge X(L) = A_0 = 0 \Rightarrow B_0 = A_0 = 0.$$

Therefore, X is trivial.

- When $\lambda > 0$, the general solution of (35.34) is

$$X(x) = A_\lambda \cos(\sqrt{\lambda}x) + B_\lambda \sin(\sqrt{\lambda}x).$$

BCs (35.45) imply that $B_\lambda = 0$ and $\cos \sqrt{\lambda}L = 0$. This implies that $\sqrt{\lambda}L = \frac{\pi}{2} + n\pi$ for $n = 0, 1, 2, \dots$. Hence, the eigenvalues and eigenfunctions are

$$\lambda_n = \left(\frac{(2n+1)\pi}{2L}\right)^2, \quad X_n(x) = A_n \cos \frac{(2n+1)\pi x}{2L}, \quad n = 0, 1, 2, \dots$$

□

Example 35.38. Write down the series solution of the PDE (35.15) together with the BCs (35.44) and the IC (35.28).

Solution. For the eigenvalues $\lambda = \frac{(2n+1)^2\pi^2}{4L^2}$, $n = 0, 1, 2, \dots$, solving (36.43) leads to $T(t)$:

$$T_n(t) = C_n e^{-\frac{(2n+1)^2\pi^2}{4L^2}kt}.$$

The general solution is

$$u(x, t) = \sum_{n=0}^{\infty} a_n e^{-\frac{(2n+1)^2\pi^2}{4L^2}kt} \cos \frac{(2n+1)\pi x}{2L}$$

where $a_n = C_n A_n$ given below is obtained from the initial condition (35.28):

$$a_n = \frac{\int_0^L f(x) \cos \frac{(2n+1)\pi x}{2L} dx}{\int_0^L \cos^2 \frac{(2n+1)\pi x}{2L} dx} = \frac{2}{L} \int_0^L f(x) \cos \frac{(2n+1)\pi x}{2L} dx.$$

□

Example 35.39 (UCCM3003 Final Exam, Apr 2012, Q2). (a) By considering the three cases: $\lambda > 0$, $\lambda = 0$, $\lambda < 0$, show that the eigenvalues and eigenfunctions of the Sturm-Liouville problem

$$X'' - \lambda X = 0, \quad X'(0) = X(\pi) = 0 \quad (\dagger)$$

are $\lambda_n = -\frac{(2n-1)^2}{4}$ and $X_n = \cos \frac{(2n-1)x}{2}$ for $n = 1, 2, 3, \dots$. (12 marks)

Solution: The answer is more or less a copy and paste of the solution of Example 35.37, with $L = \pi$ and the difference in the sign of λ . leg:rect:heat1d:mixedbc

- When $\lambda > 0$, the general solution of (†) is

$$X(x) = Ae^{\sqrt{\lambda}x} + Be^{-\sqrt{\lambda}x}.$$

From the boundary conditions, we have

$$\begin{aligned} X'(0) = 0 &\Rightarrow \sqrt{\lambda}A - \sqrt{\lambda}B = 0, \\ X(\pi) = 0 &\Rightarrow Ae^{\sqrt{\lambda}\pi} + Be^{-\sqrt{\lambda}\pi} = 0. \end{aligned}$$

Since the determinant

$$\begin{vmatrix} \sqrt{\lambda} & -\sqrt{\lambda} \\ e^{\sqrt{\lambda}\pi} & e^{-\sqrt{\lambda}\pi} \end{vmatrix} = \sqrt{\lambda}(e^{\sqrt{\lambda}\pi} + e^{-\sqrt{\lambda}\pi}) \neq 0,$$

we have $A = B = 0$ which implies X is trivial.

- When $\lambda = 0$, the general solution of (†) is

$$X(x) = A_0 + B_0x.$$

The boundary conditions le:s1-bc21 (35.45) imply

$$X'(0) = B_0 = 0 \wedge X(L) = A_0 = 0 \Rightarrow B_0 = A_0 = 0.$$

Therefore, X is trivial.

- When $\lambda < 0$, the general solution of (†) is

$$X(x) = A_\lambda \cos(\sqrt{-\lambda}x) + B_\lambda \sin(\sqrt{-\lambda}x).$$

The boundary conditions imply that $B_\lambda = 0$ and $\cos \sqrt{-\lambda}L = 0$. This implies that $\sqrt{-\lambda}\pi = n\pi - \frac{\pi}{2}$ for $n = 1, 2, \dots$. Hence, the eigenvalues and eigenfunctions are

$$\lambda_n = -\left(\frac{2n-1}{2}\right)^2, \quad X_n(x) = A_n \cos \frac{(2n-1)x}{2}, \quad n = 1, 2, \dots$$

- (b) Use the method of separation of variables (and part (a) of this question) to solve the heat equation

$$u_t = u_{xx}, \quad 0 < x < \pi, \quad t > 0$$

subject to the following conditions

$$\begin{aligned} u_x(0, t) &= u(\pi, t) = 0, \quad t > 0 \\ u(x, 0) &= 4 \cos \frac{3x}{2} - 5 \cos \frac{7x}{2}, \quad 0 < x < \pi. \end{aligned}$$

Also, find the steady-state solution for this problem. (13 marks)

Solution: The solution is more or less a copy and paste of Example 35.38 with minor differences in λ . For the eigenvalues $\lambda_n = -\frac{(2n-1)^2}{4}$, $n = 1, 2, \dots$, solving the ODE of T leads to leg:rect:heat1d:mixedbc2

$$T'_n - \lambda_n T_n = 0 \Rightarrow T_n(t) = C_n e^{-\frac{(2n-1)^2}{4}t}.$$

The general solution of u is

$$u(x, t) = \sum_{n=0}^{\infty} a_n e^{-\frac{(2n-1)^2}{4}t} \cos \frac{(2n-1)x}{2}$$

where $a_n = C_n A_n$ given below is obtained from the initial condition le:heat1d-ic (35.28):

$$a_n = \frac{\int_0^\pi \left[4 \cos \frac{3x}{2} - 5 \cos \frac{7x}{2} \right] \cos \frac{(2n-1)x}{2} dx}{\int_0^\pi \cos^2 \frac{(2n-1)x}{2} dx} = \frac{2}{\pi} \int_0^\pi \left[4 \cos \frac{3x}{2} - 5 \cos \frac{7x}{2} \right] \cos \frac{(2n-1)x}{2} dx.$$

$$= \frac{2}{\pi} \begin{cases} 4 \times \frac{\pi}{2} & n = 2 \\ -5 \times \frac{\pi}{2} & n = 4 \\ 0 & n \neq 2, 4. \end{cases} \quad \text{Hence, } u(x, t) = 4e^{-9/4t} \cos \frac{3x}{2} - 5e^{-49/4t} \cos \frac{7x}{2}.$$

Since $e^{-9/4t} \rightarrow 0$ and $e^{-49/4t} \rightarrow 0$ as $t \rightarrow \infty$, the steady-state solution is $u_\infty(x) = 0$.

35.3.5 Nonhomogeneous PDE and Homogeneous BCs

What if the PDE problem is nonhomogeneous (or sometimes called inhomogeneous)? By using the knowledge of logic (UECM1303 Discrete Mathematics Topic 1), we have the following cases:

1. PDE (35.15) is nonhomogeneous but the BCs are homogeneous.
2. PDE (35.15) is homogeneous but at least one of the BC is nonhomogeneous.
3. PDE (35.15) is nonhomogeneous and at least one of the BC is nonhomogeneous.

In this subsection, we will investigate the first case above. A PDE is said to be *(linear) nonhomogeneous* if there is a heat source F on the right-hand-side of (35.15).

$$u_t - ku_{xx} = F(x, t), \quad 0 < x < L, \quad t > 0. \quad (35.46)$$

Mathematicians have developed the *eigenfunction expansion* method, also known as *modified separation of variables*, to solve this problem. First, we need to find the eigenfunctions from the S-L problem associated with the homogeneous problem of (35.46) and the BCs. Then we expand the force F with respect to the eigenfunctions and determine the time functions $T_n(t)$ from (35.46).

We will illustrate how this method work with three examples below.

Example 35.40 (UCCM3003 Final Exam, Sep 2008, Q4). Consider the following heat equation:

$$\begin{cases} \frac{\partial u}{\partial t} = 4 \frac{\partial^2 u}{\partial x^2} + e^{-3t}, & 0 < x < 5, \quad t > 0, \\ u(0, t) = 0, \quad \frac{\partial u}{\partial x}(5, t) = 0, & t \geq 0, \\ u(x, 0) = 0, & 0 \leq x \leq 5. \end{cases}$$

Find the solution $u(x, t)$ of the heat equation using the Fourier expansion. (25 marks)

Solution: Consider the homogeneous case, the PDE and BCs can be simplified to

$$\frac{T'}{4T} = \frac{X''}{X} = -\lambda, \quad X(0) = 0, \quad X'(5) = 0. \quad [1 \text{ mark}]$$

When $\lambda \leq 0$, we only have trivial solutions. When $\lambda > 0$, $X = A \cos \sqrt{\lambda}x + B \sin \sqrt{\lambda}x$. The BCs imply $A = 0$ and when $B \neq 0$,

$$5\sqrt{\lambda} = \frac{(2n+1)\pi}{2} \Rightarrow \lambda = \left(\frac{(2n+1)\pi}{10} \right)^2, \quad n = 0, 1, 2, \dots \quad [2 \text{ marks}]$$

and $X_n(x) = B_n \sin \frac{(2n+1)\pi}{10}x. \quad [1 \text{ mark}]$

The Fourier expansion is $u = \sum_{n=0}^{\infty} T_n(t) \sin \frac{(2n+1)\pi}{10}x. \quad [2 \text{ marks}]$

Substituting the Fourier expansion to the PDE, we have $\sum_{n=0}^{\infty} T'_n(t) \sin \frac{(2n+1)\pi}{10}x = -4 \sum_{n=0}^{\infty} T_n(t) \left(\frac{(2n+1)\pi}{10} \right)^2 \sin \frac{(2n+1)\pi}{10}x + e^{-3t}. \quad [2 \text{ marks}]$

We need to find the generalised Fourier series for e^{-3t} w.r.t x :

$$e^{-3t} = e^{-3t} \sum_{n=0}^{\infty} c_n \sin \frac{(2n+1)\pi}{10}x, \quad 0 \leq x \leq 5 \quad [1 \text{ mark}]$$

$$c_n = \frac{\int_0^5 \sin \frac{(2n+1)\pi}{10}x}{\int_0^5 \sin^2 \frac{(2n+1)\pi}{10}x} = \frac{\left[-\frac{\cos \frac{(2n+1)\pi}{10}x}{\frac{(2n+1)\pi}{10}} \right]_0^5}{\frac{5}{2}} = \frac{4}{(2n+1)\pi} \quad [2+2+1=5 \text{ marks}]$$

Hence, the PDE can be written as ODEs:

$$T_n'(t) + 4\left(\frac{(2n+1)\pi}{10}\right)^2 T_n(t) = \frac{4}{(2n+1)\pi} e^{-3t}. \quad [2 \text{ marks}]$$

The initial condition implies

$$T_n(0) = 0. \quad [1 \text{ mark}]$$

Taking Laplace transform (let $T_n(s) = \mathcal{L}\{T_n(t)\}$) we have

$$sT_n(s) + 4\left(\frac{(2n+1)\pi}{10}\right)^2 T_n(s) = \frac{4}{(2n+1)\pi} \frac{1}{s+3} \quad [2 \text{ marks}]$$

$$\Rightarrow T_n(s) = \frac{4}{(2n+1)\pi} \frac{1}{(s+3)\left(s+4\left(\frac{(2n+1)\pi}{10}\right)^2\right)} \quad [1 \text{ mark}]$$

$$= \frac{4}{(2n+1)\pi} \left(\frac{\frac{1}{4\left(\frac{(2n+1)\pi}{10}\right)^2 - 3}}{s+3} + \frac{\frac{1}{3-4\left(\frac{(2n+1)\pi}{10}\right)^2}}{s+4\left(\frac{(2n+1)\pi}{10}\right)^2} \right) \quad [2 \text{ marks}]$$

$$\Rightarrow T_n(t) = \frac{4}{(2n+1)\pi(4\left(\frac{(2n+1)\pi}{10}\right)^2 - 3)} (e^{-3t} - e^{-4\left(\frac{(2n+1)\pi}{10}\right)^2 t}). \quad [2 \text{ marks}]$$

$$\text{Hence the solution } u(x, t) = \sum_{n=0}^{\infty} \frac{4 \left(e^{-3t} - e^{-\left(\frac{(2n+1)\pi}{5}\right)^2 t} \right)}{(2n+1)\pi \left(\left(\frac{(2n+1)\pi}{5}\right)^2 - 3 \right)} \sin \frac{(2n+1)\pi x}{10}. \quad [1 \text{ mark}]$$

Example 35.41 (UECM3003 Final Exam, Apr 2015, Q2(b)). Consider the following heat equation with source

$$\frac{\partial u}{\partial t} - 9 \frac{\partial^2 u}{\partial x^2} = \sin \pi x, \quad \text{for } 0 < x < 3, \quad t > 0$$

that also satisfies the boundary conditions

$$u(0, t) = 0, \quad u(3, t) = 0, \quad \text{for } t > 0.$$

Given that the eigenfunctions for the corresponding PDE and boundary conditions are $\sin \frac{n\pi x}{3}$, $n = 1, 2, 3, \dots$, find a solution to the heat equation with source when the initial condition is $u(x, 0) = 0$, for $0 < x < 3$. What is $u(x, t)$ when $t \rightarrow \infty$? (8 marks)

Solution. The eigenfunction expansion is

$$u(x, t) = \sum_{n=1}^{\infty} T_n(t) \sin \frac{n\pi x}{3}.$$

Substituting into the PDE, we obtain

$$\sum_{n=1}^{\infty} T_n'(t) \sin \frac{n\pi x}{3} + 9 \left(\frac{n\pi}{3} \right)^2 \sum_{n=1}^{\infty} T_n(t) \sin \frac{n\pi x}{3} = \sin \pi x.$$

From here, we obtain the following ODEs

$$T_n'(t) + n^2 \pi^2 T_n(t) = \begin{cases} 1, & n = 3, \\ 0, & n \neq 3. \end{cases}$$

From the initial condition $u(0, x) = 0$, we have $T_n(0) = 0$.

When $n \neq 3$, the general solution is $T_n(t) = A_n \exp(-n^2 \pi^2 t)$ and $T_n(0) = 0 \Rightarrow A_n = 0 \Rightarrow T_n(t) = 0$.

When $n = 3$, since $T_3(0) = 0$, taking Laplace transform gives

$$sT_3(s) - T_3(0) + 9\pi^2 T_3(s) = \frac{1}{s} \Rightarrow T_3(s) = \frac{1}{s(s+9\pi^2)} = \frac{\frac{1}{9\pi^2}}{s} + \frac{\frac{1}{-9\pi^2}}{s+9\pi^2}$$

Taking the inverse Laplace transform, we have

$$T_3(t) = \frac{1 - e^{-9\pi^2 t}}{9\pi^2}.$$

This means that $u(x, t) = \frac{1 - e^{-9\pi^2 t}}{9\pi^2} \sin \pi x$. □

Example 35.42. Solve the non-homogeneous problem with a nonhomogeneous heat equation (35.46), type 1 BCs (35.29) and initial condition (35.28).

Solution. The corresponding homogeneous problem has been solved in the previous section, in which the eigenfunctions are

$$X_n(x) = \sin \frac{n\pi x}{L}.$$

The eigenfunction expansion of u has the form

$$u(x, t) = \sum_{n=1}^{\infty} T_n(t) \sin \frac{n\pi x}{L}$$

for some unknown functions $T_n(t)$. By substituting the expansion above into (35.46) and expand F using the eigenfunctions, we obtain an expansion involving Fourier sine series of F below:

$$\sum_{n=1}^{\infty} \left[T'_n(t) + k \left(\frac{n\pi}{L} \right)^2 T_n(t) \right] \sin \frac{n\pi x}{L} = \sum_{n=1}^{\infty} F_n(t) \sin \frac{n\pi x}{L}$$

where

$$F_n(t) = \frac{2}{L} \int_0^L F(x, t) \sin \frac{n\pi x}{L} dx.$$

An rearrangement of terms and applying the “linear independence” rule, we obtain

$$\sum_{n=1}^{\infty} \left[T'_n(t) + k \left(\frac{n\pi}{L} \right)^2 T_n(t) - F_n(t) \right] \sin \frac{n\pi x}{L} = 0 \Rightarrow \boxed{T'_n(t) + k \left(\frac{n\pi}{L} \right)^2 T_n(t) - F_n(t) = 0} \quad (35.47)$$

Substituting the eigenfunction expansion of u into the initial condition (35.28) and expanding f into Fourier sine series, we obtain

$$\sum_{n=1}^{\infty} T_n(0) \sin \frac{n\pi x}{L} = \sum_{n=1}^{\infty} b_n \sin \frac{n\pi x}{L}, \quad b_n = \frac{2}{L} \int_0^L f(x) \sin \frac{n\pi x}{L} dx.$$

can be rearranged into

$$\sum_{n=1}^{\infty} [T_n(0) - b_n] \sin \frac{n\pi x}{L} = 0 \Rightarrow T_n(0) - b_n = 0. \quad (35.48)$$

Solving the ODE-IVP (35.47) and (35.48) using the Laplace transform we learned in UECM2023 ODE, we obtain

$$\mathcal{L}\{T_n(t)\} = \frac{b_n}{s + k(n\pi/L)^2} + \frac{\mathcal{L}\{F_n(t)\}}{s + k(n\pi/L)^2}.$$

Taking inverse Laplace transform yields:

$$T_n(t) = b_n e^{-k(n\pi/L)^2 t} + \int_0^t e^{-k(n\pi/L)^2 (t-\tau)} F_n(\tau) d\tau$$

and the solution becomes

$$u(x, t) = \sum_{n=1}^{\infty} b_n e^{-k(n\pi/L)^2 t} \sin \frac{n\pi x}{L} + \sum_{n=1}^{\infty} \left[\int_0^t e^{-k(n\pi/L)^2 (t-\tau)} F_n(\tau) d\tau \right] \sin \frac{n\pi x}{L}.$$

□

35.3.6 Homogeneous PDE and Nonhomogeneous BC

When we encounter nonhomogeneous type 1, 2 or 3 BCs:

$$c_1 u_x(0, t) + c_2 u(0, t) = f_0(t), \quad c_3 u_x(0, t) + c_4 u(0, t) = f_1(t),$$

we need to introduce either the transform $\boxed{u(x, t) = v(x, t) + A(t) + B(t)x}$ or the transform $\boxed{u(x, t) = v(x, t) + A(t)x + B(t)}$ to obtain homogeneous BCs:

$$c_1 v_x(0, t) + c_2 v(0, t) = 0, \quad c_3 v_x(0, t) + c_4 v(0, t) = 0.$$

If we attempt to separate variables directly, we shall be led to the following consequences of the nonhomogeneous boundary conditions:

$$X(0)T(t) = f_0(t), \quad \text{and} \quad X(L)T(t) = f_1(t), \quad t > 0.$$

Neither these implies anything directly about $X(0)$ or $X(L)$, and as a result we do not obtain a Sturm-Liouville problem for $X(x)$.

We will illustrate how the technique works with an example.

Example 35.43. Consider the heat equation but with nonhomogeneous type 1 boundary conditions:

$$\begin{cases} u_t = ku_{xx}, & 0 < x < L, \quad t > 0 \\ u(0, t) = f_0(t), \quad u(L, t) = f_1(t), & t > 0, \\ u(x, 0) = 0, & 0 < x < L. \end{cases} \quad (35.49)$$

where $f_0(0) = f_1(0) = 0$. Write down the series solution for u .

Solution. To reduce the problem to one with homogeneous boundary conditions, let us introduce a transform $u = v + A(t) + B(t)x$ and substitute it into (35.49), we have

$$v(0, t) + A(t) = f_0(t), \quad v(L, t) + A(t) + LB(t) = f_1(t).$$

Since we require $v(0, t) = v(L, t) = 0$, hence $A(t) = f_0(t)$, $B(t) = \frac{f_1(t) - f_0(t)}{L}$ and

$$u(x, t) = v(x, t) + \left(1 - \frac{x}{L}\right)f_0(t) + \frac{x}{L}f_1(t).$$

The problem for finding the transient solution $v(x, t)$ is then

$$\begin{cases} v_t(x, t) = kv_{xx}(x, t) + F(x, t), & 0 < x < L, \quad t > 0 \\ v(0, t) = v(L, t) = 0, & t > 0, \\ v(x, 0) = 0, & 0 < x < L. \end{cases}$$

where $F(x, t) = -\left(1 - \frac{x}{L}\right)f_0'(t) - \frac{x}{L}f_1'(t)$.

This is a heat equation problem with heat source discussed above, we obtain the solution at once by setting $f(x) = 0$:

$$v(x, t) = \sum_{n=1}^{\infty} \left[\int_0^t e^{-k(n\pi/L)^2(t-\tau)} F_n(\tau) d\tau \right] \sin \frac{n\pi x}{L}$$

where the Fourier coefficients F_n are given by

$$F_n(t) = -\frac{2}{L} \int_0^L \left[\left(1 - \frac{x}{L}\right)f_0'(t) + \frac{x}{L}f_1'(t) \right] \sin \frac{n\pi x}{L} dx = \frac{2}{n\pi} [(\cos n\pi)f_1'(t) - f_0'(t)].$$

□

What is the relationship between $v(x, t)$ and $u(x, t)$?

According to Farlow [1993], $v(x, t)$ is the *transient solution* and $\left(1 - \frac{x}{L}\right)f_0(t) + \frac{x}{L}f_1(t)$ is the *steady-state solution* of the solution $u(x, t)$. For heat equation, it is meaningful to talk of steady-state solution, that is the function u for $t \gg 1$ and a transient solution is a function that is closely related to the initial condition and tends to zero as $t \gg 1$. In general, f_0 and f_1 are constant, the steady-state solution can be obtained by taking $t \rightarrow \infty$.

Example 35.44 (UCCM3003 Final Exam, Sep 2009, Q3). Consider the flow of heat through a rod of length L . Assuming that L is long enough that we may neglect any end effects. Suppose that the heat flows only in the x direction, and at one end of the rod ($x = 0$) is fixed at 0°C ; while the other end ($x = L$) is fixed at 100°C . The temperature $u(x, t)$ at position x and time t satisfies the heat equation

$$\frac{\partial u}{\partial t} = 4 \frac{\partial^2 u}{\partial x^2}.$$

(a) Is the boundary condition homogeneous? Explain your answer. (2 marks)

Solution. No. $u(L, t) = 100$ is not homogeneous. □

- (b) Find the steady state solution $u_s(x)$ of the heat flow after a very long time. (5 marks)

Solution. The steady state solution $u_s(s)$ is going to satisfy a “Laplace equation” (to be discussed in the next section)

$$\frac{\partial^2 u_s}{\partial x^2} = 0, \quad u_s(0) = 0, \quad u_s(L) = 100$$

which has the solution $u_s = \frac{100}{L}x$. □

- (c) Hence, find $u(x, t)$ given that $u(x, 0) = 100x^2/L$. (18 marks)

Solution. Let $u = \frac{100}{L}x + v$. Substituting it into the non-homogeneous problem, we have a new homogeneous problem

$$v_t = 4v_{xx}, \quad v(0, t) = 0, \quad v(L, t) = 0, \quad v(x, 0) = \frac{100x^2}{L} - \frac{100}{L}x.$$

The general solution for this type 1 BC problem is

$$v = \sum_{n=1}^{\infty} b_n e^{-\frac{4n\pi}{L}t} \sin \frac{n\pi x}{L}.$$

When $t = 0$, we have

$$\sum_{n=1}^{\infty} b_n \sin \frac{n\pi x}{L} = \frac{100}{L}(x^2 - x) \Rightarrow b_n = \frac{200}{L^2} \int_0^L (x^2 - x) \sin \frac{n\pi x}{L} dx$$

By integration by parts,

$$\begin{aligned} b_n &= \frac{200}{L^2} \left[\frac{(2 - n^2\pi^2)(-1)^n L^3 + n^2\pi^2(-1)^n L^2}{n^3\pi^3} - \frac{2L^3}{n^3\pi^3} \right] \\ &= 200 \left[\frac{2((-1)^n - 1)L + n^2\pi^2(-1)^n(1 - L)}{n^3\pi^3} \right] \end{aligned}$$

Therefore,

$$u = \frac{100}{L}x + \sum_{n=1}^{\infty} 200 \left[\frac{2((-1)^n - 1)L + n^2\pi^2(-1)^n(1 - L)}{n^3\pi^3} \right] e^{-\frac{4n\pi}{L}t} \sin \frac{n\pi x}{L}$$

□

Example 35.45 (UECM3003 Final Exam, Apr 2014, Q2(b)). Given that a general solution to the homogeneous boundary value problem

$$\begin{cases} \frac{\partial v}{\partial t} = \frac{\partial^2 v}{\partial x^2}, & 0 < x < 1, \quad t > 0, \\ v(0, t) = 0, \quad v(1, t) = 0, & t > 0, \end{cases}$$

is $v(x, t) = \sum_{n=1}^{\infty} A_n e^{-n^2\pi^2 t} \sin n\pi x$, solve the following nonhomogeneous boundary value problem

$$\begin{cases} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, & 0 < x < 1, \quad t > 0, \\ u(0, t) = 0, \quad u(1, t) = 10, & t > 0, \\ u(x, 0) = x(11 - x), & 0 < x < 1. \end{cases} \quad (13 \text{ marks})$$

Solution. Let $u = v + A(t) + B(t)x$. From the BCs, we have

$$A(t) = 0 \wedge A(t) + B(t) = 10 \Rightarrow B(t) = 10 \Rightarrow u = v + 10x.$$

The nonhomogeneous boundary value problem for v is

$$\begin{cases} \frac{\partial v}{\partial t} = \frac{\partial^2 v}{\partial x^2}, & 0 < x < 1, \ t > 0, \\ v(0, t) = v(1, t) = 0, & t > 0, \\ v(x, 0) = x(11 - x) - 10x = x(1 - x), & 0 \leq x \leq 1. \end{cases}$$

We will calculate the coefficients A_n now:

$$\begin{aligned} A_n &= 2 \int_0^1 x(1 - x) \sin n\pi x = 2 \left[\frac{(x^2 - x) \cos n\pi x}{n\pi} + \frac{(1 - 2x) \sin n\pi x}{(n\pi)^2} - \frac{2 \cos n\pi x}{(n\pi)^3} \right]_0^1 \\ &= 2 \left[-\frac{2((-1)^n - 1)}{(n\pi)^3} \right] = \frac{4(1 - (-1)^n)}{(n\pi)^3} \end{aligned}$$

The solution is $u = 10x + \sum_{n=1}^{\infty} \frac{4(1 - (-1)^n)}{(n\pi)^3} e^{-n^2\pi^2 t} \sin n\pi x$. □

35.3.7 Nonhomogeneous PDE and BCs

When both the PDE and BCs are nonhomogeneous, we need to first transform the nonhomogeneous BCs to homogeneous BCs so that it is possible for us to apply the technique of eigenfunction expansion.

In the previous sections, we have already seen that there are $3^2 = 9$ permutations of linear homogeneous BCs. There are again 9 permutations for linear nonhomogeneous BCs. It is meaningless to enumerate all of them, instead, we will only illustrate how to combine the technique mentioned in the previous sections with just an example.

Example 35.46. Explain how to find the formal series expansion for the nonhomogeneous heat equation with a mix of type 1 BC and type 3 BC below:

$$\begin{cases} u_t = ku_{xx} + F(x, t), & 0 < x < L, \ t > 0, \\ u(0, t) = g_1(t), \ u_x(L, t) + hu(L, t) = g_2(t), & t > 0, \\ u(x, 0) = h(x), & 0 \leq x \leq L. \end{cases}$$

Solution: To change these nonzero BCs to homogeneous ones, we introduce a transform $u(x, t) = v(x, t) + A(t) + B(t)x$ where $A(t)$ and $B(t)$ are chosen so that the steady-state part $A(t) + B(t)x = S(x, t)$ satisfies the BCs of the problem. In this way, the transformed problem in $v(x, t)$ will have homogeneous BCs.

Substituting $u(x, t)$ into the BCs, we have

$$\begin{aligned} S(0, t) &= A(t) = g_1(t), \\ S_x(L, t) + hS(L, t) &= B(t) + hA(t) + hB(t)L = g_2(t) \Rightarrow B(t) = \frac{g_2(t) - hg_1(t)}{1 + hL} \end{aligned}$$

So we obtain the transform

$$u(x, t) = v(x, t) + \underbrace{\left(1 - \frac{hx}{1 + hL}\right)g_1(t) + \frac{x}{1 + hL}g_2(t)}_{S(x, t)} \quad (35.50)$$

which allows us to obtain a homogeneous boundary problem of $v(x, t)$:

$$\begin{cases} v_t = kv_{xx} + F(x, t) - S_t(x, t), & 0 < x < L, \ t > 0, \\ v(0, t) = 0, \ v_x(L, t) + hv(L, t) = 0, & t > 0, \\ u(x, 0) = h(x) - S(x, 0), & 0 \leq x \leq L. \end{cases}$$

We have our new problem with homogeneous BCs. We can solve it by eigenfunction expansions which is used in nonhomogeneous PDE with homogeneous BC above.

The solution of this problem is similar to the type 3 BCs in Section 35.3.3. We have the following SL-problem (which is a special case of (35.42), where $h_1 = 0$ and $h_2 = h$):

$$X'' + \lambda X = 0, \quad X(0) = 0, \quad X'(L) + hX(L) = 0$$

The eigenvalues λ are positive and satisfy

$$\sqrt{\lambda}h \cos \sqrt{\lambda}L + \sin \sqrt{\lambda}L = 0.$$

This equation can only be solved by using numerical methods (UECM3033, Solution of Nonlinear Equations). Theoretically, if we assume the eigenvalues to be β_n^2 ($n = 1, 2, 3, \dots$) and find the corresponding eigenfunctions $X_n(x) = B_n \sin \beta_n x$, we can obtain the eigenfunctions expansion of v as

$$v(x, t) = \sum_{n=1}^{\infty} T_n(t) \sin \beta_n x.$$

After finding $T_n(t)$, we obtain a formal series solution $u(x, t)$ by substituting into (35.50). e:transform1

35.3.8 “General” Heat Equation Problem

Heat equation is a type of parabolic equations. So parabolic equations can be regarded as general heat equations. In this section, we will study an example of general heat equations, where the separation of variables method works.

Example 35.47 (UECM3003 Final Exam, Apr 2013, Q2(a)). Chemical diffusion through a thin layer is governed by the equation

$$\frac{\partial^2 C}{\partial x^2} = \frac{\partial C}{\partial t} + C,$$

where $C(x, t)$ is the concentration in moles/cm³.

- (i) Determine the order of the given differential equation and state whether it is linear or nonlinear. (2 marks)

Solution. The order is 2. It is linear because

$$\begin{aligned} & \frac{\partial^2}{\partial x^2}(aC_1 + bC_2) - \frac{\partial}{\partial t}(aC_1 + bC_2) - (aC_1 + bC_2) \\ &= a \frac{\partial^2 C_1}{\partial x^2} + b \frac{\partial^2 C_2}{\partial x^2} - a \frac{\partial C_1}{\partial t} - b \frac{\partial C_2}{\partial t} - aC_1 - bC_2 \quad \square \\ &= a \left(\frac{\partial^2 C_1}{\partial x^2} - \frac{\partial C_1}{\partial t} - C_1 \right) + b \left(\frac{\partial^2 C_2}{\partial x^2} - \frac{\partial C_2}{\partial t} - C_2 \right) = a(0) + b(0) = 0. \end{aligned}$$

- (ii) Assume the boundary conditions are

$$C(0, t) = C(1, t) = 0, \quad t > 0,$$

and the initial concentration is given by

$$C(x, 0) = f(x), \quad 0 < x < 1.$$

Use the method of separation of variables to solve for the concentration $C(x, t)$. What happens to the concentration as $t \rightarrow \infty$? You may assume the result that the eigenvalues and the corresponding eigenfunctions of the Sturm-Liouville problem

$$X'' - \lambda X = 0, \quad X(0) = X(1) = 0$$

are

$$\lambda = \lambda_n = -n^2\pi^2, \quad X = X_n = \sin n\pi x, \quad n = 1, 2, 3, \dots \quad (14 \text{ marks})$$

Solution. Let $C = TX$. Then from the PDE, we have

$$TX'' = T'X + TX \Rightarrow \frac{X''}{X} = \frac{T'}{T} + 1 = \lambda \Rightarrow \begin{cases} X'' - \lambda X = 0 \\ T' = (\lambda - 1)T \end{cases}$$

The boundary conditions give $X(0) = X(1) = 0$ and the eigenvalues and eigenfunctions are given as

$$\lambda = -n^2\pi^2, \quad X_n = \sin n\pi x, \quad n = 1, 2, 3, \dots$$

With $\lambda = -n^2\pi^2$,

$$T_n = e^{-(n^2\pi^2+1)t} \Rightarrow C(x, t) = \sum_{n=1}^{\infty} T_n X_n = \sum_{n=1}^{\infty} b_n e^{-(n^2\pi^2+1)t} \sin n\pi x.$$

From the initial concentration, we have

$$\sum_{n=1}^{\infty} b_n \sin n\pi x = f(x) \Rightarrow b_n = 2 \int_0^1 f(x) \sin n\pi x dx.$$

When $t \rightarrow \infty$ ($t > 1$),

$$|C(x, t)| = \left| e^{-t} \sum_{n=1}^{\infty} b_n e^{-n^2 \pi^2 t} \sin n\pi x \right| \leq e^{-t} \sum_{n=1}^{\infty} |b_n e^{-n^2 \pi^2 t}| \leq e^{-t} \sum_{n=1}^{\infty} \frac{|b_n|}{2n^2} \rightarrow 0$$

□

35.4 2-D and 3-D Heat Equations

For an n -D homogeneous heat equation (38.7), we can apply the separation of variable $u = T(t)U(\mathbf{x})$ and obtain

$$T'(t)U(\mathbf{x}) - kT(t)\Delta U(\mathbf{x}) = 0 \Rightarrow \frac{T'(t)}{kT(t)} = \frac{\Delta U(\mathbf{x})}{U(\mathbf{x})} = -\lambda$$

where $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and λ is some constant.

After some rearrangement, we have an ODE $T' + k\lambda T = 0$ and a *Helmholtz equation*:

$$\Delta U + \lambda U = 0. \quad (35.51)$$

This equation will always appear when we apply separation of variables to the standard heat and wave equations in this topic as well as Topics 3 and 4. In physics, Helmholtz equation describes the “fundamental shapes” of a stretched membrane. Morse and Feshbach [1953a, 627] gave a comprehensive characterisation as to where the separation of variables method works with (35.51) in various coordinate systems.

Just as we learned from 1-D problems, we need appropriate homogeneous linear boundary conditions. When $n = 2$, there are 4 pieces in the boundary. It is possible for us to impose $3^4 = 81$ possible linear boundary conditions. When $n = 3$, the possibilities increase to $3^6 = 729$. It is impossible and also useless to list all of them. Instead we will illustrate how the separation of variables work with a few examples.

When the boundary conditions are type 1, the Laplacian in (35.51) is a positive definite operator, hence the eigenvalues are all strictly positive, i.e. $\lambda > 0$. This mathematical result may seem to be no too useful for rectangular coordinate system, but for other coordinate system, this would be a useful result.

In the previous section, we have seen that non-homogeneous problems are very difficult to solve. The eigenfunction expansions and “steady-state” transform techniques can be generalised to n -D. If the boundary conditions are non-homogeneous, we always need to transform them to become homogeneous. Then the solution of nonhomogeneous linear PDE can be expressed as the sum of eigenfunctions $X_n(\mathbf{x})$

of the corresponding homogeneous problem: $u = \sum_{n=1}^{\infty} T_n(t)X_n(\mathbf{x})$. The $T_n(t)$ will be much more difficult to calculate as we have seen in Example 35.46. Therefore, we will not deal with any non-homogeneous problem in this section.

Example 35.48 (Homogeneous Type 1 BCs). Consider a 2D heat equation (35.16) over the domain $(0, L) \times (0, M)$ for time $t > 0$ with boundary conditions and initial condition

$$\begin{aligned} u(0, y, t) &= u(L, y, t) = 0, \text{ for } 0 \leq y \leq M \text{ and } t > 0 \\ u(x, 0, t) &= u(x, M, t) = 0, \text{ for } 0 \leq x \leq L \text{ and } t > 0 \\ u(x, y, 0) &= f(x, y), \text{ for } 0 < x < L, 0 < y < M. \end{aligned}$$

Find the formal series solution to the above problem.

Solution: Let $u = TXY$, then by separation of variables, we derive the following equations from (35.16) and the homogeneous boundary conditions

$$\frac{T'}{kT} = \frac{X''}{X} + \frac{Y''}{Y} = -\lambda \bigwedge X(0) = X(L) = 0 \bigwedge Y(0) = Y(M) = 0.$$

After rearranging, we obtain three ODEs

$$T' = -k\lambda T \bigwedge \frac{X''}{X} = -\lambda - \frac{Y''}{Y} = -\mu \Rightarrow \begin{cases} X'' + \mu X = 0 \\ Y'' + (\lambda - \mu)Y = 0 \end{cases}$$

Solving the Sturm-Liouville problem for X , we obtain

$$\mu = \left(\frac{m\pi}{L}\right)^2, \quad X_m(x) = A_m \sin \frac{m\pi x}{L}, \quad m = 1, 2, \dots$$

and solving the Sturm-Liouville problem for Y , we obtain

$$\lambda - \mu = \left(\frac{n\pi}{M}\right)^2, \quad Y_n(y) = B_n \sin \frac{n\pi y}{M}, \quad n = 1, 2, \dots$$

We obtain the eigenvalues and eigenfunctions for the Helmholtz equation (35.51): e:helmholtz

$$\lambda = \left(\frac{m\pi}{L}\right)^2 + \left(\frac{n\pi}{M}\right)^2, \quad X_m(x)Y_n(y) = B_{mn} \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M}, \quad m, n = 1, 2, \dots$$

and the solution of the ODE of T :

$$T_{mn} = C_{mn} \exp\left\{-k \left[\left(\frac{m\pi}{L}\right)^2 + \left(\frac{n\pi}{M}\right)^2\right] t\right\}.$$

By the principle of superposition, the general solution is

$$u(x, y, t) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} T_{mn} X_{mn} = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} b_{mn} \exp\left\{-k \left[\left(\frac{m\pi}{L}\right)^2 + \left(\frac{n\pi}{M}\right)^2\right] t\right\} \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M}$$

From the initial condition, we have

$$\sum_{m=1}^{\infty} \sum_{n=1}^{\infty} b_{mn} \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M} = f(x, y).$$

We learned “double Fourier series” in Topic 1, by applying the “orthogonal properties”, we have

$$b_{mn} = \frac{4}{LM} \int_0^L \int_0^M f(x, y) \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M} dx dy.$$

Example 35.49 (3D Homogeneous Type 1 BCs). Consider a 3D heat equation (35.17) e:heat3d over the domain $(0, L) \times (0, M) \times (0, N)$ for time $t > 0$ with boundary conditions and initial condition

$$\begin{aligned} u(0, y, z, t) &= u(L, y, z, t) = 0, \text{ for } 0 \leq y \leq M, \ 0 \leq z \leq N, \text{ and } t > 0 \\ u(x, 0, z, t) &= u(x, M, z, t) = 0, \text{ for } 0 \leq x \leq L, \ 0 \leq z \leq N \text{ and } t > 0 \\ u(x, y, 0, t) &= u(x, y, N, t) = 0, \text{ for } 0 \leq x \leq L, \ 0 \leq y \leq M \text{ and } t > 0 \\ u(x, y, z, 0) &= f(x, y, z), \text{ for } 0 < x < L, \ 0 < y < M, \ 0 < z < N. \end{aligned}$$

Find the formal series solution to the above problem.

Solution. Let $u = TXYZ$, then by separation of variables, we derive the following equations from (35.17) e:heat3d and the homogeneous boundary conditions

$$\frac{T'}{kT} = \frac{X''}{X} + \frac{Y''}{Y} + \frac{Z''}{Z} = -\lambda \bigwedge X(0) = X(L) = 0 \bigwedge Y(0) = Y(M) = 0 \bigwedge Z(0) = Z(N) = 0.$$

After rearranging, we obtain four ODEs

$$T' = -k\lambda T \bigwedge \frac{X''}{X} = -\lambda - \frac{Z''}{Z} - \frac{Y''}{Y} = -\mu \Rightarrow \begin{cases} X'' + \mu X = 0 \\ -\lambda + \mu - \frac{Z''}{Z} = \frac{Y''}{Y} = -\nu \Rightarrow Y'' + \nu Y = 0 \\ Z'' + (\lambda - \mu - \nu)Z = 0 \end{cases}$$

Solving the Sturm-Liouville problem for X , we obtain

$$\mu = \left(\frac{m\pi}{L}\right)^2, \quad X_m(x) = \sin \frac{m\pi x}{L}, \quad m = 1, 2, \dots$$

and solving the Sturm-Liouville problem for Y , we obtain

$$\nu = \left(\frac{n\pi}{M}\right)^2, \quad Y_n(y) = \sin \frac{n\pi y}{M}, \quad n = 1, 2, \dots$$

and solving the Sturm-Liouville problem for Z , we obtain

$$\lambda - \mu - \nu = \left(\frac{p\pi}{N}\right)^2, \quad Z_p(z) = \sin \frac{p\pi z}{N}, \quad p = 1, 2, \dots$$

We obtain the eigenvalues and eigenfunctions for the Helmholtz equation (35.51): [e:helmholtz](#)

$$\lambda = \left(\frac{m\pi}{L}\right)^2 + \left(\frac{n\pi}{M}\right)^2 + \left(\frac{p\pi}{N}\right)^2, \quad X_m(x)Y_n(y)Z_p(z) = b_{mnp} \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M} \sin \frac{p\pi z}{N}$$

where $m, n, p = 1, 2, \dots$. The solution of the ODE of T is

$$T_{mn} = C_{mn} \exp\{-k \left[\left(\frac{m\pi}{L}\right)^2 + \left(\frac{n\pi}{M}\right)^2 + \left(\frac{p\pi}{N}\right)^2 \right] t\}.$$

By the principle of superposition, the general solution is

$$u(x, y, z, t) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{p=1}^{\infty} b_{mnp} \exp\{-k \left[\left(\frac{m\pi}{L}\right)^2 + \left(\frac{n\pi}{M}\right)^2 + \left(\frac{p\pi}{N}\right)^2 \right] t\} \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M} \sin \frac{p\pi z}{N}.$$

From the initial condition, we have

$$\sum_{m=1}^{\infty} \sum_{n=1}^{\infty} b_{mnp} \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M} = f(x, y).$$

By applying the “orthogonal properties” of the eigenfunctions, we have

$$b_{mnp} = \frac{8}{LMN} \int_0^L \int_0^M \int_0^N f(x, y, z) \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M} \sin \frac{p\pi z}{N} dx dy dz.$$

□

In practise, the method of separation of variable which was used by engineers before World War Two has been replaced by finite element methods, a method which requires us to transform a well-posed problem to an equivalent functional form as illustrated by the following example.

Example 35.50. Write an example using FreeFem++ (<http://www.freefem.org/ff++/>) to solve the following non-homogeneous heat equation problem numerically:

$$\begin{cases} \partial_t u - k\Delta u = f & \text{on } \Omega = (0, 1)^2, \quad t > 0 \\ k\partial_{\mathbf{n}} u + \alpha(u - u_e) = 0 & \text{on } \Gamma_2 \cup \Gamma_3 \\ u = u_e & \text{on } \Gamma_1 \cup \Gamma_4 \\ u(x, y, 0) = u_0 & \text{on } \Omega \end{cases}$$

where $u_e = 4$, $k = 1$, $\alpha = 0.001$ and $f(x, y) = \begin{cases} 20, & 0.6 \leq x \leq 0.8, \quad 0.4 \leq y \leq 0.6 \\ 0, & \text{otherwise.} \end{cases}$

Solution: First, note that $\partial\Omega = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3 \cup \Gamma_4$. In order solve this problem, we need to rewrite it into a functional using first Green’s identity and apply the forward time-difference approximation:

$$\begin{aligned} & \int_{\Omega} v \partial_t u d\mathbf{x} - k \int_{\Omega} v \Delta u d\mathbf{x} = \int_{\Omega} v f d\mathbf{x} \\ \Rightarrow & \int_{\Omega} v \partial_t u d\mathbf{x} - k \left[\int_{\partial\Omega} v \partial_{\mathbf{n}} u d\mathbf{S} - \int_{\Omega} \nabla v \cdot \nabla u d\mathbf{x} \right] - \int_{\Omega} v f d\mathbf{x} = 0 \\ \Rightarrow & \int_{\Omega} \left[v \partial_t u + k \nabla v \cdot \nabla u \right] d\mathbf{x} + \int_{\partial\Omega} (-k \partial_{\mathbf{n}} u) v d\mathbf{S} - \int_{\Omega} v f d\mathbf{x} = 0 \\ \approx & \int_{\Omega} \left[v \frac{u(t + \Delta t) - u}{\Delta t} + k \nabla v \cdot \nabla u - v f \right] d\mathbf{x} + \int_{\Gamma_2 \cup \Gamma_3} \alpha(u - u_e) v d\mathbf{S} + \int_{\Gamma_1 \cup \Gamma_4} (-k \partial_{\mathbf{n}} u) v d\mathbf{S} = 0 \end{aligned}$$

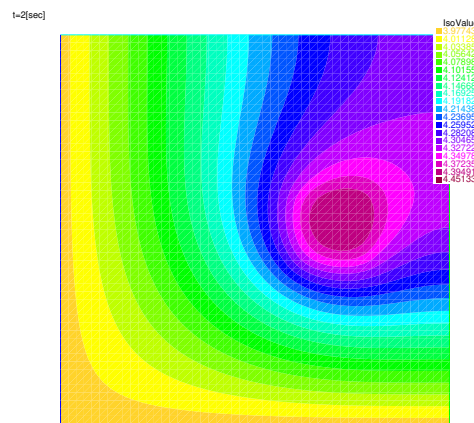
Assuming that $k\partial_{\mathbf{n}}u + \alpha(u - u_e) = 0$ on $\partial\Omega$ and we will be able to remove the last term on the left hand side above and translating the functional form into FreeFem++:

```

1  verbosity=0.;
2  real Dx=.02,Dy=.02;
3  mesh Th=square(floor(1./Dx),floor(1./Dy));
4  fespace Vh(Th,P1);
5  Vh uh,vh,uh0=10.;
6  real ue = 4., k = 1., alpha=.001, dt=0.01, Tf=2. ;
7  func f=20.*(0.6<=x & x<=.8)*(0.4<=y & y<=.6);
8  macro Grad(u)[dx(u),dy(u)]// End of Macro
9  problem HeatFunctional(uh,vh) =
10    int2d(Th)(uh*vh/dt + Grad(uh)'*Grad(vh)*k)
11    - int2d(Th)(uh0*vh/dt + f*vh) + int1d(Th,2,3)(uh*vh*alpha)
12    - int1d(Th,2,3)(ue*vh*alpha) + on(1,4,uh=ue);
13  int kk=0;
14  for (real t=0.;t<Tf;t+=dt) {
15    HeatFunctional;
16    uh0=uh;
17    if ( !(kk % 20) )
18      plot(uh,cmm="t="+t+"[sec]",dim=2,fill=true,value=true,wait=1);
19    kk+=1;
20  }
21  plot(uh,cmm="t="+Tf+"[sec]",dim=2,fill=true,value=true,ps="heat.eps");

```

The temperature distribution at $t = 2$ is numerically found to be



35.5 2-D and 3-D Laplace Equation

The 1-D Laplace equation $u_{xx} = 0$ is not a PDE according to our definition, which requires the function to have at least two variables. In the case of two variables, we have 2-D Laplace equation (35.22). Again, we can impose type 1, type 2 and type 3 boundary conditions. However, in order to apply the separation of variables, we need a pair of opposite sides to be “homogeneous”. *If the boundary condition is type 1 but not homogeneous, theoretically, we can try to find a conformal mapping to map the rectangular domain to a unit disk and apply the Poisson integral formula to find the solution* (UECM3013 Complex Analysis Topic 9). There is no “complex analysis” for 3-D Laplace equation, the more sophisticated potential theory is required to determine the existence and uniqueness of solution.

Example 35.51 (2-D, Homogeneous Type 1 BCs). Use the separation of variables to find a formal series solution for a Laplace equation (35.22) with the boundary condition

$$u(0, y) = f(y), \quad u(L, y) = g(y), \quad u(x, 0) = u(x, M) = 0. \quad (35.52)$$

Solution: By looking at the BCs, we have a pair of homogeneous type 1 BCs. Therefore, we could substitute $u = X(x)Y(y)$ into (35.22) and (35.52) to obtain two ODEs and a pair of homogeneous

boundary conditions

$$\frac{X''}{X} = -\frac{Y''}{Y} = \lambda \Rightarrow \begin{cases} X'' - \lambda X = 0, \\ Y'' + \lambda Y = 0, \quad Y(0) = Y(M) = 0. \end{cases}$$

The ODE of Y is a SL problem which we have discussed in Section 35.3.1, the eigenvalues and eigenfunctions are

$$\lambda_n = \left(\frac{n\pi}{M}\right)^2, \quad Y_n(y) = A_n \sin \frac{n\pi}{M} y, \quad n = 1, 2, \dots$$

From here, we can determine the solution of the ODE of X as

$$X_n(x) = C_n e^{\frac{n\pi}{M}x} + D_n e^{-\frac{n\pi}{M}x}.$$

By the principal of superposition

$$u = \sum_{n=1}^{\infty} X_n(x) Y_n(y) = \sum_{n=1}^{\infty} (a_n e^{\frac{n\pi}{M}x} + b_n e^{-\frac{n\pi}{M}x}) \sin \frac{n\pi y}{M} \quad (35.53)$$

where $a_n = A_n C_n$ and $b_n = A_n D_n$.

From the boundary condition at $x = 0$ and $x = L$, we have

$$\begin{aligned} \sum_{n=1}^{\infty} (a_n + b_n) \sin \frac{n\pi y}{M} &= f(y) \Rightarrow a_n + b_n = \frac{2}{M} \int_0^M f(y) \sin \frac{n\pi y}{M} dy \\ \sum_{n=1}^{\infty} (a_n e^{\frac{n\pi}{M}L} + b_n e^{-\frac{n\pi}{M}L}) \sin \frac{n\pi y}{M} &= g(y) \Rightarrow a_n e^{\frac{n\pi}{M}L} + b_n e^{-\frac{n\pi}{M}L} = \frac{2}{M} \int_0^M g(y) \sin \frac{n\pi y}{M} dy. \end{aligned}$$

Therefore,

$$\begin{aligned} a_n &= \frac{1}{M \sinh(\frac{n\pi L}{M})} \int_0^M [g(y) - e^{-\frac{n\pi L}{M}} f(y)] \sin \frac{n\pi y}{M} dy, \\ b_n &= \frac{1}{M \sinh(\frac{n\pi L}{M})} \int_0^M [e^{\frac{n\pi L}{M}} f(y) - g(y)] \sin \frac{n\pi y}{M} dy. \end{aligned}$$

Remark 35.52. Nearing [2010, Chapter 10] pointed out that we let $f(y) = 0$ and $g(y) = T_0$ be a constant. The derivative of the Fourier series (35.53) diverges. Mathematically, this means that (35.53) is not C^1 and so cannot satisfy the Laplace equation.

Example 35.53 (UECM3003 Final Exam, Apr 2015, Q2(a)). Solve the following boundary value problem by using separation of variable method

$$\begin{cases} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, & 0 < x < a, \quad 0 < y < b, \\ u(x, 0) = \frac{\partial u}{\partial y}(x, b) = 0, & 0 < x < a, \\ u(0, y) = 0, \quad u(a, y) = g(y), & 0 < y < b. \end{cases} \quad (17 \text{ marks})$$

Solution. By looking at the BCs, we have a pair of homogeneous mixed BCs. By separation of variables, we obtain two ODEs and a homogeneous boundary condition

$$\frac{X''}{X} = -\frac{Y''}{Y} = \lambda \Rightarrow \begin{cases} X'' - \lambda X = 0, \\ Y'' + \lambda Y = 0, \quad Y(0) = Y(b) = 0. \end{cases}$$

When $\lambda < 0$, the general solution of Y is

$$Y(y) = A e^{-\sqrt{-\lambda}y} + B e^{\sqrt{-\lambda}y}$$

and the solution is trivial because $A = B = 0$ according to the BCs. When $\lambda = 0$, the general solution of Y is

$$Y(y) = A + By$$

and from the BCs we know the solution is trivial

$$Y(0) = A = 0 \quad \bigwedge \quad Y'(b) = B = 0$$

When $\lambda > 0$, the general solution of Y is

$$Y(y) = A \cos \sqrt{\lambda}y + B \sin \sqrt{\lambda}y$$

and from the BCs and to have non-trivial solution, we need $B \neq 0$:

$$Y(0) = A = 0 \quad \bigwedge \quad Y'(b) = B\sqrt{\lambda} \cos \sqrt{\lambda}b = 0 \Rightarrow \sqrt{\lambda}b = n\pi - \frac{\pi}{2}, \quad n = 1, 2, \dots$$

and obtain the eigenvalues and eigenfunctions:

$$\lambda_n = \left(\frac{(2n-1)\pi}{2b}\right)^2, \quad Y_n(y) = B_n \sin \frac{(2n-1)\pi}{2b}y, \quad n = 1, 2, \dots$$

From here, we can determine the solution of the ODE of X as

$$X_n(x) = C_n e^{\frac{(2n-1)\pi}{2b}x} + D_n e^{-\frac{(2n-1)\pi}{2b}x}.$$

By the principal of superposition

$$u = \sum_{n=1}^{\infty} X_n(x)Y_n(y) = \sum_{n=1}^{\infty} \left[a_n e^{\frac{(2n-1)\pi}{2b}x} + b_n e^{-\frac{(2n-1)\pi}{2b}x} \right] \sin \frac{(2n-1)\pi}{2b}y \quad (35.54)$$

where $a_n = B_n C_n$ and $b_n = B_n D_n$.

From the boundary condition at $x = 0$ and $x = a$, we have

$$\begin{aligned} \sum_{n=1}^{\infty} (a_n + b_n) \sin \frac{(2n-1)\pi}{2b}y &= 0 \Rightarrow a_n + b_n = 0 \\ \sum_{n=1}^{\infty} \left[a_n e^{\frac{(2n-1)\pi}{2b}a} + b_n e^{-\frac{(2n-1)\pi}{2b}a} \right] \sin \frac{(2n-1)\pi}{2b}y &= g(y) \\ \Rightarrow a_n e^{\frac{(2n-1)\pi a}{2b}} + b_n e^{-\frac{(2n-1)\pi a}{2b}} &= \frac{2}{b} \int_0^b g(y) \sin \frac{(2n-1)\pi}{2b}y dy. \end{aligned}$$

Therefore,

$$a_n = -b_n = \frac{1}{b \sinh \frac{(2n-1)\pi a}{2b}} \int_0^b g(y) \sin \frac{(2n-1)\pi}{2b}y dy.$$

□

Example 35.54 (3-D, Homogeneous Type 1 BCs). Apply the separation of variables to find a formal series solution for (35.23) on $(0, \pi) \times (0, \pi) \times (0, \pi)$ with the following Dirichlet condition:

$$\begin{cases} u(x, y, z) = 0 & \text{for } x = 0, x = \pi, y = 0, y = \pi, \text{ and } z = \pi \\ u(x, y, 0) = g(x, y). \end{cases}$$

Solution. Let $u = X(x)Y(y)Z(z)$ and substitute into (35.23), we have

$$\frac{X''}{X} + \frac{Y''}{Y} = -\frac{Z''}{Z}.$$

Since the left-hand side is independent of z and the right can depend only on z , both sides must be constant:

$$\frac{X''}{X} + \frac{Y''}{Y} = -\frac{Z''}{Z} = C_1 \Rightarrow \begin{cases} Z'' + C_1 Z = 0 \\ \frac{X''}{X} = C_1 - \frac{Y''}{Y} = C_2 \end{cases} \Rightarrow \begin{cases} X'' - C_2 X = 0, \\ Y'' - (C_1 - C_2)Y = 0, \\ Z'' + C_1 Z = 0. \end{cases}$$

The homogeneous boundary conditions give

$$X(0) = X(\pi) = 0, \quad Y(0) = Y(\pi) = 0, \quad Z(\pi) = 0.$$

The SL-problem for X has the following eigenvalues and eigenfunctions:

$$C_2 = -n^2, \quad X_n = \sin nx, \quad n = 1, 2, \dots$$

The SL-problem for Y has the following eigenvalues and eigenfunctions:

$$C_1 - C_2 = -m^2, \quad Y_m = \sin my, \quad m = 1, 2, \dots$$

Finally, for $C_1 = -m^2 - n^2$, the solution to the ODE of Z is

$$\begin{cases} Z(\pi) = 0 \\ Z(z) = Ae^{\sqrt{m^2+n^2}z} + Be^{-\sqrt{m^2+n^2}z} \end{cases} \Rightarrow \begin{cases} B = -Ae^{2\sqrt{m^2+n^2}\pi} \\ Z_{mn} = \sinh \sqrt{m^2+n^2}(z-\pi). \end{cases}$$

By the principle of superposition, we obtain a general solution

$$u = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} a_{nm} \sinh \sqrt{m^2+n^2}(z-\pi) \sin nx \sin my.$$

When $z = 0$, we have

$$\sum_{n=1}^{\infty} \sum_{m=1}^{\infty} (-a_{nm}) \sinh \sqrt{m^2+n^2}\pi \sin nx \sin my = g(x, y).$$

By the orthogonality of the basis of the multiple Fourier series,

$$-a_{nm} \sinh \sqrt{m^2+n^2}\pi = \frac{4}{\pi^2} \int_0^\pi \int_0^\pi g(x, y) \sin nx \sin my dx dy.$$

□

To make sure the formal solution above is a solution to (35.23), we need to verify the convergence of the solution. See Weinberger [1995] for details.

What if all the “boundary conditions” are nonhomogeneous? In this case, we need to try to “decompose” the problem with to problems with a pair of homogeneous boundary conditions as illustrated in the following example.

Example 35.55 (2-D, Non-homogeneous Type 1 BCs). Find a formal series solution for Laplace equation (35.22) with the boundary condition

$$\begin{cases} u(0, y) = f_1(y), & u(L, y) = f_2(y), & 0 < y < M, \\ u(x, 0) = g_1(x), & u(x, M) = g_2(x), & 0 < x < L. \end{cases} \quad (35.55)$$

Solution: Let $u = v + w$ such that

$$\begin{cases} v_{xx} + v_{yy} = 0, \\ v(0, y) = v(L, y) = 0, \\ v(x, 0) = g_1(x), \quad v(x, M) = g_2(x) \end{cases} \quad \bigwedge \quad \begin{cases} w_{xx} + w_{yy} = 0, \\ w(0, y) = f_1(y), \quad w(L, y) = f_2(y), \\ w(x, 0) = w(x, M) = 0 \end{cases}$$

By using the method in Example 35.51, we can find

$$\begin{aligned} v(x, y) &= \sum_{n=1}^{\infty} \left(a_n e^{\frac{n\pi}{L}y} + b_n e^{-\frac{n\pi}{L}y} \right) \sin \frac{n\pi x}{L} \\ w(x, y) &= \sum_{n=1}^{\infty} \left(\tilde{a}_n e^{\frac{n\pi}{M}x} + \tilde{b}_n e^{-\frac{n\pi}{M}x} \right) \sin \frac{n\pi y}{M} \end{aligned}$$

where

$$\begin{aligned} a_n &= \frac{1}{L \sinh(\frac{n\pi M}{L})} \int_0^L \left[g_2(y) - e^{-\frac{n\pi M}{L}} g_1(y) \right] \sin \frac{n\pi x}{L} dx, \\ b_n &= \frac{1}{L \sinh(\frac{n\pi M}{L})} \int_0^L \left[e^{\frac{n\pi M}{L}} g_1(y) - g_2(y) \right] \sin \frac{n\pi x}{L} dx, \\ \widetilde{a}_n &= \frac{1}{M \sinh(\frac{n\pi L}{M})} \int_0^M \left[f_2(y) - e^{-\frac{n\pi L}{M}} f_1(y) \right] \sin \frac{n\pi y}{M} dy, \\ \widetilde{b}_n &= \frac{1}{M \sinh(\frac{n\pi L}{M})} \int_0^M \left[e^{\frac{n\pi L}{M}} f_1(y) - f_2(y) \right] \sin \frac{n\pi y}{M} dy, \quad n = 1, 2, \dots \end{aligned}$$

35.6 Poisson Equation

The “non-homogeneous” system corresponding to Laplace equation is the Poisson equation (defined in Topic 1 Definition 1.1.10). “Eigenfunctions expansion” generalised to the “Green’s function” method. The solution $u = G * f$ of a Poisson equation (47.4) is obtained using a slightly generalised superposition principle using the *Green’s function* G . G is characterised as a solution to $\Delta_{\mathbf{x}} G(\mathbf{x}; \xi) = \delta(\mathbf{x} - \xi)$. However, most Green’s function cannot be written down in closed form.

In some rare cases, if we find a particular solution u_p for the Poisson equation such that $\Delta u_p = f$ and then add with the general solution u_h of Laplace equation to obtain the general solution $u = u_p + u_h$ for Poisson equation. For a Poisson equation with a homogeneous type 1 BC, if f is a polynomial of degree n , we substitute a polynomial of degree $n + 2$ into it to match f . We illustrate this with an example.

Example 35.56. Find the solution of the following Poisson equation problem:

$$\begin{cases} u_{xx} + u_{yy} = -2, & 0 < x < a, \ 0 < y < b, \\ u(0, y) = u(a, y) = 0, & 0 < y < b, \\ u(x, 0) = u(x, b) = 0, & 0 < x < a. \end{cases}$$

Solution. We guess $u_p = A + Bx + Cy + Dx^2 + Exy + Fy^2$ and substitute into the Poisson equation:

$$2D + 2F = -2 \Rightarrow D + F + 1 = 0$$

Many of the values are arbitrary, we can pick $D = 0$ and $F = -1$. To maintain a pair of homogeneous boundary conditions, we need

$$\begin{cases} u_p(x, 0) = A + Bx + Dx^2 = 0 \Rightarrow A = B = D = 0 \\ u_p(x, b) = Cb + Exb - b^2 = 0 \Rightarrow E = Cb - b^2 = 0 \Rightarrow C = b. \end{cases}$$

Thus, we have $u_p = by - y^2$. By substituting $u = w + u_p$ into the above equation, we have a Laplace equation problem

$$\begin{cases} w_{xx} + w_{yy} = 0, & 0 < x < a, \ 0 < y < b, \\ w(0, y) = -u_p(0, y) = y^2 - by, \quad w(a, y) = -u_p(a, y) = y^2 - by, & 0 < y < b, \\ w(x, 0) = -u_p(x, 0) = 0, \quad w(x, b) = -u_p(x, b) = 0 & 0 < x < a. \end{cases}$$

By using the result in Example 35.55, we have

$$\begin{aligned} w(x, y) &= \sum_{n=1}^{\infty} \left(a_n e^{\frac{n\pi}{b}x} + b_n e^{-\frac{n\pi}{b}x} \right) \sin \frac{n\pi y}{b} \\ -a_n = b_n &= \frac{e^{-\frac{n\pi a}{b}} - 1}{b \sinh(\frac{n\pi a}{b})} \int_0^b (y^2 - by) \sin \frac{n\pi y}{b} dy = \frac{2(e^{-\frac{n\pi a}{b}} - 1)}{b \sinh(\frac{n\pi a}{b})} \left[\frac{((-1)^n - 1)b^3}{(n\pi)^3} \right] \\ \therefore u(x, y) &= by - y^2 + \sum_{k=1}^{\infty} \frac{8b^2}{\pi^3} \left[\frac{e^{-\frac{(2k-1)\pi a}{b}} - 1}{(2k-1)^3 \sinh(\frac{(2k-1)\pi a}{b})} \right] \sinh \frac{(2k-1)\pi x}{b} \sin \frac{(2k-1)\pi y}{b} \end{aligned}$$

□

The above solution is not very informative because it is difficult to observe how what is the temperature distribution from the infinite series. In practise, numerical methods based on functional analysis theory are required and provide better information as illustrated below.

Example 35.57. Write a script to solve the following problem numerically

$$\begin{cases} -\Delta u = f & \text{on } \Omega = (0, 1)^2, \\ u = g & \text{on } \partial\Omega. \end{cases}$$

where $f(x, y) = 1$ and $g(x, y) = 0$.

Solution: We need to transform the above problem to an equivalent weak form [Brenner and Ridgway Scott, 2008, Chapter 5] in order by using first Green's identity (Calculus III) to apply the FEM software such as FreeFEM++:

$$\begin{aligned} \int_{\Omega} v(-\Delta u) d\mathbf{x} &= \int_{\Omega} f v d\mathbf{x} \Rightarrow \int_{\partial\Omega} v \left(-\frac{\partial u}{\partial \mathbf{n}} \right) d\mathbf{S} - \int_{\Omega} \nabla v \cdot (-\nabla u) d\mathbf{x} = \int_{\Omega} f v d\mathbf{x} \\ &\Rightarrow \int_{\Omega} \nabla v \cdot \nabla u d\mathbf{x} - \int_{\Omega} f v d\mathbf{x} - \int_{\partial\Omega} v \frac{\partial u}{\partial \mathbf{n}} d\mathbf{S} \\ &\Rightarrow \int_{\Omega} \nabla v \cdot \nabla u d\mathbf{x} - \int_{\Omega} f v d\mathbf{x} \end{aligned}$$

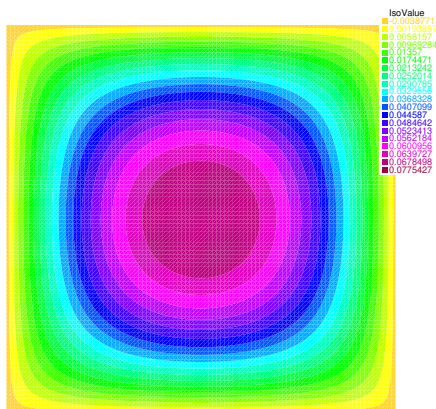
where $f(\mathbf{x}) = 1$, v is a test function on Ω such that $v(\partial\Omega) = 0$ and in 2-D, $\nabla v = (\frac{\partial v}{\partial x}, \frac{\partial v}{\partial y})$.

```

1  verbosity=0.;
2  real Dx=.01,Dy=.01;
3  mesh Th=square(floor(1./Dx),floor(1./Dy));
4  fespace Vh(Th,P1);
5  Vh uh,vh;
6  func f = 1.;
7  func g = 0.;
8  macro Grad(u) [dx(u),dy(u)]//
9  solve Poisson(uh,vh) =
10      int2d(Th) (Grad(uh)'*Grad(vh)) // bilinear form
11      - int2d(Th) (f*vh) // linear form
12      + on(1,2,3,4,uh=g); // Dirichlet BC
13  plot(uh,dim=2,fill=true,value=true,boundary=false,ps="pois.eps");

```

The distribution looks like



We can see the pattern of symmetry here but not from the formal series solution we obtain in Example 35.56.

35.7 Vibrating String — 1-D Wave Equation

Similar to the study of 1-D heat equation in Section 35.3, the problems related to the 1-D wave equation can be classified into homogeneous conditions and nonhomogeneous conditions. The solution process is more or less the same as heat equation except that T is governed by a second order ODE instead of first order equation (36.43). In the following subsections, we will assume the *initial conditions* to be

$$u(x, 0) = f(x) \quad \text{and} \quad \frac{\partial u}{\partial t}(x, 0) = g(x), \quad \text{for } x \in (0, L) \quad (35.56)$$

unless otherwise stated.

35.7.1 (Homogeneous) Type 1 BCs

Let us investigate how to find u for a wave equation (35.25) subject to type 1 BCs for both end points:

$$u(0, t) = 0 \quad \text{and} \quad u(L, t) = 0, \quad \text{for all } t > 0 \quad (35.57)$$

and to the initial conditions (35.56). This problem models a string with the ends held fixed for all time and the initial conditions give the initial shape of the string $f(x)$ and its initial velocity $g(x)$.

Example 35.58. Apply the separation of variables to find the general solution for the boundary value problem (35.25) and (38.8).

Solution:

Step 1: Separating Variables in (35.25) and (38.8)

Substitute $u(x, t) = X(x)T(t)$ into (35.25) to obtain

$$\frac{T''}{c^2 T} = \frac{X''}{X} = -\lambda \quad (35.58)$$

Thus we have two sets of ODE:

$$X'' + \lambda X = 0 \quad (35.59)$$

and

$$T'' + \lambda c^2 T = 0. \quad (35.60)$$

Step 2: Solve the associated SL problem

Our next move is to separate the variables in the boundary conditions (38.8):

$$X(0)T(t) = 0 \quad \text{and} \quad X(L)T(t) = 0, \quad \forall t > 0.$$

If $X(0) \neq 0$ or $X(L) \neq 0$, then $T(t) = 0$ and $u(x, t) = 0$, for all t . To avoid this trivial solution, we set $X(0) = 0$ and $X(L) = 0$, so we arrive at the SL problem for X :

$$X'' + \lambda X = 0, \quad X(0) = 0, \quad X(L) = 0. \quad (35.61)$$

According to what we learned in Examples 35.27, 35.28 and 35.29, we obtain the eigenvalues and eigenfunctions:

$$\lambda = \left(\frac{n\pi}{L}\right)^2, \quad X = X_n = B_n \sin \frac{n\pi x}{L}, \quad n = 1, 2, \dots$$

Step 3: Solve for $T(t)$

We now go back to (35.60) and substitute $\lambda = \left(\frac{n\pi}{L}\right)^2$ and get

$$T'' + \left(c \frac{n\pi}{L}\right)^2 T = 0.$$

The general solution of this ODE is

$$T = T_n(t) = A_n \cos \frac{cn\pi t}{L} + C_n \sin \frac{cn\pi t}{L}, \quad n = 1, 2, \dots$$

Step 4: Obtain the General Solution

Combining the solutions for X and T , we obtain an infinite set of product solutions of (35.25), all satisfying the boundary conditions (38.8):

$$u_n(x, t) = \sin \frac{n\pi x}{L} \left(a_n \cos \frac{n\pi ct}{L} + b_n \sin \frac{n\pi ct}{L} \right), \quad n = 1, 2, \dots$$

These are the *normal modes* of the wave equation.

Since all the normal modes satisfy the linear and homogeneous equation (35.25) and boundary conditions (38.8), by the *superposition principle*, any linear combination will also solve (35.25) and (38.8), hence

$$u(x, t) = \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi ct}{L} + b_n \sin \frac{n\pi ct}{L} \right) \sin \frac{n\pi x}{L} \quad (*)$$

is a solution of the boundary value problem (35.25) and (38.8) where $a_n = B_n A_n$ and $b_n = B_n C_n$ are some constants.

Example 35.59. Find u for (35.25) subject to the boundary conditions (38.8) and the initial conditions (35.56) by determining the coefficients of the general solution. Express the final expression of u in terms of the functions f and g .

Solution. The initial conditions (35.56) imply that

$$\begin{aligned} f(x) &= \sum_{n=1}^{\infty} a_n \sin \frac{n\pi x}{L} & \Rightarrow a_n &= \frac{2}{L} \int_0^L f(x) \sin \frac{n\pi x}{L} dx \\ g(x) &= \sum_{n=1}^{\infty} \frac{b_n n\pi c}{L} \sin \frac{n\pi x}{L} & \Rightarrow b_n &= \frac{2}{n\pi c} \int_0^L g(x) \sin \frac{n\pi x}{L} dx. \end{aligned}$$

where $0 < x < L$.

By apply the product rule of trigonometric functions to (*),

$$\begin{aligned} \sin \frac{n\pi x}{L} \cos \frac{n\pi ct}{L} &= \frac{1}{2} \left[\sin \frac{n\pi}{L} (x + ct) + \sin \frac{n\pi}{L} (x - ct) \right] \\ \sin \frac{n\pi x}{L} \sin \frac{n\pi ct}{L} &= \frac{1}{2} \left[\cos \frac{n\pi}{L} (x - ct) - \cos \frac{n\pi}{L} (x + ct) \right] = \frac{n\pi}{2L} \int_{x-ct}^{x+ct} \sin \frac{n\pi z}{L} dz \end{aligned}$$

we obtain

$$\begin{aligned} u(x, t) &= \sum_{n=1}^{\infty} a_n \sin \frac{n\pi x}{L} \cos \frac{n\pi ct}{L} + \sum_{n=1}^{\infty} b_n \sin \frac{n\pi x}{L} \sin \frac{n\pi ct}{L} \\ &= \sum_{n=1}^{\infty} \frac{a_n}{2} \left[\sin \frac{n\pi}{L} (x + ct) + \sin \frac{n\pi}{L} (x - ct) \right] + \sum_{n=1}^{\infty} \frac{b_n n\pi}{2L} \int_{x-ct}^{x+ct} \sin \frac{n\pi z}{L} dz \\ &= \frac{1}{2} \sum_{n=1}^{\infty} a_n \sin \frac{n\pi}{L} (x + ct) + \frac{1}{2} \sum_{n=1}^{\infty} a_n \sin \frac{n\pi}{L} (x - ct) + \frac{1}{2c} \int_{x-ct}^{x+ct} \left[\sum_{n=1}^{\infty} \frac{b_n n\pi c}{L} \sin \frac{n\pi z}{L} \right] dz \\ &= \frac{1}{2} [f(x + ct) + f(x - ct)] + \frac{1}{2c} \int_{x-ct}^{x+ct} g(z) dz. \end{aligned}$$

□

The final expression of u in terms of functions f and g in the above example is called *d'Alembert solution*. It has interesting geometric interpretations in terms of travelling waves. It can also be obtained using the *method of characteristics*.

Remark 35.60. So far, we never prove the convergence of the formal series solution. The proof of convergence requires the use of Parseval equality (from Parseval Theorem) as mentioned in Salsa [2008, 24].

Example 35.61 (UCCM3003 Final Exam, May 2010, Q2(c)). The displacement $u = u(x, t)$ of a vibrating string of length π which is released from rest with an initial displacement $u(x, 0) = 3 \sin(5x)$ is governed by the following initial-boundary value problem:

$$\begin{aligned} u_{tt} &= 4u_{xx}, & 0 < x < \pi, \quad t > 0, \\ u(0, t) &= u(\pi, t) = 0, & t > 0, \\ u(x, 0) &= 3 \sin(5x), & 0 < x < \pi, \\ u_t(x, 0) &= 0, & 0 < x < \pi. \end{aligned}$$

Use the method of separation of variables to find the solution $u(x, t)$. (13 marks)

Solution. This is a **homogeneous problem with type 1 BC** and we can apply the separation of variables to obtain

an ODE of T :

$$T'' + \lambda T = 0$$

and an SL problem of X :

$$X'' + \lambda X = 0, \quad X(0) = 0, \quad X(\pi) = 0.$$

The eigenvalues and eigenfunctions of the above SL problem is

$$\lambda = \lambda_n = n^2, \quad X = X_n(x) = B_n \sin nx, \quad n = 1, 2, \dots$$

The general solution of the ODE of T is

$$T = T_n(t) = A_n \cos 2nt + C_n \sin 2nt.$$

Therefore, by applying the principle of superposition, the general solution u is obtained:

$$u = \sum_{n=1}^{\infty} T_n X_n = \sum_{n=1}^{\infty} (a_n \cos 2nt + b_n \sin 2nt) \sin nx.$$

From the initial conditions, we have

$$\begin{aligned} \sum_{n=1}^{\infty} a_n \sin nx &= 3 \sin(5x), \\ \sum_{n=1}^{\infty} 2nb_n \sin nx &= 0. \end{aligned}$$

This implies that

$$\begin{aligned} b_n &= 0, \quad n = 1, 2, \dots \\ a_5 &= 3, \\ a_n &= 0, \quad n \neq 5 \end{aligned}$$

and so $u = 3 \cos 10t \sin 5x$. □

Example 35.62. Show that the solution of the following boundary value problem

$$\begin{aligned} u_{tt} - a^2 u_{xx} &= 0, & 0 < x < L, \quad t > 0, \\ u(0, t) &= u(L, t) = 0, & t > 0, \\ u(x, 0) &= \begin{cases} \frac{F_0(L-h)x}{T_0 L}, & 0 \leq x \leq h, \\ \frac{F_0 h}{T_0 L}(L-x), & h \leq x \leq L, \end{cases} \\ u_t(x, 0) &= 0, & 0 < x < L. \end{aligned}$$

has a (formal series) solution

$$u = \frac{2F_0 L}{T_0 \pi^2} \sum_{n=1}^{\infty} \frac{1}{n^2} \sin \frac{n\pi x}{L} \cos \frac{an\pi t}{L}.$$

Solution. This is a 1-D wave equation with type 1 boundary, according to Example 35.58, we can obtain the general solution eg: rect:wave1d:type1

$$u = \sum_{n=1}^{\infty} (a_n \cos \frac{n\pi at}{L} + b_n \sin \frac{n\pi at}{L}) \sin \frac{n\pi x}{L}.$$

From the initial conditions, we have

$$\sum_{n=1}^{\infty} a_n \sin \frac{n\pi x}{L} = \begin{cases} \frac{F_0(L-h)x}{T_0 L}, & 0 < x \leq h, \\ \frac{F_0 h}{T_0 L}(L-x), & h \leq x < L, \end{cases}$$

and

$$\sum_{n=1}^{\infty} \frac{n\pi a b_n}{L} \sin \frac{n\pi x}{L} = 0$$

The above equation implies $b_n = 0$ for $n = 1, 2, \dots$. To find a_n , we apply the property of Fourier

series from Topic 1:

$$\begin{aligned} a_n &= \frac{2}{L} \left[\int_0^h \frac{F_0(L-h)x}{T_0L} \sin \frac{n\pi x}{L} dx + \int_h^L \frac{F_0h}{T_0L} (L-x) \sin \frac{n\pi x}{L} dx \right] \\ &= \frac{2}{L} \left[\frac{F_0(L-h)}{T_0L} \left(\frac{L^2}{n^2\pi^2} \sin \frac{n\pi h}{L} - \frac{Lh}{n\pi} \cos \left(\frac{n\pi h}{L} \right) \right) \right. \\ &\quad \left. \frac{F_0h}{T_0L} \left(\frac{L(L-h)}{n\pi} \cos \frac{n\pi h}{L} + \frac{L^2}{n^2\pi^2} \sin \frac{n\pi h}{L} \right) \right] = \frac{2}{L} \left[\frac{F_0L^2}{T_0n^2\pi^2} \sin \frac{n\pi h}{L} \right] \end{aligned}$$

Hence,

$$u(x, t) = \sum_{n=1}^{\infty} \frac{2F_0L}{T_0n^2\pi^2} \sin \frac{n\pi x}{L} \cos \frac{n\pi at}{L}$$

□

35.7.2 (Homogeneous) Type 2 BCs

Similar to Section 35.3.2, we consider type 2 boundary conditions for both end points (35.38).

Example 35.63. Write down the formal series solution for the problem (35.25), (35.38) and (35.56) using the results from Section 35.3.2.

Solution. When $\lambda = 0$, $X_0(x) = A_0$ and the solution to the ODE of T is

$$T(t) = C_0 + D_0t.$$

When $\lambda = (\frac{n\pi}{L})^2 > 0$, $n = 1, 2, \dots$, $X_n(x) = A_n \cos \frac{n\pi x}{L}$ and the solution to the ODE of T is

$$T(t) = C_n \cos \frac{n\pi ct}{L} + D_n \sin \frac{n\pi ct}{L}.$$

Therefore, the general solution is

$$u(x, t) = a_0 + b_0t + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi ct}{L} + b_n \sin \frac{n\pi ct}{L} \right) \cos \frac{n\pi x}{L}$$

where $a_n = A_n C_n$ and $b_n = A_n D_n$, $n = 0, 1, 2, \dots$.

From the initial condition, we have can determined the coefficients of the general solution

$$\begin{aligned} a_0 + \sum_{n=1}^{\infty} a_n \cos \frac{n\pi x}{L} &= f(x) \Rightarrow \begin{cases} a_0 = \frac{1}{L} \int_0^L f(x) dx, \\ a_n = \frac{2}{L} \int_0^L f(x) \cos \frac{n\pi x}{L} dx, & n > 0, \end{cases} \\ b_0 + \sum_{n=1}^{\infty} \frac{n\pi cb_n}{L} \cos \frac{n\pi x}{L} &= g(x) \Rightarrow \begin{cases} b_0 = \frac{1}{L} \int_0^L g(x) dx, \\ b_n = \frac{2}{n\pi c} \int_0^L g(x) \cos \frac{n\pi x}{L} dx, & n > 0. \end{cases} \end{aligned}$$

□

35.7.3 (Homogeneous) Type 3 BCs

Similar to Section 35.3.3, the solution of the wave equation with type 3 boundary conditions for both end points

$$(u_x - h_1 u) \Big|_{x=0} = (u_x + h_2 u) \Big|_{x=L} = 0, \quad t > 0 \quad (35.62)$$

are complicated. Here, $h_1, h_2 > 0$ are constants.

Example 35.64. Write down the formal series solution for the problem (35.25), (35.62) and (35.56) using the results from Section 35.3.3.

Solution: By applying the results from Section 35.3.3, it is not difficult to obtain the general solution

$$u(x, t) = \sum_{n=1}^{\infty} (a_n \cos \beta_n c t + b_n \sin \beta_n c t) \left(\cos \beta_n x + \frac{h_1}{\beta_n} \sin \beta_n x \right).$$

From the initial conditions, we have

$$\sum_{n=1}^{\infty} a_n \left(\cos \beta_n x + \frac{h_1}{\beta_n} \sin \beta_n x \right) = f(x) \Rightarrow a_n = \frac{\int_0^L f(x) \left[\cos \beta_n x + \frac{h_1}{\beta_n} \sin \beta_n x \right] dx}{\int_0^L \left[\cos \beta_n x + \frac{h_1}{\beta_n} \sin \beta_n x \right]^2 dx}$$

$$\sum_{n=1}^{\infty} b_n \beta_n c \left(\cos \beta_n x + \frac{h_1}{\beta_n} \sin \beta_n x \right) = g(x) \Rightarrow b_n = \frac{\int_0^L g(x) \left[\cos \beta_n x + \frac{h_1}{\beta_n} \sin \beta_n x \right] dx}{\beta_n c \int_0^L \left[\cos \beta_n x + \frac{h_1}{\beta_n} \sin \beta_n x \right]^2 dx}$$

Note the expression after integrating $\int_0^L \left[\cos \beta_n x + \frac{h_1}{\beta_n} \sin \beta_n x \right]^2 dx$ is complicated because β_n satisfy a complicated formula $\tan \beta_n L = \frac{\beta_n(h_1 + h_2)}{\beta_n^2 - h_1 h_2}$.

35.7.4 (Homogeneous) Mixed BCs

Again, it is possible for us to consider a mixed of type 1, type 2, type 3 boundary conditions as in Section 35.3.4. Due to the complexity as demonstrated in subsection 35.7.3, we will limit ourselves to the following mixed boundary conditions:

$$u(0, t) = \frac{\partial u}{\partial x}(L, t) = 0, \quad t > 0. \quad (35.63)$$

Example 35.65. Find u for (35.25) subject to the boundary conditions (35.63) and the initial conditions (35.56).

Solution. The problem is homogeneous and by using the separation of variables, we obtain an ODE of T :

$$T'' + \lambda c^2 T = 0$$

and a Sturm-Liouville problem of X :

$$X'' + \lambda X = 0, \quad X(0) = X'(L) = 0.$$

When $\lambda < 0$, the general solution of X is

$$X(x) = Ae^{\sqrt{-\lambda}x} + Be^{-\sqrt{-\lambda}x}.$$

From

$$X(0) = A + B = 0$$

$$X'(L) = \sqrt{-\lambda}Ae^{\sqrt{-\lambda}L} - \sqrt{-\lambda}Be^{-\sqrt{-\lambda}L} = 0$$

we obtain $A = B = 0$ and X is trivial since $\sqrt{-\lambda}L \neq 0$ and

$$\begin{vmatrix} 1 & 1 \\ \sqrt{-\lambda}e^{\sqrt{-\lambda}L} & -\sqrt{-\lambda}e^{-\sqrt{-\lambda}L} \end{vmatrix} = -\sqrt{-\lambda}(e^{-\sqrt{-\lambda}L} - e^{\sqrt{-\lambda}L}) = \sqrt{-\lambda} \sinh(\sqrt{-\lambda}L) \neq 0$$

When $\lambda = 0$, the general solution of X is

$$X(x) = A + Bx.$$

From

$$X(0) = A = 0$$

$$X'(L) = B = 0$$

we find that X is trivial.

When $\lambda > 0$, the general solution of X is

$$X(x) = A \cos \sqrt{\lambda}x + B \sin \sqrt{\lambda}x.$$

From

$$X(0) = A = 0$$

$$X'(L) = B\sqrt{\lambda} \cos \sqrt{\lambda}L = 0$$

we obtain the eigenvalues and eigenfunctions

$$\sqrt{\lambda}L = \frac{(2n+1)\pi}{2} \Rightarrow \lambda = \left(\frac{(2n+1)\pi}{2L}\right)^2 \quad \bigwedge \quad X = X_n(x) = B_n \sin \frac{(2n+1)\pi}{2L}x$$

where $n = 0, 1, 2, \dots$. The general solution to the ODE of T is

$$T_n(t) = C_n \cos \frac{(2n+1)\pi}{2L}ct + D_n \sin \frac{(2n+1)\pi}{2L}ct.$$

Therefore, the formal series solution for u is

$$u = \sum_{n=0}^{\infty} \left[a_n \cos \frac{(2n+1)\pi}{2L}ct + b_n \sin \frac{(2n+1)\pi}{2L}ct \right] \sin \frac{(2n+1)\pi}{2L}x$$

where $a_n = B_n C_n$ and $b_n = B_n D_n$.

From the initial conditions,

$$\sum_{n=0}^{\infty} a_n \sin \frac{(2n+1)\pi}{2L}x = f(x),$$

$$\sum_{n=0}^{\infty} \frac{(2n+1)\pi c b_n}{2L} \sin \frac{(2n+1)\pi}{2L}x = g(x),$$

we obtain

$$a_n = \frac{2}{L} \int_0^L f(x) \sin \frac{(2n+1)\pi}{2L}x dx,$$

$$b_n = \frac{2L}{(2n+1)\pi c} \times \frac{2}{L} \int_0^L g(x) \sin \frac{(2n+1)\pi}{2L}x dx = \frac{4}{(2n+1)\pi c} \int_0^L g(x) \sin \frac{(2n+1)\pi}{2L}x dx$$

for $n = 0, 1, 2, \dots$. □

35.7.5 Nonhomogeneous PDE and Homogeneous BCs

When an external force is assert on the string, we have a *forced vibration* and the equation of *forced vibrations of the string* can be reduced to a nonhomogeneous wave equation after properly scaling:

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = F(x, t), \quad 0 < x < L, \quad t > 0. \quad (35.64)$$

Example 35.66. Find the formal series solution of the well-posed problem with (35.64), homogeneous type 1 BCs (35.29) and zero initial conditions (35.56) where $f(x) = g(x) = 0$.

Solution: We have homogeneous type 1 BCs, so the eigenfunction expansion is

$$u = \sum_{n=1}^{\infty} T_n(t) \sin \frac{n\pi x}{L}.$$

Substituting into the PDE, we have

$$\sum_{n=1}^{\infty} T_n''(t) \sin \frac{n\pi x}{L} + c^2 \left(\frac{n\pi}{L}\right)^2 \sum_{n=1}^{\infty} T_n(t) \sin \frac{n\pi x}{L} = \sum_{n=1}^{\infty} F_n(t) \sin \frac{n\pi x}{L}$$

where $F_n(t) = \frac{2}{L} \int_0^L F(x, t) \sin \frac{n\pi x}{L} dx$. From here, together with the initial conditions, we obtain a

sequence of ODEs

$$T_n''(t) + \left(\frac{cn\pi}{L}\right)^2 T_n(t) = F_n(t), \quad T_n(0) = 0, \quad T_n'(0) = 0$$

Using Laplace transform, we are able to obtain

$$T_n(t) = \mathcal{L}^{-1}\left\{\frac{F_n(s)}{s^2 + \left(\frac{cn\pi}{L}\right)^2}\right\} = \frac{L}{cn\pi} \int_0^t F_n(\tau - t) \sin \frac{cn\pi\tau}{L} d\tau, \quad t \geq 0.$$

Example 35.67. Find the formal series solution of the well-posed problem in Example 35.66 with $F(x, t) = -g$, where g is some constant.

Solution. We need to calculate

$$F_n(t) = \frac{2}{L} \int_0^L (-g) \sin \frac{n\pi x}{L} dx = \frac{-2g}{L} \left[-\frac{L}{n\pi} \cos \frac{n\pi x}{L} \right]_0^L = \frac{2g}{n\pi} ((-1)^n - 1)$$

and by Laplace transform, we have $F_n(s) = \frac{2g((-1)^n - 1)}{n\pi s}$ and

$$T_n(t) = \mathcal{L}^{-1}\left\{\frac{\frac{2((-1)^n - 1)g}{n\pi s}}{s^2 + \left(\frac{cn\pi}{L}\right)^2}\right\} = \frac{2gL((-1)^n - 1)}{c(n\pi)^2} \int_0^t \sin \frac{cn\pi\tau}{L} d\tau = \frac{2gL^2((-1)^n - 1)}{c^2(n\pi)^3} \left[1 - \cos \frac{cn\pi t}{L}\right].$$

□

Remark 35.68. What happens if the initial conditions (35.56) are non-zero? In this case, we can let $u = v + w$, where v is a solution of homogeneous wave equation, type 1 BCs and (35.56) given in Examples 35.58 and 35.59 (page 788); while w is a solution of Example 35.66. Therefore,

$$u = \underbrace{\sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi ct}{L} + b_n \sin \frac{n\pi ct}{L} \right) \sin \frac{n\pi x}{L}}_v + \underbrace{\sum_{n=1}^{\infty} \frac{L}{cn\pi} \sin \frac{n\pi x}{L} \int_0^t F_n(\tau - t) \sin \frac{cn\pi\tau}{L} d\tau}_w.$$

35.7.6 Homogeneous PDE and Nonhomogeneous BCs

When we encounter a problem with a homogeneous wave equation and nonhomogeneous type 1 BCs

$$\begin{cases} u_{tt} - c^2 u_{xx} = 0, \\ u(0, t) = \mu_1(t), \quad u(L, t) = \mu_2(t), \\ u(x, 0) = f(x), \quad u_t(x, 0) = g(x), \end{cases} \quad (35.65)$$

we can use a suitable transform $u = v + A(t) + B(t)x$:

$$A(t) = \mu_1(t), \quad A(t) + B(t)L = \mu_2(t) \Rightarrow B(t) = \frac{\mu_2(t) - \mu_1(t)}{L}$$

to turn (35.65) into a problem with a homogeneous BCs just like what was done in Section 35.3.6:

$$\begin{cases} v_{tt} - c^2 v_{xx} = -\left[\mu_1''(t) + \frac{\mu_2''(t) - \mu_1''(t)}{L}x\right], \\ v(0, t) = 0, \quad v(L, t) = 0, \\ v(x, 0) = f(x) - \left[\mu_1(0) + \frac{\mu_2(0) - \mu_1(0)}{L}x\right], \quad v_t(x, 0) = g(x) - \left[\mu_1'(0) + \frac{\mu_2'(0) - \mu_1'(0)}{L}x\right]. \end{cases} \quad (35.66)$$

This is a nonhomogeneous PDE and homogeneous BCs problem and we can solve it using the method mentioned in the previous section. Note that if both the boundary conditions are type 2, we choose the transform $u = v + A(t)x + B(t)x^2$ instead.

However, the right-hand-side of the PDE in (35.66) is normally zero because μ_1 and μ_2 are constants. In this case, the problem is homogeneous and separation of variables method could be applied as demonstrated in the following examples.

Example 35.69 (UCCM3003 Final Exam, Apr 2009, Q3). Consider the following boundary value problem (BVP)

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} - 9 \frac{\partial^2 u}{\partial x^2} = 0, & 0 < x < 1, t > 0, \\ u(0, t) = 0, & t > 0, \\ u(1, t) = 2, & t > 0, \\ u(x, 0) = 3 \sin 2\pi x + 2x, & 0 < x < 1, \\ \frac{\partial u}{\partial t}(x, 0) = x(1 - x), & 0 < x < 1. \end{cases}$$

- (a) Explain if the boundary conditions for the above BVP homogeneous? (2 marks)

Solution. No. It is not homogeneous since $u(1, t) \neq 0$. □

- (b) By choosing a proper definition of $w(x, t)$, show that the above BVP can be converted to the following BVP

$$\begin{cases} \frac{\partial^2 w}{\partial t^2} - 9 \frac{\partial^2 w}{\partial x^2} = 0, & 0 < x < 1, t > 0, \\ w(0, t) = w(1, t) = 0, & t > 0, \\ w(x, 0) = 3 \sin 2\pi x, & 0 < x < 1, \\ \frac{\partial w}{\partial t}(x, 0) = x - x^2, & 0 < x < 1. \end{cases} \quad (5 \text{ marks})$$

Solution. It is a nonhomogeneous type 1 problem and we choose $u = w + A(t) + B(t)x$ such that $A(t) = 0, A(t) + B(t) = 2 \Rightarrow B(t) = 2 \Rightarrow w = u - 2x$. □

- (c) Using the method of separation of variables, show that a solution $w(x, t)$ of the equation in (b) is

$$w(x, t) = 3 \cos 6\pi t \sin 2\pi x + \sum_{n=1}^{\infty} \frac{4}{3(n\pi)^4} [1 - (-1)^n] \sin 3n\pi t \sin n\pi x. \quad (16 \text{ marks})$$

Solution. According to Example 35.58, w has a general solution □

$$w(x, t) = \sum_{n=1}^{\infty} [a_n \cos 3n\pi t + b_n \sin 3n\pi t] \sin n\pi x$$

where

$$\sum_{n=1}^{\infty} a_n \sin n\pi x = 3 \sin 2\pi x \Rightarrow a_n = 3\delta_{n2}$$

$$\sum_{n=1}^{\infty} 3n\pi b_n \sin n\pi x = x - x^2 \Rightarrow b_n = \frac{2}{3n\pi} \int_0^1 (x - x^2) \sin n\pi x dx = \frac{4}{3n\pi} \left[\frac{1 - (-1)^n}{(n\pi)^3} \right]$$

for $n = 1, 2, \dots$. Hence, the solution above is obtained. □

- (d) Hence find the solution $u(x, t)$ of the original equation. (2 marks)

Solution. $u(x, t) = 2x + 3 \cos 6\pi t \sin 2\pi x + \sum_{n=1}^{\infty} \frac{4}{3(n\pi)^4} [1 - (-1)^n] \sin 3n\pi t \sin n\pi x$ □

Example 35.70 (UCCM3003 Final Exam, Apr 2011, Q2). Consider the amplitude $u = u(x, t)$ of a vibration string of length 2 that satisfies the wave equation $\frac{\partial^2 u}{\partial t^2} = 4 \frac{\partial^2 u}{\partial x^2}$, $0 < x < 2, t > 0$. Let one end of the string be fixed at $u(0, t) = 0$ and the other end of the string be fixed at $u(2, t) = 1$. Suppose the string is released with initial amplitude $u(x, 0) = x/2$ and initial velocity $\frac{\partial u}{\partial t}(x, 0) = 2\pi \sin(\pi x) - 5\pi \sin(2\pi x)$.

- (a) With an appropriate choice of $u_s(x)$, show that $w(x, t) = u(x, t) + u_s(x)$ satisfies the following boundary value problem:

$$\begin{cases} \frac{\partial^2 w}{\partial t^2} = 4 \frac{\partial^2 w}{\partial x^2}, & 0 < x < 2, t > 0, \\ w(0, t) = w(2, t) = 0, & t > 0, \\ w(x, 0) = 0, & 0 < x < 2, \\ \frac{\partial w}{\partial t}(x, 0) = 2\pi \sin(\pi x) - 5\pi \sin(2\pi x), & 0 < x < 2. \end{cases} \quad (6 \text{ marks})$$

Solution. Let $u = w + A(t) + B(t)x$. From the BCs we have

$$A(t) = 0, \quad A(t) + 2B(t) = 1 \Rightarrow u = w + \frac{x}{2} \Rightarrow \boxed{w = u - \frac{x}{2}}.$$

Substituting $u = w + \frac{x}{2}$ into the PDE, we have $u_{tt} - 4w_{xx} = 0$. When $t = 0$,

$$u(x, 0) = w(x, 0) + \frac{x}{2} = \frac{x}{2} \Rightarrow w(x, 0) = 0$$

$$u_t(x, 0) = w_t(x, 0) = 2\pi \sin(\pi x) - 5\pi \sin(2\pi x).$$

□

- (b) Find the solution $w(x, t)$ for the boundary value problem in part (a).

(17 marks)

Solution. The general solution is

$$w(x, t) = \sum_{n=1}^{\infty} \left(a_n \cos \frac{2n\pi t}{2} + b_n \sin \frac{2n\pi t}{2} \right) \sin \frac{n\pi x}{2}$$

From $w(x, 0) = 0$, $a_n = 0$. From the other initial condition, we have

$$\sum_{n=1}^{\infty} n\pi b_n \sin \frac{n\pi x}{2} = 2\pi \sin(\pi x) - 5\pi \sin(2\pi x) \Rightarrow b_2 = 1, \quad b_4 = -\frac{5}{4}, \quad b_n = 0, \quad n \neq 2, 4.$$

The solution is $w(x, t) = \sin 2\pi t \sin \pi x - \frac{5}{4} \sin 4\pi t \sin 2\pi x$.

□

- (c) Hence, find the amplitude $u(x, t)$ of the string at any time t .

(2 marks)

Solution. $u(x, t) = \frac{x}{2} + \sin 2\pi t \sin \pi x - \frac{5}{4} \sin 4\pi t \sin 2\pi x$.

□

35.7.7 Nonhomogeneous PDE and Nonhomogeneous BCs

When the PDE and BCs are nonhomogeneous, we first introduce a suitable transform $u = v + u_s$ to obtain a homogeneous problem of v . Then we apply eigenfunction expansion. The calculations are similar to those in Section 35.3.7 but much longer because there are two initial conditions for u and u_t .

Example 35.71 (UCCM3003 Final Exam, Sep 2007, Q2). The vibration of a string is described by the following partial differential equation (PDE), where $u(x, t)$ is the amplitude of the vibration, is the distance from one end of the string and is the time.

$$\begin{cases} u_{tt} - 4u_{xx} = \sin \pi x, & 0 < x < 2, \quad t > 0, \\ u(0, t) = 0, & t > 0, \\ u(2, t) = 1, & t > 0, \\ u(x, 0) = \frac{x}{2}, & 0 \leq x \leq 2, \\ u_t(x, 0) = 4 \sin \pi x, & 0 \leq x \leq 2. \end{cases}$$

- (a) Is the wave equation above homogeneous? Explain.

(2 marks)

Solution. No. The wave equation is **not** homogeneous because the boundary condition at $x = 2$ is not zero and/or there is a forcing $\sin \pi x$ in the PDE. [2 marks] □

- (b) Show that the PDE above can be transformed into the following BVP:

$$\begin{cases} v_{tt} - 4v_{xx} = \sin \pi x, & 0 < x < 2, \quad t > 0, \\ v(0, t) = 0, & t > 0, \\ v(2, t) = 0, & t > 0, \\ v(x, 0) = 0, & 0 \leq x \leq 2, \\ v_t(x, 0) = 4 \sin \pi x, & 0 \leq x \leq 2. \end{cases} \quad (4 \text{ marks})$$

Solution. By using a transform $u = v + \frac{x}{2}$, we can transform the problem to [1 mark]

$$\begin{cases} v_{tt} - 4v_{xx} = \sin \pi x, & 0 < x < 2, t > 0 \text{ (PDE)} \\ v(0, t) = v(2, t) = 0, & t > 0 \text{ (BC)} \\ v(x, 0) = 0, v_t(x, 0) = 4 \sin \pi x, & 0 \leq x \leq 2 \text{ (IC)} \end{cases} \quad [1+1+1=3 \text{ marks}]$$

□

(c) Show that the solution of the BVP in part (b) is

$$v(x, t) = \left[\frac{2 \sin(2\pi t)}{\pi} + \frac{1 - \cos(2\pi t)}{4\pi^2} \right] \sin(\pi x)$$

by using eigenfunction expansion. (17 marks)

Solution. The problem in part (b) has a forcing function but the boundary conditions are homogeneous, therefore, we can apply the eigenfunction expansions

$$v = \sum_{n=1}^{\infty} T_n(t) \sin \frac{n\pi x}{2}. \quad [1 \text{ mark}]$$

From the PDE, we have

$$\begin{aligned} \sum_{n=1}^{\infty} T_n'' \sin \frac{n\pi x}{2} - 4 \sum_{n=1}^{\infty} \left(-\left(\frac{n\pi}{2}\right)^2\right) T_n \sin \frac{n\pi x}{2} &= \sin \pi x \\ \Rightarrow \sum_{n=1}^{\infty} (T_n'' + n^2 \pi^2 T_n) \sin \frac{n\pi x}{2} &= \sin 2\left(\frac{\pi x}{2}\right). \end{aligned} \quad [3 \text{ marks}]$$

Comparing terms, we have the following ODEs

$$\begin{cases} T_2'' + 4\pi^2 T_2 = 1 \\ T_n'' + n^2 \pi^2 T_n = 0, \quad n \neq 2 \end{cases} \quad [2 \text{ marks}]$$

From the initial conditions, we have

$$\begin{cases} T_n(0) = 0, & n = 1, 2, \dots; \\ T_2'(0) = 4, \quad T_n'(0) = 0, & n \neq 2. \end{cases} \quad [1+2=3 \text{ marks}]$$

When $n = 2$, the ODE

$$T_2'' + 4\pi^2 T_2 = 1, \quad T_2(0) = 0, \quad T_2'(0) = 4$$

has a solution below using Laplace transform (UECM2023 ODE)

$$(s^2 + 4\pi^2)T_2(s) - 4 = \frac{1}{s} \Rightarrow T_2(t) = \frac{2}{\pi} \sin 2\pi t + \frac{1}{4\pi^2}(1 - \cos 2\pi t) \quad [4 \text{ marks}]$$

When $n \neq 2$, the ODE

$$T_n'' + n^2 \pi^2 T_n = 0, \quad T_n(0) = 0, \quad T_n'(0) = 0$$

has a solution below using Laplace transform (UECM2023 ODE)

$$(s^2 + n^2 \pi^2)T_n(s) = 0 \Rightarrow T_n(t) = 0. \quad [3 \text{ marks}]$$

Hence, the solution of v is

$$v(x, t) = \left[\frac{2}{\pi} \sin 2\pi t + \frac{1 - \cos 2\pi t}{4\pi^2} \right] \sin \pi x. \quad [1 \text{ mark}]$$

□

(d) Find the amplitude of the vibration $u(x, t)$. (2 marks)

Solution. $u(x, t) = \frac{x}{2} + \left[\frac{2}{\pi} \sin 2\pi t + \frac{1 - \cos 2\pi t}{4\pi^2} \right] \sin \pi x$ [2 marks] □

Remark 35.72. What happens if the coefficient c^2 of the wave equation (35.25) is changed to a function $\rho(x)$? For example,

$$u_{tt} = \frac{K}{\rho(x)} u_{xx} - g \quad (35.67)$$

where $g = 9.80 \text{ m/s}^2$ is the magnitude of the gravitational acceleration.

According to Pang [2006, 217], this is a nonhomogeneous equation, the modified separation of variables mentioned in subsections 35.3.5 and 35.7.5 still works, but we need to solve the eigenvalue problem **numerically**. Consider the solution of the homogeneous equation corresponding to (35.67)

$$u_{tt} = \frac{K}{\rho(x)} u_{xx} \Rightarrow T''X = \frac{K}{\rho(x)} TX'' \Rightarrow \begin{cases} T'' = -\lambda KT, \\ X'' = -\lambda \rho(x)X. \end{cases}$$

Together with the boundary conditions, we obtain an SL-problem and it can be solved numerically with shooting method or finite difference method (UECM3033 Numerical Methods).

In general, it is impossible to find a solution for a variable coefficient PDE. Lewy gave a non-analytic but smooth variable coefficient linear PDE which does not have any solution [Folland, 1995]. This indicates the hope to apply separation of variables to variable coefficient second order linear PDEs are impossible in general. Separation of variables cannot even help to solve a Helmholtz equation over an L-shape domain according to <http://math.stackexchange.com/questions/420868/what-kinds-of-pde-cant-be-solved-by-separation-of-variables>.

35.8 2-D and 3-D Wave Equations

Similar to Section 35.4, for an n -D homogeneous wave equation (35.24), we can apply the separation of variable $u = T(t)X(\mathbf{x})$ and obtain

$$T''(t)X(\mathbf{x}) - c^2T(t)\Delta X(\mathbf{x}) = 0 \Rightarrow \frac{T''(t)}{c^2T(t)} = \frac{\Delta X(\mathbf{x})}{X(\mathbf{x})} = -\lambda$$

where $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and λ is some constant. After some rearrangement, we have a second order ODE $T'' + c^2\lambda T = 0$ and a Helmholtz equation (35.51).

As stated in Section 35.4, when $n = 2$, there are 4 pieces in the boundary, it is possible impose 81 possible linear boundary conditions; and when $n = 3$, the possibilities increase to 729. Again, we can only limit ourselves to a few examples (but we will give more examples compare to Section 35.4). Before solving any problems, let us summarise in a table the (homogeneous) constant density physical object of shape Ω characterised by the wave equation (35.24):

Dimension	Physical Object	Domain, Ω	$\mathbf{x}$	PDE	IC
$n = 1$	String/Spring	$(0, L)$	x	(35.25)	(35.68)
$n = 2$	Membrane	$(0, L) \times (0, M)$	(x, y)	(35.26)	(35.69)
$n = 3$	Elastic body	$(0, L) \times (0, M) \times (0, N)$	(x, y, z)	(35.27)	(35.70)

where

$$u(x, 0) = f(x), \quad u_t(x, 0) = g(x) \quad (35.68)$$

$$u(x, y, 0) = f(x, y), \quad u_t(x, y, 0) = g(x, y) \quad (35.69)$$

$$u(x, y, z, 0) = f(x, y, z), \quad u_t(x, y, z, 0) = g(x, y, z) \quad (35.70)$$

Example 35.73. Write down the general solution for 2-D wave equation (35.26) over a rectangular domain $\Omega = (0, L) \times (0, M)$ if the boundary is homogeneous type 1, i.e. $u = 0$ on $\partial\Omega$. Hence, write down the formal series solution given the initial conditions (35.69).

Solution: Let $u = TXY$, by separation of variables, we derive the following equations from (35.26) and the homogeneous boundary conditions

$$\frac{T''}{c^2T} = \frac{X''}{X} + \frac{Y''}{Y} = -\lambda \bigwedge X(0) = X(L) = 0 \bigwedge Y(0) = Y(M) = 0.$$

After rearranging, we obtain two SL-problems

$$\frac{X''}{X} = -\lambda - \frac{Y''}{Y} = -\mu \Rightarrow \begin{cases} X'' + \mu X = 0, & X(0) = X(L) = 0, \\ Y'' + (\lambda - \mu)Y = 0, & Y(0) = Y(M) = 0. \end{cases}$$

Solving the Sturm-Liouville problem for X , we obtain

$$\mu = \left(\frac{m\pi}{L}\right)^2, \quad X_m(x) = A_m \sin \frac{m\pi x}{L}, \quad m = 1, 2, \dots$$

Solving the Sturm-Liouville problem for Y , we obtain

$$\lambda - \mu = \left(\frac{n\pi}{M}\right)^2, \quad Y_n(y) = B_n \sin \frac{n\pi y}{M}, \quad n = 1, 2, \dots$$

The eigenvalues and eigenfunctions of the Helmholtz equation (35.51) are

$$\lambda = \left(\frac{m\pi}{L}\right)^2 + \left(\frac{n\pi}{M}\right)^2 = \beta_{mn}^2, \quad X_m(x)Y_n(y) = a_{mn} \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M} \quad (35.71)$$

where $m, n = 1, 2, \dots$. For the eigenvalues λ , we solve for the ODE of T :

$$T_{mn}(t) = A_{mn} \cos c \sqrt{\left(\frac{m\pi}{L}\right)^2 + \left(\frac{n\pi}{M}\right)^2} t + B_{mn} \sin c \sqrt{\left(\frac{m\pi}{L}\right)^2 + \left(\frac{n\pi}{M}\right)^2} t.$$

By using the notation in (35.71) and the principle of superposition, the general solution is

$$u(x, y, t) = \sum_{m,n=1}^{\infty} T_{mn} X_m Y_n = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} (a_{mn} \cos c \beta_{mn} t + b_{mn} \sin c \beta_{mn} t) \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M}. \quad (35.72)$$

The coefficients a_{mn} and b_{mn} , $m, n = 1, 2, \dots$ are given by the double Fourier series in Topic 1:

- $a_{mn} = \frac{4}{LM} \int_0^M \int_0^L f(x, y) \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M} dx dy;$
- $b_{mn} = \frac{4}{LMc\beta_{mn}} \int_0^M \int_0^L g(x, y) \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M} dx dy;$

Example 35.74. Write down the formal series solution for Example 35.73 if $L = M$, $f(x, y) = Axy(L - x)(L - y)$ and $g(x, y) = 0$.

Solution. We just need to determine the coefficients a_{mn} and b_{mn} :

$$b_{mn} = 0 \text{ since } g = 0$$

$$\begin{aligned} a_{mn} &= \frac{4}{L^2} \int_0^L \int_0^L Axy(L-x)(L-y) \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{L} dx dy \\ &= \frac{4A}{L^2} \left[\int_0^L x(L-x) \sin \frac{m\pi x}{L} dx \right] \left[\int_0^L y(L-y) \sin \frac{n\pi y}{L} dy \right] \\ &= \frac{4A}{L^2} \left[\frac{2(1 - (-1)^m)L^3}{(m\pi)^3} \right] \left[\frac{2(1 - (-1)^n)L^3}{(n\pi)^3} \right] \end{aligned}$$

Therefore, $\beta_{mn} = \frac{\sqrt{m^2 + n^2}\pi}{L}$, $m = 2k - 1$, $n = 2\ell - 1$ and

$$\begin{aligned} u(x, y, t) &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} a_{mn} \cos c \beta_{mn} t \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{L} \\ &= \sum_{k=1}^{\infty} \sum_{\ell=1}^{\infty} \frac{64AL^4}{\pi^6} \left[\frac{\cos \frac{c\sqrt{(2k-1)^2 + (2\ell-1)^2}\pi t}{L}}{(2k-1)^3(2\ell-1)^3} \right] \sin \frac{(2k-1)\pi x}{L} \sin \frac{(2\ell-1)\pi y}{L} \end{aligned}$$

□

Example 35.75. Find the exact solution of the following problem

$$\begin{cases} u_{tt} - (u_{xx} + u_{yy}) = 0, & \Omega = (0, 1)^2, \quad t > 0, \\ u(x, y, t) = 0, & (x, y) \in \partial\Omega, \quad t > 0, \\ u(x, y, 0) = \sin(\pi x) \sin(\pi y), & \Omega, \\ u_t(x, y, 0) = 0, & \Omega. \end{cases}$$

and write a numerical script to solve it.

Solution: According to Example [35.73](#), the coefficients of the solution is characterised by $L = M = 1$, $b_{mn} = 0$ and

$$\sum_{m=1}^{\infty} \sum_{n=1}^{\infty} a_{mn} \sin m\pi x \sin n\pi y = \sin(\pi x) \sin(\pi y) \Rightarrow a_{mn} = \delta_{m1} \delta_{n1}, \quad m, n = 1, 2, \dots$$

The exact solution is $u = \cos \sqrt{2}\pi t \sin \pi x \sin \pi y$.

To solve this problem numerically, we need to rewrite it into a functional using first Green's identity and apply the centre time-difference approximation:

$$\begin{aligned} \int_{\Omega} v \partial_{tt} u d\mathbf{x} - c^2 \int_{\Omega} v \Delta u d\mathbf{x} &= 0 \quad \text{Test function } v \text{ satisfies } v(\partial\Omega) = 0 \\ \Rightarrow \int_{\Omega} v \partial_{tt} u d\mathbf{x} - c^2 \left[\int_{\partial\Omega} v \partial_{\mathbf{n}} u d\mathbf{S} - \int_{\Omega} \nabla v \cdot \nabla u d\mathbf{x} \right] &= 0 \Rightarrow \int_{\Omega} \left[v \partial_{tt} u + c^2 \nabla v \cdot \nabla u \right] d\mathbf{x} = 0 \\ \approx \int_{\Omega} \left[v \frac{u(t+\Delta t) - 2u(t) + u(t-\Delta t)}{(\Delta t)^2} + c^2 \nabla v \cdot \nabla u \right] d\mathbf{x} &= 0 \\ \Rightarrow \int_{\Omega} \left[vu(t+\Delta t) + c^2 (\Delta t)^2 \nabla v \cdot \nabla u - 2vu + vu(t-\Delta t) \right] d\mathbf{x} &= 0 \end{aligned}$$

The numerical script using FreeFem++ is:

```

1  verbosity=0;
2  load "medit"
3  load "msh3"
4  real Dx=.02,Dy=.02;
5  mesh Th=square(floor(1./Dx),floor(1./Dy));
6  fespace Vh(Th,P1);
7  func g=0.;
8  real c=1.,dt=.01,Tf=4.;
9  Vh uh,vh,uh0=sin(pi*x)*sin(pi*y),uh1=uh0+dt*g;
10 macro Grad(u)[dx(u),dy(u)]//
11 problem drum(uh,vh) =
12   int2d(Th)(uh*vh + Grad(uh)'*Grad(vh)*(c*dt)^2*.5)
13   + int2d(Th)(Grad(uh0)'*Grad(vh)*(c*dt)^2*.5)
14   - int2d(Th)(2.*uh1*vh - uh0*vh)
15   + on(1,2,3,4,uh=0);
16 int kk=0,k=0;
17 for (real t=0.;t<Tf;t+=dt) {
18   drum;
19   uh0 = uh1;
20   uh1 = uh;
21   if ( !(kk % 20)) {
22     plot(uh,cmm="t="+t,fill=true,value=true,wait=1,dim=3,boundary=false);
23     // pour visualiser la solution avec Medit
24     mesh3 TK=movemesh23(Th,transfo=[x,y,uh/2.]);
25     medit("Solution",TK);
26     savemesh(TK,"Ondes2D."+(100000+k)+".mesh");
27     // pour visualiser la solution avec Mathematica
28     { ofstream ff("Ondes2D."+(100000+k)+".txt");
29       for (int i=0;i<Th.nt;i++){
30         for (int j=0; j <3; j++){
31           ff<<Th[i][j].x << " "<< Th[i][j].y<< " "<<uh[][Vh(i,j)]<<endl;
32           ff<<Th[i][0].x << " "<< Th[i][0].y<< " "<<uh[][Vh(i,0)]<<"\n";
33         }
34       }
35       k++;
36     }

```

```

37     kk++;
38 }

```

eg:wave2d:bc2

Example 35.76. Write down the formal series solution for 2-D wave equation (35.26) over a square domain $\Omega = (0, L) \times (0, L)$ if the initial conditions are (35.69) and the boundary is homogeneous type 2, i.e.

$$u_x(0, y, t) = u_x(L, y, t) = 0, \quad u_y(x, 0, t) = u_y(x, L, t) = 0, \quad 0 < x, y < L.$$

Solution: Let $u = TXY$, by separation of variables, we derive the following equations from (35.26) and the homogeneous boundary conditions

$$\frac{T''}{c^2 T} = \frac{X''}{X} + \frac{Y''}{Y} = -\lambda \bigwedge X'(0) = X'(L) = 0 \bigwedge Y'(0) = Y'(L) = 0.$$

After rearranging, we obtain two SL-problems

$$\frac{X''}{X} = -\lambda - \frac{Y''}{Y} = -\mu \Rightarrow \begin{cases} X'' + \mu X = 0, & X'(0) = X'(L) = 0, \\ Y'' + (\lambda - \mu)Y = 0, & Y'(0) = Y'(L) = 0. \end{cases}$$

Solving the Sturm-Liouville problem for X , we obtain

$$\mu = \left(\frac{m\pi}{L}\right)^2, \quad X_m(x) = A_m \cos \frac{m\pi x}{L}, \quad m = 0, 1, 2, \dots$$

Solving the Sturm-Liouville problem for Y , we obtain

$$\lambda - \mu = \left(\frac{n\pi}{L}\right)^2, \quad Y_n(y) = B_n \cos \frac{n\pi y}{L}, \quad n = 0, 1, 2, \dots$$

The eigenvalues and eigenfunctions of the Helmholtz equation (35.51) are

$$\lambda = \frac{(m^2 + n^2)\pi^2}{L^2} = \beta_{mn}^2, \quad X_m(x)Y_n(y) = a_{mn} \cos \frac{m\pi x}{L} \cos \frac{n\pi y}{L} \quad (35.73)$$

where $m, n = 0, 1, 2, \dots$. For the eigenvalues λ , we solve for the ODE of T :

$$\lambda = 0 : T_{00}(t) = A_{00} + B_{00}t$$

$$\lambda > 0 : T_{mn}(t) = A_{mn} \cos \frac{c\sqrt{m^2 + n^2}\pi t}{L} + B_{mn} \sin \frac{c\sqrt{m^2 + n^2}\pi t}{L}.$$

By using the notation in (35.73) and the principle of superposition, the general solution is

$$\begin{aligned} u(x, y, t) = & a_{00} + b_{00}t + \sum_{m=0}^{\infty} (a_{m0} \cos c\beta_{m0}t + b_{m0} \sin c\beta_{m0}t) \cos \frac{m\pi x}{L} \\ & + \sum_{m=0}^{\infty} (a_{0n} \cos c\beta_{0n}t + b_{0n} \sin c\beta_{0n}t) \cos \frac{n\pi y}{L} \\ & + \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} (a_{mn} \cos c\beta_{mn}t + b_{mn} \sin c\beta_{mn}t) \cos \frac{m\pi x}{L} \cos \frac{n\pi y}{L}. \end{aligned}$$

Since

$$\int_0^L \int_0^L (X_m(x)Y_n(y))^2 dx dy = \frac{L^2}{4} (1 + \delta_{m0} + \delta_{n0} + \delta_{m0}\delta_{n0}),$$

where $m, n = 0, 1, 2, \dots$. For $m, n = 1, 2, \dots$, we have

$$\begin{aligned} a_{00} &= \frac{1}{L^2} \int_0^L \int_0^L f(x, y) dx dy, & a_{m0} &= \frac{2}{L^2} \int_0^L \int_0^L f(x, y) \cos \frac{m\pi x}{L} dx dy, \\ a_{0n} &= \frac{2}{L^2} \int_0^L \int_0^L f(x, y) \cos \frac{n\pi y}{L} dx dy, & a_{mn} &= \frac{2}{L^2} \int_0^L \int_0^L f(x, y) \cos \frac{m\pi x}{L} \cos \frac{n\pi y}{L} dx dy, \\ b_{00} &= \frac{1}{L^2} \int_0^L \int_0^L g(x, y) dx dy, & b_{m0} &= \frac{2}{c\beta_{m0}L^2} \int_0^L \int_0^L g(x, y) \cos \frac{m\pi x}{L} dx dy, \\ b_{0n} &= \frac{2}{c\beta_{0n}L^2} \int_0^L \int_0^L g(x, y) \cos \frac{n\pi y}{L} dx dy, \\ b_{mn} &= \frac{2}{c\beta_{mn}L^2} \int_0^L \int_0^L g(x, y) \cos \frac{m\pi x}{L} \cos \frac{n\pi y}{L} dx dy. \end{aligned}$$

Example 35.77. Write down the formal series solution for 2-D wave equation (35.26) over a square domain $\Omega = (0, L) \times (0, L)$ if the initial conditions are (35.69) and the boundary conditions are

$$u(0, y, t) = u_x(L, y, t) = 0, \quad u(x, 0, t) = u_y(x, L, t) = 0, \quad 0 < x, y < L.$$

Solution: Let $u = TXY$, by separation of variables, we derive the following equations from (35.26) and the homogeneous boundary conditions

$$\frac{T''}{c^2T} = \frac{X''}{X} + \frac{Y''}{Y} = -\lambda \bigwedge X(0) = X'(L) = 0 \bigwedge Y(0) = Y'(L) = 0.$$

After rearranging, we obtain two SL-problems

$$\frac{X''}{X} = -\lambda - \frac{Y''}{Y} = -\mu \Rightarrow \begin{cases} X'' + \mu X = 0, & X(0) = X'(L) = 0, \\ Y'' + (\lambda - \mu)Y = 0, & Y(0) = Y'(L) = 0. \end{cases}$$

Solving the Sturm-Liouville problem for X , we obtain the following eigenvalues and eigenfunctions (similar to Example 35.65 on page 791):

$$\mu = \left(\frac{(2m-1)\pi}{2L}\right)^2, \quad X_m(x) = A_m \sin \frac{(2m-1)\pi x}{2L}, \quad m = 1, 2, \dots$$

Solving the Sturm-Liouville problem for Y , we obtain

$$\lambda - \mu = \left(\frac{(2n-1)\pi}{2L}\right)^2, \quad Y_n(y) = B_n \sin \frac{(2n-1)\pi y}{2L}, \quad n = 1, 2, \dots$$

The eigenvalues and eigenfunctions of the Helmholtz equation (35.51) are

$$\boxed{\lambda = \frac{((2m-1)^2 + (2n-1)^2)\pi^2}{4L^2} = \beta_{mn}^2}, \quad X_m(x)Y_n(y) = a_{mn} \sin \frac{(2m-1)\pi x}{2L} \sin \frac{(2n-1)\pi y}{2L} \quad (35.74)$$

where $m, n = 1, 2, \dots$. For the eigenvalues λ , we solve for the ODE of T :

$$T_{mn}(t) = A_{mn} \cos \frac{c\sqrt{(2m-1)^2 + (2n-1)^2}\pi t}{2L} B_{mn} \sin \frac{c\sqrt{(2m-1)^2 + (2n-1)^2}\pi t}{2L}.$$

By using the notation in (35.74) and the principle of superposition, the general solution is

$$u(x, y, t) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} (a_{mn} \cos c\beta_{mn}t + b_{mn} \sin c\beta_{mn}t) \sin \frac{(2m-1)\pi x}{2L} \sin \frac{(2n-1)\pi y}{2L}. \quad (35.75)$$

The coefficients are given by the initial conditions (35.69): [e:wave:init2]

$$a_{mn} = \frac{4}{L^2} \int_0^L \int_0^L f(x, y) \sin \frac{(2m-1)\pi x}{2L} \sin \frac{(2n-1)\pi y}{2L} dx dy$$

$$b_{mn} = \frac{4}{L^2 \beta_{mn} c} \int_0^L \int_0^L g(x, y) \sin \frac{(2m-1)\pi x}{2L} \sin \frac{(2n-1)\pi y}{2L} dx dy$$

We will only give an example on “non-homogeneous” problem, which in general requires many steps to solve!

Example 35.78. Write down the formal series solution for a nonhomogeneous 2-D wave equation [d-nonhomog:bc1]

$$u_{tt} - c^2(u_{xx} + u_{yy}) = F(x, y, t), \quad 0 < x < L, \quad 0 < y < M, \quad t > 0 \quad (35.76) \quad \text{[e:wave2d:nonho]}$$

with homogeneous type 1 boundary conditions and initial conditions [e:wave:init2] (35.69).

Solution: Similar to the 1D case in Section [sec:waveld:nonhomog-pde] 35.7.5, the homogeneous type 1 BCs tells us that the eigenfunction expansion is

$$u(x, y, t) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} T_{mn}(t) \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M}.$$

Expand F as a multiple Fourier series:

$$F(x, y, t) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} F_{mn}(t) \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M},$$

$$F_{mn}(t) = \frac{4}{LM} \int_0^L \int_0^M F(x, y, t) \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M} dx dy.$$

From which we derive a sequence of ODEs with β_{mn}^2 defined in [e:helm:bc1:eigenval2] (35.71):

$$T_{mn}''(t) + c^2 \beta_{mn}^2 T_{mn}(t) = F_{mn}(t)$$

$$T_{mn}(0) = \frac{4}{LM} \int_0^L \int_0^M f(x, y) \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M} dx dy$$

$$T_{mn}'(0) = \frac{4}{LM} \int_0^L \int_0^M g(x, y) \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M} dx dy.$$

After solving for $T_{mn}(t)$ from the above ODE-IVP:

$$T_{mn}(t) = \mathcal{L}^{-1} \left\{ \frac{sT_{mn}(0) + T_{mn}'(0) + F_{mn}(s)}{s^2 + c^2 \beta_{mn}^2} \right\}$$

$$= T_{mn}(0) \cos c\beta_{mn}t + \frac{T_{mn}'(0)}{c\beta_{mn}} \sin c\beta_{mn}t + \frac{1}{c\beta_{mn}} \int_0^t F_{mn}(\tau - t) \sin c\beta_{mn}\tau d\tau.$$

we can obtain our formal series solution.

We will only give one 3D example because 3D in general will be very lengthy to solve.

Example 35.79 (3-D Homogeneous Type 1 BCs). Consider a 3D heat equation [e:rect:wave3d] (35.27) over the domain $(0, L) \times (0, M) \times (0, N)$ for time $t > 0$ with initial conditions [e:wave:init3] (35.70) and the following boundary conditions:

$$u(0, y, z, t) = u(L, y, z, t) = 0, \quad \text{for } 0 \leq y \leq M, \quad 0 \leq z \leq N, \quad \text{and } t > 0,$$

$$u(x, 0, z, t) = u(x, M, z, t) = 0, \quad \text{for } 0 \leq x \leq L, \quad 0 \leq z \leq N \text{ and } t > 0,$$

$$u(x, y, 0, t) = u(x, y, N, t) = 0, \quad \text{for } 0 \leq x \leq L, \quad 0 \leq y \leq M \text{ and } t > 0.$$

Find the formal series solution to the above problem.

Solution: Let $u = TXYZ$, then by separation of variables, we derive the following equations from [e:heat3d] (35.17) and the homogeneous boundary conditions

$$\frac{T''}{c^2 T} = \frac{X''}{X} + \frac{Y''}{Y} + \frac{Z''}{Z} = -\lambda \bigwedge X(0) = X(L) = 0 \bigwedge Y(0) = Y(M) = 0 \bigwedge Z(0) = Z(N) = 0.$$

After rearranging, we obtain four ODEs

$$T'' = -c^2 \lambda T \quad \bigwedge \quad \frac{X''}{X} = -\lambda - \frac{Z''}{Z} - \frac{Y''}{Y} = -\mu \Rightarrow \begin{cases} X'' + \mu X = 0 \\ -\lambda + \mu - \frac{Z''}{Z} = \frac{Y''}{Y} = -\nu \Rightarrow Y'' + \nu Y = 0 \\ Z'' + (\lambda - \mu - \nu)Z = 0 \end{cases}$$

Solving the Sturm-Liouville problem for X , we obtain

$$\mu = \left(\frac{m\pi}{L}\right)^2, \quad X_m(x) = \sin \frac{m\pi x}{L}, \quad m = 1, 2, \dots$$

and solving the Sturm-Liouville problem for Y , we obtain

$$\nu = \left(\frac{n\pi}{M}\right)^2, \quad Y_n(y) = \sin \frac{n\pi y}{M}, \quad n = 1, 2, \dots$$

and solving the Sturm-Liouville problem for Z , we obtain

$$\lambda - \mu - \nu = \left(\frac{p\pi}{N}\right)^2, \quad Z_p(z) = \sin \frac{p\pi z}{N}, \quad p = 1, 2, \dots$$

We obtain the eigenvalues and eigenfunctions for the Helmholtz equation (35.51):

$$\lambda = \left(\frac{m\pi}{L}\right)^2 + \left(\frac{n\pi}{M}\right)^2 + \left(\frac{p\pi}{N}\right)^2 = \beta_{mnp}^2, \quad X_m(x)Y_n(y)Z_p(z) = B_{mnp} \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M} \sin \frac{p\pi z}{N} \quad (35.77)$$

where $m, n, p = 1, 2, \dots$. The solution of the ODE of T is

$$T_{mn} = C_{mn} \cos(c\sqrt{\left(\frac{m\pi}{L}\right)^2 + \left(\frac{n\pi}{M}\right)^2 + \left(\frac{p\pi}{N}\right)^2}t) + D_{mn} \sin(c\sqrt{\left(\frac{m\pi}{L}\right)^2 + \left(\frac{n\pi}{M}\right)^2 + \left(\frac{p\pi}{N}\right)^2}t)$$

By using the notation in (35.77) and the principle of superposition, the general solution is

$$u(x, y, z, t) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{p=1}^{\infty} (a_{mnp} \cos c\beta_{mnp}t + b_{mnp} \sin c\beta_{mnp}t) \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M} \sin \frac{p\pi z}{N}.$$

From the initial conditions,

$$\begin{aligned} \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} a_{mnp} \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M} \sin \frac{p\pi z}{N} &= f(x, y, z), \\ \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} c\beta_{mnp} b_{mnp} \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M} \sin \frac{p\pi z}{N} &= g(x, y, z), \end{aligned}$$

by applying the “orthogonal properties” of the eigenfunctions, we have

$$\begin{aligned} a_{mnp} &= \frac{8}{LMN} \int_0^L \int_0^M \int_0^N f(x, y, z) \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M} \sin \frac{p\pi z}{N} dx dy dz \\ b_{mnp} &= \frac{8}{c\beta_{mnp}LMN} \int_0^L \int_0^M \int_0^N g(x, y, z) \sin \frac{m\pi x}{L} \sin \frac{n\pi y}{M} \sin \frac{p\pi z}{N} dx dy dz. \end{aligned}$$

Chapter 36

Boundary-Value Problems in Cylindrical Coordinates

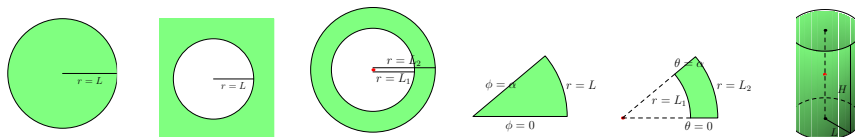
Solving BVPs in Cartesian coordinate by *separation of variables* indirectly led to Fourier series. Solving BVPs in the polar and cylindrical coordinates domains has applications in cylindrical shaped objects such as disks, pipes and optical fibres. In particular, we consider BVPs for the heat, Laplace and wave equations over disks or cylinders. Upon restating these problems in suitable coordinate systems and separating variables, we will encounter a new ODE — Bessel equation — whose solutions are called Bessel functions. The full implementation of the separation of variables method will lead us to study expansions of the solutions of heat, Laplace and wave equations problems in terms of Bessel functions in ways analogous to Fourier series expansions called *Fourier-Bessel series*.

36.1 Second Order Linear PDE in Cylindrical Coordinates

When the shapes domains are disk, wedge, annulus, cylinder, etc. (subsection [36.1.1](#)), it is more appropriate for us to use polar or cylindrical coordinates to express PDEs. The heat, Laplace and wave equations in Topic 2 are still valid on these domains. However, separation of variables will not work. To make separation of variables work, we need to formulate the Laplacian as well as heat equation, Laplace equation and wave equation in cylindrical coordinates.

36.1.1 Domains, Boundary Conditions and Initial Conditions

When the domains are of the shapes of disk, wedge, annulus, cylinder, etc. as shown below (for PDE with n variables, where $n = 1, 2$ or 3):



the PDE will have to be written in cylindrical coordinates (r, ϕ, z) for finding the solutions in terms of special functions.

Similar to Topic 2, each piece of the boundary of the domain can be imposed with a linear boundary conditions of type 1, 2 or 3. When time t occur in the PDE, initial conditions will be further imposed. For heat and wave equations problem, the initial condition is a *key* to deciding the “number of variables” required to model the problem. An initial condition is said to be *radial symmetric* if

$$u(r, 0) = f(r) \quad (36.1)$$

or

$$\frac{\partial u}{\partial t}(r, 0) = g(r) \quad (36.2)$$

where the range of r could be $0 < r < L$, $0 < a < r < b$ or $0 < L < r$. Note that f and g are depend only on the radius r but not the angle ϕ and the height z .

When there is less symmetry, the initial condition is characterised by

$$u(r, \phi, 0) = f(r, \phi) \quad (36.3)$$

or

$$\frac{\partial u}{\partial t}(r, \phi, 0) = g(r, \phi) \quad (36.4)$$

where the range of r is similar to the radial symmetric case and the range of ϕ can either be $0 \leq \phi \leq 2\pi$ or $a < \phi < b$.

When there is no symmetry, the initial condition is characterised by

$$u(r, \phi, z, 0) = f(r, \phi, z) \quad (36.5) \quad \text{e:cyl:init3a}$$

or

$$\frac{\partial u}{\partial t}(r, \phi, z, 0) = g(r, \phi, z) \quad (36.6) \quad \text{e:cyl:init3b}$$

where the range of r and ϕ are similar to the above cases and $0 < z < L$.

For circular domains, there is an additional boundary condition which is always implicitly assumed in this topic, i.e. the *periodic boundary condition*:

$$u(r, \phi) = u(r, \phi + 2\pi), \quad 0 < r < L, \quad 0 \leq \phi < 2\pi. \quad (36.7) \quad \text{e:bc:periodic}$$

For a domain that consists of the origin (or the infinity), there is an additional condition imposed on it, i.e. the *regularity condition*:

$$u \text{ is finite approaching the "limit" of the domain, i.e. } \lim_{r \rightarrow 0 \text{ or } \infty} |u| < \infty. \quad (36.8) \quad \text{e:regular}$$

36.1.2 Laplacian in Cylindrical Coordinates

Laplacian $\Delta = \nabla^2$ plays an important role in heat, Laplace and wave equations as was demonstrated in Topic 2. When we have a circular domain, such as a circle of radius 1 and a 2-D heat equation is modelled using rectangular coordinates, we have

$$\begin{cases} u_t - k(u_{xx} + u_{yy}) = 0, & x^2 + y^2 < 1, \\ u(x, y, 0) = 0, & x^2 + y^2 = 1, \\ u(x, y, t) = f(\sqrt{x^2 + y^2}) & t > 0. \end{cases}$$

After we carry out the separation of variables as in Topic 2, we will find that the BC does not allow us to derive a Sturm-Liouville problem because it offers no “separation of variables”:

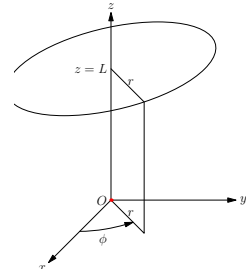
$$u(x, y, 0) = T(0)X(x)Y(y) = 0 \Rightarrow X(x)Y(\pm\sqrt{1-x^2}) = 0.$$

In order for second order PDEs with Laplacian

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}.$$

to be “separable” in circular domain, we need to find its expression in cylindrical coordinates. First, we state the relationship between Cartesian coordinates and cylindrical coordinates (Calculus III):

$$\begin{cases} x = r \cos \phi \\ y = r \sin \phi \\ z = z \end{cases} \quad \begin{cases} r = \sqrt{x^2 + y^2} \\ \phi = \arctan\left(\frac{y}{x}\right) \\ z = z \end{cases}$$



By using

$$x = r \cos \phi, \quad y = r \sin \phi, \quad r^2 = x^2 + y^2, \quad \tan \phi = \frac{y}{x}, \quad (36.9) \quad \text{e:cyl:trans}$$

and “chain rule” (Calculus II, Topic 4), the Laplacian of u in cylindrical coordinates (note that $u(x, y, z) = u(r \cos \phi, r \sin \phi, z) = \tilde{u}(r, \phi, z)$, instead of writing $\tilde{u}$, scientists normally still use u) is found to be

$$\Delta u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} + \frac{\partial^2 u}{\partial z^2} = \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} + \frac{\partial^2 u}{\partial z^2} \quad (36.10) \quad \text{e:cyl:laplacian}$$

after suitable simplification which we will illustrate below:

$$\begin{aligned} \text{(36.9)} \Rightarrow 2r \frac{\partial r}{\partial x} &= 2x \Rightarrow \frac{\partial r}{\partial x} = \frac{x}{r} = \cos \phi, \quad 2r \frac{\partial r}{\partial y} = 2y \Rightarrow \frac{\partial r}{\partial y} = \frac{y}{r} = \sin \phi; \\ \text{(36.9)} \Rightarrow \sec^2 \phi \frac{\partial \phi}{\partial x} &= -\frac{y}{x^2} = -\frac{r \sin \phi}{r^2 \cos^2 \phi} \Rightarrow \frac{\partial \phi}{\partial x} = -\frac{\sin \phi}{r} \end{aligned}$$

$$\begin{aligned}
(36.9) \Rightarrow \sec^2 \phi \frac{\partial \phi}{\partial y} &= \frac{1}{x} = \frac{1}{r \cos \phi} \Rightarrow \frac{\partial \phi}{\partial y} = \frac{\cos \phi}{r}; \\
\frac{\partial u}{\partial x} &= \frac{\partial u}{\partial r} \frac{\partial r}{\partial x} + \frac{\partial u}{\partial \phi} \frac{\partial \phi}{\partial x} + \frac{\partial u}{\partial z} \frac{\partial z}{\partial x} = \frac{\partial u}{\partial r} \cos \phi + \frac{\partial u}{\partial \phi} \left[-\frac{\sin \phi}{r} \right] \\
\frac{\partial u}{\partial y} &= \frac{\partial u}{\partial r} \frac{\partial r}{\partial y} + \frac{\partial u}{\partial \phi} \frac{\partial \phi}{\partial y} + \frac{\partial u}{\partial z} \frac{\partial z}{\partial y} = \frac{\partial u}{\partial r} \sin \phi + \frac{\partial u}{\partial \phi} \left[\frac{\cos \phi}{r} \right] \\
\frac{\partial^2 u}{\partial x^2} &= \frac{\partial}{\partial x} \left[\frac{\partial u}{\partial r} \cos \phi + \frac{\partial u}{\partial \phi} \frac{\sin \phi}{r} \right] = \frac{\partial}{\partial x} \left[\frac{\partial u}{\partial r} \cos \phi \right] + \frac{\partial}{\partial x} \left[\frac{\partial u}{\partial \phi} \frac{\sin \phi}{r} \right] \\
&= \left[u_{rr} \cos \phi - u_{r\phi} \frac{\sin \phi}{r} \right] \cos \phi + u_r \left[(-\sin \phi) \left(-\frac{\sin \phi}{r} \right) \right] - \left[u_{\phi r} \cos \phi - u_{\phi\phi} \frac{\sin \phi}{r} \right] \frac{\sin \phi}{r} + u_\phi \left[\frac{y \cdot 2x}{(x^2 + y^2)^2} \right] \\
&= u_{rr} \cos^2 \phi - u_{r\phi} \frac{\sin \phi \cos \phi}{r} + u_r \left[\frac{\sin^2 \phi}{r} \right] - u_{\phi r} \frac{\sin \phi \cos \phi}{r} + u_{\phi\phi} \frac{\sin^2 \phi}{r^2} + u_\phi \left[\frac{y \cdot 2x}{(x^2 + y^2)^2} \right] \\
\frac{\partial^2 u}{\partial y^2} &= \frac{\partial}{\partial y} \left[\frac{\partial u}{\partial r} \sin \phi + \frac{\partial u}{\partial \phi} \frac{\cos \phi}{r} \right] = \frac{\partial}{\partial y} \left[\frac{\partial u}{\partial r} \sin \phi \right] + \frac{\partial}{\partial y} \left[\frac{\partial u}{\partial \phi} \frac{\cos \phi}{r} \right] \\
&= \left[u_{rr} \sin \phi + u_{r\phi} \frac{\cos \phi}{r} \right] \sin \phi + u_r \left[(\cos \phi) \left(\frac{\cos \phi}{r} \right) \right] \\
&\quad + \left[u_{\phi r} \sin \phi + u_{\phi\phi} \frac{\cos \phi}{r} \right] \frac{\cos \phi}{r} + u_\phi \left[-\frac{x \cdot 2y}{(x^2 + y^2)^2} \right] \\
&= u_{rr} \sin^2 \phi + u_{r\phi} \frac{\sin \phi \cos \phi}{r} + u_r \left[\frac{\cos^2 \phi}{r} \right] + u_{\phi r} \frac{\sin \phi \cos \phi}{r} + u_{\phi\phi} \frac{\cos^2 \phi}{r^2} + u_\phi \left[-\frac{y \cdot 2x}{(x^2 + y^2)^2} \right]
\end{aligned}$$

Summing $\frac{\partial^2 u}{\partial x^2}$, $\frac{\partial^2 u}{\partial y^2}$ and $\frac{\partial^2 u}{\partial z^2}$ and applying trigonometric identity $\sin^2 \phi + \cos^2 \phi = 1$, we obtain (36.10). e:cyl:laplacian

36.1.3 Heat Equations

Let u be temperature (or concentration). The *heat (or diffusion) equation* $u_t - k\Delta u = 0$ we discussed in Topic 2 will have the following forms depending on the symmetry of the problem:

$$\frac{\partial u}{\partial t} - k \left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} \right) = 0, \quad t > 0; \quad (36.11) \quad \text{e:cyl:heat1d}$$

$$\frac{\partial u}{\partial t} - k \left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} \right) = 0, \quad t > 0; \quad (36.12) \quad \text{e:cyl:heat2da}$$

$$\frac{\partial u}{\partial t} - k \left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{\partial^2 u}{\partial z^2} \right) = 0, \quad t > 0; \quad (36.13) \quad \text{e:cyl:heat2db}$$

$$\frac{\partial u}{\partial t} - k \left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} + \frac{\partial^2 u}{\partial z^2} \right) = 0, \quad t > 0 \quad (36.14) \quad \text{e:cyl:heat3d}$$

where the range of r could be $0 < r < L$, $0 < a < r < b$ or $0 < L < r$; the range of ϕ could be $0 \leq \phi \leq 2\pi$ or $a < \phi < b$; and the range of z can either be $(0, H)$ or $\mathbb{R}$ depending on the “symmetry” property of the problem. With radial symmetry mentioned above, we have only one variable r and we will call (36.11) *one-dimensional* (1-D, this is not precise scientifically but we will use it for convenience). We will call both (36.12) and (36.13) two-dimensional (2-D) equations and (36.14) three-dimensional (3-D) equation. e:cyl:heat1d

For details about heat equation modellings, read the recommended text book Cengel and Ghajar [2011] by Professor Scott K. Thomas in ME 3360/5360: Heat Transfer (<http://cecs.wright.edu/~stthomas/heattransfer.html>). e:cyl:heat2da

36.1.4 Laplace and Poisson Equations

Laplace equation of “dimension” 2 can the following forms in cylindrical coordinates:

$$\Delta u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} = 0; \quad (36.15) \quad \text{e:cyl:laplace2}$$

$$\Delta u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{\partial^2 u}{\partial z^2} = 0; \quad (36.16) \quad \text{e:cyl:laplace2}$$

while Laplace equation of “dimension” 3 has the following form in cylindrical coordinates:

$$\Delta u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} + \frac{\partial^2 u}{\partial z^2} = 0. \quad (36.17) \quad \text{e:cyl:laplace3}$$

The range of r could be $0 < r < L$, $0 < a < r < b$ or $0 < L < r$; the range of ϕ could be $0 \leq \phi \leq 2\pi$ or $a < \phi < b$; and the range of z can either be $(0, H)$ or $\mathbb{R}$ depending on the “symmetry” property of the problem.

The nonhomogeneous version of Laplace equation is the *Poisson equation* $\Delta u = f(\mathbf{x})$ which describes two of the phenomena found in physics:

- When $f(\mathbf{x}) = -\rho$, a constant charge density, $u(\mathbf{x})$ describes the potential of an *electrostatic field*.
- When $f(\mathbf{x})$ is a heat source, $u(\mathbf{x})$ describes the steady state temperature due to a heat source $f(\mathbf{x})$. If $f(\mathbf{x})$ is positive at a point then heat is *generated* at that point. Negative $f(\mathbf{x})$ means that the heat is absorbed.

36.1.5 Wave Equations — Vibrating Drumhead

When the initial “wave form” on a drumhead of is “radially symmetry”, i.e. the “height” of the wave is constant for the same radius, we can use a 1-D wave equation to model the wave motion:

$$\frac{\partial^2 u}{\partial t^2} - c^2 \left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} \right) = 0, \quad t > 0. \quad (36.18) \quad \text{e:cyl:wave1d}$$

When the initial “wave form” is less symmetrical, we will need either of the 2-D equations

$$\frac{\partial^2 u}{\partial t^2} - c^2 \left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} \right) = 0, \quad t > 0; \quad (36.19) \quad \text{e:cyl:wave2da}$$

$$\frac{\partial^2 u}{\partial t^2} - c^2 \left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{\partial^2 u}{\partial z^2} \right) = 0, \quad t > 0; \quad (36.20) \quad \text{e:cyl:wave2db}$$

or a 3-D wave equation to model the wave motion:

$$\frac{\partial^2 u}{\partial t^2} - c^2 \left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} + \frac{\partial^2 u}{\partial z^2} \right) = 0, \quad t > 0. \quad (36.21) \quad \text{e:cyl:wave3d}$$

Note that the range of r could be $0 < r < L$, $0 < a < r < b$ or $0 < L < r$; the range of ϕ could be $0 \leq \phi \leq 2\pi$ or $a < \phi < b$; and the range of z can either be $(0, H)$ or $\mathbb{R}$ depending on the “symmetry” property of the problem.

A nice demonstration of the vibration of a circular drum head, characterised by [e:cyl:wave2da](#) (36.19) is found at the Joseph Gladden’s website from the Department of Physics, University of Mississippi.¹ In his demonstration, some square meshes are drawn on the membrane of the drum. The bottom of the drum is mounted on a speaker, and when the speaker is played with a sinusoidal sound signal, the membrane will vibrate and the mesh lines show the standing wave just like those of vibrating strings.

Another demonstration is from the ScienceDemonstrations from Harvard.²

36.2 Special Functions

There are two types of special functions that we need when solving PDEs in cylindrical coordinates, i.e. Bessel functions and modified Bessel functions. They are “special functions” parallel to trigonometric functions and exponential functions.

36.2.1 Bessel Equations and Bessel Functions

Bessel equation appears when we solve heat or wave equations in cylindrical coordinates. In this section, we will summarise the solution of the Bessel equation without proofs. The proofs can be found in Whittaker and Watson [1920b, Chapter XVII], Asmar [2005, 237], González-Velasco [1995, Chapter 10], Wyld [1999, Chapter 4], Koshlyakov et al. [1964], Watson [1944], Tolstov [1976], Bell [1968b].

Definition 36.1. A *Bessel equation of with parameter λ of order ν* is

$$x^2 y'' + xy' + (\lambda^2 x^2 - \nu^2)y = 0. \quad (36.22) \quad \text{e:besselp}$$

When $\lambda = 1$, we call it a *Bessel equation of order ν* :

$$x^2 y'' + xy' + (x^2 - \nu^2)y = 0. \quad (36.23) \quad \text{e:bessel}$$

Note that the range of x can either be $0 < x < L$, $0 < a < x < b < \infty$ or $0 < L < x$.

¹http://www.phy.olemiss.edu/~jgladden/other/listen_up/circular_membrane_resonances.html

²<http://sciencedemonstrations.fas.harvard.edu/>. Under the category of “Oscillations and Waves/Coupled Oscillation and Resonance”.

By a transform $t = \lambda x$, (36.22) becomes (36.23) as illustrated below:

$$\lambda \frac{dy}{dt} = \frac{dy}{dx}, \quad \lambda^2 \frac{d^2y}{dt^2} = \frac{d^2y}{dx^2} \Rightarrow t^2 \frac{d^2y}{dt^2} + t \frac{dy}{dt} + (t^2 - \nu^2)y = 0 \Leftrightarrow (36.23)$$

Therefore, in the following analysis, we will just investigate the general solution of Bessel equation of order ν (36.23).

Unlike the harmonic equation $X'' + \lambda X = 0$ we encountered in Topic 2, $x = 0$ is a regular singular point of Bessel equation (36.23). Therefore, the *method of Frobenius* from UECM2023 Topic 3, is required to determine the solution, i.e. we will need to substitute $y = \sum_{k=0}^{\infty} a_k x^{k+r}$, $a_0 \neq 0$ into (36.23) to obtain the a general solution [Asmar, 2005]

$$y(x) = a_0^{(r=\nu)} J_{\nu}(x) + a_0^{(r=-\nu)} J_{-\nu}(x)$$

Note that the above general solution is valid only when ν is *not* an integer. Here

$$J_{\nu}(x) := \sum_{k=0}^{\infty} \frac{(-1)^k}{k! \Gamma(k + \nu + 1)} \left(\frac{x}{2}\right)^{2k+\nu}$$

is called the *Bessel function of the first type (or kind)*.

To include the case when ν is an integer value, *Bessel function of the second type* (or *Weber function*) is introduced:

$$Y_{\nu}(x) := \frac{J_{\nu}(x) \cos(\nu\pi) - J_{-\nu}(x)}{\sin(\nu\pi)}.$$

where ν is non-integer. When ν is an integer value n , $Y_n(x) := \lim_{\nu \rightarrow n} Y_{\nu}(x)$. Using this definition, the general solution to Bessel equation (36.23) can be written as

$$y(x) = aJ_{\nu}(x) + bY_{\nu}(x)$$

where a and b are arbitrary constants. Note that for a fixed $\nu > 0$, $Y_{\nu}(x) \sim -\frac{1}{\pi} \Gamma(\nu) \left(\frac{x}{2}\right)^{-\nu} \rightarrow \infty$ as $x \rightarrow 0$ so it does not appear in problems with domain involving the centre $r = 0$. Here the *gamma (or factorial) function* $\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt$. It equals $(n-1)!$ when $z = n$ is an integer.

Example 36.1 (Peter V. O'Neil, p788). Write down the solution of the Bessel equation $x^2 y'' + xy' + (x^2 - \frac{1}{9})y = 0$ in terms of the Bessel function(s).

Solution. The above Bessel equation has an order $\nu = \frac{1}{3}$, hence it has a general solution of $y(x) = aJ_{1/3}(x) + bJ_{-1/3}(x)$ or $y(x) = aJ_{1/3}(x) + bY_{-1/3}(x)$. □

From the general solution of Bessel equation, the general of Bessel equation with parameter λ (36.22) can be easily obtained:

$$y(x) = aJ_{\nu}(\lambda x) + bY_{\nu}(\lambda x). \quad (36.24)$$

Before proceeding, let us recall what happens in Topic 2. From the separation of variables, we obtain at least one SL-problem (the ODE is a harmonic equation). The general solution of the SL-problem is a linear combination of sine and cosine functions. The boundary conditions are related to the zeros of sine and cosine functions.

A similar situation happens to the solution of SL-problem characterised by Bessel equations. The general solution of the SL-problem is a linear combination of Bessel functions of the first type and second type. The boundary conditions will be related the zeros of Bessel functions.

So how do the first and second order Bessel functions look like? What are their zeros? Before 1970s, the time when computers are expensive and rather complex to use (e.g. Fortran program are keyed in using punch cards), scientists and engineers have to calculate the function values and find the zeros by using suitable numerical methods and looking up in a standard handbook such as Abramowitz and Stegun [1964, 385]. Modern numerical software (such as SciPy/Python, R, Matlab) will normally include Bessel functions as part of the scientific library.

With modern plotting library and scientific library, we can write a simple program below

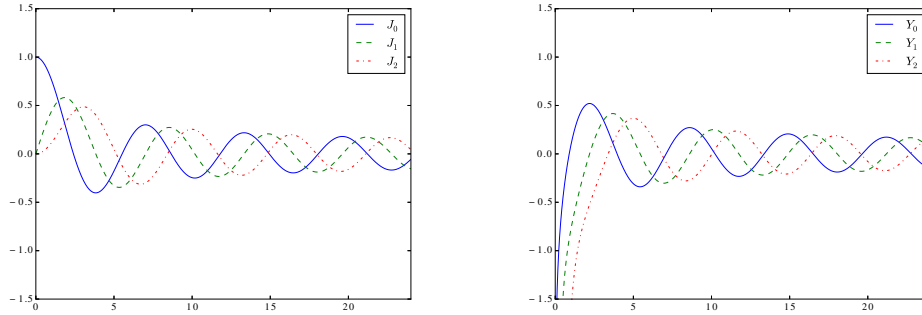
```
1 from matplotlib.pyplot import plot, show, legend, savefig, clf, axis
2 from scipy.special import j0, j1, jv, y0, y1, yv, jn_zeros
3 from numpy import arange
4
5 x = arange(0, 24, 0.002)
6 plot(x, j0(x), '--', x, j1(x), '--', x, jv(2, x), '-.')
7 legend(["$J_0$", "$J_1$", "$J_2$"])
```

```

8 axis([0,24,-1.5,1.5])
9 savefig("bessel1plt.eps")
10 clf()
11 plot(x,y0(x),'-',x,y1(x),'--',x,yv(2,x),'-.')
12 legend(["$Y_0$", "$Y_1$", "$Y_2$"])
13 axis([0,24,-1.5,1.5])
14 savefig("bessel2plt.eps")

```

to plot the Bessel functions of order 0, 1, 2 below. We can see that Bessel functions oscillate like the trigonometric functions:



Before the popularisation of computers, mathematicians are able to construct the above plots using the analytic properties of Bessel functions [Tolstov, 1976].

thm:bessel:recur

Theorem 36.2. For $x > 0$,

$$J_{\nu+1}(x) = \frac{2\nu}{x}J_{\nu}(x) - J_{\nu-1}(x), \quad xJ'_{\nu}(x) = xJ_{\nu-1}(x) - \nu J_{\nu}(x),$$

$$J'_{\nu}(x) = \frac{1}{2}[J_{\nu-1}(x) - J_{\nu+1}(x)], \quad xJ'_{\nu}(x) = \nu J_{\nu}(x) - xJ_{\nu+1}(x).$$

Theorem 36.3. $J_{\frac{1}{2}}(x) = \sqrt{\frac{2}{\pi x}} \sin x$; $J_{-\frac{1}{2}}(x) = \sqrt{\frac{2}{\pi x}} \cos x$, $x > 0$.

thm:bessel:diff

Theorem 36.4. Let $\nu \in \mathbb{R}$.

$$\frac{d}{dx} \left[x^{\nu} J_{\nu}(x) \right] = x^{\nu} J_{\nu-1}(x), \quad \frac{d}{dx} \left[\frac{J_{\nu}(x)}{x^{\nu}} \right] = -\frac{J_{\nu+1}(x)}{x^{\nu}},$$

$$\frac{d}{dx} \left[x^{\nu} Y_{\nu}(x) \right] = x^{\nu} Y_{\nu-1}(x), \quad \frac{d}{dx} \left[\frac{Y_{\nu}(x)}{x^{\nu}} \right] = -\frac{Y_{\nu+1}(x)}{x^{\nu}}.$$

The zeros of J_{ν} which are needed in later sections for the solution of SL-problem and the following are two relevant theorems [Abramowitz and Stegun, 1964, 372].

thm:zero

Theorem 36.5. When $\nu \geq -1$, the zeros of J_{ν} are all real. If $\nu < -1$ and ν is not an integer the number of complex zeros of J_{ν} is twice the integer part of $-\nu$; if the integer part of $-\nu$ is odd two of these zeros lie on the imaginary axis.

thm:zero:diff

Theorem 36.6. If $\nu \geq 0$, all zeros of J'_{ν} are real.

The following Python script

```

1 from scipy.special import j0 # Bessel function of the 1st kind of order 0
2 from scipy.special import j1 # Bessel function of the 1st kind of order 1

```

```

3 from scipy.special import jv # Bessel function of the 1st kind of real order v
4 from scipy.special import y0 # Bessel function of the 2nd kind of order 0
5 from scipy.special import y1 # Bessel function of the 2nd kind of order 1
6 from scipy.special import yv # Bessel function of the 2nd kind of real order v
7 from scipy.special import jn_zeros, jnp_zeros, jnjnp_zeros
8 from numpy import arange
9
10 print("="*70)
11 for n in range(6):
12     zs = jn_zeros(n, 7)
13     for j in zs: print("{j:8.4f} &".format(j=j), end=" ")
14     print()
15 print("="*70)
16 for n in range(6):
17     zs = jnp_zeros(n, 7)
18     for j in zs: print("{j:8.4f} &".format(j=j), end=" ")
19     print()

```

finds the first 7 zeros of J_n and J'_n for $n = 0, 1, 2, 3, 4, 5$:

$J_n \backslash i$	1	2	3	4	5	6	7
J_0	2.4048	5.5201	8.6537	11.7915	14.9309	18.0711	21.2116
J_1	3.8317	7.0156	10.1735	13.3237	16.4706	19.6159	22.7601
J_2	5.1356	8.4172	11.6198	14.7960	17.9598	21.1170	24.2701
J_3	6.3802	9.7610	13.0152	16.2235	19.4094	22.5827	25.7482
J_4	7.5883	11.0647	14.3725	17.6160	20.8269	24.0190	27.1991
J_5	8.7715	12.3386	15.7002	18.9801	22.2178	25.4303	28.6266
J'_0	3.8317	7.0156	10.1735	13.3237	16.4706	19.6159	22.7601
J'_1	1.8412	5.3314	8.5363	11.7060	14.8636	18.0155	21.1644
J'_2	3.0542	6.7061	9.9695	13.1704	16.3475	19.5129	22.6716
J'_3	4.2012	8.0152	11.3459	14.5858	17.7887	20.9725	24.1449
J'_4	5.3176	9.2824	12.6819	15.9641	19.1960	22.4010	25.5898
J'_5	6.4156	10.5199	13.9872	17.3128	20.5755	23.8036	27.0103

When $x \gg 0$, the zeros have an approximation $\mu_i^{(n)} \approx (i-1)\pi + \frac{n\pi}{2} + \frac{3}{4}$.

Apart from the application to PDE in cylindrical coordinates, Bessel functions are also applied in (a) the solution of displacement of a suspended chain $u_{tt} = gu_z + gzu_{zz}$; (b) critical length of a vertical rod $u'' + \frac{w}{EI}xu = 0$; (c) Poe's pendulum $(a+bt)\theta'' + 2b\theta' + g\theta = 0$ which can be transformed Bessel equation with $x = \frac{2}{b}\sqrt{(a+bt)g}$ and $y = \sqrt{(a+bt)}\theta$.

36.2.2 Orthogonal Property of Bessel Functions

In Topic 1, one of the most important property of trigonometric functions is the orthogonal property. Bessel functions also satisfy orthogonal property but in an unfamiliar way.

The derivation of the orthogonal property will be based on Tolstov [1976]. First, let λ and μ be two different non-negative values and consider

$$\begin{cases} x^2 y''_{\lambda} + x y'_{\lambda} + (\lambda^2 x^2 - \nu^2) y_{\lambda} = 0, \\ x^2 y''_{\mu} + x y'_{\mu} + (\mu^2 x^2 - \nu^2) y_{\mu} = 0 \end{cases} \Rightarrow \begin{cases} y_{\mu} \times \left[y''_{\lambda} + \frac{1}{x} y'_{\lambda} + \left(\lambda^2 - \frac{\nu^2}{x^2} \right) y_{\lambda} \right] = 0, \\ y_{\lambda} \times \left[y''_{\mu} + \frac{1}{x} y'_{\mu} + \left(\mu^2 - \frac{\nu^2}{x^2} \right) y_{\mu} \right] = 0 \end{cases} \quad (36.25)$$

By subtracting them, we have

$$\left[y_{\mu} y''_{\lambda} - y_{\lambda} y''_{\mu} \right] + \frac{1}{x} \left[y_{\mu} y'_{\lambda} - y_{\lambda} y'_{\mu} \right] + \left[\lambda^2 - \mu^2 \right] y_{\mu} y_{\lambda} = 0$$

Multiplying both sides by x , we can rewrite into self-adjoint form:

$$x \left[y_{\mu} y''_{\lambda} - y_{\lambda} y''_{\mu} \right] + \left[y_{\mu} y'_{\lambda} - y_{\lambda} y'_{\mu} \right] = \frac{d}{dx} \left[x (y_{\mu} y'_{\lambda} - y_{\lambda} y'_{\mu}) \right] = x \left[\mu^2 - \lambda^2 \right] y_{\mu} y_{\lambda}.$$

Integrating both sides from 0 to 1, we have

$$\left[\mu^2 - \lambda^2 \right] \int_0^1 x y_{\mu} y_{\lambda} dx = \left[x (y_{\mu} y'_{\lambda} - y_{\lambda} y'_{\mu}) \right]_0^1 = y_{\mu}(1) y'_{\lambda}(1) - y_{\lambda}(1) y'_{\mu}(1).$$

In the previous section, we learn that the Bessel function of the first kind $J_\nu(\lambda x)$ is a solution, therefore, we have

$$\begin{aligned} [\mu^2 - \lambda^2] \int_0^1 x J_\nu(\mu x) J_\nu(\lambda x) dx &= J_\nu(\mu \times 1) \frac{d}{dx} J_\nu(\lambda x) \Big|_{x=1} - J_\nu(\lambda \times 1) \frac{d}{dx} J_\nu(\mu x) \Big|_{x=1} \\ &= \lambda J_\nu(\mu) J'_\nu(\lambda) - \mu J_\nu(\lambda) J'_\nu(\mu). \end{aligned} \quad (36.26)$$

When $\lambda = \mu$, we need to take $\mu \rightarrow \lambda$, apply the L'Hôpital rule and use the Bessel equation (36.23) at $x = \lambda$:

$$\begin{aligned} \int_0^1 x [J_\nu(\lambda x)]^2 dx &= \lim_{\mu \rightarrow \lambda} \frac{\lambda J_\nu(\mu) J'_\nu(\lambda) - \mu J_\nu(\lambda) J'_\nu(\mu)}{\mu^2 - \lambda^2} \\ &= \lim_{\mu \rightarrow \lambda} \frac{\lambda J'_\nu(\mu) J'_\nu(\lambda) - J_\nu(\lambda) J'_\nu(\mu) - \mu J_\nu(\lambda) J''_\nu(\mu)}{2\mu} \\ &= \frac{1}{2\lambda} \left\{ \lambda [J'_\nu(\lambda)]^2 - J_\nu(\lambda) J'_\nu(\lambda) - \lambda J_\nu(\lambda) J''_\nu(\lambda) \right\} \\ &= \frac{1}{2\lambda} \left\{ \lambda [J'_\nu(\lambda)]^2 - J_\nu(\lambda) J'_\nu(\lambda) - \lambda J_\nu(\lambda) \left[-\frac{1}{\lambda} J'_\nu(\lambda) - \left(1 - \frac{\nu^2}{\lambda^2}\right) J_\nu(\lambda) \right] \right\} \end{aligned}$$

Simplifying the above equations leads to

$$\int_0^1 x [J_\nu(\lambda x)]^2 dx = \frac{1}{2} [J'_\nu(\lambda)]^2 + \frac{1}{2} \left(1 - \frac{\nu^2}{\lambda^2}\right) [J_\nu(\lambda)]^2. \quad (36.27)$$

When λ and μ are the roots of J_ν , we have the orthogonality under the weight x .

Theorem 36.7. If λ and μ are the roots of Bessel function J_ν where $\nu > -1$, then

$$\int_0^1 x J_\nu(\lambda x) J_\nu(\mu x) dx = \begin{cases} 0, & \lambda \neq \mu, \\ \frac{1}{2} [J_{\nu+1}(\lambda)]^2, & \lambda = \mu. \end{cases} \quad (36.28)$$

Proof. When $\lambda \neq \mu$, (36.26) becomes zero since $J_\nu(\lambda) = J_\nu(\mu) = 0$. When $\lambda = \mu$ and $J_\nu(\lambda) = 0$, (36.27) implies

$$\int_0^1 x [J_\nu(\lambda x)]^2 dx = \frac{1}{2} [J'_\nu(\lambda)]^2 = \frac{1}{2} [J_{\nu+1}(\lambda)]^2.$$

Note that Theorem 36.2 implies $\lambda J'_\nu(\lambda) - \nu J_\nu(\lambda) = -\lambda J_{\nu+1}(\lambda)$ or $J'_\nu(\lambda) = -J_{\nu+1}(\lambda)$. $\square$

Theorem 36.8. If λ and μ are the positive roots of the first order derivative of Bessel function J'_ν , then

$$\int_0^1 x J_\nu(\lambda x) J_\nu(\mu x) dx = \begin{cases} 0, & \lambda \neq \mu, \\ \frac{1}{2} \left(1 - \frac{\nu^2}{\lambda^2}\right) [J_\nu(\lambda)]^2, & \lambda = \mu. \end{cases} \quad (36.29)$$

Similar to Topic 1, we can construct Bessel's version of orthogonal basis and "Fourier series". If $\lambda_1, \dots, \lambda_n, \dots$ are positive roots of J_ν , the corresponding functions $J_\nu(\lambda_1 x), \dots, J_\nu(\lambda_n x), \dots$ are orthogonal on the interval $[0, 1]$ with respect to the weight x . Hence, any function can have an associated formal *Fourier-Bessel series* below

$$f(x) \sim c_1 J_\nu(\lambda_1 x) + \dots + c_n J_\nu(\lambda_n x) + \dots \quad (36.30)$$

where

$$c_n = \frac{\int_0^1 x f(x) J_\nu(\lambda_n x) dx}{\int_0^1 x [J_\nu(\lambda_n x)]^2 dx} = \frac{2}{J_{\nu+1}^2(\lambda_n)} \int_0^1 x f(x) J_\nu(\lambda_n x) dx. \quad (36.31)$$

If f is piecewise smooth on the open interval $(0, 1)$ then for $\nu \geq -\frac{1}{2}$, we have the following point-wise convergence

$$\sum_{n=1}^{\infty} c_n J_\nu(\lambda_n x) = \frac{f(x+0) + f(x-0)}{2}.$$

Example 36.9 (UCCM3003 Final Exam, Apr 2009, Q4(a)). Given $f(r) = 50$, $0 < r < 2$, show that $f(r)$ could be represented as a generalised Fourier series as follow

$$f(r) = \sum_{n=0}^{\infty} \frac{100}{\alpha_n J_1(\alpha_n)} J_0(\alpha_n r/2), \quad 0 < r < 2,$$

where α_n is the n -th root of the Bessel's function of order zero $J_0(\alpha_n r/2)$. (10 marks)

Solution. The Fourier-Bessel series is $f(r) \sim \sum_{n=1}^{\infty} c_n J_0(\lambda_n x)$ where $\lambda_n = \frac{\alpha_n}{2}$ and

$$\begin{aligned} c_n &= \frac{2}{2^2 J_1^2(\alpha_n)} \int_0^2 50 J_0\left(\frac{\alpha_n r}{2}\right) r dr = \frac{50}{2 J_1^2(\alpha_n)} \left(\frac{2}{\alpha_n}\right)^2 \int_0^{\alpha_n} J_0(x) x dx \\ &= \frac{100}{\alpha_n^2 J_1^2(\alpha_n)} \left[x J_1(x) \right]_0^{\alpha_n} = \frac{100}{\alpha_n J_1(\alpha_n)} \end{aligned}$$

according to Theorem 36.4 (on page 809) and $x = \frac{r}{2}$. □

Example 36.10 (UECM3003 Final Exam, Apr 2014, Q5(b)). Let α_n be the n th root of the Bessel function $J_0(x)$. Compute the Fourier-Bessel series $\sum_{n=1}^{\infty} c_n J_0(\alpha_n x)$ for $f(x) = 1 - x^2$ on the interval $[0, 1]$.

[Remark: Taken from Asmar [2005, 204]] (8 marks)

Solution. Theorem 36.4 implies that $\int x J_0(x) dx = x J_1(x) + C$ and $\int x^2 J_1(x) dx = x^2 J_2(x) + C$. The Fourier-Bessel coefficients (36.31) can be found by a suitable change of variable $t = \alpha_n x$:

$$\begin{aligned} c_n &= \frac{2}{J_1(\alpha_n)^2} \int_0^1 (1 - x^2) J_0(\alpha_n x) x dx = \frac{2}{J_1(\alpha_n)^2} \left(\int_0^1 x J_0(\alpha_n x) dx - \int_0^1 x^2 \cdot x J_0(\alpha_n x) dx \right) \\ &= \frac{2}{J_1(\alpha_n)^2} \left(\frac{1}{\alpha_n^2} \int_0^{\alpha_n} t J_0(t) dt - \frac{1}{\alpha_n^4} \int_0^{\alpha_n} t^2 \cdot t J_0(t) dt \right) \\ &= \frac{2}{J_1(\alpha_n)^2} \left(\frac{1}{\alpha_n^2} [\alpha_n J_1(\alpha_n) - 0] - \frac{1}{\alpha_n^4} \left\{ [t^2 \cdot t J_1(t)]_0^{\alpha_n} - \int_0^{\alpha_n} t J_1(t) 2t dt \right\} \right) \\ &= \frac{2}{J_1(\alpha_n)^2} \left(\frac{J_1(\alpha_n)}{\alpha_n} - \frac{J_1(\alpha_n)}{\alpha_n} + \frac{2}{\alpha_n^4} [t^2 J_2(t)]_0^{\alpha_n} \right) \\ &= \frac{2}{J_1(\alpha_n)^2} \times \frac{2 J_2(\alpha_n)}{\alpha_n^2} = \frac{2}{J_1(\alpha_n)^2} \times \frac{4 J_1(\alpha_n)}{\alpha_n^3} = \frac{8}{\alpha_n^3 J_1(\alpha_n)} \end{aligned}$$

Note that according to Theorem 36.2, $J_2(x) = \frac{2}{x} J_1(x) - J_0(x)$ and $J_0(\alpha_n) = 0$, we have $J_2(\alpha_n) = \frac{2}{\alpha_n} J_1(\alpha_n)$. Therefore, $1 - x^2 \sim \sum_{n=1}^{\infty} \frac{8}{\alpha_n^3 J_1(\alpha_n)} J_0(\alpha_n x)$. □

Example 36.11 (UECM3003 Final Exam, Apr 2015, Q4(c)). Let α_n be the n th zero of the Bessel function $J_0(x)$. Find the Fourier-Bessel series for $f(x) = x^2$ in terms of the Bessel function of the first kind $J_0(\alpha_n x)$ on the interval $[0, 1]$. (9 marks)

Solution: Theorem 36.4 implies that $\int x J_0(x) dx = x J_1(x) + C$ and $\int x^2 J_1(x) dx = x^2 J_2(x) + C$. The Fourier-Bessel coefficients (36.31) can be found by a suitable change of variable $t = \alpha_n x$:

$$\begin{aligned} c_n &= \frac{2}{J_1(\alpha_n)^2} \int_0^1 x^2 J_0(\alpha_n x) x dx = \frac{2}{J_1(\alpha_n)^2} \left(\frac{1}{\alpha_n^4} \int_0^{\alpha_n} t^2 \cdot t J_0(t) dt \right) \\ &= \frac{2}{\alpha_n^4 J_1(\alpha_n)^2} \left\{ [t^2 \cdot t J_1(t)]_0^{\alpha_n} - \int_0^{\alpha_n} t J_1(t) 2t dt \right\} \\ &= \frac{2}{\alpha_n^4 J_1(\alpha_n)^2} \left\{ \alpha_n^3 J_1(\alpha_n) - 2 [t^2 J_2(t)]_0^{\alpha_n} \right\} \\ &= \frac{2}{\alpha_n^4 J_1(\alpha_n)^2} \left\{ \alpha_n^3 J_1(\alpha_n) - 2 \alpha_n^2 J_2(\alpha_n) \right\} = \frac{2}{\alpha_n^4 J_1(\alpha_n)^2} \left\{ \alpha_n^3 J_1(\alpha_n) - 4 \alpha_n J_1(\alpha_n) \right\} \\ &= \frac{2(\alpha_n^2 - 4)}{\alpha_n^3 J_1(\alpha_n)} \end{aligned}$$

Therefore, $x^2 \sim \sum_{n=1}^{\infty} \frac{2(\alpha_n^2 - 4)}{\alpha_n^3 J_1(\alpha_n)} J_0(\alpha_n x)$.

36.2.3 Modified Bessel Functions

In this section, we will study the solutions of modified Bessel equation which appears when we apply separation of variables to Laplace equation in cylindrical coordinates.

Definition 36.2. A modified Bessel equation with parameter $\lambda(> 0)$ of order ν is

$$x^2 y'' + xy' - (\lambda^2 x^2 + \nu^2)y = 0. \quad (36.32)$$

When $\lambda = 1$, we call it a *modified Bessel equation of order ν* .

$$x^2 y'' + xy' - (x^2 + \nu^2)y = 0. \quad (36.33)$$

Note that the range of x can either be $0 < x < L$, $0 < a < x < b < \infty$ or $0 < L < x$.

By using a scaling transform $x \rightarrow \lambda x$, the modified Bessel equation with parameter λ (36.32) can be transformed to a modified Bessel equation (36.33). Therefore, the two are equivalent. By rewriting $-\lambda^2$ as $(i\lambda)^2$, the general solution of (36.33) is found to be

$$y(x) = aJ_\nu(ix) + bJ_{-\nu}(ix) \quad \text{or} \quad y(x) = aJ_\nu(ix) + bY_\nu(ix)$$

using the results of Bessel equation from the previous section.

Since they are not in real form, mathematicians introduce the following functions such that $e^{-\frac{1}{2}\nu\pi i} J_\nu(ix)$ is in real form:

$$I_\nu(x) = e^{-\frac{1}{2}\nu\pi i} J_\nu(e^{\frac{1}{2}\pi i} x) = \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(k + \nu + 1)} \left(\frac{x}{2}\right)^{2k+\nu}, \quad -\pi < \arg x \leq \frac{\pi}{2}$$

and

$$K_\nu(x) = \lim_{r \rightarrow \nu} \frac{I_r(x) \cos r\pi - I_{-r}(x)}{\sin r\pi}.$$

Definition 36.3. The functions I_ν and K_ν above are all the *modified Bessel functions of the first and second type/kind* respectively.

Hence, the general solution of the modified Bessel equation (36.33) can be written as

$$y(x) = aI_\nu(x) + bK_\nu(x) \quad (36.34)$$

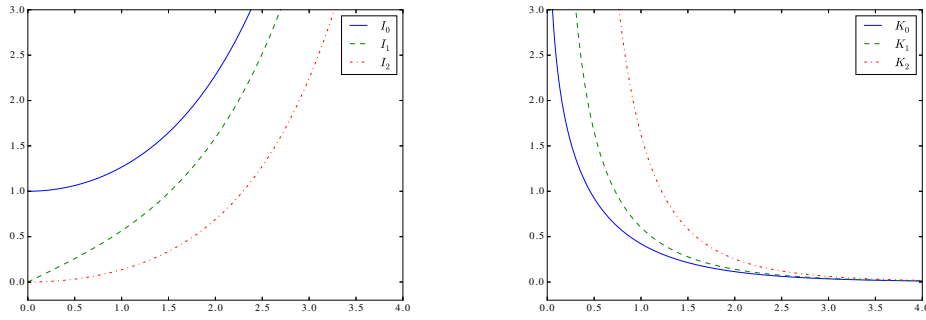
and the general solution of the modified Bessel equation with parameter λ (36.32) can be written as

$$y(x) = aI_\nu(\lambda x) + bK_\nu(\lambda x). \quad (36.35)$$

With modern plotting library and scientific library, a simple program

```
1 from matplotlib.pyplot import plot, show, legend, savefig, clf, axis
2 from scipy.special import i0, i1, iv, k0, k1, kv
3 from numpy import arange
4
5 x = arange(0, 4, 0.002)
6 plot(x, i0(x), '-', x, i1(x), '--', x, iv(2, x), '-.')
7 legend(["$I_0$", "$I_1$", "$I_2$"])
8 axis([0, 4, -0.2, 3])
9 savefig("mbessel1plt.eps")
10 clf()
11 plot(x, k0(x), '-', x, k1(x), '--', x, kv(2, x), '-.')
12 legend(["$K_0$", "$K_1$", "$K_2$"])
13 axis([0, 4, -0.2, 3])
14 savefig("mbessel2plt.eps")
```

can be used to plot the modified Bessel functions of order 0, 1, 2 easily:



From these plots we can observe their similarity with exponential functions. Just like exponential functions, they have no nice orthogonal property.

Mathematicians have developed the approximations below for us to compare Bessel functions and modified Bessel functions with trigonometric functions and exponential functions more accurately.

Function	Small x	Large x
$J_\nu(x)$	$\frac{1}{\Gamma(\nu+1)}\left(\frac{x}{2}\right)^\nu + O(x^{\nu+2})$	$\sqrt{\frac{2}{\pi x}} \cos\left(x - \frac{2\nu+1}{4}\pi\right) + O(x^{-3/2})$
$Y_\nu(x)$	$\begin{cases} \frac{2}{\pi} \ln x + O(1), & \nu = 0 \\ -\frac{\Gamma(\nu)}{\pi} \left(\frac{2}{x}\right)^\nu + \begin{cases} O(x^\nu), & \nu < 1, \\ O(x \ln \frac{1}{x}), & \nu = 1, \\ O(x^{2-\nu}), & \nu > 1. \end{cases} \end{cases}$	$\sqrt{\frac{2}{\pi x}} \sin\left(x - \frac{2\nu+1}{4}\pi\right) + O(x^{-3/2})$
$I_\nu(x)$	like $J_\nu(x)$	$\frac{1}{\sqrt{2\pi x}} e^x + O\left(\frac{e^x}{x}\right)$
$K_\nu(x)$	like $-\frac{\pi}{2} Y_\nu(x)$	$\sqrt{\frac{\pi}{2x}} e^{-x} + O\left(\frac{e^{-x}}{x}\right)$

36.3 1-D Heat Flow (in Infinite Cylinder)

In this section, we will assume the initial temperature with radial symmetry to be given by a function of “cylindrical radius”, i.e. (36.1). We shall deal with heat-flow on a disk with lateral surface or in an infinite cylinder with radial symmetry, in which case, the heat distribution is governed by (36.11) and the range of r is $0 < r < L$.

Mathematically, it is possible to discuss the range $a < r < b$ and $L < r$ with various permutations of linear boundary conditions. However, they have little engineering application or a more complicated and we will ignore them.

We have to stress here that the “1-D” heat equation can model either a disk or an infinite tube of with radial symmetry in the initial condition. So the “dimension” does not correspond to the physical dimension but is a mathematical concept related to the number of variables other than the time-variable t .

36.3.1 (Homogeneous) Type 1 BC

Let us first consider the radial flow of heat in an infinitely long circular cylinder of radius L , the lateral surface of which is kept at zero temperature. This problem is described by the PDE (36.11) and type 1 BC

$$u|_{r=L} = 0, \quad t > 0 \quad (36.36)$$

and the IC (36.1). The following example illustrates the technique to solve this problem with $L = 3$.

Example 36.12 (UCCM3003 Final Exam, Sep 2007, Q3).

(a) The temperature $u(r, t)$ of an infinite cylinder is governed by the equation

$$\begin{cases} u_t = k(u_{rr} + \frac{1}{r}u_r), & 0 < r < 3, \quad t > 0, \\ u|_{r=3} = 0, & t > 0, \\ u|_{t=0} = f(r), & 0 < r < 3. \end{cases} \quad (36.37)$$

By using separation of variables, show that the general solution of the above heat equation is

$$u(r, t) = \sum_{n=1}^{\infty} a_n e^{-(\mu_n/3)^2 kt} J_0\left(\frac{\mu_n r}{3}\right)$$

where μ_n , $n = 1, 2, \dots$ are the roots of the equation $J_0(x) = 0$. Note that an n th order Bessel equation

$$x^2 y'' + xy' + (\lambda x^2 - n^2)y = 0$$

has a general solution

$$y(x) = AJ_n(\sqrt{\lambda}x) + BY_n(\sqrt{\lambda}x)$$

where A and B are constants, J_n and Y_n are Bessel's functions of the first and second type respectively. (17 marks)

Solution: This is a homogeneous problem, let $u(r, t) = T(t)R(r) = TR$, from the heat equation, we have

$$T'R = k(TR'' + \frac{1}{r}TR') \Rightarrow \frac{T'}{kT} = \frac{R''}{R} + \frac{R'}{rR} = -\lambda$$

and we derive an ODE and singular SL-problem after rearranging the above equations and including the boundary conditions:

$$T' + k\lambda T = 0, \quad (36.38) \quad \text{e:69}$$

$$R'' + \frac{1}{r}R' + \lambda R = 0, \quad |R(0)| < \infty, \quad R(3) = 0. \quad (36.39) \quad \text{e:70}$$

where λ is an arbitrary parameter.

When $\lambda < 0$, the general solution of (36.39) is given by (36.35) with $\nu = 0$ e:70 e:mbesselp:soln

$$R(r) = aI_0(\sqrt{-\lambda}r) + bK_0(\sqrt{-\lambda}r).$$

Note the $|R(0)| < \infty$ implies that $b = 0$ while $I_0(x) > 0$ when $x > 0$ and $R(3) = 0$ implies that $a = 0$. So we have a trivial solution.

When $\lambda = 0$, the general solution of (36.39) is e:70

$$R(r) = a \ln r + b.$$

Note the $|R(0)| < \infty$ implies that $a = 0$ while $R(3) = 0$ implies that $b = 0$. So we have a trivial solution.

When $\lambda > 0$, the general solution of (36.39) is e:70

$$R(r) = aJ_0(\sqrt{\lambda}r) + bY_0(\sqrt{\lambda}r).$$

Since $Y_0(\sqrt{\lambda}r)$ becomes infinite as r approaches zero, the regularity of R at $r = 0$ implies that $b = 0$.

To have **non-trivial** solutions, we need $a \neq 0$, therefore

$$R(3) = 0 \Rightarrow J_0(\sqrt{\lambda}3) = 0 \Rightarrow \lambda_n = \left(\frac{\mu_n}{3}\right)^2, \quad n = 1, 2, \dots$$

where the μ_n 's are the positive roots of the equation $J_0(\mu) = 0$. The existence of μ_n are guaranteed by Theorem e:mb:zero 36.5.

To each eigenvalues λ_n , there corresponds an eigenfunction

$$R_n(r) = A_n J_0\left(\frac{\mu_n r}{3}\right).$$

Solving for the ODE of T , we obtain

$$T_n(t) = C_n e^{-(\mu_n/3)^2 kt}$$

Hence, the general solution is

$$u(r, t) = \sum_{n=1}^{\infty} a_n e^{-(\mu_n/3)^2 kt} J_0\left(\frac{\mu_n r}{3}\right). \quad (36.40) \quad \text{e:75}$$

(b) Find the value of a_n in part (a) and hence show that e:final:200709a3a

$$u(r, t) = \frac{2}{9} \sum_{n=1}^{\infty} \frac{J_0(\sqrt{\lambda_n}r) e^{-\lambda_n kt}}{(J_1(3\sqrt{\lambda_n}))^2} \int_0^3 \rho f(\rho) J_0(\sqrt{\lambda_n} \rho) d\rho$$

by using the orthogonal properties of Bessel function (**Remark:** Theorem [36.7](#)):

$$\int_0^c x J_n(\lambda x) J_n(\mu x) dx = 0 \quad \text{and} \quad \int_0^c x J_n(\lambda x)^2 dx = \frac{c^2}{2} (J_{n+1}(\lambda c))^2,$$

where $n \geq 0$, $c > 0$, λ and μ are two different positive roots of the Bessel function $J_n(cx)$.

(6 marks)

Solution. From the initial condition, we have

$$\sum_{n=1}^{\infty} a_n J_0\left(\frac{\mu_n r}{3}\right) = f(r). \quad (36.41)$$

This series is an expansion of the given function $f(r)$ in Bessel functions in the interval $(0, 3)$. The coefficients in the expansion (36.41) are determined from the formula

$$a_n = \frac{2}{3^2 J_1^2(\mu_n)} \int_0^3 \rho f(\rho) J_0\left(\frac{\mu_n \rho}{3}\right) d\rho. \quad (36.42)$$

Substituting the expression (36.42) for the a_n in the series (36.40), we obtain the solution of the problem (36.37) in the form

$$u(r, t) = \frac{2}{3^2} \sum_{n=1}^{\infty} \frac{J_0\left(\frac{\mu_n r}{3}\right)}{J_1^2(\mu_n)} e^{-(\mu_n/3)^2 k t} \int_0^3 \rho f(\rho) J_0\left(\frac{\mu_n \rho}{3}\right) d\rho.$$

Since $3\sqrt{\lambda_n} = \mu_n$, we obtain the values a_n for the given problem. $\square$

- (c) What is the steady-state solution of the heat problem in part (a)? (2 marks)

Solution. Since $\lambda_n > 0$ and $e^{-\lambda_n t} \rightarrow 0$ as $t \rightarrow \infty$, the steady-state solution is

$$u = 0. \quad [2 \text{ marks}]$$

$\square$

Example 36.13 (UCCM3003 Final Exam, May 2010, Q3(a)). Consider the flow of heat in an infinitely long cylinder of radius one. Let the surface of the cylinder be held at temperature zero, and the initial temperature distribution of the cylinder be a function of the radial variable r only. Then the temperature u is a function of r and t only and satisfies the heat conduction equation

$$c^2 \left[\frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} \right] = \frac{\partial u}{\partial t}, \quad 0 < r < 1, \quad t > 0,$$

and the following initial boundary conditions:

$$\begin{cases} u(r, 0) = f(r), & 0 \leq r \leq 1, \\ u(1, t) = 0, & t > 0. \end{cases}$$

- (i) Use the method of separation of variables to show that $u(r, t) = \sum_{n=1}^{\infty} A_n J_0(\lambda_n r) e^{-c^2 \lambda_n^2 t}$ where $J_0(\lambda_n) = 0$. (15 marks)

Solution. This is a homogeneous problem just like Example 36.12. By using separation of variables, we derive an ODE and singular SL-problem:

$$T' + c^2 \lambda T = 0, \quad (36.43)$$

$$R'' + \frac{1}{r} R' + \lambda R = 0, \quad |R(0)| < \infty, \quad R(1) = 0, \quad (36.44)$$

where λ is an arbitrary parameter.

When $\lambda < 0$, the general solution of (36.44) is given by (36.35) with $\nu = 0$

$$R(r) = a I_0(\sqrt{-\lambda} r) + b K_0(\sqrt{-\lambda} r).$$

Note the $|R(0)| < \infty$ implies that $b = 0$ while $I_0(x) > 0$ when $x > 0$ and $R(1) = 0$ implies that

$a = 0$. So we can only have trivial solution.
When $\lambda = 0$, the general solution of (36.44) is

$$R(r) = a \ln r + b.$$

Note the $|R(0)| < \infty$ implies that $a = 0$ while $R(1) = 0$ implies that $b = 0$. So we have a trivial solution.

When $\lambda > 0$, the general solution of (36.44) is

$$R(r) = aJ_0(\sqrt{\lambda}r) + bY_0(\sqrt{\lambda}r).$$

Since $Y_0(\sqrt{\lambda}r)$ becomes infinite as r approaches zero, the regularity of R at $r = 0$ implies that $b = 0$.

To have **non-trivial** solutions, we need $a \neq 0$, therefore

$$R(1) = 0 \Rightarrow J_0(\sqrt{\lambda}) = 0 \Rightarrow \lambda = \lambda_n^2, \quad n = 1, 2, \dots, \quad J_0(\lambda_n) = 0.$$

and the corresponding eigenfunctions are

$$R_n(r) = D_n J_0(\lambda_n r).$$

Solving for the ODE of T , we obtain

$$T_n(t) = C_n e^{-c^2 \lambda_n^2 t}$$

Hence, the general solution is $u(r, t) = \sum_{n=1}^{\infty} A_n e^{-c^2 \lambda_n^2 t} J_0(\lambda_n r)$, $A_n = C_n D_n$. □

(ii) Write down a formula for A_n . (4 marks)

Solution. $A_n = \frac{\int_0^1 J_0(\lambda_n r) f(r) r dr}{\int_0^1 [J_0(\lambda_n r)]^2 r dr} = \frac{2}{[J_1(\lambda_n)]^2} \int_0^1 J_0(\lambda_n r) f(r) r dr.$ □

Example 36.14 (UECM3003 Final Exam, Apr 2015, Q3(b)). A flat homogeneous circular disk with radius 1 is initially at a temperature of $g(r)$. The boundary of the disk is held at 0°C at all time. Let $u(r, t)$ be the radial temperature distribution of the disk. Find $u(r, t)$ and deduce what is the temperature distribution when $t \rightarrow \infty$. (14 marks)

Remark: Let $\mu_n, n = 1, 2, \dots$ be the roots of J_0 . The solution is the same as Example 36.13 with f replaced by g : leg:final:201005q3a

$$u(r, t) = \sum_{n=1}^{\infty} A_n e^{-c^2 \mu_n^2 t} J_0(\mu_n r), \quad A_n = \frac{2}{[J_1(\mu_n)]^2} \int_0^1 J_0(\mu_n r) g(r) r dr.$$

When $t \rightarrow \infty$, $u(r, t) \rightarrow 0$ since $\mu_1 > 0$ and $e^{-c^2 \mu_1^2 t} \rightarrow 0$.

Example 36.15 (UCCM3003 Final Exam, Apr 2009, Q4(b)). A flat homogeneous circular disk with radius 2 is initially at a temperature 50° . The circumference of the plate is then held at 0° at all time t . Given the heat equation is $\frac{\partial u}{\partial t} = 4\nabla^2 u$, find the time dependent temperature distribution $u(r, t)$ for the disk. What is the temperature distribution when $t \rightarrow \infty$? (15 marks)

Solution. The problem is characterised by

$$\begin{cases} u_t - 4(u_{rr} + \frac{1}{r}u_r) = 0, & 0 < r < 2, t > 0, \\ u(2, t) = 0 & t > 0, \\ u(r, 0) = 50 & 0 < r < 2. \end{cases}$$

It is radial symmetric. So using the similar solution technique in Example 36.12(a), we obtain the general solution leg:final:200904:q4b

$$u(r, t) = \sum_{n=1}^{\infty} a_n e^{-(\mu_n/2)^4 t} J_0\left(\frac{\mu_n r}{2}\right)$$

From the initial condition, we have

$$u(r, 0) = \sum_{n=1}^{\infty} a_n J_0\left(\frac{\mu_n r}{2}\right) = 50 \Rightarrow a_n = \frac{50 \int_0^2 J_0\left(\frac{\mu_n r}{2}\right) r dr}{\int_0^2 \left[J_0\left(\frac{\mu_n r}{2}\right)\right]^2 r dr} = \frac{50}{2 J_1^2(\mu_n)} \int_0^2 J_0\left(\frac{\mu_n r}{2}\right) r dr$$

Simplifying it with

$$\int_0^2 J_0\left(\frac{\mu_n r}{2}\right) r dr = \left(\frac{2}{\mu_n}\right)^2 \int_0^{\mu_n} J_0(x) x dx = \frac{4}{\mu_n^2} \left[x J_1(x)\right]_0^{\mu_n} = \frac{4}{\mu_n^2} J_1(\mu_n)$$

according to Theorem [36.4](#) and $x = \frac{\mu_n r}{2}$, the solution is

$$u(r, t) = \sum_{n=1}^{\infty} \frac{100}{\mu_n J_1(\mu_n)} e^{-(\mu_n/2)4t} J_0\left(\frac{\mu_n r}{2}\right).$$

Since $J_0(\mu_n r/2)$ is bounded and $\mu_n > 0$, so $u \rightarrow 0$ as $t \rightarrow \infty$. □

Example 36.16 (UCCM3003 Final Exam, Apr 2011, Q5(b)). The temperature $u(r, t)$ of an infinite rod with radius 2 is governed by the following heat equation:

$$\begin{cases} \frac{\partial u}{\partial t} = 4 \left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} \right), & 0 < r < 2, \quad t > 0, \\ u(r, 0) = 50, & 0 < r < 2, \\ u(2, t) = 0, & t > 0. \end{cases}$$

Given that $f(r) = 50$, $0 < r < 2$ could be expressed as a Fourier-Bessel series,

$$f(r) = \sum_{n=1}^{\infty} \frac{100}{\alpha_n J_1(\alpha_n)} J_0\left(\frac{\alpha_n r}{2}\right), \quad 0 < r < 2,$$

where α_n is the n -th root of the Bessel's function of order zero, i.e., $J_0(\alpha_n) = 0$. Find the temperature distribution $u(r, t)$ of the infinite cylinder. (15 marks)

Remark: The solution is the same as Example [36.15](#). [eq:final:200904:q4b](#)

36.3.2 (Homogeneous) Type 2 BC

Consider the radial flow of heat in an infinitely long circular cylinder of radius L , the lateral surface of which is insulated. This problem is described by the PDE [\(36.11\)](#) and type 2 BC [\(36.11\)](#)

$$\frac{\partial u}{\partial r} \Big|_{r=L} = 0, \quad t > 0 \tag{36.45}$$

and the IC [\(36.1\)](#). [e:cyl:heat1d:b](#)

Example 36.17. Use separation of variables to solve the heat equation [\(36.11\)](#) with type 2 BC [\(36.45\)](#) and IC [\(36.1\)](#). [e:cyl:heat1d](#) [e:cyl:heat1d:bc2](#)

Solution. By using separation of variables, we derive an ODE and singular SL-problem:

$$T' + c^2 \lambda T = 0, \tag{36.46}$$

$$R'' + \frac{1}{r} R' + \lambda R = 0, \quad |R(0)| < \infty, \quad R'(L) = 0. \tag{36.47}$$

When $\lambda < 0$, the general solution of [\(36.47\)](#) is given by [\(36.35\)](#) with $\nu = 0$ [e:r2](#) [e:mbessel:p:soln](#)

$$R(r) = a I_0(\sqrt{-\lambda} r) + b K_0(\sqrt{-\lambda} r).$$

Note the $|R(0)| < \infty$ implies that $b = 0$. According to Abramowitz and Stegun [1964], item 9.6.27,

$$I_0'(z) = I_1(z) \Rightarrow R'(L) = a I_1'(L) = 0 \Rightarrow a = 0.$$

So we can only have trivial solution.

When $\lambda = 0$, the general solution of (36.47) is

$$R(r) = a \ln r + b.$$

Note the $|R(0)| < \infty$ implies that $a = 0$ while

$$R'(L) = 0 \Rightarrow R(r) = b. \text{ So we have an eigenfunction and } T = C_0$$

When $\lambda > 0$, the general solution of (36.47) is

$$R(r) = aJ_0(\sqrt{\lambda}r) + bY_0(\sqrt{\lambda}r).$$

Since $Y_0(\sqrt{\lambda}r)$ becomes infinite as r approaches zero, the regularity of R at $r = 0$ implies that $b = 0$. To have **non-trivial** solutions, we need $a \neq 0$, therefore

$$R'(L) = 0 \Rightarrow \sqrt{\lambda}J'_0(\sqrt{\lambda}L) = 0 \Rightarrow \lambda = \left(\frac{\beta_n}{L}\right)^2, \quad n = 1, 2, \dots, \quad J'_0(\beta_n) = 0$$

according to Theorem 36.6. The corresponding eigenfunctions are

$$R_n(r) = D_n J_0\left(\frac{\beta_n r}{L}\right)$$

and solving for the ODE of T , we have

$$T_n(t) = C_n e^{-(\beta_n/L)^2 kt}.$$

By the principle of superposition, the general solution is

$$u(r, t) = a_0 + \sum_{n=1}^{\infty} a_n e^{-(\beta_n/L)^2 kt} J_0\left(\frac{\beta_n r}{L}\right), \quad a_n = C_n D_n.$$

From the initial condition and Theorem 36.8 with $\nu = 0$, we have for $n = 1, 2, \dots$,

$$\begin{aligned} u(r, 0) &= a_0 + \sum_{n=1}^{\infty} a_n J_0\left(\frac{\beta_n r}{L}\right) = f(r) \\ n = 0 \Rightarrow a_0 &= \frac{\int_0^L f(r) r dr}{\int_0^L r dr} = \frac{2}{L^2} \int_0^L f(r) r dr \\ n > 0 \Rightarrow a_n &= \frac{\int_0^L \left[J_0\left(\frac{\beta_n r}{L}\right) \right]^2 r dr}{\int_0^L \left[J_0(\beta_n x) \right]^2 x dx} = \frac{\int_0^L J_0\left(\frac{\beta_n r}{L}\right) f(r) r dr}{\int_0^L J_0(\beta_n x)^2 x dx} \\ &\Rightarrow a_n = \frac{1}{L^2 \times \frac{1}{2} \left(1 - \frac{0^2}{\beta_n^2}\right) [J_0(\beta_n)]^2} \int_0^L J_0\left(\frac{\beta_n r}{L}\right) f(r) r dr = \frac{2}{L^2 [J_0(\beta_n)]^2} \int_0^L J_0\left(\frac{\beta_n r}{L}\right) f(r) r dr. \end{aligned}$$

Note that $u_{\infty}(x) = a_0$. □

36.3.3 (Homogeneous) Type 3 BC

Let us now consider the case in which there is an exchange of heat between the surface of the cylinder and the surrounding medium (this is a common heat conduction problem that happens in practice), the temperature of which is assumed equal to zero. This problem is described by a type 3 boundary condition problem for the domain $r < L$:

$$\left(\frac{\partial u}{\partial r} + hu \right) \Big|_{r=L} = 0, \quad t \geq 0 \quad (36.48)$$

Similar to Topic 2, type 3 boundary condition is also complicated.

Example 36.18. Find the formal series solution for the heat equation (36.11) and the boundary condition (36.48).

Solution: This is a homogeneous problem, by separation of variable, we obtain an ODE and a

SL-problem:

$$T' + k\lambda T = 0, \quad R'' + \frac{1}{r}R' + \lambda R = 0, \quad \frac{dR}{dr}(L) + hR(L) = 0, \quad |R(0)| < \infty$$

where λ is a constant.

The general solution of R is of the form (there should be a general theory indicating that $\lambda > 0$ for type 3 boundary condition but I haven't found out):

$$R(r) = AJ_0(\sqrt{\lambda}r) + BY_0(\sqrt{\lambda}r).$$

Since $Y_0(\sqrt{\lambda}r) \rightarrow \infty$ as $r \rightarrow 0$, it follows $|R(0)| < \infty$ that $B = 0$. From the boundary condition,

$$\left(\frac{d}{dr} [AJ_0(\sqrt{\lambda}r)] + h(AJ_0(\sqrt{\lambda}r)) \right) \Big|_{r=L} = 0 \Rightarrow \sqrt{\lambda}J_0'(\sqrt{\lambda}L) + hJ_0(\sqrt{\lambda}L) = 0$$

$$\Rightarrow \lambda = \left(\frac{\mu_n}{L} \right)^2, \quad n = 1, 2, \dots$$

where the μ_n are the roots of

$$hLJ_0(z) + zJ_0'(z) = 0. \quad (36.49)$$

This equation has only real roots! However, how do we know that this equation has real roots? In the earlier sections, we don't have such a result. This result was stated in the Encyclopedia of Bessel functions — Watson [1922, 596–597]: When $\nu \geq \frac{1}{2}$, the roots of $zJ_\nu'(z) + HJ_\nu(z)$ are positive.

The eigenfunction corresponding to the eigenvalue λ is

$$R_n(r) = A_n J_0\left(\frac{\mu_n r}{L}\right).$$

Solving for the ODE of T , we obtain

$$T_n(t) = C_n e^{-(\mu_n/L)^2 kt}.$$

Hence we obtain a general solution

$$u(r, t) = \sum_{n=1}^{\infty} a_n J_0\left(\frac{\mu_n r}{L}\right) e^{-(\mu_n r/L)^2 t}, \quad (36.50)$$

where $\mu_1, \mu_2, \dots$ are the positive roots of (36.49) and each term of the series satisfies the Type 3 BC in (36.48).

Once we are able to deal with mathematical theory of finding roots of equations derived from type 3 boundary condition such as the case in the above example, the determination of coefficients of the general solution (36.50) will simply be the application of (36.27).

Example 36.19. Let n be a non-negative integer. If λ and μ are two different roots of the equation

$$hLJ_n(x) + xJ_n'(x) = 0,$$

then

$$\int_0^1 x J_n(\lambda x) J_n(\mu x) dx = 0; \quad \int_0^1 x (J_n(\lambda x))^2 dx = \frac{1}{2} \left[(J_n'(\lambda))^2 + \left(1 - \frac{n^2}{\lambda^2}\right) (J_n(\lambda))^2 \right].$$

Use this result to show that the problem involving heat equation (36.11), boundary condition (36.48) and initial condition (36.1) has a solution of the form

$$u(r, t) = \frac{2}{L^2} \sum_{n=1}^{\infty} \frac{\mu_n^2 \int_0^L \rho f(\rho) J_0\left(\frac{\mu_n \rho}{L}\right) d\rho}{(\mu_n^2 + h^2 L^2) J_0^2(\mu_n)} J_0\left(\frac{\mu_n r}{L}\right) e^{-(\mu_n r/L)^2 t}.$$

Solution. From the initial condition, we obtain

$$\sum_{n=1}^{\infty} a_n J_0\left(\frac{\mu_n r}{L}\right) = f(r).$$

Using the orthogonal property of Bessel functions, we have

$$a_n \int_0^L r \left[J_0\left(\frac{\mu_n r}{L}\right) \right]^2 dr = a_n L^2 \int_0^1 x \left[J_0(\mu_n x) \right]^2 dx = \int_0^L r f(r) J_0\left(\frac{\mu_n r}{L}\right) dr$$

by letting $x = r/L$. By applying the above formula (or (36.27))

$$\int_0^1 x \left[J_0(\mu_n x) \right]^2 dx = \frac{1}{2} \left[(J_0'(\mu_n))^2 + \left(1 - \frac{0^2}{\mu_n^2}\right) (J_0(\mu_n))^2 \right] = \frac{1}{2} \left[\frac{h^2 L^2}{\mu_n^2} (J_0(\mu_n))^2 + (J_0(\mu_n))^2 \right]$$

Hence, we have

$$a_n = \frac{2\mu_n^2}{L^2(\mu_n^2 + h^2 L^2) J_0^2(\mu_n)} \int_0^L r f(r) J_0\left(\frac{\mu_n r}{L}\right) dr.$$

Substituting this value of the coefficients a_n in the series (36.50), we have shown that the solution has the above form. $\square$

We will not proceed to discuss mixed boundary conditions for “ring”-domain due to the complexity.

36.4 2-D and 3-D Heat Equation

For a heat-flow on a disk or in an infinite cylinder, when the initial temperature does not have radial symmetry, e.g. (36.3) and (36.5), we need 2-D heat equation (36.12) and 3-D heat equation (36.14) in modelling engineering problem. The range of r will either be $0 < r < L$ or $a < r < b$, the range of ϕ can either be $[0, 2\pi]$ or $a \leq \phi \leq b$, and the range of z will either be $0 < z < \infty$ or $0 < z < H$. With various permutations of linear boundary conditions, there are more possibilities to discuss compare to Topic 2. However, many of them have little or no engineering applications and we will ignore them. We will just discuss a few examples to learn how separation of variables work.

36.4.1 In Finite Cylinder

Since 2-D problem (36.13) is a special case of 3-D problem, we will just consider a 3-D example with type 1 boundary condition. The type 2 and type 3 boundary conditions are more complicated as illustrated in the previous sections.

Example 36.20 (Type 1 BCs). Consider a problem of heat-flow in a circular cylinder of radius L and height H . Suppose that the initial temperature is given by (36.5) and that the temperature of its side and basis is held at zero. The problem can be approximated by heat equation (36.14), initial condition (36.5) and the following boundary conditions

$$\begin{cases} u(r, \phi, 0, t) = u(r, \phi, H, t) = 0, & 0 < r < L, \quad 0 \leq \phi < 2\pi, \quad t > 0, \\ u(L, \phi, z, t) = 0, & 0 \leq \phi < 2\pi, \quad 0 < z < H, \quad t > 0. \end{cases} \quad (36.51)$$

Find the solution of this problem using separation of variable.

Solution. Both the PDE and the boundary conditions are homogeneous, by applying separation of variables $u = TR\Phi Z$ to (36.14), we obtain two SL problems and two ODEs

$$\begin{aligned} \frac{T'}{kT} = \frac{R'' + \frac{R'}{r}}{R} + \frac{1}{r^2} \frac{\Phi''}{\Phi} + \frac{Z''}{Z} = -\lambda &\Rightarrow \frac{R'' + \frac{R'}{r}}{R} + \frac{1}{r^2} \frac{\Phi''}{\Phi} = -\lambda - \frac{Z''}{Z} = -\mu \\ &\Rightarrow r^2 \left(\frac{R'' + \frac{R'}{r}}{R} \right) + \mu r^2 = -\frac{\Phi''}{\Phi} = \gamma. \end{aligned}$$

Simplifying the above equations, we have

$$\begin{cases} T' = -\mu kT, \\ Z'' + (\lambda - \mu)Z = 0, & Z(0) = Z(H) = 0, \\ \Phi'' + \gamma\Phi = 0, & \Phi(0) = \Phi(2\pi), \quad \Phi'(0) = \Phi'(2\pi), \\ r^2 R'' + rR' + (\mu r^2 - \gamma)R = 0, & R(L) = 0, \quad |R(0)| < \infty. \end{cases}$$

First, let us investigate the periodic SL problem of Φ . (**Remark:** We gave the solution in Topic 1 but did not actually write down the solution steps.)

- $\boxed{\gamma < 0}$: $\Phi(\phi) = Ae^{\sqrt{-\gamma}x} + Be^{-\sqrt{-\gamma}x}$. The boundary conditions implies the following simultaneous equation

$$\begin{aligned} A(e^{2\pi\sqrt{-\gamma}} - 1) + B(e^{-2\pi\sqrt{-\gamma}} - 1) &= 0 \\ A\sqrt{-\gamma}(e^{2\pi\sqrt{-\gamma}} - 1) - B\sqrt{-\gamma}(e^{-2\pi\sqrt{-\gamma}} - 1) &= 0 \end{aligned}$$

The non-zero determinant indicates that the solution can only be trivial.

- $\boxed{\gamma = 0}$: $\Phi(\phi) = A + B\phi$. The boundary conditions implies

$$A = A + 2\pi B \Rightarrow B = 0$$

This indicates $\Phi(\phi) = A$, which is a constant function.

- $\boxed{\gamma > 0}$: $\Phi(\phi) = A \cos \sqrt{\gamma}x + B \sin \sqrt{\gamma}x$. The boundary conditions implies

$$\begin{aligned} A(\cos(2\pi\sqrt{\gamma}) - 1) + B\sin(2\pi\sqrt{\gamma}) &= 0 \\ A\sqrt{\gamma}\sin(2\pi\sqrt{\gamma}) + B\sqrt{\gamma}(1 - \cos(2\pi\sqrt{\gamma})) &= 0 \end{aligned}$$

To have non-trivial solution, we need $-(1 - \cos(2\pi\sqrt{\gamma}))^2 - \sin^2(2\pi\sqrt{\gamma}) = 0 \Rightarrow 2 - 2\cos(2\pi\sqrt{\gamma}) = 0 \Rightarrow 2\pi\sqrt{\gamma} = 2m\pi$ for $m = 0, \pm 1, \pm 2, \dots$. From this and $\lambda = 0$, the eigenvalues and eigenfunctions of the periodic SL problem of Φ are determined to be

$$\boxed{\gamma = m^2, \quad \Phi(\phi) = A_m \cos m\phi + B_m \sin m\phi, \quad m = 0, \pm 1, \pm 2, \dots}$$

With $\gamma = m^2$, the singular SL problem of R becomes Bessel equation of order m with parameter λ and has eigenvalues and eigenfunctions

$$\mu = \left(\frac{\alpha_k^{(m)}}{L}\right)^2, \quad R(r) = C_{km} J_m\left(\frac{\alpha_k^{(m)} r}{L}\right), \quad k = 1, 2, \dots, \quad m = 0, 1, 2, \dots,$$

where $\alpha_k^{(m)}$ are the roots of the Bessel function J_m .

With $\gamma = m^2$ and $\mu = \left(\frac{\alpha_k^{(m)}}{L}\right)^2$, the eigenvalues and eigenfunctions of the regular SL problem of Z are

$$\lambda - \mu = \left(\frac{n\pi}{H}\right)^2, \quad Z(z) = D_{mnk} \sin \frac{n\pi z}{H}, \quad n = 1, 2, \dots$$

and the corresponding solution of T is

$$T(t) = E_{mnk} \exp \left\{ - \left[\left(\frac{n\pi}{H}\right)^2 + \left(\frac{\alpha_k^{(m)}}{L}\right)^2 \right] kt \right\}.$$

By the principle of superposition, the general solution is found to be

$$u = \sum_{k=1}^{\infty} \sum_{m=0}^{\infty} \sum_{n=1}^{\infty} e^{-k \left[\left(\frac{\alpha_k^{(m)}}{L}\right)^2 + \frac{n^2 \pi^2}{H^2} \right] t} J_m\left(\frac{\alpha_k^{(m)} r}{L}\right) \left[A_{mnk} \cos m\phi + B_{mnk} \sin m\phi \right] \sin \frac{n\pi z}{H} \quad (36.52)$$

in which the coefficients A_{kmn} and B_{kmn} are still to be determined. Let us set $t = 0$ in the expansion (36.52). Then, taking account of the initial condition, we obtain

$$f(r, \phi, z) = \sum_{k=1}^{\infty} \sum_{m=0}^{\infty} \sum_{n=1}^{\infty} J_m\left(\frac{\alpha_k^{(m)} r}{L}\right) \left[A_{mnk} \cos m\phi + B_{mnk} \sin m\phi \right] \sin \frac{n\pi z}{H}. \quad (36.53)$$

Since the right side is an expansion of the function f in a Fourier series in $\cos m\phi$ and $\sin m\phi$, the coefficients of these trigonometric functions can be determined from the familiar formulae. Thus,

we obtain

$$\begin{aligned}\frac{1}{2\pi} \int_0^{2\pi} f(r, \phi, z) d\phi &= \sum_{k=1}^{\infty} \left[\sum_{n=1}^{\infty} A_{0nk} \sin \frac{n\pi z}{H} \right] J_0\left(\frac{\alpha_k^{(0)} r}{L}\right), \\ \frac{1}{\pi} \int_0^{2\pi} f(r, \phi, z) \cos m\phi d\phi &= \sum_{k=1}^{\infty} \left[\sum_{n=1}^{\infty} A_{mnk} \sin \frac{n\pi z}{H} \right] J_m\left(\frac{\alpha_k^{(m)} r}{L}\right), \\ \frac{1}{\pi} \int_0^{2\pi} f(r, \phi, z) \sin m\phi d\phi &= \sum_{k=1}^{\infty} \left[\sum_{n=1}^{\infty} B_{mnk} \sin \frac{n\pi z}{H} \right] J_m\left(\frac{\alpha_k^{(m)} r}{L}\right).\end{aligned}$$

Each of these equations itself is an expansion of the function treated as a function of r in a series of Bessel functions. The coefficients are given below:

$$\sum_{n=1}^{\infty} A_{0nk} \sin \frac{n\pi z}{H} = \frac{1}{\pi L^2 J_1^2(\alpha_k^{(0)})} \int_0^L \int_0^{2\pi} r f(r, \phi, z) J_0\left(\frac{\alpha_k^{(0)} r}{L}\right) dr d\phi, \quad (36.54)$$

e:104

$$\sum_{n=1}^{\infty} A_{mnk} \sin \frac{n\pi z}{H} = \frac{2}{\pi L^2 J_{n+1}^2(\alpha_k^{(m)})} \int_0^L \int_0^{2\pi} r f(r, \phi, z) J_m\left(\frac{\alpha_k^{(m)} r}{L}\right) \cos m\phi dr d\phi, \quad (36.55)$$

e:105

$$\sum_{n=1}^{\infty} B_{mnk} \sin \frac{n\pi z}{H} = \frac{2}{\pi L^2 J_{n+1}^2(\alpha_k^{(m)})} \int_0^L \int_0^{2\pi} r f(r, \phi, z) J_m\left(\frac{\alpha_k^{(m)} r}{L}\right) \sin m\phi dr d\phi. \quad (36.56)$$

e:106

Since the functions $\sin \frac{n\pi z}{H}$ ($n = 1, 2, \dots$) form an orthogonal system of functions on the interval $[0, H]$, we can obtain the coefficients A_{mnk} and B_{mnk} in the expansions (36.54)–(36.56) are given by the formulae

$$\begin{aligned}A_{0nk} &= \frac{2}{\pi H L^2 J_1^2(\alpha_k^{(0)})} \int_0^L \int_0^{2\pi} \int_0^H r f(r, \phi, z) J_0\left(\frac{\alpha_k^{(0)} r}{L}\right) \sin \frac{n\pi z}{H} dr d\phi dz, \\ A_{mnk} &= \frac{4}{\pi H L^2 J_{n+1}^2(\alpha_k^{(m)})} \int_0^L \int_0^{2\pi} \int_0^H r f(r, \phi, z) J_m\left(\frac{\alpha_k^{(m)} r}{L}\right) \cos m\phi \sin \frac{n\pi z}{H} dr d\phi dz, \\ B_{mnk} &= \frac{4}{\pi H L^2 J_{n+1}^2(\alpha_k^{(m)})} \int_0^L \int_0^{2\pi} \int_0^H r f(r, \phi, z) J_m\left(\frac{\alpha_k^{(m)} r}{L}\right) \sin m\phi \sin \frac{n\pi z}{H} dr d\phi dz.\end{aligned}$$

When we substitute these values of the coefficients into the series (36.52), we obtain the final solution to the problem. $\square$

36.4.2 In Infinite Cylinder

First, let us consider an example of infinite cylinder with some symmetry.

Example 36.21. Consider a heat equation in an infinite cylinder of radius L governed by (36.12) with type 1 homogeneous boundary condition

$$u(L, \phi, t) = 0, \quad 0 \leq \phi \leq 2\pi, \quad t \geq 0,$$

and initial condition (36.3). Use separation of variables to determine the formal series solution of this problem.

Solution. Let $u = TR\Phi$. Then the PDE can be simplified to

$$\frac{T'}{T} = \frac{R''}{R} + \frac{1}{r} \frac{R'}{R} + \frac{1}{r^2} \frac{\Phi''}{\Phi} = -\lambda.$$

Separating the variables we have three ODEs and from the boundary conditions, we have

$$\begin{cases} T' = -\lambda k T, \\ \Phi'' + \mu \Phi = 0, \\ r^2 R'' + r R' + (\lambda r^2 - \mu) R = 0, \end{cases} \quad \begin{cases} \Phi(0) = \Phi(2\pi), \quad \Phi'(0) = \Phi'(2\pi), \\ R(L) = 0, \quad |R(0)| < \infty \end{cases}$$

where λ and μ are constants to be determined.

Solving the periodic SL problem of Φ , we have

$$\mu = m^2, \quad \Phi_m(x) = A_m \cos m\phi + B_m \sin m\phi, \quad m = 0, 1, 2, \dots$$

For $\mu = m^2$, the general solution of the singular SL problem of R is

$$R(r) = C J_m(\sqrt{\lambda}r) + D Y_m(\sqrt{\lambda}r)$$

Since $|R(0)| < \infty$, $D = 0$. From the BC,

$$R(L) = C J_m(\sqrt{\lambda}L) = 0 \Rightarrow \sqrt{\lambda}L = \gamma_n^{(m)} \Rightarrow \lambda = \left(\frac{\gamma_n^{(m)}}{L}\right)^2 \quad n = 1, 2, \dots$$

where $\gamma_n^{(m)}$, $n = 1, 2, \dots$ are the zeroes of the Bessel function J_m .

Finally solving for the ODE of T , we have

$$T_{mn}(t) = E_{mn} e^{-\left(\frac{\gamma_n^{(m)}}{L}\right)^2 kt}$$

By the principle of superposition, the general solution is

$$u = \sum_{m=0}^{\infty} \sum_{n=1}^{\infty} e^{-\left(\frac{\gamma_n^{(m)}}{L}\right)^2 kt} J_m\left(\frac{\gamma_n^{(m)} r}{L}\right) (a_{mn} \cos m\phi + b_{mn} \sin m\phi)$$

where $a_{mn} = E_{mn} C A_m$, $b_{mn} = E_{mn} C B_m$ are constants.

From the initial condition, we have

$$u(r, \phi, 0) = \sum_{m=0}^{\infty} \sum_{n=1}^{\infty} J_m\left(\frac{\gamma_n^{(m)} r}{L}\right) (a_{mn} \cos m\phi + b_{mn} \sin m\phi) = f(r, \phi).$$

Using the orthogonality property we have

$$\begin{aligned} a_{mn} &= \frac{\int_0^{2\pi} \int_0^L f(r, \phi) J_m\left(\frac{\gamma_n^{(m)} r}{L}\right) \cos m\phi r dr d\phi}{\int_0^{2\pi} \int_0^L J_m^2\left(\frac{\gamma_n^{(m)} r}{L}\right) \cos^2 m\phi r dr d\phi} = \frac{\int_0^{2\pi} \int_0^L f(r, \phi) J_m\left(\frac{\gamma_n^{(m)} r}{L}\right) \cos m\phi r dr d\phi}{\int_0^{2\pi} \cos^2 m\phi d\phi \times \int_0^L J_m^2\left(\frac{\gamma_n^{(m)} r}{L}\right) r dr} \\ &= \frac{2}{\pi(1 + \delta_{m0}) \cdot L^2 J_{m+1}^2(\gamma_n^{(m)})} \int_0^{2\pi} \int_0^L f(r, \phi) J_m\left(\frac{\gamma_n^{(m)} r}{L}\right) \cos m\phi r dr d\phi \\ b_{mn} &= \frac{\int_0^{2\pi} \int_0^L f(r, \phi) J_m\left(\frac{\gamma_n^{(m)} r}{L}\right) \sin m\phi r dr d\phi}{\int_0^{2\pi} \int_0^L J_m^2\left(\frac{\gamma_n^{(m)} r}{L}\right) \sin^2 m\phi r dr d\phi} \\ &= \frac{2}{\pi L^2 J_{m+1}^2(\gamma_n^{(m)})} \int_0^{2\pi} \int_0^L f(r, \phi) J_m\left(\frac{\gamma_n^{(m)} r}{L}\right) \sin m\phi r dr d\phi \end{aligned}$$

□

It is very strange to include the infinite cylinder without any symmetry in the syllabus because when we encounter the range $\mathbb{R}$, Fourier transform (Topic 5) is required [González-Velasco, 1995, Chapter 7]. 2D and 3D problems normally arise when the initial “temperature” distribution does not have any symmetry and is a function of two variables or three variables.

Example 36.22 (Homogeneous Type 1 BC). When the temperature on the lateral surface of the cylinder is kept equal to zero, the problem is characterised by (36.14) with $z \in \mathbb{R}$, a homogeneous type 1 boundary condition

$$u(L, \phi, z, t) = 0, \quad 0 \leq \phi < 2\pi, \quad t \geq 0 \quad (36.57)$$

and an initial condition (36.5). Use “separation of variables” to solve this problem.

Solution: Similar to Example 36.20, let $u(r, \phi, z, t) = TR\Phi Z$ and substituting it into (36.14), we obtain two SL problems and two ODEs

$$\begin{aligned} \frac{T'}{a^2 T} &= \frac{R'' + \frac{R'}{r}}{R} + \frac{1}{r^2} \frac{\Phi''}{\Phi} + \frac{Z''}{Z} = -\lambda \Rightarrow \frac{R'' + \frac{R'}{r}}{R} + \frac{1}{r^2} \frac{\Phi''}{\Phi} = -\lambda - \frac{Z''}{Z} = -\mu \\ &\Rightarrow r^2 \left(\frac{R'' + \frac{R'}{r}}{R} \right) + \mu r^2 = -\frac{\Phi''}{\Phi} = \gamma. \end{aligned}$$

Simplifying the above equations, we have

$$\begin{cases} T' = -\lambda a^2 T, & \boxed{\text{Helmholtz Eqn} + \text{Type 1 BC} \Rightarrow \lambda > 0} \\ Z'' + (\lambda - \mu)Z = 0, & |Z(\infty)| < \infty, \\ \Phi'' + \gamma\Phi = 0, & \Phi(0) = \Phi(2\pi), \quad \Phi'(0) = \Phi'(2\pi), \\ r^2 R'' + rR' + (\mu r^2 - \gamma)R = 0, & R(L) = 0, \quad |R(0)| < \infty. \end{cases}$$

The eigenvalues and eigenfunctions of the periodic SL problem of Φ are

$$\gamma = m^2, \quad \Phi(\phi) = A_m \cos m\phi + B_m \sin m\phi, \quad m = 0, 1, 2, \dots$$

With $\gamma = m^2$, the singular SL problem of R becomes Bessel equation and has eigenvalues and eigenfunctions

$$\mu = \left(\frac{\mu_k^{(m)}}{L}\right)^2, \quad R(r) = C_{km} J_m\left(\frac{\mu_k^{(m)} r}{L}\right), \quad k = 1, 2, \dots, \quad m = 0, 1, 2, \dots,$$

where $\mu_k^{(m)}$ are the roots of the Bessel function J_m .

Since Z is defined over $\mathbb{R}$, we need to extend the analysis of Fourier series to make it work by generalising some of the idea to Fourier transform from Topic 5. First, the eigenvalues will no longer be “discrete”. They are continuous because when we try to find the eigenvalues, we find using the characteristic equation of constant coefficient linear ODE that when $\lambda - \mu > 0$, there is an eigenvalue (when $\lambda - \mu \leq 0$, the solution is ∞ when $z \rightarrow \infty$):

$$\lambda - \mu = \omega^2 > 0 \Rightarrow Z(z) = D \cos \omega z + E \sin \omega z$$

and so the solution of T is

$$\lambda = \omega^2 + \left(\frac{\mu_k^{(m)}}{L}\right)^2, \quad T(t) = F_{mk} \exp\left\{-\left(\left(\frac{\mu_k^{(m)}}{L}\right)^2 + \omega^2\right)kt\right\}.$$

Since there are no restrictions on the constant ω , the general solution is

$$\begin{aligned} u(r, \phi, z) = \sum_{m=0}^{\infty} \sum_{k=1}^{\infty} \int_{-\infty}^{\infty} e^{-\left(\left(\frac{\mu_k^{(m)}}{L}\right)^2 + \omega^2\right)kt} \Big\{ [A_{mk}(\omega) \cos \omega z + B_{mk}(\omega) \sin \omega z] \cos m\phi \\ + [C_{mk}(\omega) \cos \omega z + D_{mk}(\omega) \sin \omega z] \sin m\phi \Big\} J_m\left(\frac{\mu_k^{(m)} r}{L}\right) d\omega. \end{aligned} \quad (36.58)$$

e:93

To determine the functions $A_{mk}(\omega)$, etc., we substitute ^{e:93} (36.58) into the initial condition:

$$\begin{aligned} f(r, \phi, z) = \sum_{k=1}^{\infty} \int_{-\infty}^{\infty} [A_{0k}(\omega) \cos \omega z + B_{0k}(\omega) \sin \omega z] J_0\left(\frac{\mu_k^{(0)} r}{L}\right) d\omega \\ + \sum_{m=1}^{\infty} \left[\sum_{k=1}^{\infty} J_m\left(\frac{\mu_k^{(m)} r}{L}\right) \int_{-\infty}^{\infty} [A_{mk}(\omega) \cos \omega z + B_{mk}(\omega) \sin \omega z] d\omega \right] \cos m\phi \\ + \sum_{m=1}^{\infty} \left[\sum_{k=1}^{\infty} J_m\left(\frac{\mu_k^{(m)} r}{L}\right) \int_{-\infty}^{\infty} [C_{mk}(\omega) \cos \omega z + D_{mk}(\omega) \sin \omega z] d\omega \right] \sin m\phi. \end{aligned}$$

We now compare this equality with the expansion of the function $f(r, \phi, z)$, in the sine-and-cosine series over the interval $(0, 2\pi)$, in which expansion we treat f as a function of ϕ . We thus obtain

$$\begin{aligned} \frac{1}{2\pi} \int_0^{2\pi} f(r, \phi, z) d\phi &= \sum_{k=1}^{\infty} J_0\left(\frac{\mu_k^{(0)} r}{L}\right) \int_{-\infty}^{\infty} [A_{0k}(\omega) \cos \omega z + B_{0k}(\omega) \sin \omega z] d\omega, \\ \frac{1}{\pi} \int_0^{2\pi} f(r, \phi, z) \cos m\phi d\phi &= \sum_{k=1}^{\infty} J_m\left(\frac{\mu_k^{(m)} r}{L}\right) \int_{-\infty}^{\infty} [A_{mk}(\omega) \cos \omega z + B_{mk}(\omega) \sin \omega z] d\omega, \\ \frac{1}{\pi} \int_0^{2\pi} f(r, \phi, z) \sin m\phi d\phi &= \sum_{k=1}^{\infty} J_m\left(\frac{\mu_k^{(m)} r}{L}\right) \int_{-\infty}^{\infty} [C_{mk}(\omega) \cos \omega z + D_{mk}(\omega) \sin \omega z] d\omega. \end{aligned}$$

Each right-hand-side of the above equations is itself has the Fourier-Bessel series expansion

$$g(r) = \sum_{k=1}^{\infty} a_k J_m\left(\frac{\mu_k^{(m)} r}{L}\right), \quad a_k = \frac{2 \int_0^L r g(r) J_m\left(\frac{\mu_k^{(m)} r}{L}\right) dr}{L^2 J_{m+1}^2(\mu_k^{(m)})}$$

where m is a non-negative integer. Therefore we have

$$\begin{aligned} \int_{-\infty}^{\infty} [A_{0k}(\omega) \cos \omega z + B_{0k}(\omega) \sin \omega z] d\omega &= \frac{\int_0^L \int_0^{2\pi} f(r, \phi, z) J_0\left(\frac{\mu_k^{(0)} r}{L}\right) r dr d\phi}{\pi L^2 J_1^2(\mu_k^{(0)})}, \\ \int_{-\infty}^{\infty} [A_{mk}(\omega) \cos \omega z + B_{mk}(\omega) \sin \omega z] d\omega &= \frac{2 \int_0^L \int_0^{2\pi} f(r, \phi, z) J_m\left(\frac{\mu_k^{(m)} r}{L}\right) \cos(m\phi) r dr d\phi}{\pi L^2 J_{m+1}^2(\mu_k^{(m)})}, \\ \int_{-\infty}^{\infty} [C_{mk}(\omega) \cos \omega z + D_{mk}(\omega) \sin \omega z] d\omega &= \frac{2 \int_0^L \int_0^{2\pi} f(r, \phi, z) J_m\left(\frac{\mu_k^{(m)} r}{L}\right) \sin(m\phi) r dr d\phi}{\pi L^2 J_{m+1}^2(\mu_k^{(m)})}. \end{aligned}$$

The right-hand-side of the above equations integrate to some function h of z

$$\int_{-\infty}^{\infty} [a(\omega) \cos \omega z + b(\omega) \sin \omega z] d\omega = h(z),$$

By using the Fourier Integral Theorem from Topic 5 with $\omega = 2\pi\xi$, a and b can be expressed as

$$a(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} h(z) \cos \omega z dz, \quad b(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} h(z) \sin \omega z dz.$$

From these formulae, we have expressions for the desired functions $A_{0k}(\omega), \dots, D_{mk}(\omega)$, namely,

$$\begin{aligned} A_{mk}(\omega) &= \frac{1}{\delta_m \pi^2 L^2 J_{m+1}^2(\mu_k^{(m)})} \int_0^L \int_0^{2\pi} \int_{-\infty}^{\infty} f(r, \tilde{\phi}, z) \left[J_m\left(\frac{\mu_k^{(m)} r}{L}\right) \cos m\tilde{\phi} \cos \omega z \right] r dr d\tilde{\phi} dz, \\ B_{mk}(\omega) &= \frac{1}{\delta_m \pi^2 L^2 J_{m+1}^2(\mu_k^{(m)})} \int_0^L \int_0^{2\pi} \int_{-\infty}^{\infty} f(r, \tilde{\phi}, z) \left[J_m\left(\frac{\mu_k^{(m)} r}{L}\right) \cos m\tilde{\phi} \sin \omega z \right] r dr d\tilde{\phi} dz, \\ C_{mk}(\omega) &= \frac{1}{\pi^2 L^2 J_{m+1}^2(\mu_k^{(m)})} \int_0^L \int_0^{2\pi} \int_{-\infty}^{\infty} f(r, \tilde{\phi}, z) \left[J_m\left(\frac{\mu_k^{(m)} r}{L}\right) \sin m\tilde{\phi} \cos \omega z \right] r dr d\tilde{\phi} dz, \\ D_{mk}(\omega) &= \frac{1}{\pi^2 L^2 J_{m+1}^2(\mu_k^{(m)})} \int_0^L \int_0^{2\pi} \int_{-\infty}^{\infty} f(r, \tilde{\phi}, z) \left[J_m\left(\frac{\mu_k^{(m)} r}{L}\right) \sin m\tilde{\phi} \sin \omega z \right] r dr d\tilde{\phi} dz \end{aligned}$$

where $\delta_0 = 2$ and $\delta_m = 1$ for $m > 0$.

If we substitute them into the series (36.58) and carry out the integration with respect to ρ , using the integral

$$\int_{-\infty}^{\infty} e^{-\alpha^2 \rho^2} \cos \beta \rho d\rho = \frac{\sqrt{\pi}}{\alpha} e^{-\frac{\beta^2}{4\alpha^2}},$$

we obtain the solution in the form

$$\begin{aligned} u(r, \phi, z) &= \frac{1}{a\pi L^2 \sqrt{\pi t}} \sum_{m=0}^{\infty} \sum_{k=1}^{\infty} e^{-\left(\frac{\mu_k^{(m)}}{L}\right)^2 t} \frac{J_m\left(\frac{\mu_k^{(m)} r}{L}\right)}{\delta_m J_{m+1}^2(\mu_k^{(m)})} \times \\ &\quad \int_0^L \int_0^{2\pi} \int_{-\infty}^{\infty} f(\rho, \tilde{\phi}, \zeta) J_m\left(\frac{\mu_k^{(m)} \omega}{L}\right) \cos m(\phi - \tilde{\phi}) e^{\frac{(z-\zeta)^2}{4a^2 t}} \rho d\rho d\tilde{\phi} d\zeta. \end{aligned}$$

36.5 2-D and 3-D Laplace Equations

Similar to Topic 2, the Laplace equation we consider will be of “dimension” 2 or 3 because for “1-D”, we will just have an Euler equation which can be solved using the techniques from UECM2023 ODE:

$$u_{rr} + \frac{1}{r} u_r = 0 \Rightarrow u = A + B \ln r$$

where A and B are constants. However, it is impossible to discuss all possible “permutations” of boundary conditions. Instead, we will discuss some typical cases in this section for various domains mentioned in Section 36.1.1.

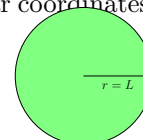
With the knowledge of complex analysis (UECM3013), we know that theoretically we can transform any simply connected domain which is not $\mathbb{C}$ to a unit circle by a conformal mapping. Therefore, the technique mentioned here can be complimented with conformal mapping to solve more complicated 2-D Laplace equation.

As mentioned in Section 36.1.1, periodic boundary condition (36.7) is implicitly assumed for “inner” disk, “outer” disk and “ring”. In addition, the regularity condition (36.8) will also be implicitly assumed throughout this section and later sections.

36.5.1 “Inner” Disk, “Outer” Disk and “Ring”

Example 36.23. (Inner Disk Type 1 BC, Weinberger [1995, 103]) Consider the Laplace equation with type 1 BC on a *circular disk* or *infinite cylinder* (which is most convenient described by polar coordinates) is given below:

$$\begin{cases} \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} = 0, & 0 < r < L, \quad 0 \leq \phi < 2\pi, \\ u(L, \phi) = f(\phi), & 0 \leq \phi < 2\pi. \end{cases}$$



Write down the formal series solution of this problem.

Solution: Suppose that $u(r, \phi) = R(r)\Phi(\phi)$, the PDE simplifies to

$$\frac{R''}{R} + \frac{1}{r} \frac{R'}{R} = -\frac{\Phi''}{\Phi} = \mu.$$

Together with the periodic BC, we obtain a periodic SL-problem of Φ and an ODE of R by the separation of variables method

$$\begin{cases} \Phi'' + \mu\Phi = 0, \\ \Phi(0) = \Phi(2\pi), \\ \Phi'(0) = \Phi'(2\pi), \end{cases} \quad \text{and} \quad \begin{cases} r^2 R'' + rR' - \mu R = 0, & r < L, \\ |R(0)| < \infty. \end{cases}$$

Solving the periodic SL problem of Φ , we obtain the eigenvalues and the corresponding eigenfunctions are

$$\mu = m^2, \quad \Phi(\phi) = A_m \cos m\phi + B_m \sin m\phi, \quad m = 0, 1, 2, \dots$$

For $\mu = 0$, we can solve the ODE for R :

$$R(r) = C_0 + D_0 \ln r.$$

The regularity condition requires $D_0 = 0$ so that $|u| < \infty$ when $r \rightarrow 0$.

For $\mu = m^2 > 0$, $m = 1, 2, \dots$, we can solve the ODE of R (which is an Euler equation):

$$R(r) = C_m r^m + D_m r^{-m}.$$

The regularity condition requires $D_m = 0$, therefore, together with the case $\mu = 0$, we have

$$R(r) = C_m r^m, \quad m = 0, 1, 2, \dots$$

Using the principle of superposition, the (formal) series solution is obtained:

$$u(r, \phi) = a_0 + \sum_{m=1}^{\infty} \{a_m r^m \cos m\phi + b_m r^m \sin m\phi\}$$

where $a_m = C_m A_m$ and $b_m = C_m B_m$ are constants.

The boundary condition $u(L, \phi) = f(\phi)$ now allows us to determine the coefficients a_m and b_m as the Fourier coefficients of the trigonometric Fourier series of $f(\phi)$ over the interval $[0, 2\pi]$. This is equivalent to the interval $[-\pi, \pi]$ since f is 2π -periodic. Therefore,

$$a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\phi) d\phi, \quad a_m = \frac{1}{\pi L^m} \int_{-\pi}^{\pi} f(\phi) \cos m\phi d\phi, \quad b_m = \frac{1}{\pi L^m} \int_{-\pi}^{\pi} f(\phi) \sin m\phi d\phi.$$

If we rewrite the trigonometric Fourier series solution to the complex form (recall Topic 1), we have

$$u(r, \phi) = a_0 + \sum_{m=1}^{\infty} \left\{ a_m r^m \frac{e^{im\phi} + e^{-im\phi}}{2} + b_m r^m \frac{ie^{-im\phi} - ie^{im\phi}}{2} \right\} = \sum_{m=-\infty}^{\infty} c_m r^{|m|} e^{im\phi}$$

where $c_0 = a_0$, $c_m = \frac{a_m - ib_m}{2}$ and $c_{-m} = \frac{a_m + ib_m}{2}$, $m = 1, 2, \dots$.

In this case, the boundary condition $u(L, \phi) = f(\phi)$ now determines the c_m as the coefficients in the complex Fourier series for $f(\phi)$ over $[-\pi, \pi]$:

$$c_m = \frac{1}{2\pi L^m} \int_{-\pi}^{\pi} f(\varphi) e^{-im\varphi} d\varphi = \frac{1}{2\pi L^m} \int_{-\pi}^{\pi} f(\varphi) e^{-im\varphi} d\varphi, \quad m = 0, \pm 1, \pm 2, \dots$$

Substituting it to $u(r, \phi)$ and transposition of summation and integration yields

$$u(r, \phi) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\sum_{m=-\infty}^{\infty} \left(\frac{r}{L} \right)^{|m|} e^{im(\phi-\varphi)} \right] f(\varphi) d\varphi.$$

But by the formula for the sum of a geometric series

$$\sum_{m=-\infty}^{\infty} \left(\frac{r}{L} \right)^{|m|} e^{im(\phi-\varphi)} = \frac{1 - \left(\frac{r}{L} \right)^2}{1 - 2 \left(\frac{r}{L} \right) \cos(\phi - \varphi) + \left(\frac{r}{L} \right)^2}$$

we have a solution in an elegant integral form called the *Poisson integral formula* (this result is also found in UECM3013 Complex Analysis, Topic 9, Application to Potential Theory):

$$u(r, \phi) = \frac{L^2 - r^2}{2\pi} \int_{-\pi}^{\pi} \frac{f(\varphi)}{L^2 - 2Lr \cos(\phi - \varphi) + r^2} d\varphi.$$

Example 36.24 (UCCM3003 Final Exam, Apr 2008, Q3). Consider the following Laplace Equation in an infinite cylinder

$$\begin{cases} \Delta u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} = 0, & 0 < r < 3, \\ u(r, \phi + 2\pi) = u(r, \phi), & 0 < r < 3, \\ u(3, \phi) = 1 + \frac{1}{2} \cos 2\phi + 9 \cos 5\phi \end{cases}$$

(a) By using separation of variables, show that the general solution is

$$u(r, \phi) = \sum_{n=0}^{\infty} r^n (a_n \cos n\phi + b_n \sin n\phi). \quad (17 \text{ marks})$$

Solution. Let $u = R\Phi$. Then the PDE can be simplified to

$$\frac{R''}{R} + \frac{1}{r} \frac{R'}{R} + \frac{1}{r^2} \frac{\Phi''}{\Phi} = 0 \quad [2 \text{ marks}]$$

Separating the variables we have

$$r^2 \frac{R''}{R} + r \frac{R'}{R} = -\frac{\Phi''}{\Phi} = \mu \quad [2 \text{ marks}]$$

where μ is a constant.

From the BC, we have

$$\Phi(2\pi) = \Phi(0), \quad \Phi'(2\pi) = \Phi'(0) \quad [2 \text{ marks}]$$

Solving the SL problem for Φ , we have (see Formula Sheet)

$$\mu = n^2, \quad \Phi(\phi) = A_n \cos n\phi + B_n \sin n\phi, \quad n = 0, 1, 2, \dots \quad [1+1+1=3 \text{ marks}]$$

For $\mu = 0$, we have

$$r^2 R'' + rR' = 0 \Rightarrow (rR')' = 0 \Rightarrow rR' = D_0 \Rightarrow R = D_0 \ln \frac{1}{r} + C_0. \quad [2 \text{ marks}]$$

For $\mu = n^2, n = 1, 2, \dots$, we have (see Formula Sheet)

$$r^2 R'' + rR' - n^2 R = 0 \Rightarrow R = C_n r^n + D_n r^{-n} \quad [2 \text{ marks}]$$

Since R needs to be well-defined for $0 < r < 3$, $D_0 = D_n = 0$ (otherwise R is singular at $r = 0$).
[2 marks]

By the principle of superposition, the general solution is

$$u(r, \phi) = \sum_{n=0}^{\infty} r^n (a_n \cos n\phi + b_n \sin n\phi) \quad [2 \text{ marks}]$$

where $a_n = A_n C_n$ and $b_n = B_n C_n$ are constants. □

- (b) Find the solution of the Laplace equation by evaluating a_n and b_n in part (a). (8 marks)

Solution. From the boundary condition, we have

$$u(3, \phi) = \sum_{n=0}^{\infty} 3^n (a_n \cos n\phi + b_n \sin n\phi) = 1 + \frac{1}{2} \cos 2\phi + 9 \cos 5\phi \quad [1 \text{ mark}]$$

Comparing coefficients, we have

$$a_0 = 1, \quad 3^2 a_2 = \frac{1}{2}, \quad 3^5 a_5 = 9; \quad b_n = 0, \quad n \geq 0, \quad a_n = 0, \quad \text{for } n \neq 0, 2, 5. \quad [1+1+1+1+1=5 \text{ marks}]$$

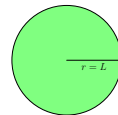
Hence, the solution is

$$u(r, \phi) = 1 + \frac{1}{18} r^2 \cos 2\phi + \frac{1}{27} r^5 \cos 5\phi. \quad [2 \text{ marks}]$$

□

Example 36.25 (Inner “Disk” Type 2 BC). The Laplace equation with type 2 BC on a disk or homogeneous infinite cylinder is given below:

$$\begin{cases} \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} = 0, & 0 < r < L, \quad 0 \leq \phi < 2\pi, \\ \frac{\partial u}{\partial r}(L, \phi) = f(\phi), & 0 \leq \phi < 2\pi. \end{cases}$$



Solution: The general solution is the same as in Example 36.23: [eq:cyl:lapl:typel:in]

$$u(r, \phi) = \sum_{m=0}^{\infty} \{a_m r^m \cos m\phi + b_m r^m \sin m\phi\}.$$

The boundary condition implies that

$$\left. \frac{\partial u}{\partial r} \right|_{r=L} = \sum_{m=1}^{\infty} \{a_m m L^{m-1} \cos m\phi + b_m m L^{m-1} \sin m\phi\} = f(\phi).$$

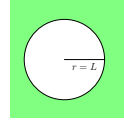
Hence a_0 is an *arbitrary constant* and the coefficients $a_m, b_m, m \geq 1$, can be obtained by applying the orthogonality of sine and cosine functions (Topic 1):

$$a_m = \frac{1}{m L^{m-1} \pi} \int_{-\pi}^{\pi} f(\phi) \cos m\phi d\phi, \quad b_m = \frac{1}{m L^{m-1} \pi} \int_{-\pi}^{\pi} f(\phi) \sin m\phi d\phi$$

[eq:outdisk:bc1]

Example 36.26 (Outer Disk Type 1 BC). Consider the Laplace equation *outside the disk* with Type 1 BC is given below:

$$\begin{cases} \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} = 0, & r > L, \quad 0 \leq \phi < 2\pi, \\ u(L, \phi) = f(\phi), & 0 \leq \phi < 2\pi. \end{cases}$$



Solution. Using the separation of variables, we can obtain

$$\begin{cases} \Phi'' + \mu\Phi = 0, \\ \Phi(0) = \Phi(2\pi), \\ \Phi'(0) = \Phi'(2\pi), \end{cases} \quad \text{and} \quad \begin{cases} r^2 R'' + rR' - \mu R = 0, & L < r, \\ |R(\infty)| < \infty. \end{cases}$$

The eigenvalues and eigenfunctions to the SL problem of Φ are

$$\mu = n^2, \quad \Phi(\phi) = A_n \cos n\phi + B_n \sin n\phi, \quad n = 0, 1, 2, \dots$$

When $\mu = 0$, the solution of the ODE of R is

$$R(r) = C_0 + D_0 \ln r$$

The regularity condition $|R(\infty)| < \infty$ requires $D_0 = 0$.

When $\mu = n^2, n = 1, 2, \dots$, the solution of the ODE of R is

$$R(r) = C_n r^n + D_n r^{-n}.$$

The regularity condition requires C_n to be zero:

$$\lim_{r \rightarrow \infty} R(r) \text{ exists} \Rightarrow C_n = 0 \Rightarrow R(r) = D_n r^{-n}.$$

By the principle of superposition, the general solution is

$$u(r, \phi) = a_0 + \sum_{n=1}^{\infty} \{a_n r^{-n} \cos n\phi + b_n r^{-n} \sin n\phi\}$$

From the boundary condition, we have

$$a_n = \frac{L^n}{\pi(1 + \delta_{n0})} \int_0^{2\pi} f(\theta) \cos n\theta d\theta, \quad b_n = \frac{L^n}{\pi} \int_0^{2\pi} f(\theta) \sin n\theta d\theta.$$

by applying the orthogonality of sine and cosine functions (Topic 1). □

Example 36.27 (UECM3003 Final Exam, Apr 2014, Q4(a)). Consider the Dirichlet problem exterior to a disk of radius a

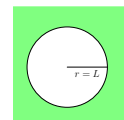
$$\begin{cases} \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} = 0, & r > a, \quad 0 \leq \phi < 2\pi, \\ u(r = a, \phi) = f(\phi), & 0 \leq \phi < 2\pi. \end{cases}$$

By separation of variables, find a general solution to the Dirichlet problem. (14 marks)

Solution: The solution is the same as Example 36.26 by replacing radius L by a . [eg: cvl:lapl:outdisk:bcl]

Example 36.28 (Outer “Disk” Type 2 BC). The Laplace equation outside a disk with *Type 2 BC* is given below:

$$\begin{cases} \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} = 0, & r > L, \quad 0 \leq \phi < 2\pi, \\ \frac{\partial u}{\partial r}(L, \phi) = f(\phi), & 0 \leq \phi < 2\pi. \end{cases}$$



Solution. The general solution is the same as Example 36.26: [eg: cvl:lapl:outdisk:bcl]

$$u(r, \phi) = \sum_{n=0}^{\infty} \{a_n r^{-n} \cos n\phi + b_n r^{-n} \sin n\phi\}$$

and the coefficients could be determined from the Type 2 BC:

$$\left. \frac{\partial u}{\partial r} \right|_{r=L} = \sum_{n=1}^{\infty} \{a_n(-n)L^{-n-1} \cos n\phi + b_n(-n)L^{-n-1} \sin n\phi\} = f(\phi).$$

Hence a_0 is an **arbitrary constant** and for $n \geq 1$,

$$a_n = \frac{-L^{n+1}}{n\pi} \int_0^{2\pi} f(\phi) \cos n\phi d\phi, \quad b_n = \frac{-L^{n+1}}{n\pi} \int_0^{2\pi} f(\phi) \sin n\phi d\phi.$$

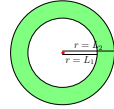
□

The boundary of a “ring” is formed from an inner circle and an outer circle, therefore, we were to impose “boundary conditions”, there will be $3^2 = 9$ possibilities. The solution is similar to the “disk” discussed above except that there is no “regularity condition” for us to make either C_n or D_n to become zero. We will illustrate this with the analysis of type 1 and 2 BCs below.

cyl:lapl:ring1

Example 36.29. The Laplace equation on a ring with *Type 1 BCs* is given below:

$$\begin{cases} \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} = 0, & A < r < B, \quad 0 \leq \phi < 2\pi, \\ u(A, \phi) = f_1(\phi), \quad u(B, \phi) = f_2(\phi), & 0 \leq \phi < 2\pi. \end{cases}$$



Solution. Suppose that $u(r, \phi) = R(r)\Phi(\phi)$, the PDE simplifies to

$$\frac{R''}{R} + \frac{1}{r} \frac{R'}{R} = -\frac{\Phi''}{\Phi} = \mu.$$

Together with the periodic BC, we obtain a periodic SL-problem of Φ and an ODE of R :

$$\begin{cases} \Phi'' + \mu\Phi = 0, & \Phi(0) = \Phi(2\pi), \quad \Phi'(0) = \Phi'(2\pi), \\ r^2 R'' + rR' - \mu R = 0. \end{cases}$$

The eigenvalues and eigenfunctions of the periodic SL problem is

$$\mu = m^2, \quad \Phi = A \cos m\phi + B \sin m\phi, \quad m = 0, 1, 2, \dots$$

When $\mu = 0$, the solution of the ODE of R is

$$R(r) = C_0 + D_0 \ln r$$

When $\mu = m^2, m = 1, 2, \dots$, the solution of the ODE of R is

$$R(r) = C_m r^m + D_m r^{-m}.$$

By the principle of superposition, the general solution is

$$u(r, \phi) = a_0 + b_0 \ln r + \sum_{m=1}^{\infty} \{(a_m r^m + b_m r^{-m}) \cos m\phi + (c_m r^m + d_m r^{-m}) \sin m\phi\}.$$

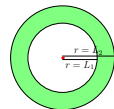
From the Type 1 BCs, we have the following simultaneous equations which allows us to determine coefficients $a_m, \dots, d_m$:

$$\begin{aligned} a_0 + b_0 \ln A &= \frac{1}{2\pi} \int_0^{2\pi} f_1(\phi) d\phi, & a_0 + b_0 \ln B &= \frac{1}{2\pi} \int_0^{2\pi} f_2(\phi) d\phi, \\ a_m A^m + b_m A^{-m} &= \frac{1}{\pi} \int_0^{2\pi} f_1(\phi) \cos m\phi d\phi, & a_m B^m + b_m B^{-m} &= \frac{1}{\pi} \int_0^{2\pi} f_2(\phi) \cos m\phi d\phi, \\ c_m A^m + d_m A^{-m} &= \frac{1}{\pi} \int_0^{2\pi} f_1(\phi) \sin m\phi d\phi, & c_m B^m + d_m B^{-m} &= \frac{1}{\pi} \int_0^{2\pi} f_2(\phi) \sin m\phi d\phi. \end{aligned}$$

□

Example 36.30. The Laplace equation on a ring with *Type 2 BCs* is given below:

$$\begin{cases} \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} = 0, & A < r < B, \quad 0 \leq \phi < 2\pi, \\ u_r(A, \phi) = f_1(\phi), \quad u_r(B, \phi) = f_2(\phi), & 0 \leq \phi < 2\pi. \end{cases}$$



Solution: The general solution can be found in exactly the same method as in Example 36.29: [eg:cyl:lapl;ring1]

$$u(r, \phi) = a_0 + b_0 \ln r + \sum_{m=1}^{\infty} \{ (a_m r^m + b_m r^{-m}) \cos m\phi + (c_m r^m + d_m r^{-m}) \sin m\phi \}.$$

From the Type 2 BCs

$$u_r(A, \phi) = \frac{b_0}{A} + \sum_{m=1}^{\infty} \{ (ma_m A^{m-1} - \frac{mb_m}{A^{m+1}}) \cos m\phi + (mc_m A^{m-1} - \frac{md_m}{A^{m+1}}) \sin m\phi \} = f_1(\phi)$$

$$u_r(B, \phi) = \frac{b_0}{B} + \sum_{m=1}^{\infty} \{ (ma_m B^{m-1} - \frac{mb_m}{B^{m+1}}) \cos m\phi + (mc_m B^{m-1} - \frac{md_m}{B^{m+1}}) \sin m\phi \} = f_2(\phi).$$

we have the following simultaneous equations which determine the values of coefficients $a_m, \dots, d_m$:

$$\frac{b_0}{A} = \frac{1}{2\pi} \int_0^{2\pi} f_1(\phi) d\phi, \quad \frac{b_0}{B} = \frac{1}{2\pi} \int_0^{2\pi} f_2(\phi) d\phi,$$

$$ma_m A^{m-1} - \frac{mb_m}{A^{m+1}} = \frac{1}{\pi} \int_0^{2\pi} f_1(\phi) \cos m\phi d\phi, \quad ma_m B^{m-1} - \frac{mb_m}{B^{m+1}} = \frac{1}{\pi} \int_0^{2\pi} f_2(\phi) \cos m\phi d\phi,$$

$$mc_m A^{m-1} - \frac{md_m}{A^{m+1}} = \frac{1}{\pi} \int_0^{2\pi} f_1(\phi) \sin m\phi d\phi, \quad mc_m B^{m-1} - \frac{md_m}{B^{m+1}} = \frac{1}{\pi} \int_0^{2\pi} f_2(\phi) \sin m\phi d\phi.$$

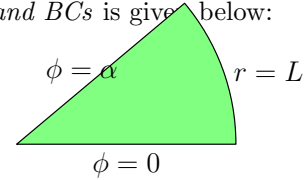
Note that a_0 is arbitrary and f_1 and f_2 need to be compatible in order for a solution to exist.

36.5.2 “Wedge” and “Sector”

A “wedge” has three pieces in the boundary while a “sector” (or a “fan”) has four pieces. Since the “wedge” contains zero, we automatically impose a “regularity condition”. The permutation of type 1, 2 and 3 boundary conditions are a lot more compare to the previous subsections. There are of little engineering interests and we will just illustrate the solution process with an example. However, we have to be a bit careful in with this type of domain because there is no more periodic boundary condition!

Example 36.31. The Laplace equation on a wedge with a mixed of Type 1 and BCs is given below:

$$\left\{ \begin{array}{l} \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} = 0, \quad 0 < r < L, \quad 0 \leq \phi < \alpha, \\ u(L, \phi) = f(\phi), \quad 0 \leq \phi \leq \alpha, \\ \frac{\partial u}{\partial \phi}(r, 0) = 0, \quad 0 \leq r \leq L \\ \frac{\partial u}{\partial \phi}(r, \alpha) = 0, \quad 0 \leq r \leq L. \end{array} \right.$$



Solution: Separation of variables and the given homogeneous boundary conditions and the implicit regularity condition lead to

$$\left\{ \begin{array}{l} \Phi'' + \mu\Phi = 0, \quad \Phi'(0) = \Phi'(\alpha) = 0, \\ r^2 R'' + rR' - \mu R = 0, \quad |R(0)| < \infty \end{array} \right.$$

Let us analyse the regular SL problem of Φ :

- When $\mu < 0$, the general solution is $\Phi = Ae^{\sqrt{-\mu}\phi} + Be^{-\sqrt{-\mu}\phi}$. The boundary conditions will imply $A = B = 0$, so we can only have trivial solution.
- When $\mu = 0$, the general solution is $\Phi = A + B\phi$. The boundary condition will imply $B = 0$. So the eigenfunction is $\Phi = A_0$.
- When $\mu > 0$, the general solution is

$$\Phi = A \cos \sqrt{\mu}\phi + B \sin \sqrt{\mu}\phi.$$

From the BCs, we have

$$\left\{ \begin{array}{l} B\sqrt{\mu} = 0, \\ -A\sqrt{\mu} \sin \sqrt{\mu}\alpha + B\sqrt{\mu} \cos \sqrt{\mu}\alpha = 0 \end{array} \right. \Rightarrow \left\{ \begin{array}{l} B = 0, \\ \sin \sqrt{\mu}\alpha = 0 \end{array} \right.$$

Hence, the eigenvalues and eigenfunctions are

$$\mu = \left(\frac{n\pi}{\alpha}\right)^2, \quad \Phi = A_n \cos \frac{n\pi}{\alpha} \phi, \quad n = 1, 2, \dots$$

When $\mu = 0$, the general solution is $R_0(r) = C_0 + D_0 \ln r$, from the regularity condition $|R_0(0)| < \infty$, $D_0 = 0$.

When $\mu = \left(\frac{n\pi}{\alpha}\right)^2$, $n = 1, 2, \dots$, $R_n(r) = C_n r^{\frac{n\pi}{\alpha}} + D_n r^{-\frac{n\pi}{\alpha}}$, from the regularity condition $|R_n(0)| < \infty$, $D_n = 0$

By the principle of superposition, the general solution is

$$u(r, \phi) = a_0 + \sum_{n=1}^{\infty} a_n r^{\frac{n\pi}{\alpha}} \cos \frac{n\pi}{\alpha} \phi, \quad a_n = C_n A_n, \quad a_n = C_n B_n.$$

From the Type 1 BC, we have

$$u(L, \phi) = \sum_{n=0}^{\infty} a_n L^{\frac{n\pi}{\alpha}} \cos \frac{n\pi}{\alpha} \phi = f(\phi) \Rightarrow a_n = \frac{\int_0^{\alpha} f(\phi) \cos \frac{n\pi}{\alpha} \phi d\phi}{L^{\frac{n\pi}{\alpha}} \int_0^{\alpha} \cos^2 \frac{n\pi}{\alpha} \phi d\phi} = \frac{2 \int_0^{\alpha} f(\phi) \cos \frac{n\pi}{\alpha} \phi d\phi}{\alpha(1 + \delta_{n0}) L^{\frac{n\pi}{\alpha}}}$$

where $n = 0, 1, \dots$.

An example on a sector of a circular annulus, i.e. the domain $a < r < b$ and $0 < \phi < \pi/2$ is given in Haberman [1987, 74].

36.5.3 Inner “Tube”

Apart from the 2-D problem that we encountered above, another type of 2-variable problem will happen on a “tube” with a different symmetry. Since the centre is included, the “regularity condition” will be implicitly assumed.

Example 36.32 (UECM3003 Final Exam, Apr 2013, Q3(a)). The steady-state temperature $u = u(r, z)$ in a solid cylinder of radius $r = 1$ with bottom $z = 0$ and top $z = 2$ satisfies the boundary value problem

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{\partial^2 u}{\partial z^2} = 0, \quad 0 < r < 1, \quad 0 < z < 2, \quad (36.59) \quad \text{e:1}$$

$$u(1, z) = 0, \quad 0 < z < 2, \quad (36.60) \quad \text{e:2}$$

$$u(r, 0) = 0, \quad 0 < r < 1. \quad (36.61) \quad \text{e:3}$$

(i) Substitute $u(r, z) = R(r)Z(z)$ in (41.39) to obtain the eigenvalue problems

$$rR'' + R' - krR = 0, \quad R(1) = 0, \quad (36.62) \quad \text{e:4}$$

$$Z'' + kZ = 0, \quad Z(0) = 0, \quad (36.63) \quad \text{e:5}$$

where k is a separation constant. (7 marks)

Solution. Let k be some constant. By separation of variables, we have

$$R''Z + \frac{1}{r}R'Z + RZ'' = 0 \Rightarrow \frac{R''}{R} + \frac{1}{r}\frac{R'}{R} = -\frac{Z''}{Z} = k \Rightarrow \begin{cases} Z'' + kZ = 0 \\ R'' + \frac{1}{r}R' - kR = 0 \end{cases}$$

From the boundary conditions, we have

$$u(1, z) = R(1)Z = 0 \Rightarrow R(1) = 0, \quad u(r, 0) = RZ(0) = 0 \Rightarrow Z(0) = 0. \quad \square$$

(ii) If the eigenvalues are known to be negative (and $k = -\lambda^2$, $\lambda > 0$), show that the gives boundary value problem has a general solution of the form

$$u(r, z) = \sum_{n=1}^{\infty} c_n J_0(\lambda_n r) \sinh(\lambda_n z)$$

where $\{\lambda_n\}_1^{\infty}$ are the positive roots of $J_0(x) = 0$. (12 marks)

Solution. The SL-problem of R is a Bessel equation of order $\nu = 0$:

$$r^2 R'' + rR' - kr^2 R = 0, \quad R(1) = 0.$$

The general solution is $R(r) = AJ_0(\sqrt{-k}r) + BY_0(\sqrt{-k}r)$ but the regularity at $r = 0$ requires $B = 0$. So

$$R(1) = AJ_0(\sqrt{-k}) = 0 \Rightarrow \sqrt{-k} = \lambda_n \Rightarrow k = -\lambda_n^2, \quad R_n = A_n J_0(\lambda_n r), \quad n = 1, 2, \dots$$

where λ_n are the zeros of J_0 . For these eigenvalues, the solution of Z is

$$Z = C_n e^{-\lambda_n z} + D_n e^{\lambda_n z} \Rightarrow Z(0) = 0 \Rightarrow C_n = -D_n \Rightarrow Z_n = 2D_n \sinh(\lambda_n z)$$

By the principle of superposition, we have

$$u(r, z) = \sum_{n=1}^{\infty} R_n Z_n = \sum_{n=1}^{\infty} c_n J_0(\lambda_n r) \sinh(\lambda_n z)$$

where $c_n = 2D_n A_n$ are constants. □

Example 36.33 (UECM3003 Final Exam, Apr 2013, Q3(b)). Given that the general solution of the boundary value problem

$$\begin{aligned} \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{\partial^2 u}{\partial z^2} &= 0, \quad 0 < r < 1, \quad 0 < z < 2, \\ u_r(1, z) &= 0, \quad 0 < z < 2, \\ u(r, 0) &= 0, \quad 0 < r < 1 \end{aligned}$$

is

$$u(r, z) = c_0 z + \sum_{n=1}^{\infty} c_n J_0(\lambda_n r) \sinh(\lambda_n z)$$

where $\{\lambda_n\}_1^{\infty}$ are the positive roots of $J'_0(x) = 0$, determine the specific solution which also satisfies the third boundary condition

$$u(r, 2) = 8, \quad 0 < r < 1. \quad (6 \text{ marks})$$

Solution: Given that the general solution is

$$u(r, z) = c_0 z + \sum_{n=1}^{\infty} c_n J_0(\lambda_n r) \sinh(\lambda_n z)$$

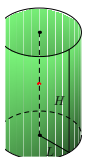
From $u(r, 2) = 8$, we have

$$2c_0 + \sum_{n=1}^{\infty} c_n J_0(\lambda_n r) \sinh(2\lambda_n) = 8 \Rightarrow \boxed{c_0 = 4, \quad c_n = 0.}$$

The specific solution is $\boxed{u(r, z) = 4z}$.

When the “temperature distribution” on boundary of a cylinder is a function of variables ϕ and z , then the Laplace equation problem we are dealing with will be “3-D”, i.e. we need three variables to characterise the problem. Although there can be many possible “permutations” of boundary conditions, we will only illustrate the application of separation of variables with the following example.

Example 36.34. [Weinberger, 1995, 149] Consider a Laplace equation with Type 1 BCs on a cylinder of radius L and height H is given below:



$$\begin{cases} \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} + \frac{\partial^2 u}{\partial z^2} = 0, & 0 < r < L, \quad 0 \leq \phi < 2\pi, \quad 0 < z < H, \\ u(r, \phi + 2\pi, z) = u(r, \phi, z), & 0 < r < L, \quad 0 \leq \phi < 2\pi, \quad 0 < z < H, \\ u(r, \phi, 0) = u(r, \phi, H) = 0, & 0 < r < L, \quad 0 \leq \phi < 2\pi, \\ u(L, \phi, z) = g(\phi, z), & 0 \leq \phi < 2\pi, \quad 0 < z < H. \end{cases} \quad (36.64)$$

e:33.1

Use the separation of variable to find the formal series solution to the above problem.

Solution. Let $u = R\Phi Z$, we obtain

$$\frac{R''}{R} + \frac{1}{r} \frac{R'}{R} + \frac{1}{r^2} \frac{\Phi''}{\Phi} = -\frac{Z''}{Z} = \mu.$$

Multiplying the left-hand-side by r^2 and transposing Φ -term gives

$$\begin{cases} Z'' + \mu Z = 0, \\ \frac{r^2 R''}{R} + \frac{r R'}{R} - \mu r^2 = -\frac{\Phi''}{\Phi} = \gamma \end{cases} \Rightarrow \begin{cases} Z'' + \mu Z = 0, \\ \Phi'' + \gamma \Phi = 0, \\ r^2 R'' + r R' - (\mu r^2 + \gamma) R = 0. \end{cases}$$

From the periodic BC and the type 1 homogeneous BC, we have

$$u(r, \phi + 2\pi, z) = u(r, \phi, z) \Rightarrow \begin{cases} u(r, 2\pi, z) = u(r, 0, z), \\ u_\phi(r, 2\pi, z) = u_\phi(r, 0, z) \end{cases} \Rightarrow \Phi(0) = \Phi(2\pi), \quad \Phi'(0) = \Phi'(2\pi),$$

$$R(r)\Phi(\phi)Z(0) = R(r)\Phi(\phi)Z(H) = 0 \Rightarrow Z(0) = Z(H) = 0.$$

Solving the regular SL problem for Z gives the eigenvalues and eigenfunctions as follows:

$$\mu = \left(\frac{n\pi}{H}\right)^2, \quad Z_n = E_n \sin \frac{n\pi z}{H}, \quad n = 1, 2, \dots$$

Next, solving the periodic SL problem of Φ gives the eigenvalues and eigenfunctions as follows:

$$\gamma = m^2, \quad \Phi_m = A_m \cos m\phi + B_m \sin m\phi, \quad m = 0, 1, 2, \dots$$

Finally, with the above eigenvalues, we obtain the **modified Bessel equation**

$$r^2 R''_{mn} + r R'_{mn} - \left[\left(\frac{n\pi}{H}\right)^2 r^2 + m^2\right] R_{mn} = 0. \quad (36.65)$$

This equation is singular at $r = 0$ and has a general solution according to Section [36.2.3](#) ^{[|sec:mbessel](#)}

$$R_{mn}(r) = \alpha_{mn} I_m\left(\frac{n\pi r}{H}\right) + \beta_{mn} K_m\left(\frac{n\pi r}{H}\right)$$

where I_m and K_m are modified Bessel functions defined in Section [36.2.3](#) ^{[|sec:mbessel](#)}

The regularity condition requires R to remain finite at $r = 0$. Therefore, $\beta_{mn} = 0$ since K_m are singular at $r = 0$ and so $R_{mn}(r) = \alpha_{mn} I_m\left(\frac{n\pi r}{H}\right)$.

The general solution of PDE problem [\(36.64\)](#) ^{[|e:33.1](#)} is

$$u(r, \phi) = \sum_{m=0}^{\infty} \sum_{n=1}^{\infty} I_m\left(\frac{n\pi r}{H}\right) \sin \frac{n\pi z}{H} (a_{mn} \cos m\phi + b_{mn} \sin m\phi) \quad (36.66)$$

[e:33.4](#)

where $a_{mn} = \alpha_{mn} E_n A_m$ and $b_{mn} = \alpha_{mn} E_n B_m$ are constants.

From the nonhomogeneous type 1 boundary condition and using the orthogonality property of the eigenfunctions, we have

$$I_m\left(\frac{n\pi L}{H}\right) a_{mn} = \frac{\int_0^H \int_0^{2\pi} g(\phi, z) \sin \frac{n\pi z}{H} \cos m\phi d\phi dz}{\int_0^H \int_0^{2\pi} \sin^2 \frac{n\pi z}{H} \cos^2 m\phi d\phi dz} = \frac{\int_0^H \int_0^{2\pi} g(\phi, z) \sin \frac{n\pi z}{H} \cos m\phi d\phi dz}{\frac{H}{2} (1 + \delta_{m0}) \pi}$$

$$I_m\left(\frac{n\pi L}{H}\right) b_{mn} = \frac{\int_0^H \int_0^{2\pi} g(\phi, z) \sin \frac{n\pi z}{H} \sin m\phi d\phi dz}{\int_0^H \int_0^{2\pi} \sin^2 \frac{n\pi z}{H} \sin^2 m\phi d\phi dz} = \frac{\int_0^H \int_0^{2\pi} g(\phi, z) \sin \frac{n\pi z}{H} \sin m\phi d\phi dz}{\frac{H}{2} \pi}.$$

This implies that

$$a_{mn} = \frac{2}{H(1 + \delta_{m0})\pi I_m(\frac{n\pi L}{H})} \int_0^H \int_0^{2\pi} g(\phi, z) \sin \frac{n\pi z}{H} \cos m\phi d\phi dz$$

$$b_{mn} = \frac{2}{H\pi I_m(\frac{n\pi L}{H})} \int_0^H \int_0^{2\pi} g(\phi, z) \sin \frac{n\pi z}{H} \sin m\phi d\phi dz.$$

The formal series solution converges uniformly for $r < L$ and this is shown in Weinberger [1995, p149]. $\square$

The more general case when all the boundary conditions are non-homogeneous is discussed in Haberman [1987, 236]. The trick is to decompose u into three problems with all except one boundary conditions are homogeneous.

36.6 Poisson Equation

The nonhomogeneous problem corresponding to Laplace equation is the Poisson equation. As mentioned in Topic 2, this is a very difficult problem involving Green's function.

Consider the Poisson equation with a nonhomogeneous boundary condition such as type 1 BC problem on an inner disk:

$$\begin{cases} \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} = -F(r, \phi), & r < L, \\ u(L, \phi) = f(\phi). \end{cases} \quad (36.67)$$

To solve it, we will need to let $u = v + w$ and split the problem into two sub-problems:

$$\begin{cases} \frac{\partial^2 v}{\partial r^2} + \frac{1}{r} \frac{\partial v}{\partial r} + \frac{1}{r^2} \frac{\partial^2 v}{\partial \phi^2} = -F(r, \phi), & r < L, \\ v(L, \phi) = 0. \end{cases} \quad \begin{cases} \frac{\partial^2 w}{\partial r^2} + \frac{1}{r} \frac{\partial w}{\partial r} + \frac{1}{r^2} \frac{\partial^2 w}{\partial \phi^2} = 0, & r < L, \\ w(L, \phi) = f(\phi). \end{cases}$$

The solution of w -problem is given in Example [36.23](#). Combining with the v -solution and we will obtain the solution u .

Example 36.35. Show that the solution of the v -problem above are characterised by ODE-BVPs by using eigenfunction expansion.

Solution: From Example [36.23](#), we know that the eigenfunctions of this problem is

$$v(r, \phi) = a_0(r) + \sum_{n=1}^{\infty} \{a_n(r) \cos n\phi + b_n(r) \sin n\phi\}$$

due to the periodic boundary condition. Substituting into the Poisson, we have

$$\begin{aligned} \left[a_0''(r) + \frac{a_0'(r)}{r} \right] + \sum_{n=1}^{\infty} \left[a_n''(r) + \frac{a_n'(r)}{r} - \frac{n^2}{r^2} a_n(r) \right] \cos n\phi \\ + \sum_{n=1}^{\infty} \left[b_n''(r) + \frac{b_n'(r)}{r} - \frac{n^2}{r^2} b_n(r) \right] \sin n\phi = -F(r, \phi) \end{aligned}$$

Since $\{\cos n\phi, \sin n\phi\}_{n=0}^{\infty}$ form an orthogonal basis over the interval $(-\pi, \pi)$, we derive the following ODEs:

$$\begin{aligned} a_0''(r) + \frac{a_0'(r)}{r} &= -\frac{1}{2\pi} \int_{-\pi}^{\pi} F(r, \phi) d\phi \\ a_n''(r) + \frac{a_n'(r)}{r} - \frac{n^2}{r^2} a_n(r) &= -\frac{1}{\pi} \int_{-\pi}^{\pi} F(r, \phi) \cos n\phi d\phi \\ b_n''(r) + \frac{b_n'(r)}{r} - \frac{n^2}{r^2} b_n(r) &= -\frac{1}{\pi} \int_{-\pi}^{\pi} F(r, \phi) \sin n\phi d\phi \end{aligned}$$

From the homogeneous type 1 boundary condition and the regularity condition, we have

$$\begin{aligned} v(L, \phi) = 0 &\Rightarrow a_0(L) = a_n(L) = b_n(L) = 0, \\ |v(0, \phi)| < \infty &\Rightarrow |a_0(0)|, |a_n(0)|, |b_n(0)| < \infty. \end{aligned}$$

The solution of the ODE-BVPs above required the Green's function method and is given in Weinberger [1995, 129].

36.7 1-D Wave Equation

The wave equation measures the displacement of a thin circular membrane (or drumhead) and can be interpreted to mean that the acceleration (or force) u_{tt} of a point on the drumhead is proportional to Δu . That is, the drumhead at a point is *accelerating upward* (force is up) if the drumhead at that point is *less* (in height) than the *average* of its neighbours. Mathematically, the vibration of a thin circular membrane with uniform density and satisfying the *radial symmetric initial conditions* (36.1) and (36.2) is governed by the “1-D” wave equation (36.18) and boundary condition

$$u(L, t) = 0, \quad t \geq 0. \quad (36.68)$$

Example 36.36. Use the separation of variables to find the formal series solution to the wave equation problem (36.18) with type 1 BC (36.68) and initial conditions (36.1) and (36.2).

Solution. Let $u(r, t) = R(r)T(t) = RT$. Then the PDE can be separated:

$$\frac{T''}{c^2 T} = \frac{R''}{R} + \frac{1}{r} \frac{R'}{R} = -\lambda \Rightarrow \begin{cases} T'' + c^2 \lambda T = 0, & t \geq 0, \\ rR'' + R' + \lambda rR = 0, & 0 < r < L. \end{cases}$$

By adding the boundary conditions, we obtain a Sturm-Liouville problem

$$r^2 R'' + rR' + \lambda r^2 R = 0, \quad R(L) = 0, \quad |R(0)| < \infty.$$

The values of λ can be classified into

- Case $\lambda < 0$: In this case, we obtain a modified Bessel equation of order $\nu = 0$ with a general solution below

$$r^2 R'' + rR' - (-\lambda)r^2 R = 0 \Rightarrow R(r) = aI_0(\sqrt{-\lambda}r) + bK_0(\sqrt{-\lambda}r) \quad 0 < r < L.$$

The regularity condition $|R(0)| < \infty$ implies $b = 0$ and $R(L) = 0$ implies $a = 0$.

- Case $\lambda = 0$: In this case, we obtain an Euler equation with a general solution below

$$r^2 R'' + rR' = 0 \Rightarrow R(r) = a + b \ln r, \quad 0 < r < L.$$

The regularity condition implies $b = 0$. $R(L) = 0$ implies $a = 0$.

- Case $\lambda > 0$: In this case, we obtain a Bessel equation of order $\nu = 0$ with a general solution below

$$r^2 R'' + rR' + \lambda r^2 R = 0 \Rightarrow R(r) = aJ_0(\sqrt{\lambda}r) + bY_0(\sqrt{\lambda}r) \quad 0 < r < L.$$

The regularity condition requires implies $b = 0$.

$$R(L) = 0 \Rightarrow a = 0 \quad \text{or} \quad J_0(\sqrt{\lambda}L) = 0.$$

For non-trivial solution, we want $a \neq 0$, from here, we obtain the eigenvalues and the eigenfunctions:

$$\lambda = \left(\frac{\alpha_n}{L}\right)^2, \quad R = A_n J_0\left(\frac{\alpha_n}{L}r\right), \quad n = 1, 2, \dots$$

where $\alpha_n, n = 1, 2, \dots$ are the roots of J_0 .

For $\lambda = \left(\frac{\alpha_n}{L}\right)^2$, we can solve for T : $T_n(t) = A_n \cos \frac{c\alpha_n}{L}t + B_n \sin \frac{c\alpha_n}{L}t$.

Thus, the general solution is

$$u(r, t) = \sum_{n=1}^{\infty} \left(a_n \cos \frac{c\alpha_n}{L}t + b_n \sin \frac{c\alpha_n}{L}t \right) J_0\left(\frac{\alpha_n}{L}r\right)$$

and the coefficients can be obtained by applying the orthogonality properties of Bessel function J_0 :

$$a_n = \frac{2}{L^2 J_1^2(\alpha_n)} \int_0^L f(r) J_0\left(\frac{\alpha_n}{L} r\right) r dr,$$

$$\frac{c\alpha_n}{L} b_n = \frac{2}{L^2 J_1^2(\alpha_n)} \int_0^L g(r) J_0\left(\frac{\alpha_n}{L} r\right) r dr \Rightarrow b_n = \frac{2}{c\alpha_n L J_1^2(\alpha_n)} \int_0^L g(r) J_0\left(\frac{\alpha_n}{L} r\right) r dr.$$

□

Example 36.37 (UCCM3003 Final Exam, Sep 2009, Q4). (a) Let $u(r, t)$ be the amplitude of the vibration for a radial symmetric drumhead of radius 1. Given that $u(r, t)$ satisfies the wave equation

$$\frac{\partial^2 u}{\partial t^2} = 4 \left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} \right), \quad 0 < r < 1, \quad t > 0,$$

and the boundary conditions $u(1, t) = 0, t > 0$, find the general solution of the above boundary value problem. (11 marks)

Solution: The wave equation is homogeneous and the boundary condition is type 1 homogeneous. Therefore, we can apply the separation of variables $u(r, t) = R(r)T(t) = RT$. The PDE can be separated:

$$\frac{T''}{4T} = \frac{R''}{R} + \frac{1}{r} \frac{R'}{R} = -\lambda \Rightarrow \begin{cases} T'' + 4\lambda T = 0, & t \geq 0, \\ rR'' + R' + \lambda rR = 0, & 0 < r < 1. \end{cases}$$

From the boundary conditions, we obtain a singular SL problem

$$r^2 R'' + rR' + \lambda r^2 R = 0, \quad R(1) = 0, \quad |R(0)| < \infty.$$

The values of λ can be classified into (just like what we did in Example [36.12](#)) [eq:final:200709q3](#)

- Case $\lambda < 0$: In this case, we obtain a modified Bessel equation of order $\nu = 0$ with a general solution below

$$r^2 R'' + rR' - (-\lambda)r^2 R = 0 \Rightarrow R(r) = aI_0(\sqrt{-\lambda}r) + bK_0(\sqrt{-\lambda}r) \quad 0 < r < 1.$$

The regularity condition $|R(0)| < \infty$ implies $b = 0$ and $R(1) = 0$ implies $a = 0$.

- Case $\lambda = 0$: In this case, we obtain an Euler equation with a general solution below

$$r^2 R'' + rR' = 0 \Rightarrow R(r) = a + b \ln r, \quad 0 < r < 1.$$

The regularity condition implies $b = 0$. $R(1) = 0$ implies $a = 0$.

- Case $\lambda > 0$: In this case, we obtain a Bessel equation of order $\nu = 0$ with a general solution below

$$r^2 R'' + rR' + \lambda r^2 R = 0 \Rightarrow R(r) = aJ_0(\sqrt{\lambda}r) + bY_0(\sqrt{\lambda}r) \quad 0 < r < 1.$$

The regularity condition requires implies $b = 0$.

$$R(1) = 0 \Rightarrow a = 0 \quad \text{or} \quad J_0(\sqrt{\lambda}) = 0.$$

For non-trivial solution, we want $a \neq 0$, from here, we obtain the eigenvalues and the eigenfunctions:

$$\lambda = \alpha_n^2, \quad R = A_n J_0(\alpha_n r), \quad n = 1, 2, \dots$$

where $\alpha_n, n = 1, 2, \dots$ are the roots of J_0 .

For $\lambda = \alpha_n^2$, we can solve for T : $T(t) = A_n \cos c\alpha_n t + B_n \sin c\alpha_n t$.

Thus, the general solution is

$$u(r, t) = \sum_{n=1}^{\infty} \left[a_n \cos c\alpha_n t + b_n \sin c\alpha_n t \right] J_0(\alpha_n r).$$

- (b) Given that at $t = 0, u(r, 0) = r^2, \frac{\partial u}{\partial t}(r, 0) = 0$, find the specific solution of the wave equation in Part (a).

Solution. $u(r, 0) = \sum_{n=1}^{\infty} a_n J_0(\alpha_n r) = r^2 \Rightarrow a_n = \frac{2}{[J_1(\alpha_n r)]^2} \int_0^1 r^2 J_0(\alpha_n r) r dr$

$$\frac{\partial u}{\partial t}(r, 0) = \sum_{n=1}^{\infty} c \alpha_n b_n J_0(\alpha_n r) = 0 \Rightarrow b_n = 0$$

If we make use of Example 36.11, we can simplify a_n and simplify our solution:

$$a_n = \frac{2(\alpha_n^2 - 4)}{\alpha_n^3 J_1(\alpha_n)} \Rightarrow u(r, t) = \sum_{n=1}^{\infty} \frac{2(\alpha_n^2 - 4)}{\alpha_n^3 J_1(\alpha_n)} J_0(\alpha_n r) \cos c \alpha_n t.$$

□

Example 36.38 (UCCM3003 Final Exam, Apr 2012, Q3). The displacement $u(r, t)$ of a circular drum-head of unit radius clamped along its circumference if its initial displacement is $u(r, 0) = 0$ and its initial velocity is $u_t(r, 0) = 2$ can be modelled by the boundary value problem:

$$\begin{cases} u_{tt} = u_{rr} + \frac{1}{r} u_r, & 0 < r < 1, t > 0, \\ u(1, t) = 0, & t > 0, \\ u(r, 0) = 0, u_t(r, 0) = 2, & 0 < r < 1. \end{cases}$$

- (a) Assume that $u(r, t) = R(r)T(t)$ and that the separation constant $k = -\lambda^2 < 0$ with $\lambda > 0$. Show that the separated differential equations are $rR'' + R' + \lambda^2 rR = 0$, $T'' + \lambda^2 T = 0$.

(5 marks)

Solution. Substituting $u = RT$ into the PDE, we have

$$RT'' = R''T + \frac{1}{r} R'T \Rightarrow \frac{T''}{T} = \frac{R'' + \frac{1}{r} R'}{R} = k.$$

This leads to an ODE $T'' - kT = 0$ and a SL problem (Bessel Equation of order 0)

$$R'' + \frac{1}{r} R - kR = 0 \Rightarrow r^2 R'' + rR - kr^2 R = 0, \quad R(1) = 0, R(0) < \infty.$$

Since $k = -\lambda^2$, the ODE and the SL problem become $\begin{cases} T'' + \lambda^2 T = 0 \\ r^2 R'' + rR + \lambda^2 r^2 R = 0. \end{cases}$

□

- (b) Write down the general solution for each of the separated equations.

(3 marks)

Solution. Solving the SL problem of R and the ODE of T , we obtain

$$R(r) = A_n J_0(\lambda_n r), \quad T(t) = C_n \cos \lambda_n t + D_n \sin \lambda_n t, \quad J_0(\lambda_n) = 0, \quad n = 1, 2, \dots$$

The general solution is $u(r, t) = \sum_{n=1}^{\infty} [a_n \cos \lambda_n t + d_n \sin \lambda_n t] J_0(\lambda_n r)$.

□

- (c) Use the result of part (b) to derive the solution $u(r, t) = 4 \sum_{n=1}^{\infty} \frac{J_0(\lambda_n r) \sin(\lambda_n t)}{\lambda_n^2 J_1(\lambda_n)}$ of the given boundary value problem where λ_n are the positive roots of $J_0(x) = 0$.

(17 marks)

Solution. $\sum_{n=1}^{\infty} a_n J_0(\lambda_n r) = 0 \Rightarrow a_n = 0$

$$\begin{aligned} \sum_{n=1}^{\infty} \lambda_n b_n J_0(\lambda_n r) &= 2 \Rightarrow b_n = \frac{2}{\lambda_n [J_1(\lambda_n)]^2} \int_0^1 2 J_0(\lambda_n r) r dr = \frac{4}{\lambda_n [J_1(\lambda_n)]^2} \cdot \frac{1}{\lambda_n^2} \int_0^{\lambda_n} J_0(x) x dx \\ &= \frac{4}{\lambda_n^3 [J_1(\lambda_n)]^2} [x J_1(x)]_0^{\lambda_n} = \frac{4}{\lambda_n^2 J_1(\lambda_n)} \text{ where } x = \lambda_n r. \end{aligned}$$

□

36.8 2-D and 3-D Wave Equations

The vibration of a thin circular membrane with uniform density with **asymmetric** initial conditions (36.3) and (36.4) is characterised by a “2-D” wave equation (36.19). The solution of the 2-D wave equation below is similar to Example 36.21 with extra steps.

Example 36.39. [Weinberger, 1995, 182] For a small transverse motion of a stretch circular membrane of radius L fixed on its boundary, i.e.

$$u(L, \phi, t) = 0, \quad t > 0. \quad (36.69) \quad \text{e:bc:type1}$$

Use separation of variables to find the solution of this problem with the initial conditions (36.3) and (36.4). e:cyl:init2a

Solution. Let $u = TR\Phi$, we have

$$\frac{T''}{c^2 T} = \frac{R''}{R} + \frac{1}{r} \frac{R'}{R} + \frac{1}{r^2} \frac{\Phi''}{\Phi} = -\lambda \Rightarrow \begin{cases} T'' = -\lambda c^2 T, \\ r^2 \frac{R''}{R} + r \frac{R'}{R} + \lambda r^2 = -\frac{\Phi''}{\Phi} = \mu. \end{cases}$$

From the boundary conditions, we have

$$\begin{cases} T'' + \lambda c^2 T = 0, & t > 0, \\ \Phi'' + \mu \Phi = 0, & \Phi(0) = \Phi(2\pi), \quad \Phi'(0) = \Phi'(2\pi), \\ r^2 R'' + r R' + (\lambda r^2 - \mu) R = 0, & R(L) = 0, \quad |R(r)|, |R'(r)| < \infty \text{ for } 0 \leq r \leq L. \end{cases}$$

The eigenvalues and eigenfunctions for the periodic SL problem of Φ are

$$\mu = m^2, \quad \Phi_m(\phi) = A_m \cos m\phi + B_m \sin m\phi, \quad m = 0, 1, 2, \dots$$

The eigenvalues and eigenfunctions for Bessel equation of R are

$$\lambda = \left(\frac{\lambda_n^{(m)}}{L}\right)^2, \quad R_n(r) = C_n J_m\left(\frac{\lambda_n^{(m)}}{L} r\right), \quad n = 1, 2, \dots$$

where $\lambda_n^{(m)}$ are the roots of the Bessel function of the first type J_m .

Now, solving for the ODE of T , we have

$$T_{mn}(t) = D_{mn} \cos \frac{\lambda_n^{(m)} ct}{L} + E_{mn} \sin \frac{\lambda_n^{(m)} ct}{L}.$$

By applying the principle of superposition, we obtain the general solution

$$\begin{aligned} u &= \sum_{m=0}^{\infty} \sum_{n=1}^{\infty} \left[D_{mn} \cos \frac{\lambda_n^{(m)} ct}{L} + E_{mn} \sin \frac{\lambda_n^{(m)} ct}{L} \right] C_n J_m\left(\frac{\lambda_n^{(m)}}{L} r\right) [A_m \cos m\phi + B_m \sin m\phi] \\ &= \sum_{m=0}^{\infty} \sum_{n=1}^{\infty} J_m\left(\frac{\lambda_n^{(m)}}{L} r\right) \left[a_{mn} \cos \frac{\lambda_n^{(m)} ct}{L} \cos m\phi + b_{mn} \cos \frac{\lambda_n^{(m)} ct}{L} \sin m\phi \right. \\ &\quad \left. + c_{mn} \sin \frac{\lambda_n^{(m)} ct}{L} \cos m\phi + d_{mn} \sin \frac{\lambda_n^{(m)} ct}{L} \sin m\phi \right] \end{aligned}$$

where $a_{mn}, \dots, d_{mn}$ are constants related to D_{mn} etc.

From the first initial condition (36.3), we can determine the coefficients by applying Theorem 36.7 e:cyl:init2a thm:bessel:orth1

$$\begin{aligned} a_{mn} &= \frac{\int_0^L \int_0^{2\pi} f(r, \phi) J_m\left(\frac{\lambda_n^{(m)}}{L} r\right) \cos m\phi r dr d\phi}{\int_0^L \left[J_m\left(\frac{\lambda_n^{(m)}}{L} r\right) \right]^2 r dr \int_0^{2\pi} \cos^2 m\phi d\phi} = \frac{\int_0^L \int_0^{2\pi} f(r, \phi) J_m\left(\frac{\lambda_n^{(m)}}{L} r\right) \cos m\phi r dr d\phi}{\pi(1 + \delta_{m0}) \times \frac{L^2}{2} [J_{m+1}(\lambda_n^{(m)})]^2}, \\ b_{mn} &= \frac{\int_0^L \int_0^{2\pi} f(r, \phi) J_m\left(\frac{\lambda_n^{(m)}}{L} r\right) \sin m\phi r dr d\phi}{\int_0^L \left[J_m\left(\frac{\lambda_n^{(m)}}{L} r\right) \right]^2 r dr \int_0^{2\pi} \sin^2 m\phi d\phi} = \frac{\int_0^L \int_0^{2\pi} f(r, \phi) J_m\left(\frac{\lambda_n^{(m)}}{L} r\right) \sin m\phi r dr d\phi}{\pi \times \frac{L^2}{2} [J_{m+1}(\lambda_n^{(m)})]^2}. \end{aligned}$$

From the second initial condition (36.4), we can determine the coefficients e:cyl:init2b

$$\begin{aligned} \frac{\lambda_n^{(m)} c}{L} c_{mn} &= \frac{\int_0^L \int_0^{2\pi} g(r, \phi) J_m\left(\frac{\lambda_n^{(m)}}{L} r\right) \cos m\phi r dr d\phi}{\int_0^L \left[J_m\left(\frac{\lambda_n^{(m)}}{L} r\right) \right]^2 r dr \int_0^{2\pi} \cos^2 m\phi d\phi} \\ \Rightarrow c_{mn} &= \frac{L}{\lambda_n^{(m)} c} \times \frac{\int_0^L \int_0^{2\pi} g(r, \phi) J_m\left(\frac{\lambda_n^{(m)}}{L} r\right) \cos m\phi r dr d\phi}{\pi(1 + \delta_{m0}) \times \frac{L^2}{2} [J_{m+1}(\lambda_n^{(m)})]^2}, \end{aligned}$$

$$\Rightarrow d_{mn} = \frac{L}{\lambda_n^{(m)} c} \times \frac{\int_0^L \int_0^{2\pi} g(r, \phi) J_m\left(\frac{\lambda_n^{(m)} r}{L}\right) \sin m\phi r dr d\phi}{\pi \times \frac{L^2}{2} [J_{m+1}(\lambda_n^{(m)})]^2}.$$

□

Remark 36.40. When f is twice continuously differentiable and $f(L, \phi) = 0$, it can be shown that the formal series above converges [Weinberger, 1995, 183].

The harmonic motions $J_m\left(\frac{\lambda_n^{(m)} r}{L}\right) \cos m\phi \cos \frac{\lambda_n^{(m)} ct}{L}$ and $J_m\left(\frac{\lambda_n^{(m)} r}{L}\right) \sin m\phi \cos \frac{\lambda_n^{(m)} ct}{L}$ are called the *normal modes* of the system. Higher vibrations for the drumhead are not pure overtones of the basic frequencies [Farlow, 1993].

Example 36.41 (UCCM3003 Final Exam, Sep 2008, Q2). Consider the following wave equation in an infinite cylinder

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2}, & 0 < r < 3, 0 < \phi < 2\pi, t > 0, \\ u(3, \phi, t) = 0 & 0 \leq \phi \leq 2\pi, t \geq 0, \\ u(r, \phi, 0) = \psi(r, \phi), & 0 \leq r \leq 3, 0 \leq \phi \leq 2\pi, \\ \frac{\partial u}{\partial t}(r, \phi, 0) = 0, & 0 \leq r \leq 3, 0 \leq \phi \leq 2\pi, \end{cases}$$

where the function $u = u(r, \phi, t)$ is 2π -periodic with respect to ϕ .

(a) By using separation of variables, show that the general solution is

$$u(r, \phi, t) = \sum_{m=0}^{\infty} \sum_{n=1}^{\infty} J_m\left(\frac{\gamma_n^{(m)} r}{3}\right) (a_{mn} \cos \frac{\gamma_n^{(m)} t}{3} \cos m\phi + b_{mn} \cos \frac{\gamma_n^{(m)} t}{3} \sin m\phi + c_{mn} \sin \frac{\gamma_n^{(m)} t}{3} \cos m\phi + d_{mn} \sin \frac{\gamma_n^{(m)} t}{3} \sin m\phi).$$

where J_m is a type 1 Bessel function of order m ; $\gamma_n^{(m)}$ are the zeroes of J_m ; a_{mn} , b_{mn} , c_{mn} and d_{mn} are constant coefficients. (16 marks)

Solution. Let $u = TR\Phi$. Then the PDE can be simplified to

$$\frac{T''}{T} = \frac{R''}{R} + \frac{1}{r} \frac{R'}{R} + \frac{1}{r^2} \frac{\Phi''}{\Phi} = -\lambda. \quad [1 \text{ mark}]$$

Separating the variables we have three ODEs

$$\begin{cases} T'' + \lambda T = 0, \\ \Phi'' + \mu \Phi = 0, \\ r^2 R'' + r R' + (\lambda r^2 - \mu) R = 0. \end{cases} \quad [3 \text{ marks}]$$

where λ and μ are constants.

From the periodicity of u w.r.t ϕ and the BC, we have

$$\Phi(2\pi) = \Phi(0), \Phi'(2\pi) = \Phi'(0), R(3) = 0. \quad [2 \text{ marks}]$$

Solving the SL problem for Φ , we have (see Formula Sheet)

$$\Phi_m(x) = A_m \cos m\phi + B_m \sin m\phi, \mu_m = m^2, m = 0, 1, 2, \dots \quad [1+1=2 \text{ marks}]$$

For $\mu_m = m^2$, the SL problem for R is

$$r^2 R'' + r R' + (\lambda r^2 - m^2) R = 0 \Rightarrow R(r) = C J_m(\sqrt{\lambda} r) + D Y_m(\sqrt{\lambda} r) \quad [1 \text{ mark}]$$

Since R needs to be well-defined for $0 < r < 3$ while Y_m is singular at $r = 0$, $D = 0$. [2 marks]

From the BC, we have

$$R(3) = C J_m(3\sqrt{\lambda}) = 0 \Rightarrow 3\sqrt{\lambda} = \gamma_n^{(m)}, \Rightarrow \lambda = \left(\frac{\gamma_n^{(m)}}{3}\right)^2 \quad n = 1, 2, \dots \quad [1 \text{ mark}]$$

where $\gamma_n^{(m)}$, $n = 1, 2, \dots$ are the zeroes of the Bessel function $J_m(x)$ [1 mark]

Finally solving for the ODE $T + (\frac{\gamma_n^{(m)}}{3})^2 T = 0$, we have

$$T_{mn}(t) = E_{mn} \cos \frac{\gamma_n^{(m)} t}{3} + F_{mn} \sin \frac{\gamma_n^{(m)} t}{3}. \quad [1 \text{ mark}]$$

The general solution is hence

$$u = \sum_{m=0}^{\infty} \sum_{n=1}^{\infty} J_m\left(\frac{\gamma_n^{(m)} r}{3}\right) (a_{mn} \cos \frac{\gamma_n^{(m)} t}{3} \cos m\phi + b_{mn} \cos \frac{\gamma_n^{(m)} t}{3} \sin m\phi \\ + c_{mn} \sin \frac{\gamma_n^{(m)} t}{3} \cos m\phi + d_{mn} \sin \frac{\gamma_n^{(m)} t}{3} \sin m\phi) \quad [2 \text{ marks}]$$

where $a_{mn} = E_{mn} A_m$, $b_{mn} = E_{mn} B_m$, $c_{mn} = F_{mn} A_m$, $d_{mn} = F_{mn} B_m$ are constants. $\square$

- (b) Find the values of the coefficients a_{mn} , b_{mn} , c_{mn} and d_{mn} in part (a). (9 marks)

Solution. From the second initial condition, we have

$$\frac{\partial u}{\partial t}(r, \theta, 0) = \sum_{n=0}^{\infty} J_m\left(\frac{\gamma_n^{(m)} r}{3}\right) \left(\frac{c_{mn} \gamma_n^{(m)}}{3} \cos m\phi + \frac{d_{mn} \gamma_n^{(m)}}{3} \sin m\phi\right) = 0 \quad [2 \text{ marks}]$$

Comparing coefficients, we have

$$\frac{c_{mn} \gamma_n^{(m)}}{3} = \frac{d_{mn} \gamma_n^{(m)}}{3} = 0 \Rightarrow c_{mn} = d_{mn} = 0 \quad [2 \text{ marks}]$$

From the first initial condition, we have

$$u(r, \theta, 0) = \sum_{m=0}^{\infty} \sum_{n=1}^{\infty} J_m\left(\frac{\gamma_n^{(m)} r}{3}\right) (a_{mn} \cos m\phi + b_{mn} \sin m\phi) = \psi(r, \phi) \quad [1 \text{ mark}]$$

Using the orthogonality property we have

$$a_{mn} = \frac{\int_0^{2\pi} \int_0^3 \psi(r, \phi) J_m\left(\frac{\gamma_n^{(m)} r}{3}\right) \cos m\phi r dr d\phi}{\int_0^{2\pi} \int_0^3 J_m^2\left(\frac{\gamma_n^{(m)} r}{3}\right) \cos^2 m\phi r dr d\phi} \quad [1 \text{ mark}]$$

$$= \frac{2}{9\pi(1 + \delta_{m0}) J_{m+1}^2(\gamma_n^{(m)})} \int_0^{2\pi} \int_0^3 \psi(r, \phi) J_m\left(\frac{\gamma_n^{(m)} r}{3}\right) \cos m\phi r dr d\phi \quad [1 \text{ mark}]$$

$$b_{mn} = \frac{\int_0^{2\pi} \int_0^3 \psi(r, \phi) J_m\left(\frac{\gamma_n^{(m)} r}{3}\right) \sin m\phi r dr d\phi}{\int_0^{2\pi} \int_0^3 J_m^2\left(\frac{\gamma_n^{(m)} r}{3}\right) \sin^2 m\phi r dr d\phi} \quad [1 \text{ mark}]$$

$$= \frac{2}{9\pi J_{m+1}^2(\gamma_n^{(m)})} \int_0^{2\pi} \int_0^3 \psi(r, \phi) J_m\left(\frac{\gamma_n^{(m)} r}{3}\right) \sin m\phi r dr d\phi \quad [1 \text{ mark}]$$

$\square$

A non-homogeneous example is given in Weinberger [1995, 185].

Chapter 37

Boundary-Value Problems in Spherical Coordinates

<https://www.youtube.com/watch?v=g4qalKCgGfc> : Spherical Harmonics and the Multipole Expansion

provides amazing explanation of spherical functions and its use in analysing the cosmos as well as the variations of height of the earth using multiple moments!!!

In this topic we turn our attention to problems in spheres and other regions, such as the region between two spheres or the region outside a sphere, for which it is natural to use spherical coordinates. For example, to find the steady-state temperature in a metallic ball whose surface is kept at a given temperature distribution, we need to solve Laplace equation inside the ball that takes the given boundary values on its surface. These and related problems can be treated by the same techniques as in the previous two topics — separation of variables and eigenfunction expansions.

Fourier series and Bessel series arose from applying the separation of variables to equations involving the Laplacian in cylindrical coordinates. Here, with equations involving the Laplacian in spherical coordinates, Legendre polynomials and Legendre functions are encountered when carrying the method of separation of variables. We will explore some examples in this topic and refer to Asmar [2005], Brown and Churchill [2008], Koshlyakov et al. [1964], Wyld [1999], O'Neil [2003], Collin [1991], Olver [2007], Boas [2006] for extra examples.

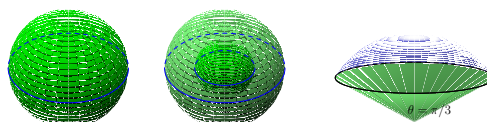
37.1 Second Order Linear PDEs in Spherical Coordinates

When the boundary of the domains consists of a sphere, it is appropriate for us to use spherical coordinates to express PDEs. The heat, Laplace and wave equations in Topic 2 need to be expressed using spherical coordinates for separation of variables to work.

37.1.1 Domains, Boundary Conditions and Initial Conditions

When the domains are a solid sphere, a spherical shell, a region outside a sphere or a portion of either of them, such as a spherical-cone. It is natural for us to express the PDE in spherical coordinates (r, ϕ, θ) , where r is the radius, $\phi \in [0, 2\pi]$ measures the *azimuthal angle* or *longitude* and $\theta \in [0, \pi]$ measures the *zenith angle* or *latitude*.

Some boundaries corresponding to the spherical domains are shown below:



Each piece of the boundary can be imposed of type 1, 2 or 3 boundary conditions. Again, we can have many possible combinations, but we will restrict ourselves to those cases which are relevant to science and engineering.

In the following sections, we will find that all except two kinds of boundary conditions are explicitly given. One of the implicitly assumed boundary condition is the regularity condition imposed on $r = 0$ and $r = \infty$ because the solution u is required to be finite in the given domain. The other is the periodic

boundary condition due to the periodicity of the solution u with respect to ϕ , i.e. $u(r, \phi + 2\pi, \theta) = u(r, \phi, \theta)$.

When there is a time variable t appearing in the PDE, apart imposing boundary conditions, we need to impose initial condition(s) to obtain a unique solution.

An initial condition in which the distribution is homogeneous has the following form

$$u(r, 0) = f(r); \quad (37.1) \quad \text{e:sph:init1a}$$

$$\frac{\partial u}{\partial t}(r, 0) = g(r) \quad (37.2) \quad \text{e:sph:init1b}$$

where $0 \leq r \leq L$ or $a < r < b$ or $L < r$.

An initial condition in which the distribution is radial symmetric has the following form

$$u(r, \theta, 0) = f(r, \theta); \quad (37.3) \quad \text{e:sph:init2a}$$

$$\frac{\partial u}{\partial t}(r, \theta, 0) = g(r, \theta) \quad (37.4) \quad \text{e:sph:init2b}$$

where $0 \leq r \leq L$ or $a < r < b$ or $L < r$; $0 \leq \theta \leq \pi$ or $a \leq \theta \leq b$.

An initial condition without any symmetry has the form

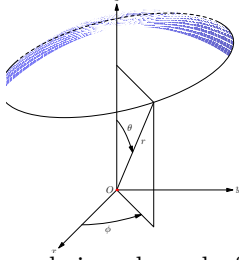
$$u(r, \phi, \theta, 0) = f(r, \phi, \theta); \quad (37.5) \quad \text{e:sph:init3a}$$

$$\frac{\partial u}{\partial t}(r, \phi, \theta, 0) = g(r, \phi, \theta) \quad (37.6) \quad \text{e:sph:init3b}$$

where $0 \leq r \leq L$ or $a < r < b$ or $L < r$; $0 \leq \theta \leq \pi$ or $a \leq \theta \leq b$; $0 \leq \phi \leq 2\pi$ or $c \leq \phi \leq d$.

37.1.2 Laplacian in Spherical Coordinates

In order for second order linear PDEs with Laplacian to be “separable” in spherical coordinates, we need to transform the Laplacian in Cartesian coordinates (x, y, z) to a Laplacian in spherical coordinates (r, ϕ, θ) through the following relation:



$$\begin{cases} x = r \cos \phi \sin \theta \\ y = r \sin \phi \sin \theta \\ z = r \cos \theta \end{cases} \quad \begin{cases} r = \sqrt{x^2 + y^2 + z^2} \\ \phi = \arctan\left(\frac{y}{x}\right) \\ \theta = \arctan\left(\frac{\sqrt{x^2 + y^2}}{z}\right) \end{cases}$$

Using chain rule and after simplifications (the details steps could be found in Asmar [2005, 196], the Laplacian of u in spherical coordinates is found to be

$$\begin{aligned} \Delta u &= \frac{1}{r^2} (r^2 u_r)_r + \frac{1}{r^2 \sin^2 \theta} u_{\phi\phi} + \frac{1}{r^2 \sin \theta} (\sin \theta u_{\theta})_{\theta} \\ &= u_{rr} + \frac{2}{r} u_r + \frac{1}{r^2 \sin^2 \theta} u_{\phi\phi} + \frac{1}{r^2} u_{\theta\theta} + \frac{\cot \theta}{r^2} u_{\theta}. \end{aligned}$$

37.1.3 Heat or Diffusion Equation

Our goal is to study heat propagation in a solid spherical body, e.g. the earth (assume that it is composed of a completely homogeneous and isotropic solid material). Let k be the diffusivity constant. A (homogeneous) heat equation (or a diffusion equation) $u_t - k\Delta u = 0$ (Topic 1) in a spherical domain with homogeneous distribution has the form

$$u_t - k(u_{rr} + \frac{2}{r} u_r) = 0. \quad (37.7) \quad \text{e:sph:heat1d}$$

A heat equation is said to be *radial symmetric* if it has a form

$$u_t - k(u_{rr} + \frac{2}{r} u_r + \frac{1}{r^2} u_{\theta\theta} + \frac{\cot \theta}{r^2} u_{\theta}) = 0. \quad (37.8) \quad \text{e:sph:heat2d}$$

A heat equation without any symmetry is

$$u_t - k(u_{rr} + \frac{2}{r} u_r + \frac{1}{r^2 \sin^2 \theta} u_{\phi\phi} + \frac{1}{r^2} u_{\theta\theta} + \frac{\cot \theta}{r^2} u_{\theta}) = 0. \quad (37.9) \quad \text{e:sph:heat3d}$$

Note that $t > 0$; $0 \leq r \leq L$ or $a < r < b$ or $L < r$; $0 \leq \theta \leq \pi$ or $a \leq \theta \leq b$; $0 \leq \phi \leq 2\pi$ or $c \leq \phi \leq d$.

37.1.4 Laplace Equation and Poisson Equation

A Laplace equation $\Delta u = 0$ with homogeneous distribution is “1-D” and becomes an ODE with a general solution

$$\frac{d^2 u}{dr^2} + \frac{2}{r} \frac{du}{dr} = 0 \Rightarrow u = A + \frac{B}{r}.$$

A Laplace equation which is *radial symmetric* has a form

$$\frac{\partial^2 u}{\partial r^2} + \frac{2}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} + \frac{\cot \theta}{r^2} \frac{\partial u}{\partial \theta} = 0, \quad (37.10) \quad \text{e:sph:lap12d}$$

while a Laplace equation without any symmetry has the form

$$\frac{\partial^2 u}{\partial r^2} + \frac{2}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 u}{\partial \phi^2} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} + \frac{\cot \theta}{r^2} \frac{\partial u}{\partial \theta} = 0. \quad (37.11) \quad \text{e:sph:lap13d}$$

Note that $0 \leq r \leq L$ or $a < r < b$ or $L < r$; $0 \leq \theta \leq \pi$ or $a \leq \theta \leq b$; $0 \leq \phi \leq 2\pi$ or $c \leq \phi \leq d$.

When the zeros on the right of (37.10) and (37.11) are replaced by non-zero functions $f(r, \theta)$ and $f(r, \phi, \theta)$ respectively, we obtain Poisson equations $\Delta u = f$ (Topic 1).

37.1.5 Wave Equation

An n -dimensional (*homogeneous*) wave equation $u_{tt} - c^2 \Delta u = 0$ (Topic 1) in a spherical domain with a homogeneous distribution has the form

$$u_{tt} - c^2(u_{rr} + \frac{2}{r}u_r) = 0. \quad (37.12) \quad \text{e:sph:wave1d}$$

It is said to be *radial symmetric* when

$$\frac{\partial^2 u}{\partial t^2} - c^2(\frac{\partial^2 u}{\partial r^2} + \frac{2}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} + \frac{\cot \theta}{r^2} \frac{\partial u}{\partial \theta}) = 0, \quad (37.13) \quad \text{e:sph:wave2d}$$

When the wave equation has no symmetry in spherical coordinates, it can be expressed in a “3-D” wave equation:

$$\frac{\partial^2 u}{\partial t^2} - c^2(\frac{\partial^2 u}{\partial r^2} + \frac{2}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 u}{\partial \phi^2} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} + \frac{\cot \theta}{r^2} \frac{\partial u}{\partial \theta}) = 0, \quad (37.14) \quad \text{e:sph:wave3d}$$

Note that $t > 0$; $0 \leq r \leq L$ or $a < r < b$ or $L < r$; $0 \leq \theta \leq \pi$ or $a \leq \theta \leq b$; $0 \leq \phi \leq 2\pi$ or $c \leq \phi \leq d$.

37.2 Special Functions

There are two types of special functions that we need when solving PDEs in cylindrical coordinates, i.e. spherical Bessel functions and associated Legendre functions. Associated Legendre functions and harmonic functions form spherical harmonics, which are used in the analysis of Helmholtz equation (and in the analysis of simple models in quantum mechanics). Spherical Bessel functions can be regarded as “trigonometric functions” in spherical coordinates.

37.2.1 Spherical Bessel Equation

Apply separation of variables to Helmholtz equation in spherical coordinate may sometimes leads to *radial equation* (which is a special case of an Euler equation):

$$r^2 R + 2rR' - n(n+1)R = 0, \quad n = 1, 2, \dots \quad (37.15) \quad \text{e:radial}$$

or sometimes leads to a *spherical Bessel equation* (with a parameter λ):

$$r^2 R'' + 2rR' + (\lambda r^2 - n(n+1))R = 0, \quad n = 0, \pm 1, \pm 2, \dots \quad (37.16) \quad \text{e:sph-bessel}$$

We have learned how to solve Euler equation in UECM2023 and the general solution of (37.15) is found to be

$$\omega(\omega - 1) + 2\omega - n(n+1) = 0 \Rightarrow \omega = n, -(n+1) \Rightarrow R(r) = Ar^n + Br^{-n-1}.$$

To solve the spherical Bessel equation (37.16), a trick is to transform (37.16) to a scaled Bessel equation of order $n + \frac{1}{2}$ using the transform $x = r$ and $y(x) = \sqrt{r}R(r)$:

$$x^2 y'' + xy' + (\lambda x^2 - (n + \frac{1}{2})^2)y = 0, \quad n = 0, \pm 1, \pm 2, \dots$$

By applying what we learn in Topic 3, we derive the general solution of the spherical Bessel equation

$$y(x) = AJ_{n+\frac{1}{2}}(\sqrt{\lambda}x) + BY_{n+\frac{1}{2}}(\sqrt{\lambda}x) \Rightarrow R(r) = A \frac{J_{n+\frac{1}{2}}(\sqrt{\lambda}r)}{\sqrt{r}} + B \frac{Y_{n+\frac{1}{2}}(\sqrt{\lambda}r)}{\sqrt{r}}$$

The two functions on the right hand side can be regarded as the Bessel functions for spherical coordinates and they are defined below.

def:sph-bess

Definition 37.1. The functions

$$j_n(x) = \sqrt{\frac{\pi}{2x}} J_{n+\frac{1}{2}}(x) \text{ and } y_n(x) = \sqrt{\frac{\pi}{2x}} Y_{n+\frac{1}{2}}(x)$$

are called the *Spherical Bessel functions of the first kind and the second kind* respectively.

With the definition above, the general solution of (37.16) can be written as

$$R(x) = \tilde{A}j_n(\sqrt{\lambda}r) + \tilde{B}y_n(\sqrt{\lambda}r)$$

where the pair $\tilde{A}, \tilde{B}$ and the pair A, B differ by just the constant $\sqrt{\pi/2}$.

Surprisingly, unlike the Bessel functions of integer order, the spherical Bessel functions are elementary functions which can be expressed in the so called *Rayleigh's formulas* [Abramowitz and Stegun, 1964, 439]:

$$j_n(x) = x^n \left(-\frac{1}{x} \frac{d}{dx} \right)^n \left(\frac{\sin x}{x} \right), \quad y_n(x) = -x^n \left(-\frac{1}{x} \frac{d}{dx} \right)^n \left(\frac{\cos x}{x} \right), \quad n = 0, 1, 2, \dots$$

In particular, $j_0(x) = \frac{\sin x}{x}$ and $y_0(x) = -\frac{\cos x}{x}$. Analytically, the spherical Bessel functions have the following approximations.

Function	Small x	Large x
$j_n(x)$	$\frac{x^n}{(2n+1)!!} + O(x^{n+2})$	$\frac{1}{x} \sin(x - \frac{n\pi}{2}) + O(x^{-2})$
$y_n(x)$	$-\frac{(2n-1)!!}{x^{n+1}} + O(x^{1-n})$	$-\frac{1}{x} \cos(x - \frac{n\pi}{2}) + O(x^{-2})$

According to the definition 37.1, the zeros of j_n and y_n are the same as the zeros of $J_{n+\frac{1}{2}}$ and $Y_{n+\frac{1}{2}}$. The orthogonal property is also characterised by the orthogonal property of Bessel functions. The following is the result of applying Theorem 3.2.8 from Topic 3, what we need to beware is that the weight function is x^2 instead of x :

hm:sph-bess:orth1

Theorem 37.1. If λ and μ are the roots of spherical Bessel function j_n , then

$$\int_0^1 x^2 j_n(\lambda x) j_n(\mu x) dx = \begin{cases} 0, & \lambda \neq \mu, \\ \frac{1}{2} [j_{n+1}(\lambda)]^2, & \lambda = \mu. \end{cases} \quad (37.17)$$

e:bessel:orth1

Proof: First, note that the roots λ and μ of j_n are also the roots of $J_{n+\frac{1}{2}}$.

$$\begin{aligned} \int_0^1 x^2 j_n(\lambda x) j_n(\mu x) dx &= \int_0^1 x^2 \sqrt{\frac{\pi}{2\lambda x}} J_{n+\frac{1}{2}}(\lambda x) \sqrt{\frac{\pi}{2\mu x}} J_{n+\frac{1}{2}}(\mu x) dx \\ &= \frac{\pi}{2\sqrt{\lambda\mu}} \int_0^1 x J_{n+\frac{1}{2}}(\lambda x) J_{n+\frac{1}{2}}(\mu x) dx = \frac{\pi}{2\sqrt{\lambda\mu}} \cdot \begin{cases} 0, & \lambda \neq \mu, \\ \frac{1}{2} [J_{n+\frac{3}{2}}(\lambda)]^2, & \lambda = \mu. \end{cases} \\ &= \begin{cases} 0, & \lambda \neq \mu, \\ \frac{\pi}{2\lambda} \cdot \frac{1}{2} [J_{n+\frac{3}{2}}(\lambda)]^2, & \lambda = \mu. \end{cases} = \begin{cases} 0, & \lambda \neq \mu, \\ \frac{1}{2} [j_{n+\frac{1}{2}}(\lambda)]^2, & \lambda = \mu. \end{cases} \end{aligned}$$

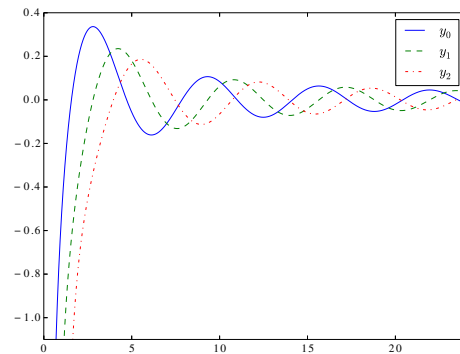
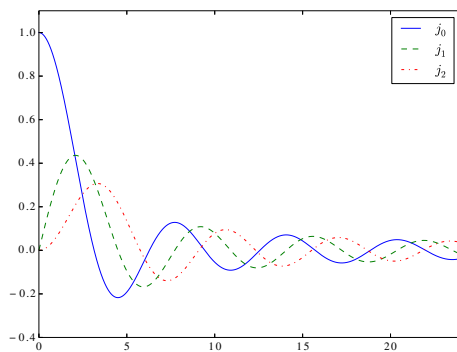
The proofs of the properties of spherical Bessel functions stated above can be found in Asmar [2005, 300–360] and Bell [1968b, Section 4.11]. The computation of the spherical Bessel functions could be obtained either using the Bessel functions or using the more specialised spherical Bessel functions which returns the values and the derivatives until order n :

```

1 from matplotlib.pyplot import plot, show, legend, savefig, clf, axis
2 from scipy.special import sph_jn, sph_yn
3 from numpy import zeros, arange, array
4
5 xarr = arange(0,24,0.002)
6 N = xarr.shape[0]
7 arr0 = zeros(N)
8 arr1 = zeros(N)
9 arr2 = zeros(N)
10 for i, x in enumerate(xarr):
11     y, yp = sph_jn(2,x)
12     arr0[i] = y[0]
13     arr1[i] = y[1]
14     arr2[i] = y[2]
15 plot(xarr,arr0,'-',xarr,arr1,'--',xarr,arr2,'-.')
16 legend(["$j_0$", "$j_1$", "$j_2$"])
17 axis([0,24,-0.4,1.1])
18 savefig("sphbes1plt.eps")
19 clf()
20 for i, x in enumerate(xarr):
21     y, yp = sph_yn(2,x)
22     arr0[i] = y[0]
23     arr1[i] = y[1]
24     arr2[i] = y[2]
25 plot(xarr,arr0,'-',xarr,arr1,'--',xarr,arr2,'-.')
26 legend(["$y_0$", "$y_1$", "$y_2$"])
27 axis([0,24,-1.1,0.4])
28 savefig("sphbes2plt.eps")

```

The spherical Bessel functions of order 0, 1, 2 are found to be:



37.2.2 Legendre Polynomials and Associated Legendre Functions

sec:legendre

In applying the separation of variables to solve a PDE in a sphere, we will obtain the following equation:

$$\sin^2 \theta \frac{d^2 \Theta}{d\theta^2} + \cos \theta \sin \theta \frac{d\Theta}{d\theta} + (\lambda \sin^2 \theta - m^2) \Theta = 0, \quad 0 \leq \theta \leq \pi. \quad (37.18)$$

e:TE

It can be shown that (37.18) can be transformed into the *m*-th order associated Legendre equation:

$$(1-x^2) \frac{d^2 y}{dx^2} - 2x \frac{dy}{dx} + \left(\lambda - \frac{m^2}{1-x^2} \right) y = 0, \quad -1 \leq x \leq 1 \quad (37.19)$$

e:assoc-legendre

with a suitable transform stated in the following example.

Example 37.2 (UCCM3003 Final Exam, Apr 2012, Q4(c)). Show that the differential equation (37.18) can be transformed into (37.19) by means of the substitution $x = \cos \theta$ and assuming $y(x) = \Theta(\theta)$. (8 marks)

e:TE
e:assoc-legendre

Proof: By introducing the change of variables $x = \cos \theta$ with $\Theta(\theta) = \Theta(\arccos x) = y(x)$. Since $0 \leq \theta \leq \pi$, we have $0 \leq \sqrt{1-x^2} = \sin \theta \leq 1$. According to the chain rule,

$$\frac{dy}{dx} = \frac{d\Theta}{d\theta} \frac{d\theta}{dx} = \frac{d\Theta}{d\theta} \frac{-1}{\sqrt{1-x^2}}, \quad (\because (\arccos x)' = \frac{-1}{\sqrt{1-x^2}})$$

$$\frac{d^2y}{dx^2} = \frac{d^2\Theta}{d\theta^2} \left(\frac{d\theta}{dx}\right)^2 + \frac{d\Theta}{d\theta} \frac{d^2\theta}{dx^2} = \frac{d^2\Theta}{d\theta^2} \frac{1}{1-x^2} + \frac{d\Theta}{d\theta} \frac{-x}{\sqrt{(1-x^2)^3}}.$$

Substitute into (37.18), we have the m -th order associated Legendre equation (37.19).

Example 37.3 (UECM3003 Final Exam, Apr 2015, Q5(a)). By using an appropriate substitution, show that the differential equation

$$\frac{d^2\Theta}{d\theta^2} + \cot \theta \frac{d\Theta}{d\theta} + \lambda\Theta = 0, \quad 0 < \theta < \pi,$$

can be transformed into Legendre equation

$$(1-x^2) \frac{d^2\Theta}{dx^2} - 2x \frac{d\Theta}{dx} + \lambda\Theta = 0, \quad -1 < x < 1. \quad (9 \text{ marks})$$

Solution. By introducing the change of variables $x = \cos \theta$. For $0 < \theta < \pi$, we have $0 < \sqrt{1-x^2} = \sin \theta < 1$ and

$$\cot \theta = \frac{\cos \theta}{\sin \theta} = \frac{x}{\sqrt{1-x^2}}.$$

According to the chain rule,

$$\frac{d\Theta}{dx} = \frac{d\Theta}{d\theta} \frac{d\theta}{dx} = \frac{d\Theta}{d\theta} \frac{-1}{\sqrt{1-x^2}} \Rightarrow \frac{d\Theta}{d\theta} = -\sqrt{1-x^2} \frac{d\Theta}{dx}$$

$$\frac{d^2\Theta}{dx^2} = \frac{d^2\Theta}{d\theta^2} \frac{1}{1-x^2} + \frac{d\Theta}{d\theta} \frac{-x}{\sqrt{(1-x^2)^3}} \Rightarrow \frac{d^2\Theta}{d\theta^2} = (1-x^2) \frac{d\Theta}{dx} + \frac{x}{\sqrt{1-x^2}} \left[-\sqrt{1-x^2} \frac{d\Theta}{dx} \right].$$

By substituting them into the original differential equation, we have

$$(1-x^2) \frac{d^2\Theta}{dx^2} - x \frac{d\Theta}{dx} + \frac{x}{\sqrt{1-x^2}} \left[-\sqrt{1-x^2} \frac{d\Theta}{dx} \right] + \lambda\Theta = 0 \Rightarrow \text{Legendre equation}$$

□

The Legendre equation in the above example can be written as when we replace Θ and its derivatives with y and its derivatives:

$$(1-x^2)y'' - 2xy' + \lambda y = 0, \quad -1 \leq x \leq 1. \quad (37.20)$$

is a special case of the associated Legendre equation (37.19) with $m = 0$. It is singular at $x = \pm 1$. To make sure y is well-defined at two end points $x = \pm 1$, we impose the boundary condition $|y(\pm 1)| < \infty$. When we write (37.20) in self-adjoint form (Topic 1, page 33) and together with the boundary condition, we obtain a singular SL-problem:

$$((1-x^2)y')' + \lambda y = 0, \quad -1 \leq x \leq 1, \quad |y(\pm 1)| < \infty. \quad (37.21)$$

Since 0 is an ordinary point, so there is a power series solution $y(x) = \sum_{n=0}^{\infty} a_n x^n$ at $x = 0$ and by substituting this series into Legendre equation (37.20), we obtain

$$\sum_{n=0}^{\infty} (n+2)(n+1)a_{n+2}x^n - \sum_{n=2}^{\infty} n(n-1)a_n x^n - \sum_{n=1}^{\infty} 2na_n x^n + \sum_{n=0}^{\infty} \lambda a_n x^n = 0.$$

Combining terms for $n \geq 2$ under one summation, we have

$$2a_2 + 6a_3x - 2a_1x + \lambda a_0 + \lambda a_1x + \sum_{n=2}^{\infty} [(n+2)(n+1)a_{n+2} - (n^2 + n - \lambda)a_n]x^n = 0.$$

The coefficients of each power of x must be zero, hence

$$2a_2 + \lambda a_0 = 0, \quad 6a_3 - 2a_1 + \lambda a_1 = 0, \quad (n+2)(n+1)a_{n+2} - (n^2 + n - \lambda)a_n = 0.$$

This implies

$$a_2 = \frac{-\lambda}{2!}a_0, \quad a_3 = \frac{2-\lambda}{3!}a_1, \quad a_{n+2} = \frac{n(n+1)-\lambda}{(n+1)(n+2)}, \quad n = 2, 3, \dots$$

Therefore, we obtain a local solution at $x = 0$:

$$y(x) = \sum_{n=0}^{\infty} a_n x^n = a_0 \left[1 - \frac{\lambda}{2!}x^2 - \frac{\lambda(6-\lambda)}{4!}x^4 - \frac{\lambda(6-\lambda)(20-\lambda)}{6!}x^6 + \dots \right] \\ + a_1 \left[x + \frac{2-\lambda}{3!}x^3 + \frac{(2-\lambda)(12-\lambda)}{5!}x^5 - \frac{(2-\lambda)(12-\lambda)(30-\lambda)}{7!}x^7 + \dots \right].$$

Amazingly, the eigenvalues and eigenfunctions (global solutions) of the singular SL problem (37.21) by choosing the “eigenvalue” $\lambda = n(n+1)$, $n = 0, 1, 2, \dots$:

- When $\lambda = 0$ and $a_1 = 0$, $y(x) = a_0$;
- When $\lambda = 2$ and $a_0 = 0$, $y(x) = a_1 x$;
- When $\lambda = 6$ and $a_1 = 0$, $y(x) = a_0(1 - 3x^2)$;
- When $\lambda = 12$ and $a_0 = 0$, $y(x) = a_1(x - \frac{5}{3}x^3)$;
- When $\lambda = 20$ and $a_1 = 0$, $y(x) = a_0(1 - 10x^2 + \frac{35}{3}x^4)$; etc.

These eigenfunctions can be normalised. After normalisation, we obtain the *Legendre polynomials* (or *Legendre functions of the first kind*), $P_n(x)$, given by the *Rodrigues formula* [Whittaker and Watson, 1920b, 303] and the expansion:

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n \quad \text{or} \quad P_n(x) = \sum_{k=0}^n \frac{(n+k)!}{(n-k)!(k!)^2 2^k} (x-1)^k.$$

In particular,

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = \frac{3x^2 - 1}{2}, \quad P_3(x) = \frac{1}{2}(5x^3 - 3x), \quad P_4(x) = \frac{35x^4 - 30x^2 + 3}{8}, \\ P_5(x) = \frac{63x^5 - 70x^3 + 15x}{8}, \quad P_6(x) = \frac{1}{16}(231x^6 - 315x^4 + 105x^2 - 5). \quad (37.22)$$

Theorem 37.4. [Asmar, 2005, 309] $(2n+1)P_n(x) = P'_{n+1}(x) - P'_{n-1}(x)$.

Theorem 37.5. [Zhang and Jin, 1996] $P_n(0) = \begin{cases} (-1)^{n/2} \frac{1 \cdot 3 \cdot 5 \cdots (n-1)}{2 \cdot 4 \cdot 6 \cdots n}, & n \text{ is even} \\ 0, & n \text{ is odd.} \end{cases}$

$$P'_n(0) = \begin{cases} (-1)^{(n-1)/2} \frac{1 \cdot 3 \cdot 5 \cdots n}{2 \cdot 4 \cdot 6 \cdots (n-1)}, & n \text{ is odd} \\ 0, & n \text{ is even.} \end{cases} \\ P_n(1) = 1; \quad P'_n(1) = \frac{n(n+1)}{2}; \quad P_n(-1) = (-1)^n; \quad P'_n(-1) = \frac{1}{2}(-1)^{n+1}n(n+1).$$

Theorem 37.6. $P_n(-x) = (-1)^n P_n(x)$, $P'_n(-x) = (-1)^{n+1} P'_n(x)$.

The above analysis could be generalised to investigate the eigenfunctions associated with (37.19), which are known as *associated Legendre functions* [Asmar, 2005]:

$$P_n^m(x) = (-1)^m (1-x^2)^{m/2} \frac{d^m}{dx^m} P_n(x), \quad n = m, m+1, \dots \quad (37.23)$$

The corresponding eigenvalues are $\lambda_n = n(n+1)$ for $n = 0, 1, 2, \dots$. It was first employed by Legendre to study the gravitational attraction of ellipsoidal bodies. The equation (37.19) is singular at $x = \pm 1$. To require y to be well-defined at $x = \pm 1$, we require that $|y(\pm 1)| < \infty$.

From the above analysis, the eigenfunctions are polynomials when $m = 0$. But when $m > 0$, the eigenfunctions (37.23) of the associated Legendre equation are not always polynomials. When $m = 2k \leq n$ is an even integer, $P_n^m(x)$ is a polynomial function, while when $m = 2k+1 \leq n$ is odd, there is an extra

factor of $\sqrt{1-x^2}$. Keep in mind that the square root is real and positive since we are restricting our attention to the interval $-1 \leq x \leq 1$.

A table of associated Legendre functions, which includes the Legendre polynomials as a special case, is stated below:

$P_n^m(x)$	$m = 0$	$m = 1$	$m = 2$	$m = 3$	$m = 4$
$n = 0$	1				
$n = 1$	x	$-\sqrt{1-x^2}$			
$n = 2$	$\frac{3x^2-1}{2}$	$-3x\sqrt{1-x^2}$	$3(1-x^2)$		
$n = 3$	$\frac{1}{2}(5x^3-3x)$	$-\left(\frac{-3+15x^2}{2}\right)\sqrt{1-x^2}$	$15x(1-x^2)$	$-15(1-x^2)^{\frac{3}{2}}$	
$n = 4$	$\frac{35x^4-30x^2+3}{8}$	$-\left(\frac{-15+35x^2}{2}\right)\sqrt{1-x^2}$	$\left(\frac{-15+105x^2}{2}\right)(1-x^2)$	$-105x(1-x^2)^{\frac{3}{2}}$	$105(1-x^2)^2$
$n = 5$	$\frac{63x^5-70x^3+15x}{8}$				

alegendre:special

Theorem 37.7. $P_n^m(0) = \begin{cases} (-1)^{(m+n)/2} \frac{1 \cdot 3 \cdot 5 \cdots (n+m-1)}{2 \cdot 4 \cdot 6 \cdots (n-m)}, & n+m \text{ is even} \\ 0, & n+m \text{ is odd.} \end{cases}$

$$\frac{d}{dx} P_n^m(0) = \begin{cases} (-1)^{(m+n-1)/2} \frac{1 \cdot 3 \cdot 5 \cdots (n+m)}{2 \cdot 4 \cdot 6 \cdots (n-m-1)}, & n+m \text{ is odd} \\ 0, & n+m \text{ is even.} \end{cases}$$

$$P_n^m(\pm 1) = (\pm 1)^n \delta_{m0}; \quad \frac{d}{dx} P_n^m(\pm 1) = \begin{cases} (\pm 1)^{n+1} \frac{n(n+1)}{2} & m = 0 \\ (\pm 1)^n \infty, & m = 1 \\ (\pm 1)^{n+1} \left(-\frac{1}{4}\right) \frac{(n+2)!}{(n-2)!}, & m = 2 \\ 0, & m > 2. \end{cases}$$

alegendre:symmetry

Theorem 37.8. $P_n^m(-x) = (-1)^{n-m} P_n^m(x)$.

Theorem 37.9. Let m be a positive integer. Then $P_n^{-m}(x) = (-1)^m \frac{(n-m)!}{(n+m)!} P_n^m(x)$.

There are many books, such as Bell [1968b, Chapter 3], Whittaker and Watson [1920b, Chapter XV], Koshlyakov et al. [1964, Chapter XV], Temme [1996b, Chapter 8], etc. characterising and proving the analytical properties of Legendre polynomials and associated Legendre functions. The analytical properties are important in the software implementation of these functions. Samuel S. Holland [1990, Section 4.6] investigated the relation between Legendre equation and Hilbert space $L^2(-1, 1)$.

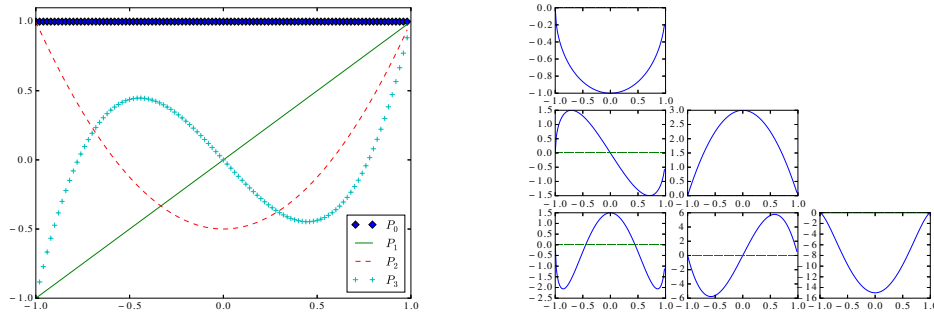
With modern plotting and scientific libraries, a simple program

```

1 from matplotlib.pyplot import plot, show, legend, savefig, clf, axis, subplot
2 from scipy.special import lpmv
3 from numpy import zeros, arange, array
4
5 x = arange(-1,1,0.02)
6 plot(x, lpmv(0,0,x), 'D', x, lpmv(0,1,x), '- ', x, lpmv(0,2,x), '-- ', x, lpmv(0,3,x), '+ ')
7 axis([-1,1,-1,1.1])
8 legend(["$P_0$", "$P_1$", "$P_2$", "$P_3$"], loc=4)
9 savefig("legendreplt.eps")
10
11 clf()
12 b = zeros(x.shape)
13 for m in range(1,4):
14     for n in range(m,4):
15         print("Processing", (m,n))
16         subplot(3,3,3*(n-1)+m)
17         plot(x, lpmv(m,n,x), '- ', x, b, ', ')
18 savefig("alegendreplt.eps")

```

can be used to plot the Legendre polynomials of order 0, 1, 2, 3 and the associated Legendre functions for $m = 1, 2, 3$ and $n = 1, 2, 3$:



The module `numpy.polynomial.legendre` implements much of the orthogonal relations above to be used in the evaluation of Gauss-Legendre quadrature (Topic 5 of UECM3033 Numerical Methods).

Returning to the associated Legendre equation (37.19) and the SL problem (37.18) with boundary condition $|\Theta(0)| < \infty$ and $|\Theta(\pi)| < \infty$. Their eigenvalues and eigenfunctions are expressed in terms of the associated Legendre functions respectively:

$$\lambda_n = n(n+1), \quad y(x) = P_n^m(x), \quad 0 \leq m \leq n; \quad (37.24)$$

$$\lambda_n = n(n+1), \quad \Theta_n^m(\theta) = P_n^m(\cos \theta), \quad 0 \leq m \leq n.$$

The eigenfunction $\Theta_n^m(\theta)$ is a trigonometric polynomial of degree n .

37.2.3 Orthogonal Property of Legendre Polynomials

As mentioned in Topic 1, the orthogonal property are valid for quite many singular SL problem, in particular, (37.21). Therefore, it is possible for us to expand a function over $[-1, 1]$ into Legendre polynomials.

The following orthogonality of Legendre polynomials P_n can be deduced using Rodrigues formula or its recurrence relation.

Theorem 37.10 (Orthogonality for Legendre Polynomials). *Let m and n be non-negative integers. Then*

$$\int_{-1}^1 P_m(x) P_n(x) dx = \frac{2}{2n+1} \delta_{mn}. \quad (37.25)$$

Using this property, any integrable function on $[-1, 1]$ can be expanded as a *Fourier-Legendre series*:

$$f(x) \sim \sum_{n=0}^{\infty} c_n P_n(x), \quad c_n = \frac{2n+1}{2} \int_{-1}^1 f(x) P_n(x) dx. \quad (37.26)$$

Remark 37.11. Sommerfeld [1949, 22] pointed out that the solution to approximating a function f given in the interval $-1 < x < 1$ by a sequence of polynomial $P_0, P_1, \dots, P_n$ of degrees $0, 1, \dots, n$ in the sense of “least square”, i.e. $\int_{-1}^1 [f(x) - (a_0 P_0(x) + \dots + a_n P_n(x))]^2 dx$ is minimum, is “Legendre polynomials”. In this case, when f is a polynomial of degree n , then $\int_{-1}^1 [f(x) - (a_0 P_0(x) + \dots + a_n P_n(x))]^2 dx = 0$ where $a_k, k = 0, 1, \dots, n$ are the “Fourier-Legendre” coefficients in (37.26).

Example 37.12 (UCCM3003 Final Exam, Apr 2012, Q4(a)).

- (i) The Legendre polynomials $P_n(x)$ are orthogonal on $[-1, 1]$ relative to the weight function $w(x) = 1$. Verify this fact for the first three Legendre polynomials:

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = \frac{1}{2}(3x^2 - 1). \quad (3 \text{ marks})$$

Proof:

$$\begin{aligned}\int_{-1}^1 P_0(x)P_1(x)dx &= \int_{-1}^1 xdx = 0 \\ \int_{-1}^1 P_0(x)P_2(x)dx &= \int_{-1}^1 \frac{1}{2}(3x^2 - 1)dx = \frac{1}{2} [x^3 - x]_{-1}^1 = 0 \\ \int_{-1}^1 P_1(x)P_2(x)dx &= \frac{1}{2} \left[\frac{3x^4}{4} - \frac{x^2}{2} \right]_{-1}^1 = \frac{1}{2} \left[\frac{1}{4} - \frac{1}{4} \right] = 0.\end{aligned}$$

(ii) Then find the first two coefficients in the expansion

$$f(x) = c_0 P_0(x) + c_1 P_1(x) + c_2 P_2(x) + \cdots,$$

where $f(x) = |x|$, $-1 < x < 1$. (5 marks)

Solution: First, note that f is even, therefore, $c_1 = 0$ because P_1 is odd (Theorem [37.6](#)). The value of c_0 is found to be

$$c_0 = \frac{1}{2} \int_{-1}^1 |x|dx = \frac{1}{2} \times 2 \int_0^1 xdx = \left[\frac{x^2}{2} \right]_0^1 = \frac{1}{2}.$$

Example 37.13 (UECM3003 Final Exam, Apr 2013, Q4(b)). Find the first three coefficients in the Fourier-Legendre expansion of

$$f(x) = |x|, \quad -1 < x < 1. \quad (7 \text{ marks})$$

Solution. The calculation of c_0 and c_1 are given in Example [37.12](#) and [eq:final:201204:q4a](#)

$$c_2 = \frac{5}{2} \int_{-1}^1 |x| \frac{3x^2 - 1}{2} dx = 5 \int_0^1 \frac{3x^3 - x}{2} dx = 5 \left[\frac{3x^4}{4} - \frac{x^2}{2} \right]_0^1 = \frac{5}{8}.$$

□

Example 37.14 (UCCM3003 Final Exam, Sep 2009, Q5(a)). Find the first two non-zero terms of the Legendre series expansion valid over the interval $(-1, 1)$ for the function

$$f(x) = \begin{cases} -1, & -1 \leq x < 0, \\ 1, & 0 \leq x < 1. \end{cases} \quad (11 \text{ marks})$$

[Remark: In exam, the formula [\(37.22\)](#) will be given]

Solution.

$$\begin{aligned}c_0 &= \frac{1}{2} \int_{-1}^1 f(x)P_0(x)dx = 0 \\ c_1 &= \frac{3}{2} \left[\int_{-1}^0 (-1)x dx + \int_0^1 x dx \right] = \frac{3}{2} \left[\frac{1}{2} + \frac{1}{2} \right] = \frac{3}{2}, \\ c_2 &= \frac{5}{2} \left[\int_{-1}^0 (-1) \frac{3x^2 - 1}{2} dx + \int_0^1 \frac{3x^2 - 1}{2} dx \right] = \frac{5}{2} (0 + 0) = 0 \\ c_3 &= \frac{7}{2} \left[\int_{-1}^0 (-1) \frac{5x^3 - 3x}{2} dx + \int_0^1 \frac{5x^3 - 3x}{2} dx \right] = \frac{7}{2} \left[\frac{-1}{8} + \frac{-1}{8} \right] = -\frac{7}{8}.\end{aligned}$$

If we take note that f is odd (removing $x = 0$) and P_{2n} are even, then we will know that $c_{2n} = 0$ and simplify the calculations above.

$$\text{So } f(x) \sim \frac{3}{2}x - \frac{7}{8} \left[\frac{5x^3 - 3x}{2} \right] + \cdots.$$

□

Example 37.15 (UCCM3003 Final Exam, Apr 2009, Q5(a)). Express $f(\theta) = 2 \cos \theta + \cos^2 \theta$ as a linear combination of associated Legendre polynomials $c_0 P_0(\cos \theta) + c_1 P_1(\cos \theta) + c_2 P_2(\cos \theta)$.

(5 marks)

Solution: Let $\cos \theta = x$. By applying the orthogonality of Legendre functions over the interval $-1 < x < 1$, we have $f(\theta) = 2x + x^2 = c_0 P_0(x) + c_1 P_1(x) + c_2 P_2(x)$ and

$$\begin{aligned} c_0 &= \frac{1}{2} \int_{-1}^1 (2x + x^2) dx = \frac{1}{2} \times \frac{2}{3} = \frac{1}{3} \\ c_1 &= \frac{3}{2} \int_{-1}^1 (2x + x^2) x dx = \frac{3}{2} \times \frac{4}{3} = 2 \\ c_2 &= \frac{5}{2} \int_{-1}^1 (2x + x^2) \frac{3x^2 - 1}{2} dx = \frac{5}{2} \times \frac{4}{15} = \frac{2}{3}. \end{aligned}$$

final:201101q3b

Example 37.16 (UCCM3003 Final Exam, Apr 2011, Q3(b)). Express $f(\theta) = 2 \cos \theta - \sin^2 \theta$ as a linear combination of the associated Legendre polynomials $P_0(\cos \theta)$, $P_1(\cos \theta)$ and $P_2(\cos \theta)$.

(5 marks)

Solution. Let $\cos \theta = x$. Then $f(\theta) = 2x - (1 - x^2) = -1 + 2x + x^2$. From

$$f(\theta) = c_0 P_0(\cos \theta) + c_1 P_1(\cos \theta) + c_2 P_2(\cos \theta) = c_0 P_0(x) + c_1 P_1(x) + c_2 P_2(x),$$

and by applying the orthogonality of Legendre polynomials over $-1 < x < 1$, we have

$$\begin{aligned} c_0 &= \frac{1}{2} \int_{-1}^1 (-1 + 2x + x^2) dx = -\frac{2}{3} \\ c_1 &= \frac{3}{2} \int_{-1}^1 (-1 + 2x + x^2) x dx = 2 \\ c_2 &= \frac{5}{2} \int_{-1}^1 (-1 + 2x + x^2) \frac{3x^2 - 1}{2} dx = \frac{2}{3}. \end{aligned}$$

$$\text{Therefore, } f(\theta) = -\frac{2}{3} + 2 \cos \theta + \frac{2}{3} \left[\frac{3 \cos^2 \theta - 1}{2} \right]. \quad \square$$

Example 37.17 (UECM3003 Final Exam, Apr 2014, Q4(c)). Write the expression $3 \cos \theta + \sin^2 \theta$ as a linear combination of Legendre polynomials $\sum_{n=0}^{\infty} c_n P_n(\cos \theta)$.

(6 marks)

Solution. Let $x = \cos \theta$, then

$$\begin{aligned} 3 \cos \theta + \sin^2 \theta &= 3x + (1 - x^2) = \sum_{n=0}^{\infty} c_n P_n(x), \quad c_n = 0, \quad n \geq 3 \\ c_0 &= \frac{1}{2} \int_{-1}^1 (1 + 3x - x^2) dx = \frac{1}{2} \times \frac{4}{3} = \frac{2}{3} \\ c_1 &= \frac{3}{2} \int_{-1}^1 (1 + 3x - x^2) x dx = \frac{3}{2} \times 2 = 3 \\ c_2 &= \frac{5}{2} \int_{-1}^1 (1 + 3x - x^2) \frac{3x^2 - 1}{2} dx = \frac{5}{2} \times \left(-\frac{4}{15}\right) = -\frac{2}{3}. \end{aligned}$$

$$\text{There } 3 \cos \theta + \sin^2 \theta = \frac{2}{3} + 3 \cos \theta - \frac{2}{3} \left[\frac{3 \cos^2 \theta - 1}{2} \right]. \quad \square$$

37.2.4 Orthogonal Property of Associated Legendre Functions

Similar to Legendre polynomials, associated Legendre functions have the following orthogonality relation.

hm:legendre:orth2

Theorem 37.18 (Orthogonality for Associated Legendre Functions). *Let k and ℓ be any integers greater or equal to m . Then*

$$\int_{-1}^1 P_k^m(x) P_\ell^m(x) dx = \frac{2}{2\ell + 1} \frac{(\ell + m)!}{(\ell - m)!} \delta_{k\ell}. \quad (37.27)$$

e:legendre:orth2

legendre:orth2

Example 37.19. Evaluate $I = \int_0^\pi \left[P_n^m(\cos \theta) \right]^2 \sin \theta d\theta$.

Solution. Let $x = \cos \theta$, then $dx = -\sin \theta d\theta$, $\cos 0 = 1$ and $\cos \pi = -1$. Therefore,

$$I = \int_1^{-1} \left[P_n^m(x) \right]^2 (-dx) = \int_{-1}^1 \left[P_n^m(x) \right]^2 dx = \frac{2}{2n+1} \frac{(n+m)!}{(n-m)!}$$

□

37.2.5 Spherical Harmonics

Spherical harmonics are the restrictions of homogeneous harmonic polynomials of three variables to the sphere S^2 [Singer, 2005]. According to Wikipedia [2016] and Press et al. [2007, Section 6.7], <http://dlmf.nist.gov/14.30>, they are defined as follows.

def:sph-harm

Definition 37.2. The (*normalised*) *spherical harmonics* in complex form is defined as

$$Y_\ell^m(\theta, \phi) := \sqrt{\frac{2\ell+1}{4\pi} \frac{(\ell-m)!}{(\ell+m)!}} e^{im\phi} P_\ell^m(\cos \theta) \quad (37.28)$$

e:sph:harmonics

where ℓ is a non-negative integer and m are integers such that $|m| \leq \ell$. The complex conjugate function of the spherical harmonics can be given by $(Y_\ell^m)^* = (-1)^m Y_\ell^{-m}$.

Remark 37.20. In the quantum mechanics, ℓ and m values of the spherical harmonics Y_ℓ^m are the *angular momentum* and *magnetic momentum numbers*. The definition of Y_ℓ^m in quantum mechanics differ from Definition 37.2 by a factor $(-1)^m$ according to Wikipedia [2016]. Wyld [1999, Chapter 3] and Tang [2007] followed the quantum mechanics' definition.

Example 37.21. [Wikipedia, 2016] Write down the analytical expression of $Y_0^0, Y_1^1, Y_1^{-1}, Y_1^0, Y_2^{-2}, Y_2^2, Y_2^{-1}, Y_2^1, Y_2^0$.

Solution:

$$\begin{aligned} Y_0^0 &= \sqrt{\frac{1}{4\pi}}, \\ Y_1^{\pm 1} &= \mp \sqrt{\frac{3}{8\pi}} \sin \theta e^{\pm i\phi}, \quad Y_1^0 = \sqrt{\frac{3}{4\pi}} \cos \theta, \\ Y_2^{\pm 2} &= \sqrt{\frac{15}{32\pi}} \sin^2 \theta e^{\pm 2i\phi}, \quad Y_2^{\pm 1} = \mp \sqrt{\frac{15}{8\pi}} \sin \theta \cos \theta e^{\pm i\phi}, \quad Y_2^0 = \sqrt{\frac{5}{16\pi}} (3 \cos^2 \theta - 1). \end{aligned}$$

def:sph-harm2

Definition 37.3. The (*normalised*) *spherical harmonics* in real form is defined as

$$\begin{aligned} Y_{\ell,m}(\phi, \theta) &:= \sqrt{\frac{2\ell+1}{2(1+\delta_{m0})\pi} \frac{(\ell-m)!}{(\ell+m)!}} P_\ell^m(\cos \theta) \cos m\phi, \\ \tilde{Y}_{\ell,m}(\phi, \theta) &:= \sqrt{\frac{2\ell+1}{2\pi} \frac{(\ell-m)!}{(\ell+m)!}} P_\ell^m(\cos \theta) \sin m\phi. \end{aligned} \quad (37.29)$$

e:sph:harmonics2

Combining the orthogonality of Legendre functions and trigonometric functions, we obtain the orthogonality of spherical harmonics.

Theorem 37.22 (Orthogonality of Spherical Harmonics). *Let m, n, ℓ, p be integers such that $|m| \leq n$ and $|p| \leq \ell$. Then for any pair of integers (n, m) and (ℓ, p) ,*

$$\begin{aligned} \langle Y_{n,m}, \tilde{Y}_{\ell,p} \rangle &= 0, \\ \langle Y_n^m, Y_\ell^p \rangle &= \delta_{n\ell} \delta_{mp}; \quad \langle Y_{n,m}, Y_{\ell,p} \rangle = \delta_{n\ell} \delta_{mp}, \\ \langle \tilde{Y}_{n,m}, \tilde{Y}_{\ell,p} \rangle &= \delta_{n\ell} \delta_{mp} \end{aligned}$$

where S^2 is the unit sphere and the “inner product” of g and h is defined as

$$\langle g, h \rangle := \iint_{S^2} gh dS = \int_0^\pi \int_0^{2\pi} g(\phi, \theta) h(\phi, \theta) \sin \theta d\theta d\phi.$$

Proof: The orthogonal part is easy to prove using the results from Topic 1. We will only show the case when the index are the same:

$$\begin{aligned} \langle Y_{\ell,m}, Y_{\ell,m} \rangle &= \frac{2\ell+1}{4\pi} \frac{(\ell-m)!}{(\ell+m)!} \int_0^\pi \int_0^{2\pi} \left[P_\ell^m(\cos \theta) \cos m\phi \right]^2 \sin \theta d\theta d\phi \\ &= \frac{2\ell+1}{4\pi} \frac{(\ell-m)!}{(\ell+m)!} \int_0^\pi \left[P_\ell^m(\cos \theta) \right]^2 \sin \theta d\theta \int_0^{2\pi} \cos^2 m\phi d\phi \\ &= \frac{2\ell+1}{2(1+\delta_{m0})\pi} \frac{(\ell-m)!}{(\ell+m)!} \times \frac{2}{2\ell+1} \frac{(\ell+m)!}{(\ell-m)!} \times \pi(1+\delta_{m0}) = 1 \end{aligned}$$

by using Example [eg:legendre:orth2](#) and the orthogonal properties of trigonometric functions from Topic 1.

The derivation of various formulas and proofs can be found in Bell [1968b, Section 3.11] and Koshlyakov et al. [1964, Chapter XXI]. With the analytical formulas, spherical harmonics can be implemented in software (in Python, it can be called by `scipy.special.sph_harm`). The spherical harmonics form a complete orthogonal system over a sphere, allowing the general solution to the Laplace equation ([37.11](#)) in a ball to be written as spherical harmonics.

Using the orthogonality of the spherical Bessel functions (Theorem [37.1](#)) and the spherical harmonics, we have the following orthogonality property [Asmar, 2005].

Theorem 37.23 (Orthogonality of the Eigenfunctions of 3-D Helmholtz Equation). *Let $\tau_k^{(n)}$, $k = 1, 2, \dots$ be the zeros of j_n , $n = 0, 1, 2, \dots$. For $k = 1, 2, \dots$, $n = 0, 1, 2, \dots$ and $|m| \leq n$, we have*

$$\int_0^L \int_0^{2\pi} \int_0^\pi j_n\left(\frac{\tau_k^{(n)} r}{L}\right) Y_{n,m}(\theta, \phi) j_n\left(\frac{\tau_{k'}^{(n')}}{L} r\right) Y_{n',m'}(\theta, \phi) r^2 \sin \theta d\theta d\phi dr = \frac{L^3}{2} j_{n+1}^2(\tau_k^{(n)}) \delta_{kk'} \delta_{nn'} \delta_{mm'}.$$

Spherical harmonics are important in quantum chemistry. It is through the “visualisation” of spherical harmonics that we start to understand more about the chemical bonds and hence the chemical properties of various molecules. One can find a comprehensive list of the graphs of spherical harmonics $r = Y_{n,m}(\phi, \theta)$ from the website <http://people.csail.mit.edu/sparis/sh/index.php?img=64>. A nicely constructed of $(Y_{n,m}(\phi, \theta), \phi, \theta)$ are given in a chapter http://www.math.umn.edu/~olver/am_/pdz.pdf from P. J. Olver’s draft on “Applied Mathematics” which is included below.

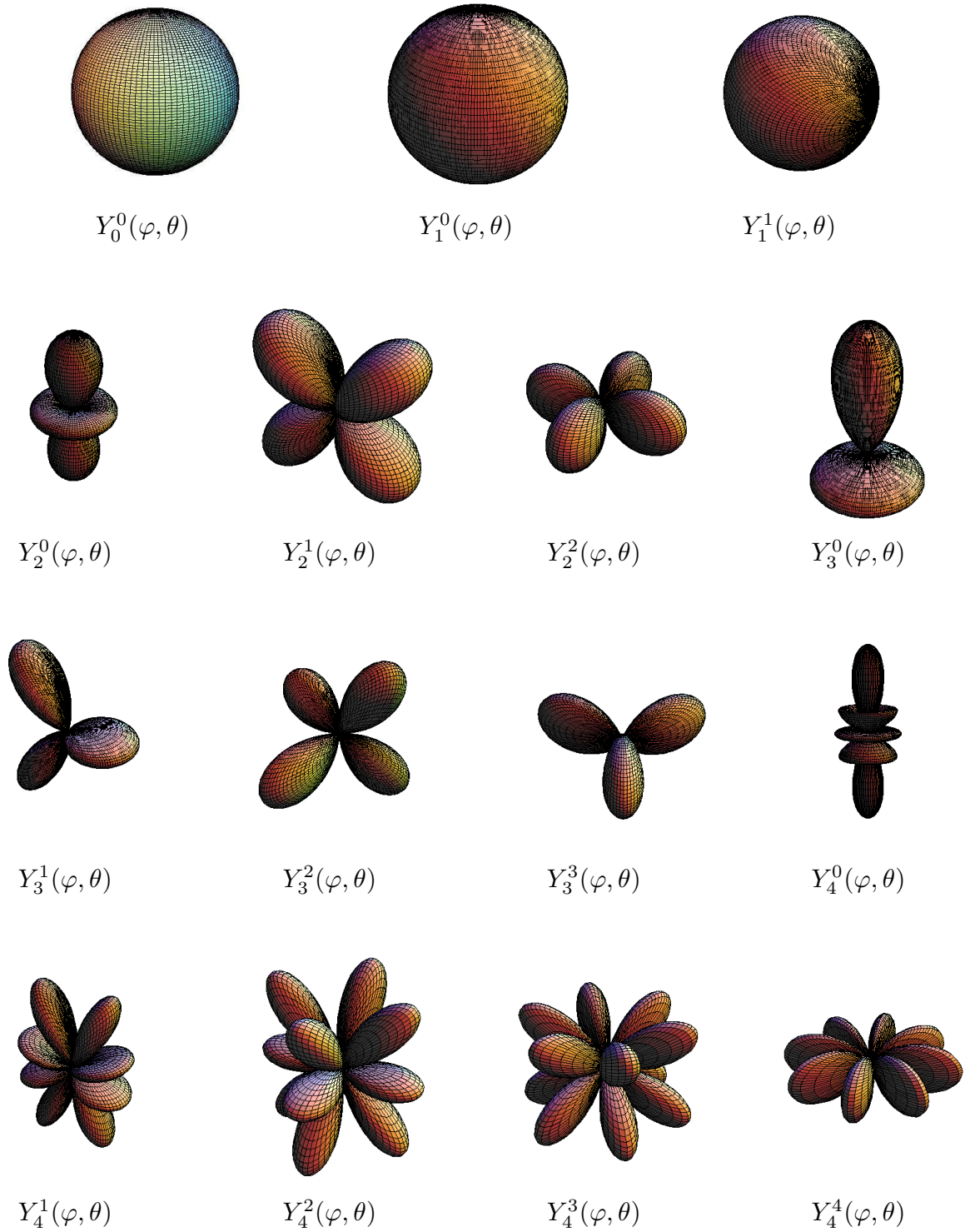


Figure 18.3. Spherical Harmonics.

37.3 1-D Heat Equation (on a Sphere with Homogeneous Distribution)

Heat equation on a sphere is used in geophysics [Wikipedia, 2016] to model the heat flow of the earth. The simplest model is a heat equation with homogeneous distribution which is characterised by (37.7). In this section, we will investigate type 1 and 3 boundary conditions which are scientifically relevant.

37.3.1 (Homogeneous) Type 1 BC

Example 37.24. Let $u(r, t)$ denote the temperatures in a solid sphere $r \leq L$, where r is the spherical coordinate, when that solid is initially at temperatures $f(r)$ and its surface $r = L$ is kept at temperature zero. The problem can be described by a heat equation (37.7), the initial condition (37.1) and a type 1 boundary condition:

$$u(L, t) = 0, \quad t > 0.$$

Find the formal series solution of this problem.

Solution: It is possible to introduce a transform $v(r, t) = ru(r, t)$ to obtain a new boundary value problem in $v(r, t)$ which has the same structure as heat equation in rectangular coordinate:

$$\begin{cases} v_t(r, t) = kv_{rr}(r, t), & 0 < r < L, \quad t > 0, \\ v(0, t) = v(L, t) = 0, & t > 0, \\ v(r, 0) = rf(r), & 0 \leq r \leq L. \end{cases}$$

Using the result from Topic 2, we find the solution of the homogeneous PDE problem of v and u are

$$v = \sum_{n=1}^{\infty} b_n e^{-\frac{n^2 \pi^2 k}{L^2} t} \sin \frac{n \pi r}{L} \Rightarrow u(r, t) = \frac{v}{r} = \frac{1}{r} \sum_{n=1}^{\infty} b_n e^{-\frac{n^2 \pi^2 k}{L^2} t} \sin \frac{n \pi r}{L},$$

where

$$b_n = \frac{2}{L} \int_0^L r f(r) \sin \frac{n \pi r}{L} dr, \quad n = 1, 2, \dots$$

Example 37.25 (UCCM3003 Final Exam, Apr 2008, Q4(a)). A solid spherical body 10 cm in radius, initially at 90°C throughout, is cooled by keeping its surface at 0°C. The cooling process is described by the following partial differential equation:

$$\begin{cases} u_t = k(u_{rr} + \frac{2}{r}u_r), & 0 \leq r < 10, \quad t > 0 \\ u|_{r=10} = 0, & t > 0, \\ u|_{t=0} = 90, & 0 \leq r \leq 10. \end{cases}$$

Show that the general solution is

$$u(x, t) = \sum_{n=1}^{\infty} b_n e^{-\frac{k n^2 \pi^2 t}{100}} \frac{\sin \frac{n \pi r}{10}}{r}.$$

(Hint: $j_0(x) = \frac{\sin x}{x}$.)

(15 marks)

Solution. Let $u = TR$. From the PDE,

$$\frac{T'}{kT} = \frac{R''}{R} + \frac{2}{r} \frac{R'}{R} = -\lambda \quad [3 \text{ marks}]$$

where $\lambda > 0$ is a constant.

The BC implies

$$R(10) = 0. \quad [1 \text{ mark}]$$

From the above equations we obtain a spherical Bessel equation (37.16) (on page 845) with $n = 0$ for R :

$$\begin{cases} r^2 R'' + 2r R' + \lambda r^2 R = 0, \\ R(10) = 0. \end{cases} \quad [1 \text{ mark}]$$

and an ODE of T :

$$T' + k\lambda T = 0 \quad [1 \text{ mark}]$$

The general solution is (from Formula sheet)

$$R(r) = A j_0(\sqrt{\lambda} r) + B y_n(\sqrt{\lambda} r) \quad [1 \text{ mark}]$$

Since $y_n(r)$ is singular at $r = 0$ and $R(r)$ needs to be well-defined at $r = 0$,

$B = 0$ [1 mark]

From $R(10) = 0$, we have

$$A j_0(\sqrt{\lambda} \times 10) = A \frac{\sin(\sqrt{\lambda} \times 10)}{\sqrt{\lambda} \times 10} = 0 \quad [1 \text{ mark}]$$

Since $\lambda > 0$ and $A \neq 0$, $\sqrt{\lambda} = \frac{n\pi}{10}$, $n = 0, 1, 2, \dots$ and $R_n(r) = A_n j_0(\frac{n\pi}{10} r)$ [2 marks]

For $\lambda = (\frac{n\pi}{10})^2$, $n = 1, 2, \dots$, the solution for T is

$$T_n(t) = C_n e^{-k(\frac{n\pi}{10})^2 t} \quad [2 \text{ marks}]$$

The general solution is

$$u(x, t) = \sum_{n=1}^{\infty} T_n(t) R_n(r) = \sum_{n=1}^{\infty} a_n e^{-k(\frac{n\pi}{10})^2 t} \frac{\sin \frac{n\pi}{10} r}{\frac{n\pi}{10} r} \quad [2 \text{ marks}]$$

where $a_n = A_n C_n$ are constants. □

Example 37.26 (UECM3003 Final Exam, Apr 2014, Q3(b)). Consider a homogeneous solid sphere of radius 1. Let the temperature be $u(r, t)$, and its initial temperature distribution be $f(r)$. It is then cooled by keeping its surface at 0°C . The cooling process is given by the boundary value problem

$$\begin{aligned} u_t &= 4(u_{rr} + \frac{2}{r}u_r), \quad 0 \leq r \leq 1, \\ u|_{r=1} &= 0, \quad t > 0. \end{aligned}$$

Show that the general solution for the BVP is

$$u(r, t) = \sum_{n=1}^{\infty} c_n j_0(\alpha_n r) e^{-4\alpha_n^2 t},$$

where j_0 is the spherical Bessel function and α_n is the n th root of j_0 . (12 marks)

Solution. By applying the separation of variables we obtain an ODE of T and a spherical Bessel equation of R :

$$\begin{cases} T' + 4\lambda T = 0, \\ r^2 R'' + 2rR' + \lambda r^2 R = 0, \quad R(1) = 0. \end{cases}$$

According the Formula Sheet, the general solution is

$$R(r) = A j_0(\sqrt{\lambda} r) + B y_n(\sqrt{\lambda} r)$$

Since y_n is singular at $r = 0$ and R needs to be well-defined at $r = 0$, $B = 0$. From $R(1) = 0$, we have

$$A j_0(\sqrt{\lambda}) = A \frac{\sin(\sqrt{\lambda})}{\sqrt{\lambda}} = 0 \quad [1 \text{ mark}]$$

Since $\lambda > 0$ and $A \neq 0$, the eigenvalues and corresponding eigenfunctions are found to be

$$\lambda = (n\pi)^2 = \alpha_n^2, \quad R_n(r) = A_n j_0(\alpha_n r), \quad n = 1, 2, \dots$$

For $\lambda = \alpha_n^2$, the solution for T is

$$T_n(t) = C_n e^{-4\alpha_n^2 t}$$

By the principle of superposition, the general solution is

$$u(x, t) = \sum_{n=1}^{\infty} T_n(t) R_n(r) = \sum_{n=1}^{\infty} a_n e^{-4\alpha_n^2 t} j_0(\alpha_n r)$$

where $a_n = A_n C_n$ are constants. □

Example 37.27 (UECM3003 Final Exam, Apr 2016, Q4(b)). The time-dependent temperature within

a radial symmetric sphere of unit radius is determined by

$$\begin{cases} \frac{\partial u}{\partial t} - \frac{1}{r^2} \frac{\partial}{\partial r} (r^2 \frac{\partial u}{\partial r}) = 0, & 0 < r < 1, \ t > 0, \\ u(1, t) = 100, & t > 0, \\ u(r, 0) = 0, & 0 < r < 1. \end{cases}$$

Topic 4,
LO6

- (i) Find the formal series solution of u . You need to write your answer clearly with justifications. (13 marks)

Solution. The problem is not homogeneous because boundary condition is constant. A transform $u = 100 + v$ will allow us to obtain a homogeneous problem

$$\begin{cases} \frac{\partial v}{\partial t} - \frac{1}{r^2} \frac{\partial}{\partial r} (r^2 \frac{\partial v}{\partial r}) = 0, & 0 < r < 1, \ t > 0, \\ u(1, t) = 100 + v(1, t) = 100 \Rightarrow v(1, t) = 0, & t > 0, \\ u(r, 0) = 100 + v(r, 0) = 0 \Rightarrow v(r, 0) = -100, & 0 < r < 1. \end{cases} \quad [2 \text{ marks}]$$

We can now apply separation of variables $v = TR$ to obtain

$$T'R - \frac{1}{r^2} (2rTR' + r^2TR'') = 0 \Rightarrow \frac{T'}{T} = \frac{2}{r} \frac{R'}{R} + \frac{1}{r^2} \frac{R''}{R} = -\lambda. \quad [2 \text{ marks}]$$

From here, we obtain an ODE $T' + \lambda T = 0$ and together with the homogeneous type 1 boundary condition, we obtain an Sturm-Liouville problem

$$r^2 R'' + 2rR' + \lambda r^2 R = 0, \quad R(1) = 0 \quad [1 \text{ mark}]$$

We identify this as a spherical Bessel equation with $n = 0$ from the Formula Sheet and it has the general solution

$$R = Aj_0(\sqrt{\lambda}r) + By_0(\sqrt{\lambda}r). \quad [1 \text{ mark}]$$

By looking at the definition of j_0 and y_0 from the Formula Sheet, we find that

$$j_0(x) = (-1)^0 \left(\frac{1}{x} \frac{d}{dx} \right)^0 \left(\frac{\sin x}{x} \right) = \frac{\sin x}{x}, \quad y_0(x) = -\frac{\cos x}{x}$$

Since y_0 is singular at $x = 0$, $B = 0$ in order to meet the implicit assume regularity condition. [1 mark]

In order to non-trivial solutions, $A \neq 0$ and this means that we can obtain the eigenvalues

$$R(1) = 0 \Rightarrow j_0(\sqrt{\lambda}) = \frac{\sin \sqrt{\lambda}}{\sqrt{\lambda}} = 0 \Rightarrow \sin \sqrt{\lambda} = 0 \Rightarrow \lambda = (n\pi)^2, \quad n = 1, 2, \dots \quad [2 \text{ marks}]$$

From the eigenvalues $\lambda = (n\pi)^2$, we can obtain the corresponding eigenfunctions, the solution of the ODE of T and the general solution is:

$$R_n = A_n \frac{\sin n\pi r}{n\pi r}, \quad T_n = C_n e^{-(n\pi)^2 t} \Rightarrow v = \sum_{n=1}^{\infty} T_n R_n = \sum_{n=1}^{\infty} a_n e^{-(n\pi)^2 t} \frac{\sin n\pi r}{n\pi r}. \quad [1 \text{ mark}]$$

The initial condition

$$v(r, 0) = \sum_{n=1}^{\infty} a_n \frac{\sin n\pi r}{n\pi r} = -100 \Rightarrow \sum_{n=1}^{\infty} \frac{a_n}{n\pi} \sin n\pi r = -100r. \quad [1 \text{ mark}]$$

By using the knowledge of Fourier series, we obtain

$$\frac{a_n}{n\pi} = 2 \int_0^1 (-100r) \sin n\pi r dr = -200 \left[-\frac{r \cos n\pi r}{n\pi} + \frac{\sin n\pi r}{(n\pi)^2} \right]_0^1 = \frac{200(-1)^n}{n\pi}. \quad [1 \text{ mark}]$$

Therefore, the formal series solution is

$$u = 100 + v = 100 + \sum_{n=1}^{\infty} 200(-1)^n e^{-(n\pi)^2 t} \frac{\sin n\pi r}{n\pi r}. \quad [1 \text{ mark}]$$

□

- (ii) Determine what happens to u when $t \rightarrow \infty$. (2 marks)

Solution. $u \rightarrow 100$ since the type 1 boundary is set to 100. [2 marks]

□

37.3.2 (Homogeneous) Type 3 BC

We will skip type 2 BC because it lacks scientific application. Type 3 BC describes the dissipation of heat by convection into a medium depends on the environment's temperature.

We learn in Section 37.3.1 that heat equation with radial symmetry can be solved by a suitable transform. This is true for type 3 boundary conditions as well. However, we learn that type 3 boundary condition is complex in Topic 2. The same complexity occur in spherical coordinates as well.

Example 37.28. ([Koshlyakov et al., 1964, 509]) Consider the case in which the temperature at an arbitrary point of the sphere depends only on the distance r of that point from the centre. If on the surface of the sphere there is an exchange of heat with the surrounding medium, the temperature of which we assume to be zero. In this case, the problem reduces to integrating the heat equation (37.7), initial condition (37.1) and a type 3 boundary condition below

$$\left(\frac{\partial u}{\partial r} + hu \right) \Big|_{r=L} = 0, \quad t > 0.$$

Find a formal series solution for this problem.

Solution: Similar to Example 37.24, we can introduce the transform $v = ru$ and reduce the original problem to a problem in “rectangular coordinate”:

$$\begin{cases} \frac{\partial v}{\partial t} = k \frac{\partial^2 v}{\partial r^2}, & 0 < r < L, \quad t > 0, \\ v|_{r=0} = 0, & t > 0, \\ \frac{\partial v}{\partial r} + (h - \frac{1}{R})v|_{r=L} = 0, & t > 0, \\ v|_{t=0} = rf(r), & 0 \leq r \leq L. \end{cases} \quad (37.30)$$

e:110

The problem is thus reduced to the problem of heat-flow in a slab/rod in Topic 2, the temperature at one end of which is held equal to zero and the other end of which exchanges heat with the surrounding medium.

This is a mixed type 1 BC and type 3 BC problem. Using separable of variables, we have the eigenvalues and corresponding eigenfunctions of the Sturm-Liouville problem to be

$$\lambda = \left(\frac{\mu_n}{L} \right)^2, \quad R_n(r) = \sin \frac{\mu_n r}{L}, \quad n = 1, 2, \dots$$

and the boundary conditions requires that μ_n to be a root of the equation

$$\mu \cos \mu + (hL - 1) \sin \mu = 0.$$

We learned that the above equation has infinitely many solutions $\mu_1, \mu_2, \dots$ in Topic 2. After finding T_n and applying the principle of superposition, we obtain a general solution

$$v(r, t) = \sum_{n=1}^{\infty} a_n e^{-(\mu_n/L)^2 t} \sin \frac{\mu_n r}{L}.$$

Substituting into the initial condition, we have

$$v(r, 0) = \sum_{n=1}^{\infty} a_n \sin \frac{\mu_n r}{L} = rf(r).$$

Multiply both sides with $\sin \frac{\mu_k r}{L}$ and integrate from 0 to L , and using the equalities

$$\int_0^L \sin \frac{\mu_n r}{L} \sin \frac{\mu_k r}{L} dr = \begin{cases} 0, & k \neq n, \\ \frac{L}{2} \frac{p(p+1) + \mu_n^2}{p^2 + \mu_n^2}, & k = n, \end{cases}$$

where $p = hL - 1$. We find the coefficients a_n from the formula

$$a_n \left(\frac{L}{2} \frac{p(p+1) + \mu_n^2}{p^2 + \mu_n^2} \right) = \int_0^L r f(r) \sin \frac{\mu_n r}{L} dr.$$

Substituting this expression back to $v(x, t)$ and remembering $v = ru$, we obtain the solution of the problem in the form

$$u(r, t) = \frac{2}{Lr} \sum_{n=1}^{\infty} \frac{p^2 + \mu_n^2}{p(p+1) + \mu_n^2} e^{-(\mu_n a/L)^2 t} \sin \frac{\mu_n r}{L} \int_0^L \rho f(\rho) \sin \frac{\mu_n \rho}{L} d\rho.$$

Remark 37.29. The derivation of $\int_0^L \sin^2 \frac{\mu_n x}{L} dx = \frac{L}{2} \left[\frac{hL(hL-1) + \mu_n^2}{(hL-1)^2 + \mu_n^2} \right]$ should be stated in Topic 2. However, it is only through the technique mentioned in Topic 3 that we can prove it: Let λ and μ be any positive real numbers and let $X = \sin \sqrt{\lambda}x$, $Y = \sin \sqrt{\mu}x$. Then

$$\begin{cases} X'' + \lambda X = 0 \\ Y'' + \mu Y = 0 \end{cases} \Rightarrow X''Y - XY'' + (\lambda - \mu)XY = 0 \Rightarrow (\mu - \lambda)XY = (X'Y - XY')'.$$

Therefore

$$\begin{aligned} (\mu - \lambda) \int_0^1 \sin \sqrt{\lambda}x \sin \sqrt{\mu}x dx &= \left[\sqrt{\lambda} \cos \sqrt{\lambda}x \sin \sqrt{\mu}x - \sqrt{\mu} \sin \sqrt{\lambda}x \cos \sqrt{\mu}x \right]_0^1 \\ &= \sqrt{\lambda} \cos \sqrt{\lambda} \sin \sqrt{\mu} - \sqrt{\mu} \sin \sqrt{\lambda} \cos \sqrt{\mu} \end{aligned}$$

When $\mu \rightarrow \lambda$, we have

$$\begin{aligned} \int_0^1 \sin^2 \sqrt{\lambda}x dx &= \lim_{\mu \rightarrow \lambda} \frac{\sqrt{\lambda} \cos \sqrt{\lambda} \sin \sqrt{\mu} - \sqrt{\mu} \sin \sqrt{\lambda} \cos \sqrt{\mu}}{\mu - \lambda} \\ &= \lim_{\mu \rightarrow \lambda} \left[\sqrt{\lambda} \cos \sqrt{\lambda} \cdot \frac{\cos \sqrt{\mu}}{2\sqrt{\mu}} - \sin \sqrt{\lambda} \left(\frac{1}{2\sqrt{\mu}} \cdot \cos \sqrt{\mu} + \sqrt{\mu} \cdot (-\sin \sqrt{\mu}) \cdot \frac{1}{2\sqrt{\mu}} \right) \right] \\ &= \frac{\cos^2 \sqrt{\lambda}}{2} - \sin \sqrt{\lambda} \left(\frac{\cos \sqrt{\lambda}}{2\sqrt{\lambda}} - \frac{\sin \sqrt{\lambda}}{2} \right) = \frac{1}{2} - \frac{\sin \sqrt{\lambda} \cos \sqrt{\lambda}}{2\sqrt{\lambda}} \end{aligned}$$

For $\lambda = \mu_n^2$ and a change of variable $x \rightarrow \frac{x}{L}$, we have

$$\begin{aligned} \frac{1}{L} \int_0^L \sin^2 \frac{\mu_n x}{L} dx &= \frac{1}{2} - \frac{\sin 2\mu_n}{4\mu_n} = \frac{1}{2} - \frac{2 \tan \mu_n}{4\mu_n(1 + \tan^2 \mu_n)} = \frac{1}{2} - \frac{\frac{\mu_n}{1-hL}}{2\mu_n(1 + (\frac{\mu_n}{1-hL})^2)} \\ \Rightarrow \int_0^L \sin^2 \frac{\mu_n x}{L} dx &= \frac{L}{2} \left[1 - \frac{1-hL}{(1-hL)^2 + \mu_n^2} \right] = \frac{L}{2} \left[\frac{(hL-1)hL + \mu_n^2}{(hL-1)^2 + \mu_n^2} \right] \end{aligned}$$

where $\tan \mu_n = \frac{\mu_n}{1-hL}$ and $\sin 2A = \frac{2 \tan A}{1 + \tan^2 A}$.

37.3.3 Nonhomogeneous PDE and Homogeneous BC

When the PDE is non-homogeneous but the boundary conditions are homogeneous, we can apply the eigenfunction expansion we learned in Topic 2. However, it is possible to find the solution quickly if we know a suitable transform as illustrated below.

Example 37.30 (UCCM3003 Final Exam, Sep 2007, Q4(a)). A solid sphere $r \leq 1$ is initially at temperature zero and its surface is constantly kept at that temperature. The temperature distribution function $u = u(r, t)$ satisfies the following non-homogeneous heat equation:

$$\frac{\partial u}{\partial t} = \frac{1}{r} \frac{\partial^2}{\partial r^2} (ru) + 1, \quad 0 < r < 1, \quad t > 0.$$

- (i) Transform the PDE into the following form by using $u = U + \frac{1-r^2}{6}$:

$$\begin{cases} U_t = U_{rr} + \frac{2}{r}U_r, & 0 < r < 1, t > 0, \\ U(1, t) = 0, & t > 0, \\ U(r, 0) = \frac{r^2-1}{6}, & 0 \leq r \leq 1. \end{cases} \quad (4 \text{ marks})$$

Solution: Since

$$u_r = U_r - \frac{r}{3}, \quad u_{rr} = U_{rr} - \frac{1}{3}, \quad u_t = U_t \quad [1.5 \text{ marks}]$$

we have

$$U_t = U_{rr} - \frac{1}{3} + \frac{2}{r}U_r - \frac{2}{3} + 1 = U_{rr} + \frac{2}{r}U_r. \quad [0.5 \text{ mark}]$$

$$u(1, t) = U(1, t) + \frac{1-1^2}{6} \Rightarrow U(1, t) = 0. \quad [1 \text{ mark}]$$

$$u(r, 0) = U(r, 0) + \frac{r-r^2}{6} = 0 \Rightarrow \frac{r^2-1}{6}. \quad [1 \text{ mark}]$$

- (ii) Solve the equation in part (i) by letting $v(r, t) = rU(r, t)$ given that the Fourier sine series of $\frac{r^2-1}{6}$ for the interval $[0, 1]$ is $\sum_{n=1}^{\infty} \frac{2(-1)^n}{n^3\pi^3} \sin n\pi r$. (11 marks)

Solution: Note that

$$v = rU \Rightarrow v_t = rU_t, \quad v_r = U + rU_r, \quad v_{rr} = 2U_r + rU_{rr} \quad [2 \text{ marks}]$$

Therefore

$$\begin{cases} U_t = U_{rr} + \frac{2}{r}U_r \Rightarrow v_t = v_{rr} \\ U(1, t) = 0 \Rightarrow v(1, t) = 0 \\ |U(0, t)| < \infty \Rightarrow v(0, t) = 0 \times U(0, t) = 0 \\ U(r, 0) = \frac{r^2-1}{6} \Rightarrow v(r, 0) = \frac{r^3-r}{6} \end{cases} \quad [4 \text{ marks}]$$

From the above SL problem, it is easy to obtain the formal series solution

$$v(r, t) = \sum_{n=1}^{\infty} b_n e^{-(n\pi)^2 t} \sin n\pi r. \quad [3 \text{ marks}]$$

From the initial condition above, we can determine the values of b_n :

$$b_n = 2 \int_0^1 \frac{r^3-r}{6} \sin n\pi r dr = \frac{2(-1)^n}{(n\pi)^3}, \quad n = 1, 2, \dots$$

We have the formal series solution

$$v = \sum_{n=1}^{\infty} \frac{2(-1)^n}{(n\pi)^3} e^{-n^2\pi^2 t} \sin n\pi r.$$

- (iii) Show that $u(r, t) = \frac{1}{r} \left[\frac{r(1-r^2)}{6} + \frac{2}{\pi^3} \sum_{n=1}^{\infty} \frac{(-1)^n}{n^3} e^{-n^2\pi^2 t} \sin n\pi r \right]$. (2 marks)

Solution: $u(r, t) = U + \frac{1-r^2}{6} = \frac{1}{r}v + \frac{1-r^2}{6} = RHS. \quad [2 \text{ marks}]$

37.4 2-D and 3-D Heat Equations (on a Sphere with Non-Homogeneous Distribution)

To solve non-homogeneous distribution, we may need to deal with “2-D” heat equation (37.9) or “3-D” heat equation (37.9). Since 2-D is a special case of 3-D, we will limit ourselves to 3-D case, where

separation of variables $u = TU$ leads to a first order ODE of T and a 3-D Helmholtz equation. We will illustrate this with a homogeneous type 1 BCs problem.

Example 37.31. Consider the heat equation (37.9) on a solid spherical ball of $r = L$ with an initial condition (37.5) subject to homogeneous Dirichlet BCs

$$u(L, \phi, \theta, t) = 0, \quad 0 \leq \phi < 2\pi, \quad 0 \leq \theta \leq \pi, \quad t > 0.$$

Find the formal series solution for this problem.

Solution. By the separation of variables, we will obtain a ODE of T :

$$T' + k\lambda T = 0$$

and a Helmholtz equation

$$\Delta U = -\lambda U \Rightarrow \frac{R''}{R} + \frac{2}{r} \frac{R'}{R} + \frac{1}{r^2 \sin^2 \theta} \frac{\Phi''}{\Phi} + \frac{1}{r^2} \frac{\Theta''}{\Theta} + \frac{\cot \theta}{r^2} \frac{\Theta'}{\Theta} = -\lambda$$

By separating for Φ , we obtain

$$r^2 \sin^2 \theta \left(\frac{R''}{R} + \frac{2}{r} \frac{R'}{R} \right) + \sin^2 \theta \left(\frac{\Theta''}{\Theta} + \cot \theta \frac{\Theta'}{\Theta} \right) + \lambda r^2 \sin^2 \theta = -\frac{\Phi''}{\Phi} = \mu$$

and separating R and Θ , we obtain

$$r^2 \left(\frac{R''}{R} + \frac{2}{r} \frac{R'}{R} \right) + \frac{\Theta''}{\Theta} + \cot \theta \frac{\Theta'}{\Theta} + \lambda r^2 = \frac{\mu}{\sin^2 \theta} \Rightarrow r^2 \frac{R''}{R} + 2r \frac{R'}{R} + \lambda r^2 = \frac{\mu}{\sin^2 \theta} - \frac{\Theta''}{\Theta} - \cot \theta \frac{\Theta'}{\Theta} = \gamma.$$

With all the implicit and explicit boundary conditions, we obtain the following SL problems:

$$\begin{cases} \Phi'' + \mu \Phi = 0, & \Phi(0) = \Phi(2\pi), \quad \Phi'(0) = \Phi'(2\pi), \\ r^2 R'' + 2r R' + (\lambda r^2 - \gamma) R = 0, & |R(0)| < \infty, \quad R(L) = 0, \\ \sin^2 \theta \Theta'' + \cos \theta \sin \theta \Theta' + (\gamma \sin^2 \theta - \mu) \Theta = 0, & |\Theta(0)| < \infty, \quad |\Theta(\pi)| < \infty. \end{cases}$$

According to the result from Topic 1, the periodic SL problem of Φ is:

$$\mu = m^2, \quad \Phi_m(\phi) = A_m \cos m\phi + B_m \sin m\phi, \quad m = 0, 1, 2, \dots$$

The singular SL problem of Θ has the following eigenvalues and eigenfunctions using the results from Section 37.2.2:

$$\gamma = n(n+1), \quad \Theta_n^m(\theta) = P_n^m(\cos \theta), \quad n = m, m+1, \dots$$

where $P_n^m(x)$ is the associated Legendre function.

The singular SL problem of R (with $\gamma = n(n+1)$) has the following eigen-solutions according to Section 37.2.1:

$$\lambda = \left(\frac{\tau_k^{(n)}}{L} \right)^2, \quad R_k(r) = j_n \left(\frac{\tau_k^{(n)}}{L} r \right), \quad k = 1, 2, \dots;$$

where j_n is the n^{th} order spherical Bessel function defined in Section 37.2.1 and $\tau_k^{(n)}$ are the roots of $j_n(\tau) = 0$.

Re-assembling the individual eigenfunctions and applying the principle of superposition, we have the general solution

$$u(r, \phi, \theta, t) = \sum_{k=1}^{\infty} \sum_{m=0}^{\infty} \sum_{n=m}^{\infty} e^{-(\tau_k^{(n)}/L)^2 kt} j_n \left(\frac{\tau_k^{(n)}}{L} r \right) P_n^m(\cos \theta) (a_{kmn} \cos m\phi + b_{kmn} \sin m\phi).$$

From the initial conditions and by applying Theorem 37.23, we obtain

$$a_{knm} = \frac{\int_0^L \int_0^{2\pi} \int_0^\pi f(r, \phi, \theta) j_n \left(\frac{\tau_k^{(n)}}{L} r \right) P_n^m(\cos \theta) \cos m\phi r^2 \sin \theta d\theta d\phi dr}{\int_0^L \int_0^{2\pi} \int_0^\pi \left[j_n \left(\frac{\tau_k^{(n)}}{L} r \right) P_n^m(\cos \theta) \cos m\phi \right]^2 r^2 \sin \theta d\theta d\phi dr}$$

$$\begin{aligned}
&= \frac{\int_0^L \int_0^{2\pi} \int_0^\pi f(r, \phi, \theta) j_n\left(\frac{\tau_k^{(n)} r}{L}\right) P_n^m(\cos \theta) \cos m\phi r^2 \sin \theta d\theta d\phi dr}{\frac{L^3}{2} \left[j_{n+1}(\alpha_k^{(n+\frac{1}{2})})\right]^2 \cdot \frac{2}{2n+1} \frac{(n+m)!}{(n-m)!} \cdot \pi(1 + \delta_{m0})} \\
b_{knm} &= \frac{\int_0^L \int_0^{2\pi} \int_0^\pi f(r, \phi, \theta) j_n\left(\frac{\tau_k^{(n)} r}{L}\right) P_n^m(\cos \theta) \sin m\phi r^2 \sin \theta d\theta d\phi dr}{\int_0^L \int_0^{2\pi} \int_0^\pi \left[j_n\left(\frac{\tau_k^{(n)} r}{L}\right) \tilde{Y}_{n,m}(\phi, \theta)\right]^2 r^2 \sin \theta d\theta d\phi dr} \\
&= \frac{\int_0^L \int_0^{2\pi} \int_0^\pi f(r, \phi, \theta) j_n\left(\frac{\tau_k^{(n)} r}{L}\right) P_n^m(\cos \theta) \sin m\phi r^2 \sin \theta d\theta d\phi dr}{\frac{L^3}{2} \left[j_{n+1}(\alpha_k^{(n+\frac{1}{2})})\right]^2 \cdot \frac{2}{2n+1} \frac{(n+m)!}{(n-m)!} \cdot \pi}
\end{aligned}$$

where $\alpha_k^{(n+\frac{1}{2})}$ are the zeros of the Bessel function $J_{n+\frac{1}{2}}$. □

A nonhomogeneous example is given in Asmar [2005, 296].

37.5 2-D and 3-D Laplace Equations

In contrast to heat problem, where we limit ourselves to the heat conduction in a sphere, Laplace equation can be used to describe electrostatic field which is defined inside a sphere or outside a sphere [Jackson, 1999] or special fluid flows outside a sphere [Lamb, 1953].

The domain is a sphere and the problem we study is 3-D. However, when the boundary is characterised by electric charge distribution which is radial symmetric, it is possible for us to use two-variable function and the Laplace equation will need only two variables to characterise. In this case, we say the Laplace equation to be 2-D. If there is no symmetry, three variables are required and we say the Laplace equation to be 3-D. Limited examples are given in this section, more examples are found in Tang [2007, Section 6.4] and Morse and Feshbach [1953b, Chapter 10].

Note that for a Laplace equation, if the boundary is constant, then the solution is also a constant according to maximum principle.

37.5.1 BVP in a Sphere

Laplace equation in a sphere can be used to describe the steady-state temperature distribution in a solid sphere. Inside the sphere of radius L , the temperature u satisfies Laplace equation. If the surface of the sphere is held at temperature with symmetry, the physical problem will be a “2-D” problem.

We begin by solving 2-D examples.

Example 37.32 (UCCM3003 Final Exam, May 2010, Q4). The steady-state temperature in a sphere of radius $r = 5$ is governed by the equation

$$\frac{\partial^2 u}{\partial r^2} + \frac{2}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} + \frac{\cot \theta}{r^2} \frac{\partial u}{\partial \theta} = 0, \quad 0 < r < 5, \quad 0 < \theta < \pi.$$

- (a) Show that if $u(r, \theta) = R(r)\Theta(\theta)$ and the separation constant is λ^2 , then R and Θ satisfy ordinary differential equations of the form:

$$\begin{aligned}
r^2 R'' + 2r R' - \lambda^2 R &= 0, \\
\sin \theta \Theta'' + \cos \theta \Theta' + \lambda^2 \sin \theta \Theta &= 0.
\end{aligned}$$

The first of these equations is a Cauchy-Euler equation, while the second is related to Legendre's equation. (7 marks)

Proof: By separation of variables $u = R\Theta$, we have

$$\left[\frac{1}{r^2} \frac{d}{dr} (r^2 R') \right] \Theta + R \left[\frac{1}{r^2 \sin \theta} \frac{d}{d\theta} (\sin \theta \Theta') \right] = 0 \Rightarrow \frac{R''}{R} + \frac{2}{r} \frac{R'}{R} + \frac{\cot \theta}{r^2} \frac{\Theta'}{\Theta} + \frac{1}{r^2} \frac{\Theta''}{\Theta} = 0$$

Now separating the “variables”, we obtain

$$\frac{r^2 R''}{R} + \frac{2r R'}{R} = -\frac{\Theta''}{\Theta} - \cot \theta \frac{\Theta'}{\Theta} = \lambda \Rightarrow \begin{cases} r^2 R'' + 2r R' - \lambda R = 0, \\ \sin \theta \Theta'' + \cos \theta \Theta' + \lambda \sin \theta \Theta = 0. \end{cases}$$

- (b) Show that if a new independent variable is defined by $x = \cos \theta$, then the equation for Θ in part (a) becomes

$$(1 - x^2) \frac{d^2 \Theta}{dx^2} - 2x \frac{d\Theta}{dx} + \lambda^2 \Theta = 0, \quad -1 \leq x \leq 1.$$

Note that this is Legendre's equation.

(7 marks)

Proof: This was shown in Example [37.3](#) and we can copy the proof from there.

- (c) Find $u(r, \theta)$ is $u(5, \theta) = \cos \theta$, $0 \leq \theta \leq \pi$.

(11 marks)

You may assume the following facts:

- The only solutions of the equation in part (b) that are bounded for $-1 \leq x \leq 1$ are the Legendre polynomials $P_n(x) = P_n(\cos \theta)$. Furthermore, this boundedness condition can be satisfied if and only if $\lambda^2 = n(n+1)$ for $n = 0, 1, 2, \dots$.
- The roots of the auxiliary equation for the Cauchy-Euler equation in part (a) are n and $-(n+1)$.
- $P_1(\cos \theta) = \cos \theta$.

Solution. The facts are telling us that the general solution has the form

$$u(r, \theta) = \sum_{n=0}^{\infty} (a_n r^n + b_n r^{-n-1}) P_n(\cos \theta).$$

However, since $0 < r < 5$ and when $r \rightarrow 0$, $u(r, \theta)$ is required to be bounded, therefore $b_n = 0$. When $r = 5$, by orthogonality relation, $a_n = 0$ when $n \geq 2$ and so

$$\sum_{n=0}^{\infty} a_n 5^n P_n(\cos \theta) = a_0 + 5a_1 \cos \theta = \cos \theta, \quad 0 < \theta < \pi$$

implies $a_0 = 0$, $a_1 = \frac{1}{5}$ and the solution is $u(r, \theta) = \frac{r}{5} \cos \theta$. □

Example 37.33 (UCCM3003 Final Exam, Apr 2011, Q4(a)). Show that $u(r, \theta) = \frac{2}{5}rP_1(\cos \theta) - \frac{2}{5}r^3P_3(\cos \theta)$ is a solution to the boundary value problem in polar coordinates (**Remark:** this past year exam question set wrongly, it should be spherical coordinates. Note that $0 \leq \theta \leq \pi$ instead the one stated below):

$$\begin{cases} \nabla^2 u(r, \theta) = 0, & 0 < \theta < 2\pi, \quad r < 1, \\ u(1, \theta) = \cos \theta - \cos^3 \theta, & 0 < \theta < 2\pi. \end{cases} \quad (6 \text{ marks})$$

Solution. By looking up in the formula sheet for Legendre function, we verify that the boundary condition is satisfied:

$$\begin{aligned} u(1, \theta) &= \frac{2}{5}P_1(\cos \theta) - \frac{2}{5}P_3(\cos \theta) = \frac{2}{5} \left[\cos \theta - \frac{1}{2}(5 \cos^3 \theta - 3 \cos \theta) \right] \\ &= \frac{2}{5} \left[\frac{5}{2} \cos \theta - \frac{5}{2} \cos^3 \theta \right] = RHS. \end{aligned}$$

And now we verify that the Laplace equation is satisfied:

$$\begin{aligned} \frac{1}{r^2} \frac{\partial}{\partial r} \left[r^2 \frac{\partial u}{\partial r} \right] &= \frac{1}{r^2} \frac{\partial}{\partial r} \left[r^2 \left(\frac{2}{5}P_1(\cos \theta) - \frac{6}{5}r^2P_3(\cos \theta) \right) \right] = \frac{1}{r^2} \left[\frac{4r}{5}P_1(\cos \theta) - \frac{24r^3}{5}P_3(\cos \theta) \right] \\ &= \frac{4}{5r^2} \left[r \cos \theta - 6r^3 \left(\frac{5}{2} \cos^3 \theta - \frac{3}{2} \cos \theta \right) \right] = \frac{4}{5} \left[\frac{\cos \theta}{r} - r(15 \cos^3 \theta - 9 \cos \theta) \right] \\ \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left[\sin \theta \frac{\partial u}{\partial \theta} \right] &= \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left[\sin \theta \left(\frac{2}{5}r(-\sin \theta) - \frac{1}{5}r^3(15 \cos^2 \theta(-\sin \theta) + 3 \sin \theta) \right) \right] \\ &= \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left[-\frac{2r}{5} \sin^2 \theta - \frac{r^3}{5} (15 \cos^2 \theta(-\sin^2 \theta) + 3 \sin^2 \theta) \right] \\ &= \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left[-\frac{2r}{5} \sin^2 \theta - \frac{r^3}{5} (-12 \sin^2 \theta + 15 \sin^4 \theta) \right] \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{r^2 \sin \theta} \left[-\frac{4r}{5} \sin \theta \cos \theta - \frac{r^3}{5} \left(-24 \sin \theta \cos \theta + 60 \sin^3 \theta \cos \theta \right) \right] \\
&= -\frac{4}{5r} \cos \theta - \frac{4r}{5} \left(-6 \cos \theta + 15 \sin^2 \theta \cos \theta \right) \\
&= -\frac{5r}{4} \cos \theta - \frac{4r}{5} \left(9 \cos \theta - 15 \cos^3 \theta \right) \\
\text{Hence, } \frac{1}{r^2} \frac{\partial}{\partial r} \left[r^2 \frac{\partial u}{\partial r} \right] + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left[\sin \theta \frac{\partial u}{\partial \theta} \right] &= 0 \quad \square
\end{aligned}$$

Example 37.34 (UCCM3003 Final Exam, Apr 2012, Q4(b)). The interior Dirichlet problem

$$\begin{aligned}
\frac{\partial^2 u}{\partial r^2} + \frac{2}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} + \frac{\cot \theta}{r^2} \frac{\partial u}{\partial \theta} &= 0, \quad 0 < r < 1, \quad 0 < \theta < \pi, \\
u(1, \theta) &= f(\theta), \quad 0 < \theta < \pi
\end{aligned}$$

has a solution of the form

$$u(r, \theta) = \sum_{n=0}^{\infty} A_n r^n P_n(\cos \theta).$$

Solve for $u(r, \theta)$ if $f(\theta) = 2 - 2 \cos^2 \theta$. (9 marks)

Solution. Let $x = \cos \theta$. When $r = 1$,

$$\sum_{n=0}^{\infty} A_n P_n(\cos \theta) = 2 - 2 \cos^2 \theta \Rightarrow \sum_{n=0}^{\infty} A_n P_n(x) = 2 - 2x^2$$

Since f is even over $-1 < x < 1$ and P_{2n+1} are odd functions, $A_1 = 0$. Since the degree of the polynomial is 2, $A_n = 0$, $n \geq 3$. We need to calculate

$$\begin{aligned}
A_0 &= \frac{1}{2} \int_{-1}^1 (2 - 2x^2) dx = \frac{1}{2} \times \frac{8}{3} = \frac{4}{3} \\
A_2 &= \frac{5}{2} \int_{-1}^1 (2 - 2x^2) \frac{3x^2 - 1}{2} dx = \frac{5}{2} \times \left(-\frac{8}{15} \right) = -\frac{4}{3}.
\end{aligned}$$

Therefore, $u(r, \theta) = \frac{4}{3} - \frac{4}{3} r^2 P_2(\cos \theta)$. □

Example 37.35 (UCCM3003 Final Exam, Sep 2009, Q5(a)). Consider the radial symmetric Laplace equation defined inside a sphere of radius L .

$$\frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial u}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial u}{\partial \theta} \right) = 0, \quad 0 < r < L, \quad 0 < \theta < \pi.$$

Given that $u(L, \theta) = 2 \cos^2 \theta$, $0 < \theta < \pi$, and $u(r, \theta)$ is finite in $0 < r < L$, find $u(r, \theta)$.

(14 marks)

Solution. By separation of variables $u = R\Theta$, we have

$$\left[\frac{1}{r^2} \frac{d}{dr} (r^2 R') \right] \Theta + R \left[\frac{1}{r^2 \sin \theta} \frac{d}{d\theta} (\sin \theta \Theta') \right] = 0 \Rightarrow \frac{R''}{R} + \frac{2}{r} \frac{R'}{R} + \frac{\cot \theta}{r^2} \frac{\Theta'}{\Theta} + \frac{1}{r^2} \frac{\Theta''}{\Theta} = 0$$

Now separating the “variables” and together with regularity condition

$$\frac{r^2 R''}{R} + \frac{2rR'}{R} = -\frac{\Theta''}{\Theta} - \cot \theta \frac{\Theta'}{\Theta} = \mu \Rightarrow \begin{cases} r^2 R'' + 2rR' - \mu R = 0, & |R(0)| < \infty, \\ \Theta'' + \cot \theta \Theta' + \mu \Theta = 0, & |\Theta(0)| < \infty, |\Theta(\pi)| < \infty. \end{cases}$$

The singular SL problem of Θ is related to the Legendre equation with regular eigenfunctions

$$\mu = n(n+1), \quad \Theta_n = A_n P_n(\cos \theta), \quad n = 0, 1, 2, \dots$$

where P_n are Legendre polynomials. From here, we can solve the radial equation with a general solution

$$R_n = Br^n + Cr^{-n-1}, \quad |R(0)| < \infty \Rightarrow C = 0.$$

By the principle of superposition, $u = \sum_{n=0}^{\infty} a_n r^n P_n(\cos \theta)$.

From the boundary condition and by using (37.22) le:legendre:6

$$\begin{aligned}\sum_{n=0}^{\infty} a_n L^n P_n(\cos \theta) &= 2 \cos^2 \theta a_0 + a_1 L P_1(\cos \theta) + a_2 L^2 P_2(\cos \theta) \\ &= a_0 + a_1 L \cos \theta + a_2 L^2 \frac{3 \cos^2 \theta - 1}{2} \\ &= a_0 - \frac{a_2 L^2}{2} + a_1 L \cos \theta + \frac{a_2 3 L^2 \cos^2 \theta}{2} = 2 \cos^2 \theta \\ \Rightarrow a_1 &= 0, \quad a_2 = \frac{4}{3 L^2}, \quad a_0 = \frac{2}{3}.\end{aligned}$$

The solution is $u = \frac{2}{3} + \frac{4}{3} \left(\frac{r}{L}\right)^2 \frac{3 \cos^2 \theta - 1}{2}$. □

Example 37.36. [Asmar, 2005, 277] Find the steady-state temperature in a sphere of radius L , given that the upper hemisphere is kept at 100° and the lower one is kept at 0° .

Solution: This problem is described by 2-D Laplace equation since the temperature distribution on the boundary is symmetrical. The general solution is $u = \sum_{\ell=0}^{\infty} c_\ell r^\ell P_\ell(\cos \theta)$.

When the surface of the upper half is held at 100° and the surface of the lower half at 0° , the coefficients c_n can be determined when $r = L$, i.e.

$$u|_{r=L} = \sum_{n=0}^{\infty} c_n L^n P_n(\cos \theta) = \begin{cases} 100, & 0 < \theta < \frac{\pi}{2} \text{ (i.e. } 0 < \cos \theta < 1), \\ 0, & \frac{\pi}{2} < \theta < \pi \text{ (i.e. } -1 < \cos \theta < 0). \end{cases} \quad (37.31)$$

e:7.12

Let $x = \cos \theta$, $u|_{r=L} = \sum_{n=0}^{\infty} c_n L^n P_n(x) = \begin{cases} 100, & 0 < x < 1, \\ 0, & -1 < x < 0. \end{cases}$

When $n = 0$, $c_0 = \frac{1}{2} \int_0^1 100 P_0(x) dx = 50$.

When $n \geq 1$, by using Theorems 37.10, 37.4 and 37.5, we have thm:legendre:thm:legendre:special

$$\begin{aligned}L^n c_n &= \frac{2n+1}{2} \int_0^1 100 P_n(x) dx = 50(2n+1) \int_0^1 \frac{P'_{n+1}(x) - P'_{n-1}(x)}{2n+1} dx \\ &= 50 [P_{n+1}(1) - P_{n+1}(0) - (P_{n-1}(1) - P_{n-1}(0))] \\ &= \begin{cases} 0, & n \text{ is even} \\ (1 - (-1)^{(n+1)/2} \frac{1 \cdot 3 \cdot 5 \cdots n}{2 \cdot 4 \cdot 6 \cdots (n+1)}) - (1 - (-1)^{(n-1)/2} \frac{1 \cdot 3 \cdot 5 \cdots (n-2)}{2 \cdot 4 \cdot 6 \cdots (n-1)}), & n \text{ is odd} \end{cases}\end{aligned}$$

Therefore, $u(r, \theta) = 50 + 25 \sum_{n=0}^{\infty} \frac{(4n+3)(-1)^n (2n)!}{2^{2n} (n!)^2 (n+1)} \left(\frac{r}{L}\right)^{2n+1} P_{2n+1}(\cos \theta)$.

If the surface of the sphere is held at temperature $f(\phi, \theta)$, i.e. without symmetry, the physical problem will be a “3-D” Dirichlet problem:

$$\begin{cases} \frac{\partial^2 u}{\partial r^2} + \frac{2}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 u}{\partial \phi^2} + \frac{\cot \theta}{r^2} \frac{\partial u}{\partial \theta} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} = 0, & 0 < r < L, \quad 0 \leq \phi < 2\pi, \quad 0 \leq \theta \leq \pi, \\ u(L, \phi, \theta) = f(\phi, \theta), & 0 \leq \phi < 2\pi, \quad 0 \leq \theta \leq \pi. \end{cases} \quad (37.32)$$

e:dirichlet:in

Note that ϕ is periodic, i.e. $u(r, \phi, \theta) = u(r, \phi + 2\pi, \theta)$ and is implicitly assumed.

Example 37.37. Use the separation of variables $u = R(r)\Phi(\phi)\Theta(\theta)$ to find a formal series solution for the Dirichlet problem (37.32). e:dirichlet:insphere

Solution. If $u = R\Phi\Theta$, then after multiplying by $\frac{r^2}{R\Phi\Theta}$ (37.32) becomes e:dirichlet:insphere

$$\frac{r^2 R'' + 2r R'}{R} + \frac{\Phi''}{\Phi \sin^2 \theta} + \frac{\Theta''}{\Theta} + \cot \theta \frac{\Theta'}{\Theta} = 0. \quad (37.33)$$

e:7.3

Multiplying (37.33) by $\sin^2 \theta$, the middle term becomes a function of ϕ only and the other terms do not contain ϕ . Thus we obtain the periodic SL problem of Φ :

$$\Phi'' + \mu \Phi = 0, \quad \Phi(0) = \Phi(2\pi), \quad \Phi'(0) = \Phi'(2\pi)$$

and its eigen-solutions are

$$\mu = m^2, \quad \Phi = A_m \sin m\phi + B_m \cos m\phi, \quad m = 0, 1, 2, \dots$$

After separating Φ -term, equation (37.33) can now be written as

$$\begin{aligned} -\frac{\Phi''}{\Phi} &= \frac{r^2 R'' + 2rR'}{R} \sin^2 \theta + \frac{\Theta''}{\Theta} \sin^2 \theta + \sin \theta \cos \theta \frac{\Theta'}{\Theta} = \mu = m^2 \\ \Rightarrow \frac{\Theta''}{\Theta} + \cot \theta \frac{\Theta'}{\Theta} - \frac{m^2}{\sin^2 \theta} &= \frac{r^2 R'' + 2rR'}{R} = \gamma \end{aligned} \quad (37.34)$$

and this simplifies to

$$\Theta'' + \cot \theta \Theta' + \frac{m^2}{\sin^2 \theta} \Theta + \mu \Theta = 0 \Rightarrow \sin^2 \theta \Theta'' + \sin \theta \cos \theta \Theta' + (\gamma \sin^2 \theta - m^2) \Theta = 0, \quad (37.35)$$

$$r^2 R'' + 2rR' - \gamma R = 0. \quad (37.36)$$

Equation (37.35) can be transformed into an m -th order associated Legendre equation (37.19) with $x = \cos \theta$. In Section 37.2.2, we learned that this is a singular SL problem and the eigenvalues and eigenfunctions are

$$\gamma = \ell(\ell + 1), \quad \Theta = C_{m\ell} P_\ell^m(\cos \theta), \quad 0 \leq \ell \leq m.$$

Putting $\gamma = \ell(\ell + 1)$ into (37.36), we obtain an Euler-type equation (Section 37.2.1) and its solution are

$$R = D_\ell r^\ell + E_\ell r^{-\ell-1}.$$

Since we are interested in the interior of the sphere, the regularity condition requires us to discard the $r^{-\ell-1}$ term because they become infinite at the origin.

By the principle of superposition, the general solution is

$$u = \sum_{m=0}^{\infty} \sum_{\ell=m}^{\infty} r^\ell P_\ell^m(\cos \theta) (a_{\ell m} \sin m\phi + b_{\ell m} \cos m\phi), \quad (37.37)$$

where $a_{\ell m}$ and $b_{\ell m}$ are multiplication of constants.

From the initial condition, the surface temperature at $r = L$ were given by a function $f(\phi, \theta)$:

$$\sum_{m=0}^{\infty} \sum_{\ell=m}^{\infty} L^\ell P_\ell^m(\cos \theta) (a_{\ell m} \sin m\phi + b_{\ell m} \cos m\phi) = f(\phi, \theta).$$

$$\begin{aligned} \text{Hence, } a_{\ell m} &= \frac{\int_0^{2\pi} \int_0^\pi f(\phi, \theta) \cos(m\phi) P_\ell^m(\cos \theta) \sin \theta d\theta d\phi}{\int_0^{2\pi} \int_0^\pi (\cos^2 m\phi) (P_\ell^m(\cos \theta))^2 \sin \theta d\theta d\phi} \\ &= \frac{(2\ell + 1)(\ell - m)!}{2\pi(1 + \delta_{m0})(\ell + m)!} \int_0^{2\pi} \int_0^\pi f(\phi, \theta) \cos(m\phi) P_\ell^m(\cos \theta) \sin \theta d\theta d\phi, \\ b_{\ell m} &= \frac{\int_0^{2\pi} \int_0^\pi f(\phi, \theta) \sin(m\phi) P_\ell^m(\cos \theta) \sin \theta d\theta d\phi}{\int_0^{2\pi} \int_0^\pi (\sin^2 m\phi) (P_\ell^m(\cos \theta))^2 \sin \theta d\theta d\phi} \\ &= \frac{(2\ell + 1)(\ell - m)!}{2\pi(\ell + m)!} \int_0^{2\pi} \int_0^\pi f(\phi, \theta) \sin(m\phi) P_\ell^m(\cos \theta) \sin \theta d\theta d\phi. \end{aligned}$$

□

37.5.2 BVP exterior to a Sphere

The analysis of the Laplace equation exterior to a sphere is very similar to the situation internal to a sphere, the only difference is that we require $R(r) < \infty$ for all $r > L$. Hence $D = 0$ and the general solution is [Koshlyakov et al., 1964, 381]

$$u(r, \theta, \phi) = \sum_{m=0}^{\infty} \sum_{n=m}^{\infty} \frac{1}{r^{n+1}} P_n^m(\cos \theta) (a_{nm} \sin m\phi + b_{nm} \cos m\phi),$$

where a_{nm} and b_{nm} are constants which depend on the boundary condition.

Example 37.38 (UCCM3003 Final Exam, Sep 2008, Q3(a)). Consider the Laplace equation exterior to

a sphere of radius 13 below:

$$\begin{cases} \frac{\partial^2 u}{\partial r^2} + \frac{2}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} + \frac{\cot \theta}{r^2} \frac{\partial u}{\partial \theta} = 0, & r > 13, \quad 0 \leq \theta \leq \pi, \\ u(13, \theta) = \sin \theta, & 0 \leq \theta \leq \pi. \end{cases}$$

Show that the general solution is

$$u(r, \theta) = \sum_{n=0}^{\infty} b_n r^{-n-1} P_n(\cos \theta). \quad (13 \text{ marks})$$

Solution: Let $u = R\Theta$. From the PDE,

$$\frac{R''}{R} + \frac{2}{r} \frac{R'}{R} + \frac{1}{r^2} \frac{\Theta''}{\Theta} + \frac{\cot \theta}{r^2} \frac{\Theta'}{\Theta} = 0 \quad [2 \text{ marks}]$$

Separating variables leads to

$$\frac{r^2 R''}{R} + \frac{2r R'}{R} = -\frac{\Theta''}{\Theta} - \cot \theta \frac{\Theta'}{\Theta} = \gamma \quad [2 \text{ marks}]$$

which can be written as two ODEs:

$$\begin{cases} r^2 R'' + 2r R' - \gamma R = 0, \\ \Theta'' + \cot \theta \Theta' + \gamma \Theta = 0. \end{cases} \quad [2 \text{ marks}]$$

Transforming the ODE of Θ to Legendre equation (by letting $x = \cos \theta$) and we can solve for the eigenvalues and eigenfunctions (from Formula sheet):

$$\gamma = n(n+1), \quad \Theta_n(\theta) = C_n P_n(\cos \theta), \quad n = 0, 1, 2, \dots \quad [1+1+1=3 \text{ marks}]$$

For $\gamma = n(n+1)$, the ODE of R is an Euler equation with the following general solution (from Formula sheet)

$$R(r) = Ar^n + Br^{-n-1}. \quad [1 \text{ mark}]$$

Since u is required to be regular as $r \rightarrow \infty$, hence $A = 0$ and the regular solution for R is

$$R_n(r) = B_n r^{-n-1}. \quad [1 \text{ mark}]$$

By using the superposition principle, we obtain the general solution

$$u(r, \theta) = \sum_{n=0}^{\infty} R_n(r) \Theta_n(\theta) = \sum_{n=0}^{\infty} b_n r^{-n-1} P_n(\cos \theta) \quad [2 \text{ marks}]$$

where $b_n = B_n C_n$, $n = 0, 1, \dots$ are constants.

nal:200904:q5b

Example 37.39 (UCCM3003 Final Exam, Apr 2009, Q5(b)). The steady state temperature distribution exterior to a sphere with radial symmetry is given by the following Laplace equation and boundary conditions

$$\begin{cases} \nabla^2 u(r, \theta) = \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial u}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial u}{\partial \theta} \right) = 0, & r > a, \quad 0 \leq \theta \leq \pi, \\ u(a, \theta) = 2 \cos \theta + \cos^2 \theta, & 0 \leq \theta \leq \pi, \\ \lim_{r \rightarrow \infty} u(r, \theta) = 0, & 0 \leq \theta \leq \pi. \end{cases}$$

(i) Show that a general solution to the steady state temperature distribution is

$$u(r, \theta) = \sum_{n=0}^{\infty} \frac{A_n}{r^{n+1}} P_n(\cos \theta). \quad (12 \text{ marks})$$

Solution. Let $u = R\Theta$. From the PDE,

$$\frac{1}{r^2} \frac{d}{dr} (r^2 R) \Theta + \frac{1}{r^2 \sin \theta} \frac{d}{d\theta} (\sin \theta \frac{d\Theta}{d\theta}) R = 0$$

which expands to

$$\left[R'' + \frac{2}{r} R' \right] \Theta + \left[\frac{1}{r^2} \Theta'' + \frac{\cos \theta}{r^2 \sin \theta} \Theta' \right] R = 0.$$

Separating variables leads to

$$\frac{r^2 R''}{R} + \frac{2r R'}{R} = -\frac{\Theta''}{\Theta} - \cot \theta \frac{\Theta'}{\Theta} = \gamma$$

which can be written as an ODE and a singular SL problem:

$$\begin{cases} r^2 R'' + 2r R' - \gamma R = 0, \\ \Theta'' + \cot \theta \Theta' + \gamma \Theta = 0, \quad |\Theta(0)| < \infty, \quad |\Theta(\pi)| < \infty. \end{cases}$$

Transforming the ODE of Θ to Legendre equation (by letting $x = \cos \theta$) and we can solve for the eigenvalues and eigenfunctions (from Formula sheet):

$$\gamma = n(n+1), \quad \Theta_n(\theta) = C_n P_n(\cos \theta), \quad n = 0, 1, 2, \dots$$

For $\gamma = n(n+1)$, the ODE of R is an Euler equation with the following general solution (from Formula sheet)

$$R(r) = Ar^n + Br^{-n-1}.$$

Since $u \rightarrow 0$ as $r \rightarrow \infty$, hence $A = 0$ and so

$$R_n(r) = B_n r^{-n-1}.$$

By using the superposition principle, we obtain the general solution

$$u(r, \theta) = \sum_{n=0}^{\infty} R_n(r) \Theta_n(\theta) = \sum_{n=0}^{\infty} b_n r^{-n-1} P_n(\cos \theta)$$

where $b_n = B_n C_n$, $n = 0, 1, \dots$ are constants. □

(ii) Find the steady-state temperature distribution $u(r, \theta)$.

(8 marks).

Solution. From the boundary condition

$$u(a, \theta) = \sum_{n=0}^{\infty} b_n a^{-n-1} P_n(\cos \theta) = 2 \cos \theta + \cos^2 \theta$$

we obtain (by using the Fourier-Legendre series (37.26)):

$$\begin{aligned} \frac{b_0}{a} &= \frac{1}{2} \int_{-1}^1 [2x + x^2] P_0(x) dx = \frac{1}{2} \times \frac{2}{3} \Rightarrow b_0 = \frac{1}{3} a \\ \frac{b_1}{a^2} &= \frac{3}{2} \int_{-1}^1 [2x + x^2] P_1(x) dx = \frac{3}{2} \times \frac{4}{3} \Rightarrow b_1 = 2a^2 \\ \frac{b_2}{a^3} &= \frac{5}{2} \int_{-1}^1 [2x + x^2] P_2(x) dx = \frac{5}{2} \times \frac{4}{15} \Rightarrow b_2 = \frac{2}{3} a^3 \\ \frac{b_n}{a^{n+1}} &= 0 \Rightarrow b_n = 0, \quad n \geq 3. \end{aligned}$$

The steady-state temperature distribution is

$$u(r, \theta) = \frac{a}{3r} + \frac{2a^2}{r^2} \cos \theta + \frac{2a^3}{3r^3} \left[\frac{3 \cos^2 \theta - 1}{2} \right].$$

□

Example 37.40 (UECM3003 Final Exam, Apr 2013, Q4(a)). You are given that the substitution $u(r, \theta) = R(r)\Theta(\theta)$ in the exterior Dirichlet problem of (37.10), $r > 1$, $0 < \theta < \pi$, leads to the equations

$$r^2 R'' + 2r R' - \lambda R = 0 \tag{37.38}$$

$$\sin \theta \Theta'' + \cos \theta \Theta' + \lambda \sin \theta \Theta = 0 \tag{37.39}$$

- (i) If the eigenvalues and eigenfunctions of (37.39) are

$$\lambda = \lambda_n = n(n+1), \quad \Theta = \Theta_n = P_n(\cos \theta), \quad n = 0, 1, 2, \dots,$$

show that the corresponding eigenfunctions (with $\lambda = n(n+1)$) of (37.38) are

$$R_n = r^{-(n+1)}. \quad (10 \text{ marks})$$

Solution. When $\lambda = n(n+1)$, $n = 0, 1, 2, \dots$, we obtain an Euler equation

$$r^2 R'' + 2rR' - n(n+1)R = 0 \Rightarrow \omega(\omega-1) + 2\omega - n(n+1) = 0 \Rightarrow \omega^2 + \omega - n(n+1) = 0$$

After factorisation, we obtain the general solution:

$$(\omega - n)(\omega + n + 1) = 0 \Rightarrow \omega = n, -(n+1) \Rightarrow R = Ar^n + Br^{-n-1}.$$

In order for R to be finite when $r \rightarrow \infty$, $A = 0$. Therefore, $R = Br^{-n-1}$ is the well-defined solution for this problem. $\square$

- (ii) Hence, write down a solution of the exterior Dirichlet problem if

$$u(1, \theta) = 3 + 4 \cos \theta, \quad 0 < \theta < \pi.$$

[Hint: $P_0(\cos \theta) = 1$, $P_1(\cos \theta) = \cos \theta$.]

(8 marks)

Solution.

$$u(1, \theta) = \sum_{n=0}^{\infty} B_n P_n(\cos \theta) = 3 + 4 \cos \theta, \quad 0 < \theta < \pi.$$

Let $x = \cos \theta$, then

$$\sum_{n=0}^{\infty} B_n P_n(x) = B_0 + B_1 x = 3 + 4x, \quad -1 < x < 1.$$

This implies $B_0 = 3$, $B_1 = 4$, $B_n = 0$, $n \geq 2$. Therefore, the solution $u(r, \theta) = \frac{3}{r} + \frac{4 \cos \theta}{r^2}$. $\square$

Example 37.41 (UCCM3003 Final Exam, Apr 2011, Q3(c)). By using the result in Example 37.16, solve the following boundary value problem:

$$\begin{cases} \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial u}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial u}{\partial \theta} \right) = 0, & r > 1, \quad 0 \leq \theta \leq \pi, \\ u(1, \theta) = 2 \cos \theta - \sin^2 \theta, & 0 \leq \theta \leq \pi, \\ \lim_{r \rightarrow \infty} u(r, \theta) = 0, & 0 \leq \theta \leq \pi. \end{cases}$$

You could assume without proof that the differential equation

$$\Theta''(\theta) + \cot(\theta)\Theta'(\theta) + \lambda\Theta(\theta) = 0, \quad 0 \leq \theta \leq \pi$$

could be transformed to

$$(1 - s^2)\Theta''(s) - 2s\Theta'(s) + \lambda\Theta(s) = 0, \quad -1 \leq s \leq 1$$

where $s = \cos \theta$.

(14 marks)

Solution. From Example 37.39, the general solution is found to be

$$u(r, \theta) = \sum_{n=0}^{\infty} b_n r^{-n-1} P_n(\cos \theta).$$

From the boundary condition

$$u(1, \theta) = \sum_{n=0}^{\infty} b_n P_n(\cos \theta) = 2 \cos \theta - \sin^2 \theta = 2 \cos \theta - \frac{1 - \cos 2\theta}{2} \Rightarrow \sum_{n=0}^{\infty} b_n P_n(s) = -\frac{1}{2} + 2s + \frac{1}{2}s^2$$

By using the Fourier-Legendre series ([le:series:legendre](#) (37.26)), we have

$$\begin{aligned} b_0 &= \frac{1}{2} \int_{-1}^1 \left[-\frac{1}{2} + 2s + \frac{1}{2}s^2 \right] P_0(s) ds = \frac{1}{2} \times \left(-\frac{2}{3} \right) = -\frac{1}{3} \\ b_1 &= \frac{3}{2} \int_{-1}^1 \left[-\frac{1}{2} + 2s + \frac{1}{2}s^2 \right] P_1(s) ds = \frac{3}{2} \times \frac{4}{3} = 2 \\ b_2 &= \frac{5}{2} \int_{-1}^1 \left[-\frac{1}{2} + 2s + \frac{1}{2}s^2 \right] P_2(s) ds = \frac{5}{2} \times \frac{2}{15} = \frac{1}{3} \\ b_n &= 0, \quad n \geq 3. \end{aligned}$$

The solution is $u(r, \theta) = -\frac{1}{3r} + \frac{2}{r^2} \cos \theta + \frac{1}{3r^3} \left[\frac{3 \cos^2 \theta - 1}{2} \right]$. □

Example 37.42 (UECM3003 Final Exam, Apr 2015, Q3(c)). The exterior Dirichlet problem

$$\begin{cases} \frac{\partial^2 u}{\partial r^2} + \frac{2}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} + \frac{\cot \theta}{r^2} \frac{\partial u}{\partial \theta} = 0, & r > 1, \quad 0 < \theta < \pi, \\ u(1, \theta) = f(\theta), & 0 < \theta < \pi \end{cases}$$

has a solution of a form $u(r, \theta) = \sum_{n=0}^{\infty} A_n r^{-n-1} P_n(\cos \theta)$, where P_n is the Legendre polynomial. Solve for $u(r, \theta)$ if $f(\theta) = 1 + 2 \cos(\theta) - \sin^2(\theta)$. (5 marks)

Solution. When $r = 1$ and let $x = \cos \theta$

$$\sum_{n=0}^{\infty} A_n P_n(\cos \theta) = 1 + 2 \cos(\theta) - \sin^2(\theta) = 2 \cos \theta + \cos^2 \theta \Rightarrow \sum_{n=0}^{\infty} A_n P_n(x) = 2x + x^2$$

From which we obtain $A_n = 0, n \geq 3$ and

$$\begin{aligned} A_0 &= \frac{1}{2} \int_{-1}^1 2x + x^2 dx = \frac{1}{2} \times \frac{2}{3} = \frac{1}{3} \\ A_1 &= \frac{3}{2} \int_{-1}^1 (2x + x^2)x dx = \frac{3}{2} \times \frac{4}{3} = 2 \\ A_2 &= \frac{5}{2} \int_{-1}^1 (2x + x^2) \frac{3x^2 - 1}{2} dx = \frac{5}{2} \times \frac{4}{15} = \frac{2}{3}. \end{aligned}$$

Therefore, $u = \frac{1}{3r} + \frac{2}{r^2} P_1(\cos \theta) + \frac{2}{3r^3} P_2(\cos \theta)$. □

37.5.3 BVP on a Hemisphere

Another domain which may sometimes be of interest to scientists is the hemisphere. The following example studies Laplace equation on a hemisphere.

Example 37.43 (UECM3003 Final Exam, Apr 2016, Q3(b)). The base $r < 1, \theta = \pi/2$ of a solid hemisphere $r \leq 1, 0 \leq \theta \leq \pi/2$ is insulated; and the flux of heat inward through the hemispherical surface is kept at prescribed values $f(\theta)$. The boundary value problem for steady temperatures $u(r, \theta)$ in the hemisphere consists of Laplace equation

$$\frac{1}{r^2} \frac{\partial}{\partial r} (r^2 \frac{\partial u}{\partial r}) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} (\sin \theta \frac{\partial u}{\partial \theta}) = 0, \quad r \leq 1, \quad 0 \leq \theta \leq \frac{\pi}{2}.$$

The condition of insulation at the base and flux condition are given by the following equations:

$$\begin{cases} \frac{\partial u}{\partial \theta}(r, \frac{\pi}{2}) = 0, & 0 < r < 1 \\ C \frac{\partial u}{\partial r}(1, \theta) = f(\theta), & 0 < \theta < \frac{\pi}{2} \end{cases}$$

where C is thermal conductivity. In order that temperatures be steady, we assume that the values $f(\theta)$ are such that the resultant rate of flow through the hemispherical surface is zero, i.e. $\int_0^{\pi/2} f(\theta) \sin \theta d\theta = 0$. Assume that f is piecewise smooth on $0 < \theta < \pi/2$ and that u satisfies the usual continuity conditions when $0 \leq r < 1$ and $0 < \theta < \pi/2$. Find the formal series solution u for the given problem.

Solution: Let $u = R\Theta$. The PDE becomes

$$\frac{2}{r} \frac{\partial u}{\partial r} + \frac{\partial^2 u}{\partial r^2} + \frac{\cos \theta}{r^2 \sin \theta} \frac{\partial u}{\partial \theta} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} = \frac{2}{r} R' \Theta + R'' \Theta + \frac{\cot \theta}{r^2} R \Theta' + \frac{1}{r^2} R \Theta'' = 0. \quad [1 \text{ mark}]$$

By dividing with $R\Theta$, we have

$$\frac{2}{r} \frac{R'}{R} + \frac{R''}{R} + \frac{\cot \theta}{r^2} \frac{\Theta'}{\Theta} + \frac{1}{r^2} \frac{\Theta''}{\Theta} = 0. \quad [1 \text{ mark}]$$

By multiplying with r^2 , it is possible for us to separate functions of different variables:

$$r^2 \frac{R''}{R} + 2r \frac{R'}{R} = -\frac{\Theta''}{\Theta} - \cot \theta \frac{\Theta'}{\Theta} = \lambda \quad [1 \text{ mark}]$$

Together with the homogeneous boundary condition, we can obtain two ODEs below

$$\begin{cases} r^2 R'' + 2r R' + \lambda R = 0, \\ \Theta'' + \cot \theta \Theta' + \lambda \Theta = 0, \quad \Theta'(\frac{\pi}{2}) = 0. \end{cases} \quad [1 \text{ mark}]$$

The Sturm-Liouville problem of Θ can be transformed to a Legendre equation by a change of variable $x = \cos \theta$. Let P_n be the Legendre polynomial, the general solution is

$$\Theta = A_n P_n(\cos \theta) \quad [1 \text{ mark}]$$

but

$$\Theta'(\frac{\pi}{2}) = A_n P'_n(\cos \frac{\pi}{2})(-\sin \frac{\pi}{2}) = 0 \Rightarrow P'_n(0) = 0 \Rightarrow n \text{ is even.} \quad [1 \text{ mark}]$$

Therefore, the eigenvalues and corresponding eigen-functions of the Sturm-Liouville problem are

$$\lambda = 2k(2k+1), \quad \Theta = A_{2k} P_{2k}(\cos \theta), \quad k = 0, 1, 2, \dots \quad [1 \text{ mark}]$$

For $\lambda = 2k(2k+1)$, the Euler equation for R has the solution $R = C_k r^{2k} + D_k r^{-2k-1}$. But the regularity condition requires $D_k = 0$ (otherwise, we will have a singularity at $r = 0$), therefore,

$$R = C_k r^{2k}, \quad 0 < r < 1. \quad [1 \text{ mark}]$$

The formal series representation of the general solution is by the principle of superposition:

$$u = \sum_{k=0}^{\infty} a_k r^{2k} P_{2k}(\cos \theta). \quad [1 \text{ mark}]$$

Now, by the orthogonality of Legendre polynomials

$$\frac{\partial u}{\partial r}(1, \theta) = \sum_{k=0}^{\infty} 2k a_k P_{2k}(\cos \theta) = \frac{1}{C} f(\theta) \Rightarrow a_k = \frac{\int_0^{\pi/2} f(\theta) P_{2k}(\cos \theta) \sin \theta d\theta}{2kC \int_0^{\pi/2} [P_{2k}(\cos \theta)]^2 \sin \theta d\theta} \quad [1 \text{ mark}]$$

Remark 37.44. The constant a_0 of the formal series solution in above example is the temperature at the origin $r = 0$. Solutions of such problems with just type 2 conditions are determined only up to such an arbitrary additive constant because all the boundary conditions prescribe only values of derivatives of the harmonic functions.

We are not including any example of Poisson equation in spherical coordinate, an example can be found in Riley et al. [2006, 744].

37.6 1-D Wave Equation

There are many kinds of (uniform) wave that can be discovered from wave equation. One-dimensional waves are called *plane waves*; two-dimensional waves are called *cylindrical waves*; three-dimensional waves are called *spherical waves*. We have seen the first two types of waves discussed in topics 2 and 3. In this section we will discuss the “spherical waves” [Koshlyakov et al., 1964, Chapter VI], [Morse and Feshbach, 1953b, Chapter 11], [Koshlyakov et al., 1964, p. 202].

Let us assume that the region of initial oscillations is a sphere of radius L , from which oscillations are propagated uniformly in all directions. If we place the coordinate origin at the centre of this sphere, and denote r as the distance from the origin to the variable point (r, ϕ, θ) . Then, the initial conditions of the oscillations will be of the form (37.1) and (37.2). The PDE that characterise this is (37.12).

Example 37.45. Use the separation of variables to solve the BVP with PDE (37.12), initial conditions (37.1), (37.2) and a type 1 boundary condition:

$$u(L, t) = 0, \quad t > 0.$$

Solution. Let $u = TR$. From the PDE,

$$\frac{T''}{c^2 T} = \frac{R''}{R} + \frac{2}{r} \frac{R'}{R} = -\lambda$$

where $\lambda > 0$ is a constant.

Together with the boundary conditions, we have

$$\begin{cases} r^2 R'' + 2rR' + \lambda r^2 R = 0, & R(L) = 0, \quad |R(0)| < \infty \\ T'' + c^2 \lambda T = 0 \end{cases}$$

The general solution to the ODE of R is $R(r) = A j_0(\sqrt{\lambda} r) + B y_n(\sqrt{\lambda} r)$.

Since $|R(0)| < \infty$, $B = 0$, otherwise we have singularity at $r = 0$ due to y_n . To have non-trivial solution, $A \neq 0$, therefore

$$R(L) = 0 \Rightarrow j_0(\sqrt{\lambda} L) = \frac{\sin(\sqrt{\lambda} L)}{\sqrt{\lambda} L} = 0 \Rightarrow \lambda = \left(\frac{n\pi}{L}\right)^2, \quad n = 1, 2, \dots$$

For $\lambda = \left(\frac{n\pi}{L}\right)^2$, the solution for ODE of T is $T = C_n \cos\left(\frac{n\pi c t}{L}\right) + D_n \sin\left(\frac{n\pi c t}{L}\right)$.

By principle of superposition, we have

$$u(t, r) = \sum_{n=1}^{\infty} \left[a_n \cos\left(\frac{n\pi c t}{L}\right) + b_n \sin\left(\frac{n\pi c t}{L}\right) \right] j_0\left(\frac{n\pi r}{L}\right)$$

where $a_n = A_n C_n$ and $b_n = A_n D_n$ are constants.

From the initial conditions, we have

$$a_n = \frac{2}{L} \int_0^L f(r) j_0\left(\frac{n\pi r}{L}\right) r^2 dr, \quad b_n = \frac{L}{n\pi c} \times \frac{2}{L} \int_0^L g(r) j_0\left(\frac{n\pi r}{L}\right) r^2 dr.$$

□

37.7 2-D and 3-D Wave Equations

The three-dimensional wave equation

$$u_{tt} = c^2 \Delta u$$

in which c denotes the velocity of wave. It can be electromagnetic waves (light, radio, X-rays, etc.) in a homogeneous medium, including empty space. While the electric and magnetic vector fields $\mathbf{E}$ and $\mathbf{B}$ are intrinsically coupled by the more complicated system of Maxwell's equations, each individual component satisfies the wave equation.

The wave equation also models certain classes of *vibrations of a uniform solid body*. The solutions $u(x, y, z, t)$ represents a scalar-valued displacement of the body at time t and position $(x, y, z) \in \Omega \subset \mathbb{R}^3$. For example, u might represent the radial displacement of the body.

Let us study an example on the radial vibrations of a solid ball.

Example 37.46. Let u represent the radial displacement of the “atom” that is situated at position (x, y, z) when the ball is at rest. The radial vibration of a solid ball is modelled by a type 1 BVP:

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} = k \Delta u, & 0 < r < L, \quad 0 \leq \phi < 2\pi, \quad 0 \leq \theta \leq \pi, \quad t > 0, \\ u(L, \phi, \theta, t) = 0, & 0 \leq \phi < 2\pi, \quad 0 \leq \theta \leq \pi, \quad t > 0 \\ u(r, \phi, \theta, 0) = f(r, \phi, \theta), & 0 \leq r \leq L, \quad 0 \leq \phi < 2\pi, \quad 0 \leq \theta \leq \pi, \\ u_t(r, \phi, \theta, 0) = g(r, \phi, \theta), & 0 \leq r \leq L, \quad 0 \leq \phi < 2\pi, \quad 0 \leq \theta \leq \pi. \end{cases}$$

Use separation of variables to determine the formal series solution of this BVP.

Solution: Similar to Example [eg:sph:heat3d:type1](#), by using separation of variables, we obtain an ODE of T and SL problems related to the Helmholtz equation subject to homogeneous type 1 boundary condition:

$$\begin{cases} T'' + \lambda c^2 T = 0, \\ \Phi'' + \mu \Phi = 0, & \Phi(0) = \Phi(2\pi), \quad \Phi'(0) = \Phi'(2\pi), \\ r^2 R'' + 2rR' + (\lambda r^2 - \gamma)R = 0, & |R(0)| < \infty, \quad R(L) = 0, \\ \sin^2 \theta \Theta'' + \cos \theta \sin \theta \Theta' + (\gamma \sin^2 \theta - \mu)\Theta = 0, & |\Theta(0)| < \infty, \quad |\Theta(\pi)| < \infty. \end{cases}$$

The eigenvalues and corresponding eigenfunctions of the periodic SL problem of Φ are

$$\mu = m^2, \quad \Phi_m(\phi) = A_m \cos m\phi + B_m \sin m\phi, \quad m = 0, 1, 2, \dots$$

The singular SL problem of Θ has the following eigenvalues and eigenfunctions

$$\gamma = n(n+1), \quad \Theta_n^m(\theta) = P_n^m(\cos \theta), \quad n = m, m+1, \dots$$

With $\gamma = n(n+1)$, the singular SL problem of R has the following eigen-solutions according to Section [37.2.1](#):

$$\lambda = \left(\frac{\tau_k^{(n)}}{L} \right)^2, \quad R_k(r) = j_n\left(\frac{\tau_k^{(n)}}{L} r\right), \quad k = 1, 2, \dots;$$

where j_n is the n^{th} order spherical Bessel function and $\tau_k^{(n)}$ are the roots of j_n .

The general solution of the ODE of T is

$$T = C_{kn} \cos \frac{\tau_k^{(n)}}{L} ct + D_{kn} \sin \frac{\tau_k^{(n)}}{L} ct$$

and the general solution of the homogeneous wave equation and type 1 boundary condition is

$$u = \sum_{k=1}^{\infty} \sum_{m=0}^{\infty} \sum_{n=m}^{\infty} \left\{ a_{kmn} \cos\left(\frac{c\tau_k^{(n)}}{L} t\right) j_n\left(\frac{\tau_k^{(n)}}{L} r\right) Y_{n,m}(\theta, \phi) + b_{kmn} \cos\left(\frac{c\tau_k^{(n)}}{L} t\right) j_n\left(\frac{\tau_k^{(n)}}{L} r\right) \tilde{Y}_{n,m}(\theta, \phi) \right. \\ \left. c_{kmn} \sin\left(\frac{c\tau_k^{(n)}}{L} t\right) j_n\left(\frac{\tau_k^{(n)}}{L} r\right) Y_{n,m}(\theta, \phi) + d_{kmn} \sin\left(\frac{c\tau_k^{(n)}}{L} t\right) j_n\left(\frac{\tau_k^{(n)}}{L} r\right) \tilde{Y}_{n,m}(\theta, \phi) \right\}.$$

From the initial condition, we can apply the orthogonality theorem [thm:orth:3dhelmholtz 37.23](#) to obtain

$$\begin{aligned} a_{kmn} &= \frac{\int_0^L \int_0^{2\pi} \int_0^\pi f(r, \phi, \theta) j_n\left(\frac{\tau_k^{(n)}}{L} r\right) Y_{n,m}(\theta, \phi) r^2 \sin \theta d\theta d\phi dr}{\int_0^L \int_0^{2\pi} \int_0^\pi j_n\left(\frac{\tau_k^{(n)}}{L} r\right)^2 Y_{n,m}(\theta, \phi)^2 r^2 \sin \theta d\theta d\phi dr} \\ b_{kmn} &= \frac{\int_0^L \int_0^{2\pi} \int_0^\pi f(r, \phi, \theta) j_n\left(\frac{\tau_k^{(n)}}{L} r\right) \tilde{Y}_{n,m}(\theta, \phi) r^2 \sin \theta d\theta d\phi dr}{\int_0^L \int_0^{2\pi} \int_0^\pi j_n\left(\frac{\tau_k^{(n)}}{L} r\right)^2 \tilde{Y}_{n,m}(\theta, \phi)^2 r^2 \sin \theta d\theta d\phi dr} \\ c_{kmn} &= \frac{L}{c\tau_k^{(n)}} \frac{\int_0^L \int_0^{2\pi} \int_0^\pi g(r, \phi, \theta) j_n\left(\frac{\tau_k^{(n)}}{L} r\right) Y_{n,m}(\theta, \phi) r^2 \sin \theta d\theta d\phi dr}{\int_0^L \int_0^{2\pi} \int_0^\pi j_n\left(\frac{\tau_k^{(n)}}{L} r\right)^2 Y_{n,m}(\theta, \phi)^2 r^2 \sin \theta d\theta d\phi dr} \\ d_{kmn} &= \frac{L}{c\tau_k^{(n)}} \frac{\int_0^L \int_0^{2\pi} \int_0^\pi g(r, \phi, \theta) j_n\left(\frac{\tau_k^{(n)}}{L} r\right) \tilde{Y}_{n,m}(\theta, \phi) r^2 \sin \theta d\theta d\phi dr}{\int_0^L \int_0^{2\pi} \int_0^\pi j_n\left(\frac{\tau_k^{(n)}}{L} r\right)^2 \tilde{Y}_{n,m}(\theta, \phi)^2 r^2 \sin \theta d\theta d\phi dr}. \end{aligned}$$

The fundamental vibrational frequencies of a solid ball are

$$\omega_{n,k} = c\sqrt{\gamma_k^{(n)}} = c\tau_k^{(n)}, \quad n = 0, 1, 2, \dots, \quad k = 1, 2, \dots$$

For each distinct frequency, i.e. when n and k are fixed, there are a total of $2m+1$ since $0 \leq m \leq n$ independent vibrational modes. For more discussion of wave equation, see Olver [2007].

Chapter 38

Fourier Transforms and Applications

https://en.wikipedia.org/wiki/Fourier_transform

Laplace transform and Fourier transform are two commonly used *transform methods* for transforming linear PDE over an *infinite domain* (as opposed to *finite domain* discussed in Topics 2–4) to linear ODE problems. These are called the *global Cauchy problem* (with $n = 1$) [Salsa, 2008, p. 68]. This topic investigates only the application of Fourier transform. The proofs and advanced applications can be found in Duffy [2004], Bracewell [2000], Vladimirov [1979], Tihonov and Samarskiĭ [1977], Šolín [2006].

The domain of real physical problems are never infinite, but by introducing a mathematical model with infinite extent, we are able to determine problems in situations in which the influence of actual boundaries is negligible [Haberman, 1987, p. 335]. Transform methods are important because they provide a unifying mathematical approach to the study of electrical networks, devices for energy conversion and control, antennas, and other components of electrical systems, as well as to complete linear systems and to many other physical systems (which occur in geophysics, seismology, oceanography, planetary surfaces, civil engineering) and devices in biomedicine and remote sensing.

The numerical method related to Fourier transform is the *spectral methods*. If one wants to solve an ODE or PDE to high accuracy on a simple domain, and if the data defining the problem are smooth, then spectral methods are usually the best tool [Trefethen, 2000]. However, it should be noted that modern boundary value problems are largely solved by using finite element method based on functional analysis ideas [Daya Reddy, 1998].

38.1 Fourier Transform Basics

A *Fourier transform* is a *linear operator* (a linear transform for function spaces) that maps functions to other functions [Asmar, 2005, Tolstov, 1976, Reed and Simon, 1975, Butkov, 1968]:

$$\mathcal{F} : L^1(\mathbb{R}^n) \rightarrow \text{Set of Uniformly Continuous Functions},$$
$$f(\mathbf{x}) \mapsto \mathcal{F}\{f(x)\} := \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} f(\mathbf{x}) e^{-i\omega \cdot \mathbf{x}} d\mathbf{x} = F(\omega) \quad (38.1) \quad \text{e:ft:v1}$$

where $L^1(\mathbb{R}^n)$ are the set of Lebesgue integrable functions over $\mathbb{R}^n$, i.e. f satisfies $\int_{\mathbb{R}^n} |f(\mathbf{x})| d\mathbf{x} < \infty$. If $F \in L^1(\mathbb{R}^n)$, the *inverse Fourier transform* is defined as

$$\mathcal{F}^{-1}\{F(\omega)\} = \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} F(\omega) e^{i\omega \cdot \mathbf{x}} d\omega = f(\mathbf{x}).$$

In signal processing and some modern mathematical textbooks [Marks II, 2009, Stein and Shakarchi, 2003, Rahman, 2011, Folland, 1995, Lighthill, 1958], the Fourier transform and the inverse Fourier transform are defined below.

Definition 38.1. Let $f \in L^1(\mathbb{R}^n)$. The Fourier transform of f and the corresponding inverse Fourier transform are defined as

$$\begin{aligned} \hat{f} &= \mathcal{F}\{f\} := \int_{\mathbb{R}^n} f(\mathbf{x}) e^{-i2\pi\xi \cdot \mathbf{x}} d^n \mathbf{x} = F(\xi), \\ \check{F} &= \mathcal{F}^{-1}\{F\} := \int_{\mathbb{R}^n} F(\xi) e^{i2\pi\xi \cdot \mathbf{x}} d^n \xi = f(\mathbf{x}). \end{aligned} \quad (38.2) \quad \text{e:ft}$$

We will be using this version of the definition with $n = 1$. There are other conventions for defining the Fourier transform of a *complex-valued Lebesgue integrable function* $f(x)$.

Example 38.1 (UCCM3003 Final Exam, Apr 2009, Q6(a)). Find the Fourier transform of the following function

$$f(x) = \begin{cases} x^2, & -\pi < x < \pi, \\ 0, & \text{otherwise} \end{cases}$$

using the definition (e:ft:48.2).

(6 marks)

$$\begin{aligned} \text{Solution. } F(\xi) &= \int_{-\pi}^{\pi} x^2 e^{-2\pi i \xi x} dx = 2 \int_0^{\pi} x^2 \cos(-2\pi \xi x) dx \\ &= 2 \left[\frac{x^2 \sin(2\pi \xi x)}{2\pi \xi} + \frac{2x \cos(2\pi \xi x)}{(2\pi \xi)^2} - \frac{2 \sin(2\pi \xi x)}{(2\pi \xi)^3} \right]_0^{\pi} \\ &= 2 \left[\frac{\pi^2 \sin(2\pi^2 \xi)}{2\pi \xi} + \frac{2\pi \cos(2\pi^2 \xi)}{(2\pi \xi)^2} - \frac{2 \sin(2\pi^2 \xi)}{(2\pi \xi)^3} \right] \end{aligned}$$

□

From a physicist's perspective, the Fourier transform of the "signal" $f(\mathbf{x})$ gives the "frequency distribution" $F(\xi)$. The transform variable ξ represents *ordinary frequencies* (in hertz) and the variable ω , which is 2π times of ξ , is defined as the "angular frequency". The factor 2π appearing in the transform formulae may be lumped with ξ to yield the following version, which is used in some mathematical physics textbooks such as Erdelyi [1954], Mallat [1988], etc.:

$$\begin{aligned} F(\omega) &= \int_{-\infty}^{\infty} f(x) e^{-i\omega x} dx, \\ f(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{i\omega x} d\omega. \end{aligned} \tag{38.3}$$

e:ft:v2

Fourier transform is generally used to solve PDE with "(two-sided) infinite" boundary condition, for "semi-infinite" boundary condition, Fourier sine and cosine transforms are used instead. This is similar to the Type 1 and 2 boundary conditions we encountered in Topic 2.

Fourier:sine-cosine

Definition 38.2. The *Fourier sine transform* with its inverse are

$$F_s(\xi) = 2 \int_0^{\infty} f(x) \sin(2\pi \xi x) dx, \quad f(x) = 2 \int_0^{\infty} F_s(\xi) \sin(2\pi \xi x) d\xi$$

and *Fourier cosine transform* with its inverse are

$$F_c(\xi) = 2 \int_0^{\infty} f(x) \cos(2\pi \xi x) dx, \quad f(x) = 2 \int_0^{\infty} F_c(\xi) \cos(2\pi \xi x) d\xi.$$

Example 38.2 (UCCM3003 Final Exam, Apr 2012, Q5(b)). Use the definition to show that the Fourier sine transform of

$$f(x) = \begin{cases} 1, & 0 < x < 3, \\ 0, & x \geq 3 \end{cases}$$

$$\text{is } \sqrt{\frac{2}{\pi}} \frac{1 - \cos 3\omega}{\omega}.$$

(4 marks)

Proof: Note that in the previous semester, the following definition of Fourier sine transform corresponding to (e:ft:vi:38.1) was used:

$$F_s(\omega) := \sqrt{\frac{2}{\pi}} \int_0^{\infty} f(x) \sin \omega x dx = \sqrt{\frac{2}{\pi}} \int_0^3 \sin \omega x dx = \sqrt{\frac{2}{\pi}} \left[\frac{-\cos \omega x}{\omega} \right]_0^3 = \sqrt{\frac{2}{\pi}} \left[\frac{1 - \cos 3\omega}{\omega} \right].$$

Example 38.3 (UECM3003 Final Exam, Apr 2013, Q5(a)). Find the Fourier sine transform of $f(x) = \begin{cases} 1, & 0 < x < 1, \\ 0, & x \geq 1. \end{cases}$

(4 marks)

Solution: Using Definition [38.2](#), we have

$$F_s(\omega) := 2 \int_0^\infty f(x) \sin(2\pi\xi x) dx = 2 \int_0^1 \sin(2\pi\xi x) dx = 2 \left[\frac{-\cos 2\pi\xi x}{2\pi\xi} \right]_0^1 = \frac{2(1 - \cos 2\pi\xi)}{2\pi\xi}$$

Example 38.4 (UECM3003 Final Exam, Apr 2015, Q5(b)). Let f and f' be a function and its derivatives that vanish when $x \rightarrow \infty$. The Fourier sine transform of $f : (0, \infty) \rightarrow \mathbb{R}$ is defined as

$$F_s(\omega) = \mathcal{F}_s[f(x)] = \sqrt{\frac{2}{\pi}} \int_0^\infty f(x) \sin \omega x dx.$$

Prove that $\mathcal{F}_s[f''(x)] = \sqrt{\frac{2}{\pi}} \omega f(0) - \omega^2 F_s(\omega)$. (5 marks)

Proof: By making use of $\sin 0 = 0$ and $f(x) \rightarrow 0, f'(x) \rightarrow 0$ as $x \rightarrow \infty$, we have

$$\begin{aligned} \mathcal{F}_s[f''(x)] &= \sqrt{\frac{2}{\pi}} \int_0^\infty f''(x) \sin \omega x dx = \sqrt{\frac{2}{\pi}} \left\{ f'(x) \sin \omega x \Big|_0^\infty - \omega \int_0^\infty f'(x) \cos \omega x dx \right\} \\ &= -\omega \sqrt{\frac{2}{\pi}} \int_0^\infty f'(x) \cos \omega x dx = -\omega \sqrt{\frac{2}{\pi}} \left\{ f(x) \cos \omega x \Big|_0^\infty + \omega \int_0^\infty f(x) \sin \omega x dx \right\} \\ &= -\omega \sqrt{\frac{2}{\pi}} \left\{ -f(0) + \omega \int_0^\infty f(x) \sin \omega x dx \right\} = \sqrt{\frac{2}{\pi}} \omega f(0) - \omega^2 F_s(\omega) \end{aligned}$$

The inverse Fourier transforms in [\(38.1\)](#), [\(48.2\)](#), [\(38.3\)](#) and the Fourier sine, cosine transforms are obtained based on the following statement.

Theorem 38.5 (Fourier Integral Theorem). *If f is a piecewise continuous function and x is a continuous point of f , then*

$$f(x) = \int_{-\infty}^\infty \int_{-\infty}^\infty f(t) e^{i2\pi\xi(x-t)} dt d\xi = \frac{1}{2\pi} \int_{-\infty}^\infty \int_{-\infty}^\infty f(t) e^{i\omega(x-t)} dt d\omega. \quad (38.4)$$

If x is discontinuous, then we need to replace $f(x)$ with $\frac{f(x-0) + f(x+0)}{2}$.

Let x be continuous. When f is even and odd, we have

$$f(x) = 4 \int_0^\infty \left[\int_0^\infty f(t) \cos(2\pi\xi t) dt \right] \cos(2\pi\xi x) d\xi \quad \text{and} \quad (38.5)$$

$$f(x) = 4 \int_0^\infty \left[\int_0^\infty f(t) \sin(2\pi\xi t) dt \right] \sin(2\pi\xi x) d\xi \quad \text{respectively.} \quad (38.6)$$

Proof: According to Körner [1988, Chapter 60] and Dym and McKean [1972, Chapter 2], the equality (38.4) can be regarded as the limit of complex Fourier series for $L \rightarrow \infty$ (with a rigorous proof requiring the estimation on uniform convergence).

We will derive the Fourier integrals in terms of sine and cosine functions based on (38.4) and assume the point is real continuous:

$$\begin{aligned} f(x) &= \int_{-\infty}^\infty \int_{-\infty}^\infty f(t) \left[(\cos(2\pi\xi x) + i \sin(2\pi\xi x)) (\cos(2\pi\xi t) - i \sin(2\pi\xi t)) \right] dt d\xi \\ &= \int_{-\infty}^\infty \int_{-\infty}^\infty f(t) \left[(\cos(2\pi\xi x) \cos(2\pi\xi t) + \sin(2\pi\xi x) \sin(2\pi\xi t)) \right] dt d\xi \\ &\quad + i \int_{-\infty}^\infty \int_{-\infty}^\infty f(t) \left[\sin(2\pi\xi x) \cos(2\pi\xi t) - \cos(2\pi\xi x) \sin(2\pi\xi t) \right] dt d\xi \end{aligned}$$

Since f is real, the imaginary part is zero and we have

$$f(x) = \int_{-\infty}^\infty \int_{-\infty}^\infty f(t) \cos(2\pi\xi x) \cos(2\pi\xi t) dt d\xi + \int_{-\infty}^\infty \int_{-\infty}^\infty f(t) \sin(2\pi\xi x) \sin(2\pi\xi t) dt d\xi$$

When f is even, the right part becomes zero and we have

$$\begin{aligned} f(x) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(t) \cos(2\pi\xi x) \cos(2\pi\xi t) dt d\xi = 2 \int_{-\infty}^{\infty} \cos(2\pi\xi x) \left[\int_0^{\infty} \underbrace{f(t) \cos(2\pi\xi t)}_{\text{even in } t} dt \right] d\xi \\ &= 4 \int_0^{\infty} \cos(2\pi\xi x) \left[\underbrace{\int_0^{\infty} f(t) \cos(2\pi\xi t) dt}_{\text{even in } \xi} \right] d\xi \end{aligned}$$

When f is odd, the left part becomes zero and we have

$$\begin{aligned} f(x) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(t) \sin(2\pi\xi x) \sin(2\pi\xi t) dt d\xi = 2 \int_{-\infty}^{\infty} \sin(2\pi\xi x) \left[\int_0^{\infty} f(t) \sin(2\pi\xi t) dt \right] d\xi \\ &= 4 \int_0^{\infty} \sin(2\pi\xi x) \left[\int_0^{\infty} f(t) \sin(2\pi\xi t) dt \right] d\xi \end{aligned}$$

During the solution of PDE, convolution arises naturally. It is defined below.

Definition 38.3. Given two piecewise-continuous function f and g , the *convolution* of the two functions is denoted and defined as

$$(f * g)(x) := \int_{-\infty}^{\infty} f(\tau)g(x - \tau)d\tau$$

It can be shown that convolution is commutative, $f * g = g * f$, associative, $(f * g) * h = f * (g * h)$ and distributive, $f * (g + h) = f * g + f * h$.

Theorem 38.6 (Convolution Theorem). If $f, g \in L^1(\mathbb{R})$, $\mathcal{F}\{f(x)\} = F(\xi)$, $\mathcal{F}\{g(x)\} = G(\xi) \in L^1(\mathbb{R})$ then

1. $\mathcal{F}\{f * g\} = F(\xi)G(\xi); \quad \mathcal{F}\{f \cdot g\} = F(\xi) * G(\xi);$
2. $\mathcal{F}^{-1}\{F * G\} = f(x)g(x); \quad \mathcal{F}^{-1}\{F \cdot G\} = f(x) * g(x).$

Example 38.7 (UECM3003 Final Exam, Apr 2015, Q4(a)). Suppose that f is integrable and even, and $g(x) = \cos(ax)$ where a is a constant. Show that the convolution

$$f * g(x) = \cos(ax)F(a)$$

where F is the Fourier transform (38.3) of f .

(6 marks)

Proof: In the following, $f(-y) = f(y)$ since f is even:

$$\begin{aligned} f * g(x) &= \int_{-\infty}^{\infty} f(y) \cos(a(y - x)) dy = \int_{-\infty}^{\infty} f(y) \frac{e^{ia(y-x)} + e^{-ia(y-x)}}{2} dy \\ &= \frac{1}{2} \left[e^{-iax} \int_{-\infty}^{\infty} e^{ia y} f(y) dy + e^{iax} \int_{-\infty}^{\infty} e^{-ia y} f(y) dy \right] \\ &= \frac{1}{2} \left[e^{-iax} \int_{\infty}^{-\infty} e^{ia(-y)} f(-y) (-dy) + e^{iax} F(a) \right] \\ &= \frac{1}{2} \left[e^{-iax} \int_{-\infty}^{\infty} e^{-ia y} f(y) dy + e^{iax} F(a) \right] = \frac{1}{2} [e^{-iax} F(a) + e^{iax} F(a)] = \cos(ax)F(a) \end{aligned}$$

38.2 Properties of Fourier Transform when $n = 1$

By using the definition of Fourier transform, it is possible to derive the basic properties stated below.

Theorem 38.8. If $f, g \in L^1(\mathbb{R})$, $\mathcal{F}\{f(x)\} = F(\xi)$, $\mathcal{F}\{g(x)\} = G(\xi)$ according to (48.2), $\mathcal{F}\{f(x)\} = F_1(\omega)$, $\mathcal{F}\{g(x)\} = G_1(\omega)$ according to (38.3), then we can derive the following properties:

Name	Function	Fourier Transform (48.2)	Fourier Transform (38.3)
Definition	$f(x)$	$F(\xi) = \int_{-\infty}^{\infty} f(x)e^{-2\pi i \xi x} dx$	$F_1(\omega) = \int_{-\infty}^{\infty} f(x)e^{-i\omega x} dx$
1. Linearity	$af(x) + bg(x)$	$aF(\xi) + bG(\xi)$	$aF_1(\xi) + bG_1(\xi)$
2. "Time"-shift	$f(x - a)$	$e^{-2\pi i \xi a} F(\xi)$	$e^{-i\omega a} F_1(\omega)$
3. Frequency-shift/ Modulation	$e^{i\alpha x} f(x)$	$F(\xi - \frac{\alpha}{2\pi})$	$F_1(\omega - \alpha)$
4. Scaling	$f(kx)$, $k \neq 0$	$\frac{1}{ k } F(\frac{\xi}{k})$	$\frac{1}{ k } F_1(\frac{\omega}{k})$
5. Duality/ Conjugation	$F(x)$ or $F_1(x)$	$f(-\xi)$	$2\pi f(-\omega)$
6. "Time"-differentiation	$\frac{d^n}{dx^n} f(x)$	$(2\pi i \xi)^n F(\xi)$	$(i\omega)^n F_1(\omega)$
7. Frequency-differentiation	$x^n f(x)$	$\left(\frac{i}{2\pi}\right)^n \frac{d^n}{d\xi^n} F(\xi)$	$i^n \frac{d^n F_1(\omega)}{d\omega^n}$
8. Convolution	$(f * g)(x)$	$F(\xi)G(\xi)$	$F(\omega)G(\omega)$
9. Multiplication	$f(x)g(x)$	$(F * G)(\xi)$	$\frac{1}{2\pi} (F * G)(\omega)$

Proof: Let $y = kx$.

1. A definite integral is linear.
2. By a change of variable $y = x - a$, we have

$$\int_{-\infty}^{\infty} f(x - a)e^{-2\pi i \xi x} dx = e^{-2\pi i \xi a} \int_{-\infty}^{\infty} f(y)e^{-2\pi i \xi y} dy = e^{-2\pi i \xi a} F(\xi).$$

$$3. \int_{-\infty}^{\infty} e^{i\alpha x} f(x)e^{-2\pi i \xi x} dx = \int_{-\infty}^{\infty} f(x)e^{-2\pi i (\xi - \frac{\alpha}{2\pi})x} dx = F(\xi - \frac{\alpha}{2\pi}).$$

$$4. \text{ When } k > 0, \int_{-\infty}^{\infty} f(kx)e^{-2\pi i \xi x} dx = \int_{-\infty}^{\infty} f(y)e^{-2\pi i \xi y/k} \frac{1}{k} dy = \frac{1}{k} \int_{-\infty}^{\infty} f(y)e^{-2\pi i (\xi/k)y} dy = \frac{1}{k} F(\xi/k)$$

$$\text{When } k < 0, \int_{-\infty}^{\infty} f(kx)e^{-2\pi i \xi x} dx = \int_{\infty}^{-\infty} f(y)e^{-2\pi i \xi y/k} \frac{1}{k} dy = -\frac{1}{k} \int_{-\infty}^{\infty} f(y)e^{-2\pi i (\xi/k)y} dy = -\frac{1}{k} F(\xi/k)$$

5. By definition,

$$\begin{aligned} \int_{-\infty}^{\infty} F(y)e^{-2\pi i \xi y} dy &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f(x)e^{-2\pi i y x} dx \right] e^{-2\pi i \xi y} dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x)e^{2\pi i y(-\xi-x)} dx dy = f(-\xi) \text{ by Theorem 38.5} \end{aligned}$$

6. We need to note that if $f \in L^1(\mathbb{R})$ and f is differentiable n -times, then $f(\pm\infty) = \dots = \frac{d^{n-1}}{dx^{n-1}} f(\pm\infty) = 0$ and so by applying integration by part, we have

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{d^n}{dx^n} f(x)e^{-2\pi i \xi x} dx &= e^{-2\pi i \xi x} \frac{d^{n-1}}{dx^{n-1}} f(x) \Big|_{-\infty}^{\infty} - (-2\pi i \xi) \int_{-\infty}^{\infty} \frac{d^{n-1}}{dx^{n-1}} f(x)e^{-2\pi i \xi x} dx \\ &= 2\pi i \xi \int_{-\infty}^{\infty} \frac{d^{n-1}}{dx^{n-1}} f(x)e^{-2\pi i \xi x} dx \\ &= (2\pi i \xi)^n \int_{-\infty}^{\infty} f(x)e^{-2\pi i \xi x} dx = (2\pi i \xi)^n F(\xi) \end{aligned}$$

$$\begin{aligned}
7. \quad \frac{d^n}{d\xi^n} \int_{-\infty}^{\infty} f(x) e^{-2\pi i \xi x} dx &= \int_{-\infty}^{\infty} f(x) \frac{d^n}{d\xi^n} e^{-2\pi i \xi x} dx = \int_{-\infty}^{\infty} f(x) (-2\pi i)^n e^{-2\pi i \xi x} dx \\
&= (-2\pi i)^n \int_{-\infty}^{\infty} x^n f(x) e^{-2\pi i \xi x} dx
\end{aligned}$$

Therefore, we have $\mathcal{F}\{x^n f(x)\} = \frac{1}{(-2\pi i)^n} \mathcal{F}\{f(x)\}$.

8. According to Folland [1995], the proof is an application of Fubini's theorem (which guarantees the exchange of integration variables):

$$\begin{aligned}
\mathcal{F}\{(f * g)(x)\} &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f(x-y) g(y) dy \right] e^{-2\pi i \xi x} dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x-y) e^{-2\pi i \xi (x-y)} g(y) e^{-2\pi i \xi y} dx dy \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f(x-y) e^{-2\pi i \xi (x-y)} dx \right] g(y) e^{-2\pi i \xi y} dy \\
&= F(\xi) \int_{-\infty}^{\infty} g(y) e^{-2\pi i \xi y} dy = F(\xi) G(\xi)
\end{aligned}$$

9. We need to apply the Fubini's theorem again and also $\mathcal{F}^{-1}\{F(\xi)\} = f(x)$.

$$\begin{aligned}
\mathcal{F}\{f(x)g(x)\} &= \int_{-\infty}^{\infty} f(x)g(x) e^{-2\pi i \xi x} dx = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} F(\lambda) e^{2\pi i \lambda x} d\lambda \right] g(x) e^{-2\pi i \xi x} dx \\
&= \int_{-\infty}^{\infty} F(\lambda) \left[\int_{-\infty}^{\infty} g(x) e^{2\pi i \lambda x} e^{-2\pi i \xi x} dx \right] d\lambda \\
&= \int_{-\infty}^{\infty} F(\lambda) \left[\int_{-\infty}^{\infty} g(x) e^{-2\pi i (\xi - \lambda)x} dx \right] d\lambda = \int_{-\infty}^{\infty} F(\lambda) G(\xi - \lambda) d\lambda \\
&= (F * G)(\xi)
\end{aligned}$$

Remark 38.9. The properties of general n case can be found in Folland [1995]. The above properties indicates that if f decays very rapidly at infinity, then its Fourier transform is “smooth” and if f is “smooth” with rapidly decaying derivatives, then its Fourier transform rapidly decays.

The next few theorems illustrate the properties of Fourier transform related to “smoothness”, “duality” and “energy” for general n case.

Theorem 38.10. Suppose that $f \in L^1(\mathbb{R}^n)$. Then $\mathcal{F}\{f\}$ is a bounded, continuous function and

$$\|\mathcal{F}\{f\}\|_{C^0} \leq \|f\|_{L^1}.$$

Theorem 38.11. If $f, g \in L^1$, then

$$\int_{\mathbb{R}^n} \widehat{f}(\mathbf{x}) g(\mathbf{x}) d^n \mathbf{x} = \int_{\mathbb{R}^n} f(\mathbf{x}) \widehat{g}(\mathbf{x}) d^n \mathbf{x}.$$

Theorem 38.12 (Fourier Inversion Theorem). Suppose that $f : \mathbb{R}^n \rightarrow \mathbb{C}$ is a continuous function, that $f \in L^1$ and $\widehat{f} \in L^1$. Then

$$\mathcal{F}^{-1}\{\widehat{f}\} = \widehat{\mathcal{F}^{-1}\{f\}} = f.$$

In physics, the notion of energy is defined by the integration of the square of a function $\int_a^b |f(x)|^2 dx$. All functions f with “finite energy” $\int_{-\infty}^{\infty} |f(x)|^2 dx < \infty$ is a set $\mathcal{L}^2(\mathbb{R})$ which has Hilbert space structure (extension of Euclidean space to infinite dimension case).

Theorem 38.13 (Plancherel Theorem). Suppose $f, g : \mathbb{R}^n \rightarrow \mathbb{C}$ are continuous functions and if $f, g \in L^1(\mathbb{R}^n) \cap L^2(\mathbb{R}^n)$ and $\widehat{f}, \widehat{g} \in L^1$. Then $\widehat{f}, \widehat{g} \in L^2$ and

$$\int_{\mathbb{R}^n} f(\mathbf{x}) \overline{g(\mathbf{x})} d^n \mathbf{x} = \int_{\mathbb{R}^n} \widehat{f}(\xi) \overline{\widehat{g}(\xi)} d^n \xi.$$

In particular, we have

$$\int_{\mathbb{R}^n} |f(x)|^2 d^n \mathbf{x} = \int_{\mathbb{R}^n} |F(\xi)|^2 d^n \xi.$$

where $|f(\mathbf{x})|^2 = f(\mathbf{x})\overline{f(\mathbf{x})}$, $\overline{f(\mathbf{x})}$ is the complex conjugate of $f(\mathbf{x})$.

We can think of the above theorem as the “Parseval Theorem” (Topic 1) for Fourier transform.

Example 38.14. (UCCM3003 Final Exam, May 2010, Q5(a); UECM3003 Final Exam, Apr 2014, Q1(b). Both with modification) Let

$$f(x) = \begin{cases} -1, & -1 < x < 0, \\ 1, & 0 < x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Use the definition to show that the Fourier transform of f is

$$F(\omega) = \frac{i(\cos(2\pi\xi) - 1)}{\pi\xi} \quad (\text{original: } i\sqrt{\frac{2}{\pi}} \frac{\cos \omega - 1}{\omega}).$$

Hence, use the First Shifting Theorem to deduce the Fourier transform of

$$g(x) = \begin{cases} -e^{2ix}, & -1 < x < 0, \\ e^{2ix}, & 0 < x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Solution.

$$\begin{aligned} F(\xi) &= \int_{-1}^1 f(x)(\cos(2\pi\xi x) - i\sin(2\pi\xi x))dx = -2i \int_0^1 \sin(2\pi\xi x)dx \\ &= -2i \left[\frac{-\cos(2\pi\xi x)}{2\pi\xi} \right]_0^1 = 2i \left[\frac{\cos(2\pi\xi) - 1}{2\pi\xi} \right] \\ \mathcal{F}\{g(x)\} &= \mathcal{F}\{e^{2ix}f(x)\} = F\left(\xi - \frac{2}{\pi}\right) = \frac{i(\cos(2\pi(\xi - \frac{1}{\pi})) - 1)}{\pi(\xi - \frac{1}{\pi})} = \frac{i(\cos(2\pi\xi - 2) - 1)}{\pi\xi - 1} \end{aligned}$$

□

Some special L^1 functions which have nice analytic form for their Fourier transform are defined below [Bracewell, 2000, 91].

Definition 38.4. $\Pi(ax) = \begin{cases} 1, & |x| < \frac{1}{2a}, \\ 0, & |x| > \frac{1}{2a}. \end{cases}$ is called a *rectangular function*.

Definition 38.5. $\text{sinc}(ax) := \frac{\sin(ax)}{ax}$ is called a *sinc function*.

Definition 38.6. $\Lambda(ax) = \begin{cases} 1 - |ax| & |ax| < 1 \\ 0 & |ax| > 1 \end{cases}$ is called a *triangular function*. Note that $\Lambda(x) = \Pi * \Pi(x)$.

Definition 38.7. Let $\alpha > 0$. $x \mapsto e^{-\alpha x^2}$ is called a *Gaussian function*.

Definition 38.8. Let $a > 0$. The *one-side exponential decay* and *two-side exponential decay* are defined as

$$x \mapsto e^{-ax}H(x) \quad \text{and} \quad x \mapsto e^{-a|x|}$$

respectively. Here $H(x) := \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$ is the Heaviside step function.

By using the definition and the properties of Fourier transform, we can derive the following transform table (assuming $a > 0$ and $\alpha > 0$):

Function	Fourier Transform (48.2)	Fourier Transform (38.3)
$\Pi(ax)$	$\frac{1}{ a } \text{sinc}\left(\frac{\xi}{a}\right)$	$\frac{1}{ a } \text{sinc}\left(\frac{\omega}{2\pi a}\right)$
$\text{sinc}(ax)$	$\frac{1}{ a } \Pi\left(\frac{\xi}{a}\right)$	$\frac{1}{ a } \Pi\left(\frac{\omega}{2\pi a}\right)$,
$\Lambda(ax)$	$\frac{1}{ a } \text{sinc}^2\left(\frac{\xi}{a}\right)$	$\frac{1}{ a } \text{sinc}^2\left(\frac{\omega}{2\pi a}\right)$
$\text{sinc}^2(ax)$	$\frac{1}{ a } \Lambda\left(\frac{\xi}{a}\right)$	$\frac{1}{ a } \Lambda\left(\frac{\omega}{2\pi a}\right)$
$e^{-\alpha x^2}$	$\sqrt{\frac{\pi}{\alpha}} e^{-\frac{\pi^2 \xi^2}{\alpha}}$	$\sqrt{\frac{\pi}{\alpha}} e^{-\frac{\omega^2}{4\alpha}}$
$e^{-ax}H(x)$	$\frac{1}{a + 2\pi i\xi}$	$\frac{1}{a + i\omega}$
$e^{-a x }$	$\frac{2a}{a^2 + (2\pi\xi)^2}$	$\frac{2a}{a^2 + \omega^2}$

Example 38.15 (UCCM3003 Final Exam, Apr 2009, Q6(b)). Show that the Fourier transform of $f(x) = \frac{1}{a^2 + x^2}$, $a > 0$ is $F(\omega) = \frac{\sqrt{2\pi}}{2a} e^{-a|\omega|}$. [Assume that the Fourier transform is defined by (38.1)] (4 marks)

Solution.

$$\mathcal{F}\{e^{-a|x|}\} = \frac{2a}{a^2 + (2\pi\xi)^2} \Rightarrow \mathcal{F}\{e^{-a|2\pi x|}\} = \frac{1}{2\pi} \frac{2a}{a^2 + (2\pi(\frac{\xi}{2\pi}))^2} = \frac{a}{\pi} \frac{1}{a^2 + \xi^2}$$

By duality in Theorem 38.8,

$$\frac{a}{\pi} \mathcal{F}\left\{\frac{1}{a^2 + x^2}\right\} = e^{-a|2\pi(-\xi)|} \Rightarrow \mathcal{F}\left\{\frac{1}{a^2 + x^2}\right\} = \frac{\pi}{a} e^{-a|2\pi\xi|} \Rightarrow F_1(\omega) = \frac{1}{\sqrt{2\pi}} \frac{\pi}{a} e^{-a|\omega|}$$

□

It is very difficult to apply complex analysis to find the inverse Fourier transform in (48.2), performing table look-up to find the inverse Fourier transform seems to be the best choice!

Example 38.16 (UCCM3003 Final Exam, Apr 2012, Q5(c) with modification). Find the inverse Fourier transform of the function $F(\xi) = \frac{e^{-4\pi i\xi}}{2 + 8\pi i\xi}$. (5 marks)

Solution:

$$\begin{aligned} \mathcal{F}^{-1}\left\{e^{-4\pi i\xi} \cdot \frac{1}{2 + 8\pi i\xi}\right\} &= \mathcal{F}^{-1}\left\{\frac{1}{2 + 8\pi i\xi}\right\} \Big|_{x \rightarrow x-2} = \frac{1}{4} \mathcal{F}^{-1}\left\{\frac{1}{\frac{1}{2} + 2\pi i\xi}\right\} \Big|_{x \rightarrow x-2} \\ &= \frac{1}{4} e^{-\frac{1}{2}x} H(x) \Big|_{x \rightarrow x-2} = \frac{1}{4} e^{-\frac{1}{2}(x-2)} H(x-2) \end{aligned}$$

Example 38.17 (UECM3003 Final Exam, Apr 2013, Q5(c)). Find the inverse Fourier transform of $F(\omega) = \frac{2 + 2i\omega}{15 + 8i\omega - \omega^2}$.

[Hint: Factor the denominator and use partial fractions.]

(6 marks)

Solution. Let $s = i\omega$. Then

$$\frac{2+2s}{15+8s+s^2} = \frac{2+2s}{(s+3)(s+5)} = \frac{\frac{2+2(-3)}{(-3)+5}}{s+3} + \frac{\frac{2+2(-5)}{(-5)+3}}{s+5} = \frac{-2}{s+3} + \frac{4}{s+5}$$

and according to the table in page [tbl:ft:spec-fn](#) 883,

$$\mathcal{F}^{-1}\left\{\frac{2+2i\omega}{15+8i\omega-\omega^2}\right\} = \mathcal{F}^{-1}\left\{\frac{-2}{i\omega+3} + \frac{4}{i\omega+5}\right\} = 4e^{-5x}H(x) - 2e^{-3x}H(x).$$

□

Example 38.18 (UECM3003 Final Exam, Apr 2014, Q3(c)). Solve the following differential equation $y' - 3y = H(t)e^{-3t}$, where $H(t)$ is the Heaviside step function. (5 marks)

[**Remark:** Dr Goh YK did not specify what technique to use. Is Laplace transform or integrating factor allowed?]

Solution.

$$(2\pi i\xi)Y(\xi) - 3Y(\xi) = \frac{1}{3+2\pi i\xi} \Rightarrow Y(\xi) = \frac{1}{(2\pi i\xi+3)(2\pi i\xi-3)}$$

Let $s = 2\pi i\xi$. Then $\frac{1}{(s+3)(s-3)} = \frac{1}{6} \left[\frac{1}{s-3} - \frac{1}{s+3} \right]$ and

$$y(t) = \frac{1}{6} \mathcal{F}^{-1}\left\{\frac{1}{2\pi i\xi-3} - \frac{1}{2\pi i\xi+3}\right\} = \frac{1}{6} \left[e^{3t}H(t) - e^{-3t}H(t) \right].$$

□

38.3 Fourier Transform for Generalised Functions

A circuit expert finds it obvious that every waveform has a spectrum, and the antenna designer is confident that every antenna has a radiation pattern. But mathematically there are some functions without Fourier transforms. For instance, the harmonic function $\sin x$ which plays an important role in Fourier series; step function, $H(x)$ which is used in Laplace transform and signal analysis. These functions are physically impossible since they are defined to be none zero over and infinite time interval. However, being able to perform calculation with them would be very convenient. The trouble is that Fourier transform does not exist since they are not L^1 functions.

In order to obtain the Fourier transform for constant and periodic functions, we have to introduce the concept of generalised functions. The Fourier transform can no longer be expressed in integral form but needs to be consider as a *linear operator over a generalised function space*.

Without going into the details of the theory which can be found in Bleistein [1984] and many functional analysis text books, we list some of the Fourier transform as follows. They are all closely related to the Dirac delta function $\delta(x)$ which can be regarded as a “sequence of functions” which has a “limit” as follows

$$\delta(x) = \begin{cases} \infty, & x = 0, \\ 0, & x \neq 0. \end{cases}$$

This generalised function is “even”, i.e. $\delta(-x) = \delta(x)$.

Fourier transform can be generalised from a linear operator over $L^1(\mathbb{R})$ to a more general space of distributions which includes the Dirac delta function, etc. which satisfies scaling, duality, time-/frequency-differentiation/shift and convolution properties.

Generalised Function	Fourier Transform (48.2)	Fourier Transform (38.3)
$\delta(x - a)$	$e^{-2\pi i \xi a}$	$e^{-i \omega a}$
e^{iax}	$\delta(\xi - \frac{a}{2\pi})$	$2\pi \delta(\omega - a)$ (by duality)
$\cos(ax)$	$\frac{\delta(\xi - \frac{a}{2\pi}) + \delta(\xi + \frac{a}{2\pi})}{2}$	$\pi(\delta(\omega - a) + \delta(\omega + a))$
$\sin(ax)$	$\frac{\delta(\xi - \frac{a}{2\pi}) - \delta(\xi + \frac{a}{2\pi})}{2i}$	$\frac{\pi}{i}(\delta(\omega - a) - \delta(\omega + a))$
x^n	$(\frac{i}{2\pi})^n \delta^{(n)}(\xi)$	$i^n 2\pi \delta^{(n)}(\omega)$
$\text{sgn}(x)$	$\frac{1}{i\pi \xi}$	$\frac{2}{i\omega}$
$\frac{1}{x}$	$-i\pi \text{sgn}(\xi)$	$-i\pi \text{sgn}(\omega)$
$\sum_{n=-\infty}^{\infty} \delta(x - nT)$	$\frac{1}{T} \sum_{n=-\infty}^{\infty} \delta(\xi - \frac{n}{T})$	$\frac{2\pi}{T} \sum_{n=-\infty}^{\infty} \delta(\omega - \frac{2\pi n}{T})$
$\int_{-\infty}^x f(z) dz$	$\frac{1}{2\pi i \xi} F(\xi) + F(0) \delta(\xi)$	$\frac{1}{i\omega} F(\omega) + F(0) \delta(\omega)$

Proof: It is extremely difficult to get “generalised function” right because they are **not** functions. Rahman [2011] is a good introductory book on Fourier transform for generalised function. In the book the Fourier transform of $\text{sgn}(x)$ is given as follows:

$$\begin{aligned}
 \mathcal{F}\{\text{sgn}(x)\} &= \mathcal{F}\{\lim_{a \rightarrow 0} e^{-a|x|} \text{sgn}(x)\} = \lim_{a \rightarrow 0} \mathcal{F}\{e^{-a|x|} \text{sgn}(x)\} = \lim_{a \rightarrow 0} \int_{-\infty}^{\infty} e^{-a|x|} \text{sgn}(x) e^{-2\pi i \xi x} dx \\
 &= \lim_{a \rightarrow 0} \left[- \int_{-\infty}^0 e^{-a(-x)} e^{-2\pi i \xi x} dx + \int_0^{\infty} e^{-ax} e^{-2\pi i \xi x} dx \right] \\
 &= \lim_{a \rightarrow 0} \left\{ - \left[\frac{e^{(a-2\pi i \xi)x}}{a-2\pi i \xi} \right]_{-\infty}^0 + \left[\frac{e^{(-a-2\pi i \xi)x}}{-a-2\pi i \xi} \right]_0^{\infty} \right\} \\
 &= \lim_{a \rightarrow 0} \left\{ - \frac{1}{a-2\pi i \xi} + \frac{-1}{-a-2\pi i \xi} \right\} = \frac{1}{2\pi i \xi} - \frac{1}{-2\pi i \xi} = \frac{1}{\pi i \xi}
 \end{aligned}$$

The Fourier transform of an “integral” is given as follows:

$$\begin{aligned}
 \int_{-\infty}^x f(z) dz &= \int_{-\infty}^x \left[\int_{-\infty}^{\infty} F(\xi) e^{2\pi i \xi z} d\xi \right] dz = \int_{-\infty}^{\infty} F(\xi) \left[\int_{-\infty}^x e^{2\pi i \xi z} dz \right] d\xi \\
 &= \int_{-\infty}^{\infty} F(\xi) \left[\frac{e^{2\pi i \xi z}}{2\pi i \xi} - \lim_{z \rightarrow \infty} \frac{e^{2\pi i \xi z}}{2\pi i \xi} \right] d\xi = \int_{-\infty}^{\infty} \frac{F(\xi)}{2\pi i \xi} e^{2\pi i \xi z} d\xi + \int_{-\infty}^{\infty} F(\xi) \delta(\xi) d\xi \\
 &= \mathcal{F}^{-1} \left\{ \frac{F(\xi)}{2\pi i \xi} \right\} - F(0)
 \end{aligned}$$

Taking Fourier transform on both sides, we have

$$\mathcal{F}\left\{\int_{-\infty}^x f(z) dz\right\} = \frac{F(\xi)}{2\pi i \xi} + F(0) \delta(\xi).$$

Example 38.19 (UECM3003 Final Exam, Apr 2014, Q4(a)). Find the Fourier transform of $f(x) = \cos^2 3x$. (5 marks)

$$\begin{aligned}
 \text{Solution: } \mathcal{F}\{\cos^2 3x\} &= \mathcal{F}\left\{\frac{1 + \cos 6x}{2}\right\} = \frac{1}{2} [\mathcal{F}\{1\} + \mathcal{F}\{\cos 6x\}] \\
 &= \frac{1}{2} \left[\delta(\xi) + \frac{1}{2} (\delta(\xi - \frac{6}{2\pi}) + \delta(\xi + \frac{6}{2\pi})) \right] \\
 &= \frac{1}{2} \delta(\xi) + \frac{1}{4} \delta(\xi - \frac{3}{\pi}) + \frac{1}{4} \delta(\xi + \frac{3}{\pi})
 \end{aligned}$$

Note that although $\mathcal{F}\{1\}$ is not given in the table, but we know that $e^{i0x} = 1$, so $\mathcal{F}\{1\} = \delta(\xi - \frac{0}{2\pi}) = \delta(\xi)$.

Example 38.20 (UCCM3003 Final Exam, May 2010, Q6(a)). Find the Fourier transform of the following expressions:

$$(i) \ f(x) = \begin{cases} 1 - |x|, & |x| < 1, \\ 0, & \text{otherwise.} \end{cases} \quad (5 \text{ marks})$$

Solution. $\mathcal{F}\{f(x)\} = \mathcal{F}\{\Lambda(x)\} = \text{sinc}^2(\xi) = \left(\frac{\sin \xi}{\xi}\right)^2$ □

(ii) $g(x) = \sin 2x$, $-\infty < x < \infty$.

(5 marks)

Solution. $\mathcal{F}\{\sin 2x\} = \mathcal{F}\left\{\frac{e^{i2x} - e^{-i2x}}{2i}\right\} = \frac{1}{2i} \left[\delta\left(\xi - \frac{1}{\pi}\right) - \delta\left(\xi + \frac{1}{\pi}\right) \right]$ □

Example 38.21. Evaluate $\mathcal{F}\{H(x)\}$.

Solution.

$$\mathcal{F}\{H(x)\} = \frac{1}{2} \mathcal{F}\{1 + \text{sgn}(x)\} = \frac{1}{2} \left[\delta(\xi) + \frac{1}{i\pi\xi} \right]$$

□

Example 38.22. ([Bracewell, 2000, 152]) Show that $\mathcal{F}\left\{\frac{\cos \pi x}{\pi(x - \frac{1}{2})}\right\} = -e^{i\pi\xi} \Pi(\xi)$.

Solution.
$$\begin{aligned} \mathcal{F}\left\{\frac{\cos \pi x}{\pi(x - \frac{1}{2})}\right\} &= \frac{1}{2\pi} \left[\mathcal{F}\left\{e^{i\pi x} \frac{1}{x - \frac{1}{2}}\right\} + \mathcal{F}\left\{e^{i(-\pi)x} \frac{1}{x - \frac{1}{2}}\right\} \right] \\ &= \frac{1}{2\pi} \left[\mathcal{F}\left\{\frac{1}{x - \frac{1}{2}}\right\}\Big|_{\xi \rightarrow \xi - \frac{1}{2}} + \mathcal{F}\left\{\frac{1}{x - \frac{1}{2}}\right\}\Big|_{\xi \rightarrow \xi + \frac{1}{2}} \right] \\ &= \frac{1}{2\pi} \left[e^{-i\pi\xi} \mathcal{F}\left\{\frac{1}{x}\right\}\Big|_{\xi \rightarrow \xi - \frac{1}{2}} + e^{-i\pi\xi} \mathcal{F}\left\{\frac{1}{x}\right\}\Big|_{\xi \rightarrow \xi + \frac{1}{2}} \right] \\ &= \frac{1}{2\pi} \left[e^{-i\pi\xi} (-i\pi \text{sgn}(\xi))\Big|_{\xi \rightarrow \xi - \frac{1}{2}} + e^{-i\pi\xi} (-i\pi \text{sgn}(\xi))\Big|_{\xi \rightarrow \xi + \frac{1}{2}} \right] \\ &= \frac{-i}{2} \left[e^{-i\pi(\xi - \frac{1}{2})} \text{sgn}(\xi - \frac{1}{2}) + e^{-i\pi(\xi + \frac{1}{2})} \text{sgn}(\xi + \frac{1}{2}) \right] \\ &= \frac{-ie^{-i\pi\xi}}{2} \left[e^{i\pi/2} \text{sgn}(\xi - \frac{1}{2}) + e^{-i\pi/2} \text{sgn}(\xi + \frac{1}{2}) \right] \\ &= \frac{-ie^{-i\pi\xi}}{2} \left[i \text{sgn}(\xi - \frac{1}{2}) - i \text{sgn}(\xi + \frac{1}{2}) \right] \\ &= \frac{e^{-i\pi\xi}}{2} \left[\text{sgn}(\xi - \frac{1}{2}) - \text{sgn}(\xi + \frac{1}{2}) \right] = \frac{e^{-i\pi\xi}}{2} (-2\Pi(\xi)) = \text{RHS} \end{aligned}$$
 □

Example 38.23. Determine the Fourier transform of $f(x) = \frac{1}{(1+x^2)^2}$.

Solution. Note that $\mathcal{F}\{\exp(-2\pi|x|)\} = \frac{4\pi}{(2\pi)^2 + (2\pi\xi)^2} \Rightarrow \mathcal{F}\{\pi e^{-2\pi|x|}\} = \frac{1}{1+\xi^2}$ By duality (Theorem 38.8), we have

$$\mathcal{F}\left\{\frac{1}{1+x^2}\right\} = \pi e^{-2\pi|\xi|} = \pi e^{-2\pi|\xi|}$$

$$\begin{aligned} \mathcal{F}\left\{\frac{1}{(1+x^2)^2}\right\} &= \pi e^{-2\pi|\xi|} * \pi e^{-2\pi|\xi|} = \pi^2 \int_{-\infty}^{\infty} e^{-2\pi|\tau-\xi|} e^{-2\pi|\tau|} d\tau = \pi^2 \int_{-\infty}^{\infty} e^{-2\pi(|\tau-\xi|+|\tau|)} d\tau \\ &= \pi^2 \begin{cases} \int_{-\infty}^0 e^{-2\pi(-\tau+\xi-\tau)} d\tau + \int_0^{\xi} e^{-2\pi(\xi-\tau+\tau)} d\tau + \int_{\xi}^{\infty} e^{-2\pi(\tau-\xi+\tau)} d\tau & \xi \geq 0 \\ \int_{-\infty}^{\xi} e^{-2\pi(-\tau+\xi-\tau)} d\tau + \int_{\xi}^0 e^{-2\pi(\tau-\xi-\tau)} d\tau + \int_0^{\infty} e^{-2\pi(\tau-\xi+\tau)} d\tau & \xi < 0 \end{cases} \\ &= \pi^2 \times \begin{cases} \left[\frac{e^{-2\pi(\xi-2\tau)}}{4\pi} \right]_{-\infty}^0 + \left[e^{-2\pi\xi} \cdot \tau \right]_0^{\xi} + \left[\frac{e^{-2\pi(2\tau-\xi)}}{-4\pi} \right]_{\xi}^{\infty} & \xi \geq 0 \\ \left[\frac{e^{-2\pi(\xi-2\tau)}}{4\pi} \right]_{-\infty}^{\xi} + \left[e^{-2\pi(-\xi)} \cdot \tau \right]_{\xi}^0 + \left[\frac{e^{-2\pi(2\tau-\xi)}}{-4\pi} \right]_0^{\infty} & \xi < 0 \end{cases} \\ &= \pi^2 \times \begin{cases} \frac{e^{-2\pi\xi}}{4\pi} + e^{-2\pi\xi} \cdot \xi + \left[0 - \frac{e^{-2\pi\xi}}{-4\pi} \right] & \xi \geq 0 \\ \frac{e^{-2\pi(-\xi)}}{4\pi} + e^{-2\pi(-\xi)} \cdot (-\xi) + \left[0 - \frac{e^{-2\pi(-\xi)}}{-4\pi} \right] & \xi < 0 \end{cases} \\ &= \pi^2 \times \begin{cases} \frac{e^{-2\pi\xi}}{2\pi} + \xi e^{-2\pi\xi} & \xi \geq 0 \\ \frac{e^{-2\pi(-\xi)}}{2\pi} + (-\xi) e^{-2\pi(-\xi)} & \xi < 0 \end{cases} = \frac{\pi e^{-2\pi|\xi|}}{2} + \pi^2 |\xi| e^{-2\pi|\xi|} \end{aligned}$$

□

38.4 Heat Equation

Having learned Fourier transform and its inverse, we are ready to see how we can use them to solve the heat, wave and Laplace equations in infinite and semi-infinite domains starting from this section. There are other methods to solve

The process of solving PDE using Fourier transform and other integral transform methods are shown below. Integral transform can transform a PDE of n variables to a PDE of $n - 1$ or less variables which is easier to solve most of the time. These transform techniques are closely related to complex analysis [Duffy, 2004]. When $n = 1$, the PDE will be transformed to an ODE problem. The inverse Fourier transform is required to get back a representation for u . However, inverse Fourier transform is the hardest because it sometimes requires clever algebra and calculus manipulations.

A simple application of solving the heat equation by Fourier transform is illustrated in the following problem.

Example 38.24 (UCCM3003 Final Exam, Sep 2007, Q4(b)). Consider a heat equation

$$\frac{\partial u}{\partial t} = k \frac{\partial^2 u}{\partial x^2}, \quad -\infty < x < \infty, \quad t > 0, \quad (38.7)$$

subject to the boundary conditions (closely related to $u(x, \cdot) \in L^1(\mathbb{R})$):

$$\lim_{|x| \rightarrow \infty} u(x, t) = 0, \quad (\text{implicitly, } u_x \rightarrow 0 \text{ as } |x| \rightarrow \infty) \quad (38.8)$$

and an initial condition

$$u(x, 0) = e^{-\beta^2 x^2}, \quad -\infty < x < \infty. \quad (38.9)$$

Find the analytic solution $u(x, t)$. (8 marks)

Solution: Before we apply Fourier transform, we need to check the domain. This is a problem with $(-\infty, \infty)$ as domain. So we can take Fourier transform with respect to x of (38.7) to obtain

$$\mathcal{F}\left\{\frac{\partial u}{\partial t}\right\} = \frac{d}{dt} \mathcal{F}\{u(x, t)\} = k \mathcal{F}\left\{\frac{\partial^2 u}{\partial x^2}\right\} = k(2\pi i \xi)^2 U(\xi, t) = -k(2\pi \xi)^2 U(\xi, t) \quad (38.10)$$

according to the “time”-differentiation property in Theorem 38.8. Together with the initial condition, we obtain an ODE-IVP

$$\frac{dU}{dt} + k(2\pi \xi)^2 U(\xi, t) = 0, \quad U(\xi, 0) = \mathcal{F}\{e^{-\beta^2 x^2}\} = \frac{\sqrt{\pi}}{\beta} e^{-\frac{\pi^2 \xi^2}{\beta^2}} \quad (38.11)$$

where $U(\xi, t) = \mathcal{F}\{u(x, t)\}$.

The solution to Equation (38.11) is

$$U(\xi, t) = \frac{\sqrt{\pi}}{\beta} e^{-\frac{\pi^2 \xi^2}{\beta^2}} e^{-(2\pi \xi)^2 kt} = \frac{\sqrt{\pi}}{\beta} \exp\left\{-\pi^2 \xi^2 \left(\frac{1}{\beta^2} + 4kt\right)\right\}$$

Therefore,

$$\begin{aligned} u(x, t) &= \mathcal{F}^{-1}\{U\} = \frac{1}{\beta} \frac{1}{\sqrt{\frac{1}{\beta^2} + 4kt}} \mathcal{F}^{-1}\left\{\frac{\sqrt{\pi}}{\sqrt{\frac{1}{\beta^2} + 4kt}} \exp\left(-\pi^2 \xi^2 \times \frac{1}{\frac{1}{\beta^2} + 4kt}\right)\right\} \\ &= \frac{1}{\sqrt{1 + 4k\beta^2 t}} \exp\left\{-\frac{1}{\beta^2 + 4kt} x^2\right\} = \frac{1}{\sqrt{1 + 4k\beta^2 t}} \exp\left\{-\frac{\beta^2 x^2}{1 + 4k\beta^2 t}\right\} \end{aligned}$$

Example 38.25. Consider a heat equation

$$\frac{\partial u}{\partial t} = 9 \frac{\partial^2 u}{\partial x^2}, \quad -\infty < x < \infty, \quad t > 0,$$

subject to the boundary condition $\lim_{|x| \rightarrow \infty} u(x, t) = 0$ and the initial condition

$$u(x, 0) = e^{-4x^2}, \quad -\infty < x < \infty.$$

Find the analytic solution $u(x, t)$ and determine what happens when $t \rightarrow \infty$. (10 marks)

Solution. Taking the Fourier transform with respect to x of the heat equation, we have

$$\mathcal{F}\left\{\frac{\partial u}{\partial t}\right\} = \frac{d}{dt}\mathcal{F}\{u(x, t)\} = 9\mathcal{F}\left\{\frac{\partial^2 u}{\partial x^2}\right\} = 9(2\pi i\xi)^2 U(\xi, t) = -36\pi^2 \xi U(\xi, t) \quad [2 \text{ marks}]$$

according to the “time”-differentiation property. Together with the initial condition, we obtain an ODE-IVP

$$\frac{dU}{dt} + 36\pi^2 \xi^2 U(\xi, t) = 0, \quad U(\xi, 0) = \mathcal{F}\{e^{-4x^2}\} = \frac{\sqrt{\pi}}{2} e^{-\frac{\pi^2 \xi^2}{4}} \quad [1+2=3 \text{ marks}]$$

where $U(\xi, t) = \mathcal{F}\{u(x, t)\}$.

The solution to the ODE-IVP above is

$$U(\xi, t) = \frac{\sqrt{\pi}}{2} e^{-\frac{\pi^2 \xi^2}{4}} e^{-36\pi^2 \xi^2 t} = \frac{\sqrt{\pi}}{2} \exp\left\{-\left(\frac{1}{4} + 36t\right)\pi^2 \xi^2 t\right\}. \quad [2 \text{ marks}]$$

Therefore,

$$u(x, t) = \mathcal{F}^{-1}\{U\} = \frac{\sqrt{\pi}}{2} \mathcal{F}^{-1}\left\{\exp\left\{-\left(\frac{1}{4} + 36t\right)\pi^2 \xi^2 t\right\}\right\} \quad [1 \text{ mark}]$$

$$= \frac{1}{\sqrt{1+144t}} \exp\left\{-\frac{4x^2}{1+144t}\right\} \quad [1 \text{ mark}]$$

When $t \rightarrow \infty$, $u(x, t) \rightarrow 0$ [1 mark] □

In general if $u(x, 0) = \varphi(x)$ and $u, \varphi, u_x, \varphi_x \rightarrow 0$ as $|x| \rightarrow \infty$, then

$$u(x, t) = \mathcal{F}^{-1}\{\hat{\varphi} e^{-(2\pi\xi)^2 kt}\} = \varphi(t) * \frac{1}{2\sqrt{k\pi t}} \exp\left\{-\frac{x^2}{4kt}\right\}.$$

Example 38.26 (UCCM3003 Final Exam, Apr 2011, Q5(a)). By taking an appropriate Fourier Transform, solve the following heat equation with infinite boundary conditions:

$$\begin{cases} \frac{\partial u}{\partial t} = 16 \frac{\partial^2 u}{\partial x^2}, & -\infty < x < \infty, \quad t > 0, \\ u, \frac{\partial u}{\partial x} \rightarrow 0 \text{ as } |x| \rightarrow \infty, & t > 0, \\ u(x, 0) = f(x), & -\infty < x < \infty \end{cases} \quad (10 \text{ marks})$$

Solution. The steps are similar to those in Example 38.24. The final answer is eq:final:200705:g4b

$$u(x, t) = f(x) * \frac{1}{2\sqrt{kt\pi}} e^{-\frac{1}{4kt}x^2} = \frac{1}{8\sqrt{\pi t}} \int_{-\infty}^{\infty} \exp\left\{-\frac{(x-s)^2}{64t}\right\} f(s) ds.$$

□

Note that when we apply Fourier transform to solve infinite domain, the boundary conditions “ $u, \frac{\partial u}{\partial x} \rightarrow 0$ as $|x| \rightarrow \infty$ ” are always **implicitly assumed**, as was demonstrated in the following examples. e:heat

Does the separation of variables still work? Well, sort of. First, note that the heat equation (38.7) and the boundary conditions (38.8) are homogeneous, so the “principle of superposition” should work. After the separation of variables $u = TX$, we obtain an “sort of” SL problem of X :

$$\frac{T'}{a^2 T} = \frac{X''}{X} = -\lambda \Rightarrow X'' + \lambda X = 0.$$

When $\lambda < 0$, the general solution $X = Ae^{\sqrt{-\lambda}x} + Be^{-\sqrt{-\lambda}x}$ tends to infinity when $|x| \rightarrow \infty$ and we can only have trivial solution. When $\lambda = (2\pi\xi)^2 \geq 0$ (contrast with the finite case, $\lambda = (\frac{n\pi}{L})^2$, $n = 1, 2, \dots$ which is discrete), we can have “bounded solutions”, i.e.

$$X = A(\xi) \cos 2\pi\xi x + B(\xi) \sin 2\pi\xi x.$$

By the principle of superposition, the general solutions “should” be

$$u = \int_0^\infty T_\xi X_\xi d\xi = \int_0^\infty e^{-a^2(2\pi\xi)^2 t} [A(\xi) \cos 2\pi\xi x + B(\xi) \sin 2\pi\xi x] d\xi$$

because ξ is “continuous” and takes values in $\mathbb{R}$. When $t = 0$, the initial condition implies

$$\int_0^\infty [A(\xi) \cos 2\pi\xi x + B(\xi) \sin 2\pi\xi x] d\xi = \int_0^\infty \left[\frac{A(\xi) - iB(\xi)}{2} e^{i2\pi\xi x} + \frac{A(\xi) + iB(\xi)}{2} e^{-i2\pi\xi x} \right] d\xi = \varphi(x) \quad (38.12)$$

e:sv:init

By $\xi \rightarrow -\xi$, part of the integral can be written as

$$\int_0^\infty \frac{A(\xi) + iB(\xi)}{2} e^{-i2\pi\xi x} d\xi = \int_{-\infty}^0 \frac{A(-\xi) + iB(-\xi)}{2} e^{i2\pi\xi x} d\xi.$$

We want to **choose** $A(\xi)$ and $B(\xi)$ so that Fourier Integral Theorem 38.5 could be applied. This could be achieved by having (with a change of variable $t \rightarrow -t$)

$$\begin{aligned} A(-\xi) &= A(\xi) = 2 \int_{-\infty}^\infty \varphi(t) \cos 2\pi\xi t dt = \int_{-\infty}^\infty \varphi(-t) e^{-i2\pi\xi t} dt + \int_{-\infty}^\infty \varphi(t) e^{-i2\pi\xi t} dt, \\ B(-\xi) &= -B(\xi), \quad B(\xi) = 2 \int_{-\infty}^\infty \varphi(t) \sin 2\pi\xi t dt = -i \left[\int_{-\infty}^\infty \varphi(-t) e^{-i2\pi\xi t} dt - \int_{-\infty}^\infty \varphi(t) e^{-i2\pi\xi t} dt \right]. \end{aligned}$$

This implies

$$\frac{A(-\xi) + iB(-\xi)}{2} = \frac{A(\xi) - iB(\xi)}{2} = \int_{-\infty}^\infty \varphi(t) e^{-i2\pi\xi t} dt$$

and equation (38.12) becomes

$$\int_0^\infty \left[\frac{A(\xi) - iB(\xi)}{2} e^{i2\pi\xi x} + \frac{A(\xi) + iB(\xi)}{2} e^{-i2\pi\xi x} \right] d\xi = \int_{-\infty}^\infty \left[\int_{-\infty}^\infty \varphi(x) e^{-i2\pi\xi x} dx \right] e^{i2\pi\xi x} d\xi = \varphi(x)$$

More rigorous theory is related to functional analysis which are introduced in Samuel S. Holland [1990], Showalter [1994], etc.

Example 38.27 (UCCM3003 Final Exam, Sep 2009, Q1(b)). Show that

$$u(x, t) = \frac{1}{2\sqrt{\pi kt}} \int_{-\infty}^\infty \exp\left(-\frac{(x-s)^2}{4kt}\right) f(s) ds \quad (*)$$

is a solution to the heat equation, $\frac{\partial u}{\partial t} = k \frac{\partial^2 u}{\partial x^2}$. (7 marks)

Proof: As mentioned in UECM2023 ODE, to verify that a function is a solution, we need to substitute it into the equation. It is the same case here.

$$\begin{aligned} u_t &= \frac{d}{dt} \left(\frac{1}{2\sqrt{\pi kt}} \right) \int_{-\infty}^\infty \exp\left(-\frac{(x-s)^2}{4kt}\right) f(s) ds + \frac{1}{2\sqrt{\pi kt}} \int_{-\infty}^\infty \frac{d}{dt} \left(\exp\left(-\frac{(x-s)^2}{4kt}\right) \right) f(s) ds \\ &\quad (\text{a rigorous proof requires the notion of uniform convergence which should be taught in Calculus 3 or Elementary Real Analysis}) \\ &= -\frac{1}{2} \times \frac{1}{2\sqrt{\pi k}} t^{-3/2} \int_{-\infty}^\infty \exp\left(-\frac{(x-s)^2}{4kt}\right) f(s) ds + \\ &\quad \int_{-\infty}^\infty \exp\left(-\frac{(x-s)^2}{4kt}\right) \left(-\frac{(x-s)^2}{4k} \times t^{-2} \times (-1) \right) f(s) ds \\ &= -\frac{1}{4\sqrt{\pi kt}} \times \frac{1}{t} \int_{-\infty}^\infty \exp\left(-\frac{(x-s)^2}{4kt}\right) f(s) ds + \frac{1}{2\sqrt{\pi kt}} \int_{-\infty}^\infty \frac{(x-s)^2}{4kt^2} \exp\left(-\frac{(x-s)^2}{4kt}\right) f(s) ds \\ u_x &= \frac{1}{2\sqrt{\pi kt}} \left[\int_{-\infty}^\infty e^{-\frac{(x-s)^2}{4kt}} \left(\frac{-2(x-s)}{4kt} \right) f(s) ds \right] \\ u_{xx} &= \frac{1}{2\sqrt{\pi kt}} \left[\int_{-\infty}^\infty \left(e^{-\frac{(x-s)^2}{4kt}} \left(\frac{-2(x-s)}{4kt} \right)^2 + e^{-\frac{(x-s)^2}{4kt}} \left(\frac{-2}{4kt} \right) \right) f(s) ds \right] \\ &= \frac{1}{2\sqrt{\pi kt}} \int_{-\infty}^\infty e^{-\frac{(x-s)^2}{4kt}} \frac{(x-s)^2}{4k^2 t^2} f(s) ds - \frac{1}{4\sqrt{\pi kt}} \int_{-\infty}^\infty e^{-\frac{(x-s)^2}{4kt}} \frac{1}{kt} f(s) ds \\ &= \frac{1}{k} \left[\frac{1}{2\sqrt{\pi kt}} \int_{-\infty}^\infty \frac{(x-s)^2}{4kt^2} e^{-\frac{(x-s)^2}{4kt}} f(s) ds - \frac{1}{4\sqrt{\pi kt}} \times \frac{1}{t} \int_{-\infty}^\infty e^{-\frac{(x-s)^2}{4kt}} f(s) ds \right] = \frac{1}{k} u_t. \quad \square \end{aligned}$$

Remark: Strauss [1992, Section 2.4] avoid the differentiation and use a different approach to show that (*) is a solution.

Example 38.28 (UECM3003 Final Exam, Apr 2014, Q5(b)). Consider the heat flow in an infinite long homogeneous bar modelled by

$$\frac{\partial u}{\partial t} = k \frac{\partial^2 u}{\partial x^2}, \quad -\infty < x < \infty, \quad t > 0, \quad u(x, 0) = f(x) = \begin{cases} 1, & 0 \leq x \leq 1, \\ -1, & -1 \leq x < 0, \\ 0, & |x| > 1. \end{cases}$$

Find the solution to the above problem using Fourier transform method. (10 marks)

Solution. Taking Fourier transform with respect to x , we transform the PDE to an ODE

$$\frac{d}{dt}U(\xi, t) = k(2\pi i\xi)^2 U(\xi, t) = -4k\pi^2 \xi^2 U(\xi, t)$$

The initial condition becomes (is the following calculation necessary?)

$$\begin{aligned} U(\xi, 0) &= \int_{-1}^0 (-1)e^{-2\pi i\xi x} dx + \int_0^1 e^{-2\pi i\xi x} dx = -\frac{1 - e^{-2\pi i\xi}}{-2\pi i\xi} + \frac{e^{-2\pi i\xi} - 1}{-2\pi i\xi} \\ &= \frac{2 - e^{2\pi i\xi} - e^{-2\pi i\xi}}{2\pi i\xi} = \frac{2 - 2\cos(2\pi\xi)}{2\pi i\xi} = \frac{1 - \cos(2\pi\xi)}{\pi i\xi} \end{aligned}$$

The solution of the ODE-IVP is

$$\begin{aligned} U(\xi, t) &= \frac{1 - \cos(2\pi\xi)}{\pi i\xi} e^{-4k\pi^2 \xi^2 t} \\ \Rightarrow u(x, t) &= f(x) * \frac{1}{2\sqrt{k\pi t}} e^{-\frac{x^2}{4kt}} = \frac{1}{2\sqrt{k\pi t}} \left[\int_0^1 e^{-\frac{(x-s)^2}{4kt}} ds - \int_{-1}^0 e^{-\frac{(x-s)^2}{4kt}} ds \right] \end{aligned}$$

□

Example 38.29 (UECM3003 Final Exam, Apr 2015, Q4(b)). Find a solution to the following boundary value problem

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2}, \quad -\infty < x < \infty, \quad t \geq 0, \\ u(x, 0) &= 3 + 2\cos 3x, \quad -\infty < x < \infty. \end{aligned}$$

[Hint: $\int_{-\infty}^{\infty} e^{-\omega^2} d\omega = \sqrt{\pi}$.]

(10 marks)

Solution: The solution is

$$\begin{aligned} u(x, t) &= (3 + 2\cos 3x) * \frac{1}{2\sqrt{\pi t}} e^{-\frac{x^2}{4t}} \\ &= 3 \int_{-\infty}^{\infty} \frac{1}{2\sqrt{\pi t}} e^{-\frac{s^2}{4t}} ds + 2 \int_{-\infty}^{\infty} \frac{1}{2\sqrt{\pi t}} e^{-\frac{s^2}{4t}} \cos 3(x-s) ds \\ &= 3 \frac{\sqrt{\pi}}{\sqrt{\pi}} + \frac{1}{\sqrt{\pi t}} \int_{-\infty}^{\infty} e^{-\frac{s^2}{4t}} \cos 3(x-s) ds = 3 + \frac{1}{\sqrt{\pi t}} \int_{-\infty}^{\infty} e^{-\frac{s^2}{4t}} \cos 3(x-s) ds \end{aligned}$$

When the domain is $\mathbb{R}^n$, there is a general formula for the solution of n -dimensional heat equation in convolution form given below.

Theorem 38.30 (The solution of global Cauchy problem). Assume that $g(\mathbf{x})$ is a continuous function on $\mathbb{R}^n$ that verifies the bounds $|g(\mathbf{x})| \leq ae^{b|\mathbf{x}|^2}$, where $a, b > 0$ are constants. For the homogeneous heat equation

$$\begin{aligned} u_t - k\Delta u &= 0, \quad (\mathbf{x}, t) \in \mathbb{R}^n \times (0, \infty), \\ u(\mathbf{x}, 0) &= g(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n, \end{aligned} \tag{38.13}$$

there is a solution $u(\mathbf{x}, t)$ for $0 \leq t < \frac{1}{4kb}$ in the following form

$$u(\mathbf{x}, t) = \frac{1}{(4\pi kt)^{n/2}} \int_{\mathbb{R}^n} e^{-\frac{|\mathbf{x}-\xi|^2}{4kt}} g(\xi) d\xi = \frac{1}{(4\pi kt)^{n/2}} e^{-\frac{|\mathbf{x}|^2}{4kt}} * g(x). \tag{38.14}$$

Here, $\frac{1}{(4\pi kt)^{n/2}} e^{-\frac{|\mathbf{x}|^2}{4kt}}$ is called the fundamental solution of (38.13). The solution u is of regularity $C^\infty(\mathbb{R}^n \times (0, \frac{1}{4kb}))$. For each compact subinterval $[0, T'] \subset [0, \frac{1}{4kb})$, there exist constants $A, B > 0$ (depending on the compact subinterval) such that

$$|u(\mathbf{x}, t)| \leq Ae^{B|\mathbf{x}|^2}, \quad \forall (\mathbf{x}, t) \in \mathbb{R}^n \times [0, T']. \tag{38.15}$$

The solution u is the unique solution in the class of functions verifying a bound of the form (38.15).

e:heat:cauchy

e:heat:cauchy:so

e:heat:cauchy:bo

e:heat:cauchy:bound

When the domain is semi-infinite, i.e. $0 < x < \infty$, we can no longer apply Fourier transform directly. Instead, we need to determine the type of the linear boundary condition and extend the problem to infinite case by either odd extension or even extension [Strauss, 1992, Section 3.1] as in Topic 1 to functions defined over half interval $[0, L]$. This is equivalent to applying either a Fourier sine transform or a Fourier cosine transform. Alternatively, we can apply the Fourier Integral Theorem 38.5.

Example 38.31. (Haberman [1987, p. 360]) Show that

$$u(x, t) = \frac{1}{2\sqrt{k\pi t}} \int_0^\infty \phi(y) \left[e^{-\frac{(x-y)^2}{4kt}} - e^{-\frac{(x+y)^2}{4kt}} \right] dy$$

is a solution of the following boundary value problem

$$\begin{cases} u_t - ku_{xx} = 0, & 0 < x < \infty, t \geq 0, \\ u(0, t) = 0, & 0 < x < \infty, \\ u(x, 0) = \phi(x), & 0 < x < \infty. \end{cases}$$

Solution. Since we have a type 1 boundary condition, we perform **odd extension** to u and ϕ to obtain u^o and ϕ^o . Now taking Fourier transform and let $\mathcal{F}\{u^o\} = U(\xi, t)$, we have

$$\frac{d}{dt}U(\xi, t) = k(2\pi i\xi)^2 U(\xi, t) = -k(2\pi\xi)^2 U(\xi, t) \Rightarrow U(\xi, t) = Ae^{-k(2\pi\xi)^2 t}.$$

When $t = 0$, from the initial condition, we have

$$U(\xi, 0) = A = \mathcal{F}\{\phi^o\} \Rightarrow U(\xi, t) = \mathcal{F}\{\phi^o\} e^{-k(2\pi\xi)^2 t}.$$

Taking inverse Fourier transform and applying Table 38.2 and (??),

$$\begin{aligned} u &= \phi^o * \mathcal{F}^{-1}\left\{e^{-\frac{\pi^2 \xi^2}{4kt}}\right\} = \phi^o * \sqrt{\frac{1}{4kt}} e^{-\frac{1}{4kt}x^2} = \phi^o * \frac{1}{2\sqrt{k\pi t}} e^{-\frac{x^2}{4kt}} \\ &= \frac{1}{2\sqrt{k\pi t}} \int_{-\infty}^\infty \phi^o(y) e^{-\frac{(x-y)^2}{4kt}} dy = \frac{1}{2\sqrt{k\pi t}} \left[\int_{-\infty}^0 (-\phi(-y)) e^{-\frac{(x-y)^2}{4kt}} dy + \int_0^\infty \phi(y) e^{-\frac{(x-y)^2}{4kt}} dy \right] \\ &= \frac{1}{2\sqrt{k\pi t}} \left[\int_0^\infty \phi(y) e^{-\frac{(x-y)^2}{4kt}} dy - \int_{-\infty}^0 \phi(\tau) e^{-\frac{(x+\tau)^2}{4kt}} d(-\tau) \right] \quad (\tau = -y) \\ &= \frac{1}{2\sqrt{k\pi t}} \left[\int_0^\infty \phi(y) e^{-\frac{(x-y)^2}{4kt}} dy - \int_0^\infty \phi(y) e^{-\frac{(x+y)^2}{4kt}} dy \right] \quad (\tau \rightarrow y) \\ &= \frac{1}{2\sqrt{k\pi t}} \int_0^\infty \phi(y) \left[e^{-\frac{(x-y)^2}{4kt}} - e^{-\frac{(x+y)^2}{4kt}} \right] dy. \end{aligned}$$

□

Example 38.32. (Strauss [1992, Section 3.1]) Solve the following boundary value problem

$$\begin{cases} u_t - ku_{xx} = 0, & 0 < x < \infty, t \geq 0, \\ u_x(0, t) = 0, & 0 < x < \infty, \\ u(x, 0) = \phi(x), & 0 < x < \infty. \end{cases}$$

Solution. Since we have a type 2 boundary condition, we perform **even extension** to u and ϕ to obtain u^e and ϕ^e . Now taking Fourier transform and let $\mathcal{F}\{u^e\} = U(\xi, t)$, we have

$$\frac{d}{dt}U(\xi, t) = k(2\pi i\xi)^2 U(\xi, t) = -k(2\pi\xi)^2 U(\xi, t) \Rightarrow U(\xi, t) = Ae^{-k(2\pi\xi)^2 t}.$$

When $t = 0$, from the initial condition, we have

$$U(\xi, 0) = A = \mathcal{F}\{\phi^e\} \Rightarrow U(\xi, t) = \mathcal{F}\{\phi^e\} e^{-k(2\pi\xi)^2 t}.$$

Taking inverse Fourier transform and applying Table 38.2 and (77), [tbl:ft:spacefrconvolution](#)

$$\begin{aligned}
 u &= \phi^e * \frac{1}{2\sqrt{k\pi t}} e^{-\frac{x^2}{4kt}} = \frac{1}{2\sqrt{k\pi t}} \int_{-\infty}^{\infty} \phi^e(y) e^{-\frac{(x-y)^2}{4kt}} dy \\
 &= \frac{1}{2\sqrt{k\pi t}} \left[\int_{-\infty}^0 \phi(-y) e^{-\frac{(x-y)^2}{4kt}} dy + \int_0^{\infty} \phi(y) e^{-\frac{(x-y)^2}{4kt}} dy \right] \\
 &= \frac{1}{2\sqrt{k\pi t}} \left[\int_{\infty}^0 \phi(\tau) e^{-\frac{(x+\tau)^2}{4kt}} d(-\tau) + \int_0^{\infty} \phi(y) e^{-\frac{(x-y)^2}{4kt}} dy \right] \quad (\tau = -y) \\
 &= \frac{1}{2\sqrt{k\pi t}} \left[\int_0^{\infty} \phi(y) e^{-\frac{(x-y)^2}{4kt}} dy + \int_0^{\infty} \phi(y) e^{-\frac{(x+y)^2}{4kt}} dy \right] \quad (\tau \rightarrow y) \\
 &= \frac{1}{2\sqrt{k\pi t}} \int_0^{\infty} \phi(y) \left[e^{-\frac{(x-y)^2}{4kt}} + e^{-\frac{(x+y)^2}{4kt}} \right] dy.
 \end{aligned}$$

□

Final:201301q5b

Example 38.33 (UECM3003 Final Exam, Apr 2013, Q5(b)). Show that the solution of the boundary value problem

$$\begin{cases} u_t = 4u_{xx}, & 0 < x < \infty, t \geq 0, \\ u_x(0, t) = 0, & 0 < x < \infty, \\ u(x, 0) = e^{-2x}, & 0 < x < \infty, \end{cases}$$

can be written as $u(x, t) = \frac{4}{\pi} \int_0^{\infty} \frac{e^{-4\omega^2 t}}{4 + \omega^2} \cos \omega x d\omega$.

[Hint: $\mathcal{F}_c\{e^{-ax}\} = \sqrt{\frac{2}{\pi}} \frac{a}{a^2 + \omega^2}$.] (15 marks)

Solution: Since we have a type 2 boundary condition, we perform **even extension** to u and e^{-2x} to obtain u^e and $e^{-2|x|}$. Now taking Fourier transform and let $\mathcal{F}\{u^e\} = U(\xi, t)$, we have

$$\frac{d}{dt} U(\xi, t) = 4(2\pi i \xi)^2 U(\xi, t) = -4(2\pi \xi)^2 U(\xi, t) \Rightarrow U(\xi, t) = A e^{-4(2\pi \xi)^2 t}.$$

From the initial condition, we have

$$U(\xi, 0) = A = \mathcal{F}\{e^{-2|x|}\} = \frac{4}{4 + (2\pi \xi)^2} \Rightarrow U(\xi, t) = \frac{4}{4 + (2\pi \xi)^2} e^{-4(2\pi \xi)^2 t}.$$

Taking inverse Fourier transform and with a change of variable $\omega = 2\pi \xi$,

$$\begin{aligned}
 u(x, t) &= \int_{-\infty}^{\infty} \frac{4}{4 + (2\pi \xi)^2} e^{-4(2\pi \xi)^2 t} e^{2\pi i \xi x} d\xi = \int_{-\infty}^{\infty} \frac{4}{4 + (2\pi \xi)^2} e^{-4(2\pi \xi)^2 t} \cos(2\pi i \xi x) d\xi \\
 &= 2 \int_0^{\infty} \frac{4}{4 + \omega^2} e^{-4\omega^2 t} \cos \omega x \frac{d\omega}{2\pi} = \frac{4}{\pi} \int_0^{\infty} \frac{e^{-4\omega^2 t}}{4 + \omega^2} \cos \omega x d\omega.
 \end{aligned}$$

Alternative Solution: Since we have a type 2 boundary condition, we will use the **Fourier cosine transform** on the heat equation

$$\frac{d}{dt} \int_0^{\infty} u(x, t) \cos(2\pi \xi x) dx = 4 \left[\underbrace{u_x \cos(2\pi \xi x)}_{=0} \Big|_0^{\infty} + (2\pi \xi) \int_0^{\infty} u_x \sin(2\pi \xi x) dx \right]$$

since $u_x(0, t) = 0$ and $u_x \rightarrow 0$ as $x \rightarrow \infty$. Applying integration by part, we have

$$\frac{d}{dt} \int_0^{\infty} u \cos(2\pi \xi x) dx = 4(2\pi \xi) \left[u \sin(2\pi \xi x) \Big|_0^{\infty} - (2\pi \xi) \int_0^{\infty} u \cos(2\pi \xi x) dx \right]$$

which simplifies to $\frac{d}{dt} U_c(\xi, t) = -16\pi^2 \xi^2 U_c(\xi, t)$

if we let $U_c(\xi, t) = \int_0^\infty u \cos(2\pi\xi x) dx$. The solution of the ODE is

$$\begin{aligned} U_c(\xi, t) &= A(\xi) e^{-16\pi^2 \xi^2 t} \\ \Rightarrow U_c(\xi, 0) &= A(\xi) = \int_0^\infty e^{-2x} \cos(2\pi\xi x) dx = \frac{e^{-2x}}{4 + (2\pi\xi)^2} (-2 \cos 2\pi\xi x + 2\pi\xi \sin 2\pi\xi x) \Big|_0^\infty \\ &= 0 - \frac{-2}{4 + (2\pi\xi)^2} \quad \text{since} \quad \int e^{ax} \cos bx dx = \frac{e^{ax}}{a^2 + b^2} (a \cos bx + b \sin bx) \end{aligned}$$

By the Fourier Integral Theorem and let $\omega = 2\pi\xi$, we have

$$u(x, t) = 4 \int_0^\infty \frac{2}{4 + (2\pi\xi)^2} e^{-16\pi^2 \xi^2 t} \cos 2\pi\xi x d\xi = \frac{4}{\pi} \int_0^\infty \frac{e^{-4\omega^2 t}}{4 + \omega^2} \cos \omega x d\omega.$$

For non-homogeneous heat equation, we can use the *Duhamel's principle* [Logan, 2015, Section 2.5] to reduce the non-homogeneous PDE to homogeneous PDE so that Fourier transform could be applied.

38.5 Laplace Equation

Fourier transform is a very popular method for solving Laplace equation when the domain is of infinite extent in one direction. In the case of semi-infinite domains, an important question is whether we should use Fourier cosine or sine transform. The answer lies in the nature of the boundary conditions (type 1 or type 2). The choice must be such that repeated integration by parts can be evaluated with known quantities.

In the following examples involving infinite domains, we are using Fourier transform (or Fourier Integral Theorem (38.5)) instead of the Fourier sine or cosine transforms as in Duffy [2004]. When there is type 1 or 2 BCs, odd or even extension will be introduced before the application of Fourier transform. The method used here is probably more convenient compare to looking up in Fourier sine and cosine transforms tables.

Example 38.34 (UCCM3003 Final Exam, Apr 2009, Q6(c), with modification). Let $u(x, y)$ be the solution to the Laplace's equation $\nabla^2 u = 0$ in a domain $-\infty < x < \infty$, $y > 0$, the boundary conditions $u(x, 0) = g(x)$ and as $x^2 + y^2 \rightarrow \infty$, $u \rightarrow 0$.

- (i) Show that $U(\xi, y)$, the Fourier transform of $u(x, y)$ with respect to x , is $U(\xi, y) = G(\xi) e^{-2\pi|\xi|y}$. (8 marks)

Proof: Note that $u \rightarrow 0$ as $x^2 + y^2 \rightarrow \infty$ is the condition where we could apply the Fourier transform.

Since the domain of x is $\mathbb{R}$, we will take the Fourier transform with respect to x and let $\mathcal{F}\{u\} = \int_{-\infty}^\infty u(x, y) e^{-2\pi i \xi x} dx =: U(\xi, y)$, then

$$\begin{aligned} \mathcal{F}\{u_{xx} + u_{yy}\} &= \mathcal{F}\{0\} \Rightarrow (2\pi i \xi)^2 U(\xi, y) + \frac{d^2 y}{dy^2} U(\xi, y) = 0 \\ \mathcal{F}\{u(x, 0)\} &= \mathcal{F}\{g(x)\} \Rightarrow U(\xi, 0) = G(\xi) \end{aligned}$$

To have $u \rightarrow 0$ as $x \rightarrow \pm\infty$, we choose the general solution of the $U(\xi, y)$ -ODE to be

$$\frac{d^2 y}{dy^2} U(\xi, y) - 4\pi^2 \xi^2 U(\xi, y) = 0 \Rightarrow U(\xi, y) = A e^{2\pi|\xi|y} + B e^{-2\pi|\xi|y}$$

Because $U(\xi, y)$ **needs to well-defined** for $y > 0$, $A = 0$. Therefore,

$$U(\xi, 0) = B e^0 = G(\xi) \Rightarrow U(\xi, y) = G(\xi) e^{-2\pi|\xi|y}.$$

- (ii) Show that a solution of the above Laplace equation is

$$u(x, y) = \frac{y}{\pi} \int_{-\infty}^\infty \frac{g(t)}{y^2 + (x - t)^2} dt. \quad (7 \text{ marks})$$

Proof: The solution could be obtained by taking inverse Fourier transform:

$$\begin{aligned} u(x, y) &= \mathcal{F}^{-1}\{U(\xi, y)\} = \mathcal{F}^{-1}\{G(\xi)e^{-2\pi|\xi|y}\} \\ &= \mathcal{F}^{-1}\{G(\xi)\} * \mathcal{F}^{-1}\{e^{-(2\pi y)|-\xi|}\} = g(x) * \frac{2(2\pi y)}{(2\pi y)^2 + (2\pi x)^2} = g(x) * \frac{y}{\pi(y^2 + x^2)} \\ &= \int_{-\infty}^{\infty} g(\tau) \cdot \frac{y}{\pi(y^2 + (x - \tau)^2)} d\tau = \frac{y}{\pi} \int_{-\infty}^{\infty} \frac{g(\tau)}{y^2 + (x - \tau)^2} d\tau \end{aligned}$$

Example 38.35. (Haberman [1987, Section 9.7.2]) Solve the Laplace equation on a semi-infinite strip $(0, L) \times (0, \infty)$:

$$u_{xx} + u_{yy} = 0, \quad 0 < x < L, \quad 0 < y < \infty$$

subject to the boundary conditions

$$u(0, y) = g_1(y), \quad u(L, y) = g_2(y), \quad u(x, 0) = f(x).$$

Here, $g_1(x) \rightarrow 0$ and $g_2(y) \rightarrow 0$ as $y \rightarrow \infty$.

Solution: Compare to Example 38.34, the domain of this problem is semi-infinite, hence we have to either use Fourier sine transform or apply odd extension. Because the “formula sheet” only includes the Fourier transform, we will apply the method of odd extension in this example.

Similar to Example 2.5.5 (Topic 2), we need to decompose the problem to the situation where all except a pair of boundary conditions are homogeneous. In this case, we need to let $u = v + w$, $(x, y) \in (0, L) \times (0, \infty)$, such that

$$\begin{cases} v_{xx} + v_{yy} = 0, \\ v(0, y) = g_1(y), \quad v(L, y) = g_2(y), \\ v(x, 0) = 0. \end{cases} \quad \text{and} \quad \begin{cases} w_{xx} + w_{yy} = 0, \\ w(0, y) = 0, \quad w(L, y) = 0, \\ w(x, 0) = f(x). \end{cases}$$

By applying the odd-extension to v , g_1 and g_2 so that $v^o(x, -y) = -v^o(x, y)$, $g_1^o(-y) = -g_1^o(y)$, $g_2^o(-y) = -g_2^o(y)$ and let $\mathcal{F}\{v^o(x, y)\} = V(x, \xi)$, $\mathcal{F}\{g_i^o(y)\} = G_i^o(\xi)$. Then

$$\frac{d^2}{dx^2} V(x, \xi) + (2\pi i \xi)^2 V(x, \xi) = \frac{d^2}{dx^2} V(x, \xi) - 4\pi^2 \xi^2 V(x, \xi) = 0.$$

It has a general solution $V(x, \xi) = Ae^{2\pi \xi x} + Be^{-2\pi \xi x}$.

From the non-homogeneous boundary conditions, we have

$$V(0, \xi) = A + B = G_1^o(\xi), \quad V(L, \xi) = Ae^{2\pi \xi L} + Be^{-2\pi \xi L} = G_2^o(\xi).$$

By solving the above simultaneous equation, we have

$$A = \frac{\begin{vmatrix} G_1^o(\xi) & 1 \\ G_2^o(\xi) & e^{-2\pi \xi L} \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ e^{2\pi \xi L} & e^{-2\pi \xi L} \end{vmatrix}} = \frac{G_1^o(\xi)e^{-2\pi \xi L} - G_2^o(\xi)}{-2 \sinh(2\pi \xi L)}, \quad B = \frac{\begin{vmatrix} 1 & G_1^o(\xi) \\ e^{2\pi \xi L} & G_2^o(\xi) \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ e^{2\pi \xi L} & e^{-2\pi \xi L} \end{vmatrix}} = \frac{G_2^o(\xi) - G_1^o(\xi)e^{2\pi \xi L}}{-2 \sinh(2\pi \xi L)}.$$

Hence,

$$\begin{aligned} V(x, \xi) &= \frac{G_2^o(\xi) - G_1^o(\xi)e^{-2\pi \xi L}}{2 \sinh(2\pi \xi L)} e^{2\pi \xi x} + \frac{G_1^o(\xi)e^{2\pi \xi L} - G_2^o(\xi)}{2 \sinh(2\pi \xi L)} e^{-2\pi \xi x} \\ &= G_1^o(\xi) \frac{e^{2\pi \xi(L-x)} - e^{-2\pi \xi(L-x)}}{2 \sinh(2\pi \xi L)} + G_2^o(\xi) \frac{e^{2\pi \xi x} - e^{-2\pi \xi x}}{2 \sinh(2\pi \xi L)} \\ &= G_1^o(\xi) \frac{\sinh(2\pi \xi(L-x))}{\sinh(2\pi \xi L)} + G_2^o(\xi) \frac{\sinh(2\pi \xi x)}{\sinh(2\pi \xi L)} \end{aligned}$$

Both G_1^o and G_2^o need to be odd and for $j = 1, 2$,

$$\begin{aligned} G_j^o(\xi) &= \int_{-\infty}^{\infty} g_j^o(y) e^{-2\pi i \xi y} dy = \int_{-\infty}^{\infty} g_j^o(y) (-i) \sin(2\pi i \xi y) dy \\ &= 2(-i) \int_0^{\infty} g_j(y) \sin(2\pi i \xi y) dy = (-i) G_j(\xi), \quad 0 < \xi < \infty \end{aligned}$$

where G_j are the Fourier sine transform (Definition [38.2](#)) of g_j , $j = 1, 2$. Hence,

$$\begin{aligned} v(x, y) &= \int_{-\infty}^{\infty} \left[\frac{\sinh(2\pi\xi(L-x))}{\sinh(2\pi\xi L)} G_1^o(\xi) + \frac{\sinh(2\pi\xi x)}{\sinh(2\pi\xi L)} G_2^o(\xi) \right] e^{2\pi i \xi y} d\xi \\ &= 2 \int_0^{\infty} \left[\frac{\sinh(2\pi\xi(L-x))}{\sinh(2\pi\xi L)} (-i) G_1(\xi) + \frac{\sinh(2\pi\xi x)}{\sinh(2\pi\xi L)} (-i) G_2(\xi) \right] i \sin(2\pi\xi y) d\xi \\ &= 2 \int_0^{\infty} \left[\frac{\sinh(2\pi\xi(L-x))}{\sinh(2\pi\xi L)} G_1(\xi) + \frac{\sinh(2\pi\xi x)}{\sinh(2\pi\xi L)} G_2(\xi) \right] \sin(2\pi\xi y) d\xi. \end{aligned}$$

For the w -problem, the x -direction is finite, therefore, we can apply the separation of variables $u = XY$ to obtain the eigenvalues and eigenfunctions for X as in Topic 2:

$$\begin{cases} \frac{X''}{X} = -\frac{Y''}{Y} = -\lambda \\ X(0) = X(L) = 0 \end{cases} \Rightarrow \lambda = \left(\frac{n\pi}{L}\right)^2, \quad X = A_n \sin \frac{n\pi x}{L}, \quad n = 1, 2, \dots$$

Solving for Y , we have $Y = C_n e^{n\pi y/L} + D_n e^{-n\pi y/L}$. Therefore, the general solution is

$$w(x, y) = \sum_{n=1}^{\infty} \left[a_n e^{n\pi y/L} + b_n e^{-n\pi y/L} \right] \sin \frac{n\pi x}{L}.$$

The requirement that $w \rightarrow 0$ as $y \rightarrow \infty$ requires that $a_n = 0$ and

$$\begin{aligned} w(x, 0) &= \sum_{n=1}^{\infty} b_n \sin \frac{n\pi x}{L} = f(x) \Rightarrow b_n = \frac{2}{L} \int_0^L f(x) \sin \frac{n\pi x}{L} dx. \\ w(x, y) &= \sum_{n=1}^{\infty} \frac{2}{L} e^{-n\pi y/L} \sin \frac{n\pi x}{L} \int_0^L f(x) \sin \frac{n\pi x}{L} dx. \end{aligned}$$

Example 38.36 (UECM3003 Final Exam, Apr 2015, Q5(c)). Solve the Laplace's equation on a semi-infinite strip

$$u_{xx} + u_{yy} = 0, \quad 0 < x < \infty, \quad 0 < y < b$$

subject to the boundary conditions $u(0, y) = 0$, $\lim_{x \rightarrow \infty} u(x, y) = 0$, $0 < y < b$, and $u(x, 0) = 0$, $u(x, b) = e^{-x}$, $0 < x < \infty$. (11 marks)

Solution. By using the result from Example [38.35](#), we have

$$u(x, y) = 2 \int_0^{\infty} \left[\frac{\sinh(2\pi\xi(b-y))}{\sinh(2\pi\xi b)} G_1(\xi) + \frac{\sinh(2\pi\xi y)}{\sinh(2\pi\xi b)} G_2(\xi) \right] \sin(2\pi\xi x) d\xi$$

where $g_1(x) = 0$, $g_2(x) = e^{-x}$ and so $G_1(\xi) = 0$ and according to [Bronshtein et al., 2007, 1094]

$$\begin{aligned} G_2(\xi) &= 2 \int_0^{\infty} e^{-x} \sin 2\pi\xi x dx = 2 \left[\frac{e^{-x}}{(-1)^2 + (2\pi\xi)^2} ((-1) \sin 2\pi\xi x - 2\pi\xi \cos 2\pi\xi x) \right]_0^{\infty} \\ &= 2 \left[0 - \frac{(-0 - 2\pi\xi)}{1 + (2\pi\xi)^2} \right] = \frac{4\pi\xi}{1 + (2\pi\xi)^2} \end{aligned}$$

Therefore, the solution is $u(x, y) = 4 \int_0^{\infty} \left[\frac{\sinh(2\pi\xi y)}{\sinh(2\pi\xi b)} \frac{2\pi\xi}{1 + (2\pi\xi)^2} \right] \sin(2\pi\xi x) d\xi$. □

Example 38.37. Consider a Laplace equation on the semi-infinite strip

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \quad 0 < x < \infty, \quad 0 < y < b, \quad (38.16) \quad \text{e:3.4.1}$$

subject to the boundary conditions

$$u_x(0, y) = 0, \quad \lim_{x \rightarrow \infty} u(x, y) = \lim_{x \rightarrow \infty} u_x(x, y) = 0, \quad 0 < y < b, \quad (38.17) \quad \text{e:3.4.2}$$

and

$$u_y(x, 0) = 0, \quad u(x, b) = f(x), \quad 0 < x < \infty. \quad (38.18) \quad \text{e:3.4.3}$$

Solution:

to introduce even extension $\bar{u}$, f^e for the type 2 boundary condition (similar to the use of Fourier cosine series for even-extended function in Topic 1). In this case, (38.16)–(38.17) becomes

$$\begin{cases} \bar{u}_{xx} + \bar{u}_{yy} = 0 & -\infty < x < \infty, \quad 0 < y < b, \\ \bar{u}(-x, y) = \bar{u}(x, y), \quad \bar{u}, \bar{u}_x \rightarrow 0 \text{ as } |x| \rightarrow \infty, & 0 < y < b, \\ \bar{u}_y(x, 0) = 0, \quad \bar{u}(x, b) = f^e(x), & -\infty < x < \infty. \end{cases}$$

Taking Fourier transform of the PDE with respect to x and let $\mathcal{F}\{\bar{u}\} = U(\xi, y)$, we have

$$\frac{d^2}{dy^2} U(\xi, y) + (2\pi i \xi)^2 U(\xi, y) = 0 \Rightarrow U(\xi, y) = Ae^{2\pi \xi y} + Be^{-2\pi \xi y}$$

Taking Fourier transform of the boundary conditions with respect to x , we have

$$\begin{cases} \frac{d}{dy} U(\xi, 0) = 0 \Rightarrow 2\pi \xi A e^{2\pi \xi 0} + (-2\pi \xi) B e^{-2\pi \xi 0} = 0 \Rightarrow A = B \\ U(\xi, b) = \mathcal{F}\{f^e(x)\} =: F^e(\xi) \Rightarrow 2A \cosh(2\pi \xi b) = F^e(\xi) \end{cases}$$

Therefore,

$$U(\xi, y) = \frac{F^e(\xi) \cosh(2\pi \xi y)}{\cosh(2\pi \xi b)} \Rightarrow u(x, y) = \mathcal{F}^{-1}\left\{\frac{F^e(\xi)}{\cosh(2\pi \xi b)} \cosh(2\pi \xi y)\right\}$$

Note that $F^e(\xi)$ and hyperbolic cosine functions are even with respect to ξ , therefore,

$$\begin{aligned} F^e(\xi) &= \int_{-\infty}^{\infty} f^e(x) e^{-2\pi i \xi x} dx = 2 \int_0^{\infty} f^e(x) \cos(2\pi \xi x) dx = F_c(\xi) \text{ for } 0 < \xi < \infty, \\ u(x, y) &= \int_{-\infty}^{\infty} \frac{F^e(\xi) \cosh(2\pi \xi y)}{\cosh(2\pi \xi b)} e^{2\pi i \xi x} d\xi = \int_{-\infty}^{\infty} \frac{F^e(\xi) \cosh(2\pi \xi y)}{\cosh(2\pi \xi b)} \cos(2\pi \xi x) d\xi \\ &= 2 \int_0^{\infty} \frac{F_c(\xi) \cosh(2\pi \xi y)}{\cosh(2\pi \xi b)} \cos(2\pi \xi x) d\xi \end{aligned}$$

Alternative Solution: An alternative technique will be to apply the Fourier Integral Theorem (38.5). First, we notice that the domain is semi-infinite in x -direction, so the Fourier transform will be applied to “variable x ”. By looking at the homogeneous boundary conditions, a Fourier cosine transform of x is required. Following Duffy [2004, Example 3.4.1, p. 251], we begin by multiplying (38.16) with $\cos(2\pi \xi x)$ and integrating from 0 to ∞ . This yields

$$\int_0^{\infty} \frac{\partial^2 u}{\partial y^2} \cos(2\pi \xi x) dx + \int_0^{\infty} \frac{\partial^2 u}{\partial x^2} \cos(2\pi \xi x) dx = 0 \quad (38.19)$$

or

$$\frac{d^2}{dy^2} \int_0^{\infty} u \cos(2\pi \xi x) dx + \int_0^{\infty} \frac{\partial^2 u}{\partial x^2} \cos(2\pi \xi x) dx = 0 \quad (38.20)$$

Integrating the u_{xx} term of (38.20) twice by parts,

$$\begin{aligned} \int_0^{\infty} \frac{\partial^2 u}{\partial x^2} \cos(2\pi \xi x) dx &= \frac{\partial u}{\partial x} \cos(2\pi \xi x) \Big|_{x=0}^{x=\infty} + 2\pi \xi \int_0^{\infty} \frac{\partial u}{\partial x} \sin(2\pi \xi x) dx \\ &= 2\pi \xi \left[u \sin(2\pi \xi x) \Big|_{x=0}^{x=\infty} - 2\pi \xi \int_0^{\infty} u \cos(2\pi \xi x) dx \right] \\ &= -(2\pi \xi)^2 \int_0^{\infty} u \cos(2\pi \xi x) dx \end{aligned} \quad (38.21)$$

Let $U(\xi, y) = \int_0^{\infty} u(x, y) \cos(2\pi \xi x) dx$. From (38.20) and (38.21) we have

$$\frac{d^2}{dy^2} U(\xi, y) - (2\pi \xi)^2 U(\xi, y) = 0, \quad 0 < y < b,$$

and from (38.18) and Definition 38.2, we have

$$\frac{d}{dy}U(\xi, 0) = 0, \quad U(\xi, b) = \int_0^\infty f(x) \cos(2\pi\xi x) dx =: \frac{1}{2}F_c(\xi). \quad (38.22)$$

The solution of the ODE-IVP is $U(\xi, y) = A \cosh 2\pi\xi y + B \sinh 2\pi\xi y$.

From the conditions (38.22),

$$U_y(\xi, 0) = 2\pi\xi A \sinh 2\pi\xi 0 + 2\pi\xi B \cosh 2\pi\xi 0 = 0 \Rightarrow B = 0; \quad \text{and}$$

$$U(\xi, b) = A \cosh 2\pi\xi b = \frac{1}{2}F_c(\xi) \Rightarrow A = \frac{F_c(\xi)}{2 \cosh 2\pi\xi b} \Rightarrow U(\xi, y) = \frac{F_c(\xi) \cosh(2\pi\xi y)}{2 \cosh(2\pi\xi b)}.$$

Therefore, by (38.5) in Fourier integral theorem, we have

$$u(x, y) = 2 \int_0^\infty \left[F_c(\xi) \frac{\cosh 2\pi\xi y}{\cosh 2\pi\xi b} \right] \cos 2\pi\xi x d\xi.$$

Remark 38.38. If we use the Fourier cosine transform corresponding to (38.1), then

$$u(x, y) = \sqrt{\frac{2}{\pi}} \mathcal{F}_c^{-1}\{U\} = \frac{2}{\pi} \int_0^\infty F(\omega) \frac{\cosh(\omega y)}{\cosh(\omega b)} \cos \omega x d\omega.$$

Example 38.39 (UCCM3003 Final Exam, May 2010, Q5(b)). Use the Fourier (cosine) transform method to show that

$$u(x, y) = \frac{2}{\pi} \int_0^\infty \frac{\sinh \omega y \cos \omega x}{\omega(1 + \omega^2) \cosh \omega \pi} d\omega$$

is a solution of the following boundary value problem

$$\begin{cases} u_{xx} + u_{yy} = 0, & x > 0, \quad 0 < y < \pi, \\ u_x(0, y) = 0, & 0 < y < \pi, \\ u(x, 0) = 0, \quad u_y(x, \pi) = e^{-x} & x > 0. \end{cases}$$

Solution. We have semi-infinite in x -direction. By applying the even-extension to u and $f_2(x) = e^{-x}$, we obtain $u^e(-x, y) = u^e(x, y)$ and $f_2^e(x) = e^{-|x|}$. Let $\mathcal{F}\{u^e(x, y)\} = U(\xi, y)$. Then by taking Fourier transform on variable x , we have

$$\mathcal{F}\{u_{xx}^e\} + \mathcal{F}\{u_{yy}^e\} = 0 \Rightarrow \frac{d^2}{dy^2}U(\xi, y) + (2\pi i\xi)^2 U(\xi, y) = 0.$$

It has a general solution $U(\xi, y) = Ae^{2\pi\xi y} + Be^{-2\pi\xi y}$.

From the boundary conditions, we have

$$\begin{cases} U(\xi, 0) = 0 \Rightarrow A + B = 0 \\ \frac{d}{dy}U(\xi, \pi) = \mathcal{F}\{e^{-|x|}\} \Rightarrow 2\pi\xi Ae^{2\pi\xi \cdot \pi} - 2\pi\xi(-A)e^{-2\pi\xi \cdot \pi} = \frac{2}{1 + (2\pi\xi)^2} \end{cases}$$

and so $U(\xi, y) = \frac{2 \sinh(2\pi\xi y)}{2\pi\xi \cosh(2\pi^2\xi) [1 + (2\pi\xi)^2]}$

Note that $\omega = 2\pi\xi$, $U(\xi, y)$ is even in ξ and taking inverse Fourier transform gives

$$\begin{aligned} u(x, y) &= \int_{-\infty}^\infty \frac{2 \sinh(2\pi\xi y)}{2\pi\xi \cosh(2\pi^2\xi) [1 + (2\pi\xi)^2]} e^{2\pi i\xi x} d\xi \\ &= 2 \int_0^\infty \frac{2 \sinh(2\pi\xi y)}{2\pi\xi \cosh(2\pi^2\xi) [1 + (2\pi\xi)^2]} \cos(2\pi\xi x) d\xi \\ &= 2 \int_0^\infty \frac{2 \sinh(\omega y)}{\omega \cosh(\omega \pi) [1 + \omega^2]} \cos(\omega x) \frac{d\omega}{2\pi} = \frac{2}{\pi} \int_0^\infty \frac{\sinh(\omega y) \cos(\omega x)}{\omega [1 + \omega^2] \cosh(\omega \pi)} d\omega. \end{aligned}$$

□

For linear PDE problems in either cylinder or sphere with infinite or semi-infinite domain, sometimes, it is possible to solve them using Fourier transform as illustrated by the following example.

Example 38.40. (Duffy [2004, Example 3.4.2, p. 252])

Consider Laplace equation inside a semi-infinite cylinder of radius a .

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial u}{\partial r} \right) + \frac{\partial^2 u}{\partial z^2} = 0, \quad 0 \leq r < a, \quad 0 < z < \infty, \quad (38.23) \quad \text{e:3.4.12}$$

subject to the boundary conditions

$$\lim_{r \rightarrow 0} |u(r, z)| < \infty, \quad u(a, z) = 0, \quad 0 < z < \infty, \quad (38.24) \quad \text{e:3.4.13}$$

$$u(r, 0) = T_0, \quad \lim_{z \rightarrow \infty} u(r, z) = 0, \quad 0 \leq r < a. \quad (38.25) \quad \text{e:3.4.14}$$

Solution: We begin by introducing a new dependent variable $v(r, z)$, where $u(r, z) = T_0 + v(r, z)$. Thus, rewriting (38.23)–(38.25) in terms of $v(r, z)$,

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial v}{\partial r} \right) + \frac{\partial^2 v}{\partial z^2} = 0, \quad 0 \leq r < a, \quad 0 < z < \infty, \quad (38.26) \quad \text{e:3.4.15}$$

subject to the boundary conditions (38.25) (with u replaced by v) and

$$\lim_{r \rightarrow 0} |v(r, z)| < \infty, \quad v(a, z) = -T_0, \quad 0 < z < \infty, \quad (38.27) \quad \text{e:3.4.16}$$

$$v(r, 0) = 0, \quad \lim_{z \rightarrow \infty} v(r, z) = 0, \quad 0 \leq r < a. \quad (38.28) \quad \text{e:3.4.17}$$

Equation (38.26)–(38.28) can now be solved using Fourier sine transforms. Multiplying Equation (38.26) by $\sin \omega z$ and integrating from 0 to ∞ , we find

$$\int_0^\infty \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial v}{\partial r} \right) \sin \omega z dz + \int_0^\infty \frac{\partial^2 v}{\partial z^2} \sin \omega z dz = 0,$$

Integrating the v_{zz} term twice by parts,

$$\int_0^\infty \frac{\partial^2 v}{\partial z^2} \sin \omega z dz = \frac{\partial v}{\partial z} \sin \omega z \Big|_{z=0}^{z=\infty} - \omega v(r, z) \cos \omega z \Big|_{z=0}^{z=\infty} - \omega^2 \int_0^\infty v(r, z) \sin \omega z dz.$$

Because $v(r, 0) = 0$ and $v_z(r, z) \rightarrow 0$, $v(r, z) \rightarrow 0$ as $z \rightarrow \infty$, we have

$$\frac{1}{r} \frac{d}{dr} \left\{ r \frac{dV(r, \omega)}{dr} \right\} - \omega^2 V(r, \omega) = 0, \quad 0 \leq r < a,$$

with $V(r, \omega) := \int_0^\infty v(r, z) \sin \omega z dz$, $\lim_{r \rightarrow 0} |V(r, \omega)| < \infty$ and $V(a, \omega) = -\frac{T_0}{\omega}$,

since $\frac{2}{\pi} \int_0^\infty \frac{\sin \omega x}{\omega} d\omega = 1$ for $x > 0$.

The solution to the BVP of V is $V(r, \omega) = \frac{T_0}{\omega} \frac{I_0(\omega r)}{I_0(\omega a)}$.

Therefore, $u(r, z) = T_0 - \frac{2T_0}{\pi} \int_0^\infty \frac{I_0(\omega r)}{I_0(\omega a)} \frac{\sin(\omega z)}{\omega} d\omega$.

38.6 Wave Equation

First, let us consider the wave equation over $\mathbb{R}$ (a kind of Cauchy problem) and then we will try to solve some special cases. Similar to Section 38.4, it is possible to apply the separation of variables to find the solution [O’Neil, 2008, Section 4.10], however, the derivation is much shorter by using Fourier transform.

Example 38.41. (Duffy [2004, p. 236, Example 3.2.1.]) Consider a wave equation described by the following mathematical equations:

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \quad -\infty < x < \infty, \quad t > 0, \quad (38.29) \quad \text{e:3.2.1}$$

subject to the boundary conditions $\lim_{|x| \rightarrow \infty} u(x, t) = \lim_{|x| \rightarrow \infty} \frac{\partial u}{\partial x}(x, t) = 0$ and the initial conditions

$$u(x, 0) = f(x) \text{ and } u_t(x, 0) = g(x), \quad -\infty < x < \infty. \quad (38.30) \quad \text{e:3.2.3}$$

Solution: Let $U(\xi, t) = \mathcal{F}\{u(x, t)\}$. Taking the Fourier transform of (38.29), we have e:3.2.1
38.29

$$\mathcal{F}\left\{\frac{\partial^2 u}{\partial t^2}\right\} = \frac{d^2}{dt^2}U(\xi, t) = c^2 \mathcal{F}\left\{\frac{\partial^2 u}{\partial x^2}\right\} = c^2(2\pi i\xi)^2 U(\xi, t) = -c^2(2\pi\xi)^2 U(\xi, t) \quad (38.31)$$

and so

$$U_{tt} + 4c^2\pi^2\xi^2 U(\xi, t) = 0 \Rightarrow U = A \cos 2\pi\xi ct + B \sin 2\pi\xi ct. \quad (38.32)$$

From the initial conditions (38.30) e:3.2.3
38.30

$$U(\xi, 0) = A = \mathcal{F}\{f(x)\} = F(\xi) \quad \text{and} \quad U'(\xi, 0) = 2\pi\xi cB = \mathcal{F}\{g(x)\} = G(\xi).$$

The solution to the above ODE-IVP is

$$U(\xi, t) = F(\xi)\left(\frac{e^{2\pi i\xi ct} + e^{-2\pi i\xi ct}}{2}\right) + \frac{G(\xi)}{2\pi\xi c}\left(\frac{e^{2\pi i\xi ct} - e^{-2\pi i\xi ct}}{2i}\right). \quad (38.33)$$

Taking inverse Fourier transform and applying “time”-shift from Theorem 38.8 and the table from page 884 (Fourier transform of integral), we have thm:ft:property
ft:gen-in

$$u(x, t) = \frac{f(x+ct) + f(x-ct)}{2} + \frac{1}{2c} \int_{x-ct}^{x+ct} g(\xi) d\xi.$$

This classic result is known as *D'Alembert's formula* which we have encountered in Topic 2, p. 49, Example 2.7.2. It gives the wave motions within a homogeneous domain of infinite extent. The main difference between finite and infinite cases are in the former, f and g are periodic whereas the later, f and g are not periodic over $\mathbb{R}$.

Example 38.42 (UCCM3003 Final Exam, Apr 2008, Q4(b)). Solve the following wave equation by using Fourier Transform:

$$\begin{cases} u_{tt} = 16u_{xx}, & -\infty < x < \infty, \quad t > 0, \\ u, \quad u_x \rightarrow 0 \text{ as } x \rightarrow \infty, & t > 0, \\ u|_{t=0} = f(x), \quad u_t|_{t=0} = 0, & -\infty < x < \infty. \end{cases}$$

Solution. The BC implies that Fourier Transform could be applied on the PDE:

$$\frac{d^2}{dt^2}U(\xi, t) = 16(2\pi i\xi)^2 U(\xi, t) = -16(2\pi\xi)^2 U(\xi, t) \quad [2 \text{ marks}]$$

where $U(\xi, t) = \mathcal{F}\{u(x, t)\}$.

Taking Fourier Transform on the ICs, we have

$$U(\xi, 0) = \mathcal{F}\{f(x)\}, \quad \frac{d}{dt}U(\xi, 0) = 0. \quad [1 \text{ mark}]$$

Solving the ODE for U , we have the general solution

$$U(\xi, t) = A \cos 8\pi\xi t + B(\xi) \sin 8\pi\xi t. \quad [1 \text{ mark}]$$

From the ICs, we have

$$U(\xi, 0) = A = \mathcal{F}\{f(x)\} \quad [1 \text{ mark}]$$

and

$$\frac{d}{dt}U(\xi, 0) = 8\pi\xi B = 0. \quad [1 \text{ mark}]$$

Hence

$$U(\xi, t) = \mathcal{F}\{f(x)\} \cos 8\pi\xi t. \quad [2 \text{ marks}]$$

Taking the inverse Fourier transform, we have

$$u(x, t) = \mathcal{F}^{-1}\left\{\frac{e^{i8\pi\xi t} + e^{-i8\pi\xi t}}{2} \mathcal{F}\{f(x)\}\right\} = \frac{f(x+4t) + f(x-4t)}{2}. \quad [2 \text{ marks}]$$

□

Example 38.43 (UCCM3003 Final Exam, Sep 2008, Q3(b)). Solve the following wave equation by using

Fourier Transform:

$$\begin{cases} \frac{\partial^2 u}{\partial t^2}(x, t) = 4 \frac{\partial^2 u}{\partial x^2}(x, t), & -\infty < x < \infty, \quad t > 0, \\ \lim_{x \rightarrow \pm\infty} u(x, t) = 0, & t > 0, \\ \lim_{x \rightarrow \pm\infty} \frac{\partial u}{\partial x}(x, t) = 0, & t > 0, \\ u(x, 0) = 0, & -\infty < x < \infty, \\ \frac{\partial u}{\partial t}(x, 0) = e^{-x^2}, & -\infty < x < \infty. \end{cases} \quad (12 \text{ marks})$$

Solution. The BC implies that Fourier Transform could be applied to the PDE to obtain:

$$\frac{d^2}{dt^2} U(\xi, t) = 4(2\pi i \xi)^2 U(\xi, t) = -16(\pi \xi)^2 U(\xi, t) \quad [2 \text{ marks}]$$

where $U(\xi, t) = \mathcal{F}\{u(x, t)\}$.

Taking Fourier Transform on the ICs, we have

$$U(\xi, 0) = 0 \quad [1 \text{ mark}]$$

$$\frac{d}{dt} U(\xi, 0) = \mathcal{F}\{e^{-x^2}\} = \sqrt{\pi} e^{-(\pi \xi)^2} \quad [1 \text{ mark}]$$

Solving the ODE for U , we have the general solution

$$U(\xi, t) = A \cos 4\pi \xi t + B \sin 4\pi \xi t. \quad [1 \text{ mark}]$$

From the ICs, we have

$$U(\omega, 0) = A = 0 \quad [1 \text{ mark}]$$

$$\frac{d}{dt} U(\omega, 0) = 4\pi \xi B = \sqrt{\pi} e^{-(\pi \xi)^2}. \quad [1 \text{ mark}]$$

Hence

$$U(\xi, t) = \frac{\sqrt{\pi}}{4\pi \xi} e^{-(\pi \xi)^2} \sin 4\pi \xi t. \quad [2 \text{ marks}]$$

Taking the inverse Fourier transform, we have

$$u(x, t) = \mathcal{F}^{-1} \left\{ \frac{e^{4\pi i \xi t} - e^{-4\pi i \xi t}}{2i} \times \frac{\sqrt{\pi}}{4\pi \xi} e^{-(\pi \xi)^2} \right\} \quad [1 \text{ mark}]$$

$$= \frac{\sqrt{\pi}}{4} \left[\mathcal{F}^{-1} \left\{ e^{4\pi i \xi t} \frac{e^{-(\pi \xi)^2}}{2\pi i \xi} \right\} - \mathcal{F}^{-1} \left\{ e^{-4\pi i \xi t} \frac{e^{-(\pi \xi)^2}}{2\pi i \xi} \right\} \right] \quad [1 \text{ mark}]$$

$$= \frac{\sqrt{\pi}}{4} \left[\mathcal{F}^{-1} \left\{ \frac{e^{-(\pi \xi)^2}}{2\pi i \xi} \right\} \Big|_{x \rightarrow x+2t} - \mathcal{F}^{-1} \left\{ \frac{e^{-(\pi \xi)^2}}{2\pi i \xi} \right\} \Big|_{x \rightarrow x-2t} \right]$$

$$= \frac{\sqrt{\pi}}{4} \left[\int_0^x \frac{e^{-u^2}}{\sqrt{\pi}} du \Big|_{x \rightarrow x+2t} - \int_0^x \frac{e^{-u^2}}{\sqrt{\pi}} du \Big|_{x \rightarrow x-2t} \right]$$

$$= \frac{1}{4} \left[\int_0^{x+2t} e^{-u^2} du - \int_0^{x-2t} e^{-u^2} du \right] = \frac{1}{4} \int_{x-2t}^{x+2t} e^{-u^2} du. \quad [2 \text{ marks}]$$

□

Example 38.44 (UCCM3003 Final Exam, May 2009, Q6(b)). Show that a solution to the following initial value problem

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \quad -\infty < x < \infty, \quad t > 0,$$

$$u(x, 0) = 0, \quad \frac{\partial u}{\partial t}(x, 0) = \begin{cases} 1 - |x|, & -1 < x < 1, \\ 0, & |x| > 1, \end{cases}$$

$$\text{is } u(x, t) = \frac{2}{c\pi} \int_0^\infty \frac{1 - \cos \omega}{\omega^3} \cos(\omega x) \sin(\omega ct) d\omega. \quad (15 \text{ marks})$$

Solution. Let $U(\xi, t) = \mathcal{F}\{u(x, t)\}$. Taking Fourier transform, the PDE and the initial conditions becomes becomes

$$U_{tt} = c^2 (2\pi i \xi)^2 U = -(2\pi \xi c)^2 U,$$

$$U(\xi, 0) = 0, \quad U_t(\xi, 0) = \int_{-1}^1 (1 - |x|) e^{2\pi i \xi x} dx = 2 \left[\frac{1 - \cos(2\pi \xi)}{(2\pi \xi)^2} \right].$$

The solution of this ODE-IVP is $U(\xi, t) = \frac{2(1 - \cos(2\pi\xi))}{(2\pi\xi)^3 c} \sin(2\pi\xi ct)$.

U is even in ξ , therefore, taking inverse, we have

$$\begin{aligned} u(x, t) &= \int_{-\infty}^{\infty} \frac{2(1 - \cos(2\pi\xi))}{(2\pi\xi)^3 c} \sin(2\pi\xi ct) e^{2\pi i \xi x} d\xi \\ &= \frac{4}{c} \int_0^{\infty} \frac{1 - \cos(2\pi\xi)}{(2\pi\xi)^3} \sin(2\pi\xi ct) \cos(2\pi\xi x) d\xi. \end{aligned}$$

□

Example 38.45 (UCCM3003 Final Exam, Apr 2012, Q5(a)). Use the Fourier transform method to show that

$$u(x, t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{20}{25 + \omega^2} \cos(3\omega t) e^{i\omega x} d\omega$$

is the solution of the following boundary value problem:

$$\begin{cases} u_{tt} = 9u_{xx}, & -\infty < x < \infty, t \geq 0, \\ u(x, 0) = 4e^{-5|x|}, & -\infty < x < \infty, \\ u_t(x, 0) = 0, & -\infty < x < \infty. \end{cases} \quad (16 \text{ marks})$$

Solution:

$$U_{tt} = 9(2\pi i \xi)^2 U = -(6\pi \xi)^2 U, \quad U(\xi, 0) = 4 \times \frac{10}{5^2 + (2\pi \xi)^2}, \quad U_t(\xi, 0) = 0.$$

Solving the ODE-IVP, we have the solution $U(\xi, t) = \frac{40}{5^2 + (2\pi \xi)^2} \cos 6\pi \xi t$.

Taking inverse Fourier transform, we have

$$u(x, t) = \int_{-\infty}^{\infty} \frac{40}{5^2 + (2\pi \xi)^2} \cos 6\pi \xi t e^{2\pi i \xi x} d\xi = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{20}{25 + \omega^2} \cos 3\omega t e^{i\omega x} d\omega$$

where $\omega = 2\pi \xi$.

Now, let us consider the wave equation over the half-line $(0, \infty)$.

jhu:cwang:q3

Example 38.46 (UECM3003 Final Exam, Apr 2016, Q5(b)). Use the Fourier transform to find the analytic solution for the wave equation on the half-line. **Clearly explain each step.**

$$\begin{cases} u_{tt} = u_{xx}, & 0 < x < \infty, 0 < t < \infty, \\ u_x(0, t) = 0, & 0 < t < \infty, \\ u(x, 0) = x^2 e^{-x}, \quad u_t(x, 0) = 0, & 0 < x < \infty. \end{cases} \quad (13 \text{ marks})$$

Solution. Since the boundary condition is type 2, we will be using the Fourier cosine transform on the wave equation:

$$\frac{d^2}{dt^2} \int_0^{\infty} u \cos 2\pi \xi x dx = \int_0^{\infty} u_{xx} \cos 2\pi \xi x dx = u_x \cos 2\pi \xi x \Big|_0^{\infty} - (-2\pi \xi) \int_0^{\infty} u_x \sin 2\pi \xi x dx. \quad [2 \text{ marks}]$$

Since $u_x(0, t) = 0$ and $u_x \rightarrow 0$ as $x \rightarrow \infty$, we can use the above expression and apply the integration by parts again:

$$\frac{d^2}{dt^2} \int_0^{\infty} u \cos 2\pi \xi x dx = 2\pi \xi \left[u \sin 2\pi \xi \Big|_0^{\infty} - 2\pi \xi \int_0^{\infty} u \cos 2\pi \xi x dx \right] \quad [1 \text{ mark}]$$

Let $U = U(\xi, t) = \int_0^{\infty} u \cos 2\pi \xi x dx$. Since $u \rightarrow 0$ as $x \rightarrow \infty$ and $\sin 0 = 0$, the above ODE can be simplified to a second order linear ODE with a general solution below

$$\frac{d^2 U}{dt^2} + 4\pi^2 \xi^2 U = 0 \Rightarrow U = A(\xi) \cos 2\pi \xi t + B(\xi) \sin 2\pi \xi t. \quad [2 \text{ marks}]$$

From the initial conditions, we have

$$A(\xi) = \int_0^\infty x^2 e^{-x} \cos 2\pi\xi x dx$$

$$2\pi\xi B(\xi) = 0 \Rightarrow B(\xi) = 0. \quad [2 \text{ marks}]$$

Since we are only given with the Fourier transform, we will rewrite $A(\xi)$ as follows:

$$\begin{aligned} \int_0^\infty x^2 e^{-x} \cos 2\pi\xi x dx &= \frac{1}{2} \int_{-\infty}^\infty \underbrace{x^2 e^{-|x|}}_{\text{even}} \cos 2\pi\xi x dx \\ &= \frac{1}{2} \int_{-\infty}^\infty x^2 e^{-|x|} e^{-i2\pi\xi x} dx = \frac{1}{2} \mathcal{F}\{x^2 e^{-|x|}\} \end{aligned} \quad [1 \text{ mark}]$$

By Fourier Integral Theorem,

$$u(x, t) = 4 \int_0^\infty \left[\int_0^\infty u(\tilde{x}, t) \cos 2\pi\xi \tilde{x} d\tilde{x} \right] \cos 2\pi\xi x d\xi \quad [1 \text{ mark}]$$

$$= 4 \int_0^\infty U(\xi, t) \cos 2\pi\xi x d\xi = 4 \int_0^\infty \left[\frac{1}{2} \mathcal{F}\{x^2 e^{-|x|}\} \cos 2\pi\xi t \right] \cos 2\pi\xi x d\xi \quad [1 \text{ mark}]$$

$$= 2 \int_0^\infty \underbrace{\left[\mathcal{F}\{x^2 e^{-|x|}\} \cos 2\pi\xi t \right]}_{\text{even w.r.t } \xi} \cos 2\pi\xi x d\xi = \int_{-\infty}^\infty \left[\mathcal{F}\{x^2 e^{-|x|}\} \cos 2\pi\xi t \right] e^{i2\pi\xi x} d\xi$$

$$= \mathcal{F}^{-1} \left\{ \mathcal{F}\{x^2 e^{-|x|}\} \cos 2\pi\xi t \right\} = \frac{1}{2} \mathcal{F}^{-1} \left\{ (e^{i2\pi\xi t} + e^{-i2\pi\xi t}) \mathcal{F}\{x^2 e^{-|x|}\} \right\} \quad [1 \text{ mark}]$$

$$= \frac{1}{2} \left[\mathcal{F}^{-1} \{ \mathcal{F}\{x^2 e^{-|x|}\} \} \Big|_{x \rightarrow x+t} + \mathcal{F}^{-1} \{ \mathcal{F}\{x^2 e^{-|x|}\} \} \Big|_{x \rightarrow x-t} \right] \quad [1 \text{ mark}]$$

$$= \frac{1}{2} \left[(x+t)^2 e^{-|x+t|} + (x-t)^2 e^{-|x-t|} \right] \quad [1 \text{ mark}]$$

□

Realistic wave equations are more complicated. The following example illustrated this point.

Example 38.47. (Telegraph Equation, González-Velasco [1995, 280]).

The non-homogeneous equation of telegraphy is given below:

$$\frac{\partial^2 u}{\partial x^2} = LC \frac{\partial^2 u}{\partial t^2} + (RC + GL) \frac{\partial u}{\partial t} + GRu.$$

where u is the voltage or current of the conducting material; the constants C , R , L , G are capacitance, resistance, self-inductance and leakage respectively. Assume $R = 0$, this can be simplified to the following form by change the constants and introducing a source $f(x)$:

$$\frac{\partial^2 u}{\partial x^2} - \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} - \frac{4\pi\sigma}{c^2} \frac{\partial u}{\partial t} = \frac{4\pi}{c} f(x), \quad -\infty < x < \infty, \quad t > 0,$$

where the initial condition are $u(x, 0) = u_t(x, 0) = 0$.

Solution: According to Duffy [2004], steps to solve this problem are:

Step 1. Assuming that f has a Fourier transform and $U(\omega, t) = \mathcal{F}\{u(x, t)\}$, we can reduce the PDE to the IVP:

$$\begin{aligned} \frac{d^2 U(\omega, t)}{dt^2} + 4\pi\sigma \frac{dU(\omega, t)}{dt} + \omega^2 c^2 U(\omega, t) &= -4\pi c F(\omega), \\ U(\omega, 0) = U'(\omega, 0) &= 0, \end{aligned}$$

where $F(\omega)$ is the Fourier transform of $f(x)$.

Step 2. Use Laplace transform to show that the solution to the IVP is

$$U(\omega, t) = -4\pi c F(\omega) \int_0^t e^{-2\pi\sigma(t-\tau)} \frac{\sin[(t-\tau)\sqrt{\omega^2 c^2 - 4\pi^2 \sigma^2}]}{\sqrt{\omega^2 c^2 - 4\pi^2 \sigma^2}} d\tau.$$

Step 3. The solution $u(x, t)$ is obtained by using inverse Fourier transform

$$\begin{aligned} u(x, t) &= -2 \int_0^t e^{-2\pi\sigma(t-\tau)} \left\{ \int_{-\infty}^{\infty} F(\omega) \frac{c \sin[(t-\tau)\sqrt{\omega^2 c^2 - 4\pi^2 \sigma^2}]}{\sqrt{\omega^2 c^2 - 4\pi^2 \sigma^2}} e^{i\omega x} d\omega \right\} d\tau \\ &= -4 \int_0^t \left[\int_{-\infty}^{\infty} f(\chi) I(|x-\chi|, t-\tau) d\chi \right] e^{-2\pi\sigma(t-\tau)} d\tau, \end{aligned}$$

where

$$I(x, t) = \int_0^{\infty} \frac{\sin \left[ct \sqrt{\omega^2 - \left(\frac{2\pi\sigma}{c} \right)^2} \right]}{\sqrt{\omega^2 - \left(\frac{2\pi\sigma}{c} \right)^2}} \cos(\omega x) d\omega = \frac{\pi}{2} I_0 \left[2\pi\sigma \sqrt{t^2 - \left(\frac{x}{c} \right)^2} \right] H\left(t - \frac{x}{c}\right).$$

With appropriate transform, it is possible to apply Fourier transform to spherical problems as well. Let us calculate the sound waves radiated by a sphere of radius a whose surface expands radially with an impulsive acceleration $v_0 \delta(t)$. This problem has a number of applications; one of them would be the sound waves generated by an exploding depth charge.

Example 38.48. (Duffy [2004, Example 3.2.2, p. 238]) Consider sound waves radiated by a sphere of radius a whose surface expands radially with an impulsive acceleration $v_0 \delta(t)$. Assume radial symmetry, the corresponding wave equation is

$$\frac{\partial^2 u}{\partial t^2} = \frac{c^2}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial u}{\partial r} \right) \quad (38.34) \quad \text{e:3.2.17}$$

subject to the boundary condition

$$\left. \frac{\partial u}{\partial r} \right|_{r=a} = -\rho v_0 \delta(t) \quad (38.35) \quad \text{e:3.2.18}$$

where $u(r, t)$ is the pressure field, c is the speed of sound and ρ is the average density of the fluid.

Solution: Assuming that the pressure field possess a Fourier transform, we may re-express it by the Fourier integral

$$u(r, t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} U(r, \omega) e^{i\omega t} d\omega. \quad (38.36) \quad \text{e:3.2.19}$$

Substituting into (38.34), it becomes e:3.2.17

$$\frac{1}{r^2} \frac{d}{dr} \left(r^2 \frac{dU}{dr} \right) + k^2 U = 0, \quad (38.37) \quad \text{e:3.2.20}$$

and

$$\frac{dU(a, \omega)}{dr} = -\rho v_0 \quad (38.38) \quad \text{e:3.2.21}$$

with $k = \frac{\omega}{c}$. The most general solution of (38.37) is e:3.2.20

$$U(r, \omega) = A(\omega) \frac{e^{ikr}}{r} + B(\omega) \frac{e^{-ikr}}{r} \quad (38.39) \quad \text{e:3.2.22}$$

and

$$u(r, t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[A(\omega) \frac{e^{i\omega t + ikr}}{r} + B(\omega) \frac{e^{i\omega t - ikr}}{r} \right] d\omega. \quad (38.40) \quad \text{e:3.2.23}$$

At this point, we note that the first term on the right side of (38.40) e:3.2.23 represents an inward propagating wave while the second term is an outward propagating wave. Because there is no source of energy at infinity, the inward propagating wave is *non-physical* and we discard it. This boundary condition often referred to as the *Sommerfeld radiation condition* because Sommerfeld used it first in his study of electromagnetic waves on the surface of a conducting earth.

Upon substituting (38.39) into (38.38) with $A(\omega) = 0$, e:3.2.22 e:3.2.21

$$U(r, \omega) = \frac{\rho a^2 v_0}{2\pi r} \frac{e^{-i\omega(r-a)/c}}{1 + i\omega a/c},$$

or

$$u(r, t) = \frac{\rho a^2 v_0}{2\pi r} \int_{-\infty}^{\infty} \frac{e^{-i\omega(r-a)/c}}{1 + i\omega a/c} d\omega = \frac{\rho a c v_0}{r} \exp \left[-\frac{c}{a} \left(t - \frac{r-a}{c} \right) \right] H\left(t - \frac{r-a}{c}\right) \quad (38.41)$$

e:3.2.25

by using residue theorem on the infinite semicircle in the lower and upper half-plane for $t < \frac{r-a}{c}$ and $t > \frac{r-a}{c}$ respectively.

Part VI

Numerical Methods

Chapter 39

Computer Algebra with Reduce

<https://arxiv.org/abs/cs/0105033>: Lectures on Reduce and Maple at UAM-I, Mexico 1

Reduce has allfac, exp, mcd, lcm, ratpri, pri, nat turned on by default.

switch	description	e.g.
* allfac	factorise simple factors	$2x + 2 \rightarrow 2(x + 1)$
div	divide by the denominator	$(x^2 + 2)/x \rightarrow x + 2/x$
* exp	expand all expressions	$(x + 1)(x - 1) \rightarrow x^2 - 1$
* mcd	make (common) denominators	$x^{-1} \rightarrow 1/x$
* lcm	cancel least common multiples	
gcd	cancel greatest common divisor	
rat	display as polynomial in factor	$\frac{x+1}{x} \rightarrow 1 + x^{-1}$
* ratpri	display rationals as fraction $1/x \rightarrow \frac{1}{x}$	
* pri	dominates allfac, div, rat, revpri	
revpri	display polynomials in opposite order $x^2 + x + 1 \rightarrow 1 + x + x^2$	
rounded	calculate with floats	$1/3 \rightarrow 0.333333333333$
complex	simplify complex expressions	$1/i \rightarrow -i$
nero	don't display zero results $0 \rightarrow$	
* nat	display in Reduce input format	$\frac{x^2}{3} \rightarrow x * 2/3$
msg	suppress messages	
fort	display in Fortran format	
tex	display in TeX format	

The terminator decides whether the Reduce's response (same as Maxima) is displayed (; terminator) or not (\$ terminator).

Reserved constants: i infinity nil pi t

```

1  %% line comment in reduce
2  % refer to https://en.wikipedia.org/wiki/Reduce_(computer_algebra_system)#Comments
3  % file "red1"
4
5  off factor, exp, mcd, ratpri$
6  (x+1)**3/3 - x; %-> 1/3*(x + 1)**3 - x
7  on mcd$ ws; %-> ((x + 1)**3 - 3*x)/3
8  on ratpri$ ws; %->
9
10 on exp$ ws; %-> (x**3 + 3*x**2 + 1)/3
11 factor x$on rat$ ws; %-> x**3/3 + x**2 + 1/3
12 remfac x$off exp$
13 z+a; %-> a + z
14 order z,a$ ws; %-> z + a
15 order nil$ ws; %-> a + z
16
17 off mcd$
18 my_name123 := (x**2-1)/(x+1); %-> my_name123 := (x + 1)**(-1)*(x**2 - 1)
19 on mcd$ my_name123; %-> x - 1
20 clear my_name123;
21
22 a:=b$ a; %-> b
23 b:=1$ a; %-> 1
24 a:=a$b:=2$ a; %-> 1

```

```

25 clear a,b;
26
27 /* define alias=expression; */
28 define isprime = primep;
29 for x:=0:20 do if isprime(x) then write x," is a prime!";
30 % throws 2 3 5 7 11 13 17 19
31 define is_assigned_to = :=;
32 x is_assigned_to 5; %-> x := 5
33
34 /* Substitutions */
35 f := cos(x)$ x:=0$ f; clear x$ %-> 1
36 f := cos(x)$ let { x=>0 }$ f; clear x$ %-> 1
37 f := cos(x)$ f where { x=>0 }; %-> 1
38 f := cos(x)$ sub(x=0,f); %-> 1
39
40 /* Rule: expr => expr [when boolean]
41 formal names are marked with a twiddle such as ~x
42 */
43 sin(x)**2 + cos(x)**2; %-> cos(x)**2 + sin(x)**2
44 trig_rules:={
45   cos(~x)*cos(~y) => (1/2)*(cos(x+y)+cos(x-y)),
46   sin(~x)**2 + cos(~x)**2 => 1 }$
47 let trig_rules$
48 sin(x+1)**2 + cos(x+1)**2; %-> 1
49 cos(pi/3-1) * cos(1); %-> (2*cos((pi - 6)/3) + 1)/4
50 showrules cos; % shows ALL rules associated with cos!
51 clearrules trig_rules;
52 showrules log;
53
54 /* Operator */
55 operator eps$ % declares a new operator
56 antisymmetric eps$ % declares eps to be totally antisymmetric
57 let { eps(0,1,2,3) => 1 }$ % defines the properties of this operator
58 eps(3,1,2,0); eps(2,0,1,1); %-> -1 and 0
59 showrules eps; %-> {eps(3,2,1,0) => 1}$
60 clear my_sin$ operator my_sin$
61 my_sin_rules:={
62   my_sin(~x) => - my_sin(-x) when numberp(x/pi) and x/pi<0,
63   my_sin(~x) => my_sin(x - 2*pi) when numberp(x/pi) and x/pi>2,
64   my_sin(~x) => - my_sin(x - pi) when numberp(x/pi) and x/pi>1,
65   my_sin(~x) => my_sin(pi - x) when numberp(x/pi) and x/pi>1/2,
66   my_sin(0) => 0,
67   my_sin(pi/6) => 1/2,
68   my_sin(pi/4) => 1/2*sqrt(2),
69   my_sin(pi/3) => 1/2*sqrt(3),
70   my_sin(pi/2) => 1
71 }$
72 let my_sin_rules$
73 my_sin(-124*pi/6); %-> - sqrt(3)/2
74 clearrules my_sin_rules;
75
76 /* algebra solver: solve(expr list,name list); */
77 solve(x**3-3*x**2-61*x+63,x); %-> {x=9,x=1,x=-7} which are the nulls of this polynomial
78 solve({x+y-9,-2*y**2+3*x},{x,y}); %-> { {y= - 9/2,x=27/2} , {y=3,x=6} }
79
80 /*
81 https://www.aramis.admin.ch/Texte/?projectid=6557&Sprache=en-US
82 http://www-sop.inria.fr/cafe/Manuel.Bronstein/cathode/software.html
83 https://reduce-algebra.sourceforge.io/manual/manuale167.html#x190-113100020.41
84 https://reduce-algebra.sourceforge.io/extra-docs/crack.pdf
85 https://reduce-algebra.sourceforge.io/extra-docs/conlaw.pdf
86 Reduce has no build-in routines to solve differential equations.
87 However, there exist packages (Crack and ODEsolve) for this problem by
88 the CATHODE 2 project (Computer Algebra Tools for Handling Ordinary
89 Differential Equations) which develops routines to handle
90 differential equation systems also for Reduce. The members of
91 this group are M. MacCallum (London), V. Fairén (Madrid),

```

```

92 E. Tournier (Grenoble), Th. Mulders (Zurich),
93 M. van der Puf (Groningen), F. Schwarz (Bonn),
94 L. Brenig (Brussel).
95 */
96
97
98 /* General commands for maths??? */
99 for l:=1:50 sum l; %-> 1275
100 for l:=1:100 product l; % gives the huge number 100!
101 int(sin(x)**2,x); %-> x/2 + (-cos(x)*sin(x))/2
102 depend (f,x)$ df(f,x); int(f,x); %-> df(f,x) int(f,x)
103 nodepend (f,x)$df(f,x);int(f,x); %-> 0 x*f
104 matrix m(3,2)$ % declare matrix
105 m:=mat((1,2,3),(2,1,2),(3,2,1))$ % construct matrix
106 det m; %->8 determinante
107 1/m; % gives the inverse
108 matrix v(3,1)$ % declare a vector
109 v:=mat((1),(2),(3))$
110 1/m * v; %->(1,0,0) matrix multiplication
111 procedure fac1(k)$
112 for l:=1:k product l$ % defines a new factorial procedure:
113 fac1(16); %-> 20922789888000 which is 16!
114 operator fac2$
115 let { fac2(1) => 1 ,
116 fac2(~n) => n*fac2(n-1) when numberp n and n > 1}$
117 fac2(16); %-> 20922789888000
118 write "next "prime > 500" is ",nextprime(500)$ %-> next "prime > 500" is 503
119 showtime$ % milli seconds since last showtime
120 in "my_file"$ % executes all lines in "my_file" as input
121 quit$ % exits Reduce
122
123 /* Running files with Reduce and packages */
124 reduce < sample.rei > sample.reo
125
126
127 Reduce packages:
128 ALGINT integration for functions involving roots (James H. Davenport)
129 ARNUM algebraic numbers (Eberhard Schrfer)
130 ASSIST useful utilities for various applications (Hubert Caprasse)
131 AVECTOR vector algebra (David Harper)
132 CALI package for computational commutative algebra (Hans-Gert Graebe)
133 CAMAL calculations in celestial mechanics (John Fitch)
134 CHANGEVAR transformation of variables in differential equations (G. oluk)
135 COMPACT condensing of expressions with polynomial side relations (Anthony C. Hearn)
136 CRACK package for solving overdetermined systems of PDEs or ODEs (Andreas Brand, Thomas Wolf)
137 CVIT Dirac gamma matrices (V.Ilyin, A.Kryukov, A.Rodionov, A.Taranov)
138 DESIR differential equations and singularities (C. Dicrescenzo, F. Richard-Jung, E. Tournier)
139 EXCALC calculus for differential geometry (Eberhard Schrfer)
140 FIDE code generation for finite difference schemes (Richard Liska)
141 GENTRAN code generation in FORTRAN, RATFOR, C (Barbara Gates)
142 GNUPLOT display of functions and surfaces (Herbert Melenk)
143 GROEBNER computation in multivariate polynomial ideals (Herbert Melenk, H.Michael Mller, Winfried
144 HEPHYS high energy physics (Anthony C. Hearn)
145 IDEALS Arithmetic for polynomial ideals (Herbert Melenk)
146 LAPLACE Laplace and inverse Laplace transform (C. Kazasov et al.)
147 LIE functions for the classification of real n-dimensional Lie algebras (Carsten, Franziska Schbe
148 LIMITS a package for finding limits (Stanley L. Kameny)
149 LININEQ linear inequalities and linear programming (Herbert Melenk)
150 NUMERIC solving numerical problems using rounded mode (Herbert Melenk)
151 ODESOLVE ordinary differential equations (Malcolm MacCallum et al.)
152 ORTHOVEC calculus for scalar and vector quantities (J.W. Eastwood)
153 PHYSOP additional support for non-commuting quantities (Mathias Warns)
154 PM general algebraic pattern matcher (Kevin McIsaac)
155 REACTEQN manipulation of chemical reaction systems (Herbert Melenk)
156 RLFI, TRI TeX and LaTeX output (Richard Liska, Ladislav Drska, Werner Antweiler)
157 ROOTS roots of polynomials (Stanley L. Kameny)
158 SCOPE optimization of numerical programs (J. A. van Hulzen)

```

```

159 SPDE symmetry analysis for partial differential equations (Fritz Schwarz)
160 SPECFN package for special functions (Chris Cannam et al.)
161 SPECFN2 package for special special functions (Victor Adamchik, Winfried Neun)
162 SYMMETRY symmetry-adapted bases and block diagonal forms of symmetric matrices (Karin Gatermann)
163 SUM sum and product of series (Fuji Kako)
164 TAYLOR multivariate Taylor series (Rainer Schopf)
165 TPS univariate Taylor series with indefinite order (Alan Barnes, Julian Padget)
166 WU Wu Algorithm for polynomial systems (Russell Bradford)
167
168 Generating a Julia set
169 % file "red3"
170 on comp$
171 procedure julia(c,s,file)$ begin
172   on complex,rounded$
173     precision(6)$
174     l:={}
175     for x:=-3/2*s:3/2*s do << write x$ for y:=-3/2*s:3/2*s do <<
176       z:=x/s+i*y/s$
177       j:=0$ repeat z:=z**2+c until
178         abs(repart(z))>2 or abs(impart(z))>2 or (j:=j+1)=50$
179       if j=50 then l:={x,y}.l$
180     >> >>$
181   off complex,rounded$
182   out file$
183   for each point in l do
184     write first(point)," ",second(point)$
185   shut file$
186 end$
187 showtime$
188 julia(-0.11+0.67*i,100,"l2.red")$
189 showtime;
190 % time: 780730 ms
191 %in gnuplot:
192 %set data style dots
193 %set nokey; set noxtics; set noytics; set size square; set noborder
194 %plot "l2.red"
195 %set terminal postscript
196 %set output "julia1.ps"
197 %replot
198
199 /* Export / import */
200
201 Storing results is very important for large problems. Reduce offers no extra command for
202 storing expressions but it is quite easy to write a little procedure that stores e.g. an arbitrary
203 list of names together with their values. For this we write appropriate assignments into an
204 extra file with the nat switch turned off. Such a procedure could read
205 procedure store(filename,namelist,valuelist)$ begin
206   off nat$
207   out filename$
208   for each name in namelist do <<
209     write name,"=",first valuelist$
210     valuelist:=rest valuelist$
211   >>$
212   write "end"$
213   shut filename$
214   on nat$
215 end$
216 With this procedure defined we can execute the lines
217 f:=sin(x)$ g:=cos(x)$
218 store("FandG","f","g",f,g)$
219 clear f,g$
220 in "FandG"$
221 f;g;
222 where, the filename "FandG", the name list {"f","g"}, and their values {f,g} is passed
223 to the procedure store. This produces the file "FandG" in the current directory reading
224 f:=sin(x)$
225 g:=cos(x)$

```

```

226 end$
227 such that the in command reassigns their values to f and g.
228 2.4.2 TeX
229 To export algebraic expression to TeX you need to load the package tri. This packages
230 provides two switches tex and texbreak which causes Reduce to format any output in
231 TeX syntax (texbreak also breaks the line in long equations). Example:
232 1: load_package tri;
233 *** global 'metricu!*' cannot become fluid
234 *** global 'indx1!*' cannot become fluid
235 % TeX-REDUCE-Interface 0.50
236 2.4. EXPORT / IMPORT 25
237 % set greek asserted
238 % set lowercase asserted
239 % \tolerance 10
240 % \hsize=150mm
241 2: depend(f,x);
242 3: df(f,x);
243 $$
244 \frac{\partial f}{\partial x}
245 \partial x
246 $$
247 For some TeX-symbols Reduce uses own macros which can be included in TeX with an
248 \input{tridefs} command, (You can find the tridefs.tex file at [5].) To export all the
249 output of a Reduce file in one file in TeX format one could add the lines
250 load_package tri$
251 off msg$
252 on texbreak$
253 out "sample.out.tex"$
254 at the beginning of the Reduce file and the following lines at the end
255 shut "sample.out.tex"$
256 quit$
257 Turning off the msg switch prevents errors. For further information on the tri package see
258 http://www.uni-koeln.de/REDUCE/3.6/doc/tri.ps
259
260 To display functions with gnuplot and to save them as postscripts you need to load the
261 package gnuplot. This package provides the commands plot, gnuplot, plotreset, and
262 plotshow. With plot you will automatically open a gnuplot window that displays your
263 graph. Example:
264 load package gnuplot$
265 plot(cos(x)*cos(y),x=(-pi..pi),y=(-pi..pi),contour);
266 will open a window that looks like the one in figure 2.2
267 The syntax of this plot command is similar to that in gnuplot. The command
268 gnuplot(cmd,param1,param2,...);
269 2.4. EXPORT / IMPORT 27
270 -4 -3 -2 -1 0 1 2 3 4
271 y -4 -3 -2 -1 0 1 2 3 4
272 x
273 -1
274 -0.5
275 0
276 0.5
277 1
278 z
279 Figure 2.2: The gnuplot window and the postscript produced by gnuplot
280 is supposed to execute all other gnuplot commands (which it does not always!) The pa-
281 rameters of the gnuplot command com have to be separated by commas (not by blanks as
282 in gnuplot). Example:
283 gnuplot(set,logscale,y)$
284 plot(x**2);
285 will produce a logarithmic plot of x2. Finally, plotting into a file in postscript format is
286 achieved by adding the options terminal=postscript and output=filename to the plot
287 command. Example:
288 plot(cos(x)*cos(y),x=(-pi..pi),y=(-pi..pi),contour,terminal=postscript,output="gntest.eps");
289 produces the postscript displayed in figure 2.2. For more information on the gnuplot
290 package see
291 http://www.zib.de/Symbolik/reduce/moredocs/gnuplot.pdf

```


Chapter 40

Introduction to Numerical Methods

The numerical methods are based on the textbook Burden, R. L. and Faires, J. D., 2015. Numerical analysis. 10th ed. Boston, MA: Brooks/Cole Cengage Learning.

Many of the numerical methods mentioned in NumericalRecipes3e-CUP2007 are missing.

Many engineering, scientific, statistical and financial problems are either modelled by a system of ODEs (Chapter 33) or/and PDEs (Chapter 34). Majority of the ODE and PDE models do not have analytical solutions and one has to use numerical methods to find an approximation to the solution.

Burden and Faires [2015] and Cheney and Kincaid [2012] are the standard textbooks for numerical methods. The website <http://www.ma.utexas.edu/CNA/NMC7/index.html> provides supplementary materials to the textbook Cheney and Kincaid [2012]. Similar textbooks are Conte and de Boor [1981], Gerald and Wheatley [2004], Wood [2000], etc.

40.1 Overview

sec:intro

Numerical method, numerical analysis, scientific computing are subjects with different emphasis that study how to solve *continuous model* problems in science and engineering using computers. Mathematicians, scientists and engineers have developed a software framework to use numerical methods to solve problems. The software framework can be summarised in the following diagram [Thijssen, 1999]:

Computer Aided Design, Modelling and Simulation					Log Scale
Quantum Dynamics	Material Engineering Molecular Dynamics	Mechanical, Electrical, ... Classical Mechanics & Physics	Engineering Solar System	Cosmology	
Topic 8	PDE Solvers: FDM (Topic 8), FVM (UEME4213 CFD), FEM (UEME4343 Comp Methods in Solid Mech), Spectral Method				
UECM3093 Topic 4, 5, 7	Optimisation Libraries (Operational Research)		ODE-BVP Solver (Topic 7)	Integration (Topic 5) Interpolation (Topic 4)	
Topic 3 & 6	Nonlinear Equation Solver (Topic 3): PETSc/SNES, Sundials/KINSOL		ODE-DAE-IVP Solver (Topic 6): DASSL, DASPK, Sundials/CVODE+IDA		
Topic 2	Linear Algebra Libraries: BLAS, LAPACK, Armadillo (C++11), SuiteSparse, ...				
Topic 1	C Math Library <code><math.h></code> , Fortran library, <code>java.lang.Math</code> , CUDA, OpenCL, GSL, Cernlib, ROOT, ...				
UECS3043	Networking and Parallel Processing: MPI, Clusters, Cloud Computing, ...				
	System Library: GNU C Library, BSD C Library, Windows API				
UECS2103	Operating System: Linux, *BSD, MacOS/X, Windows				
UEEA2283 UECS1013	Hardware: CPU (Intel, AMD, MIPS, Xeon, ...)		GPU (Nvidia, AMD, Intel)		

where

- UECS1013 = Introduction to Computer Organisation and Architecture (Software Engineering)
- UEEA2283 = Computer Architecture (Electrical and Communication Engineering), old code: UEEA4653
- UECS2103 = Operating Systems
- UECS3043 = Parallel Processing
- UECM3093 = Operations Research II

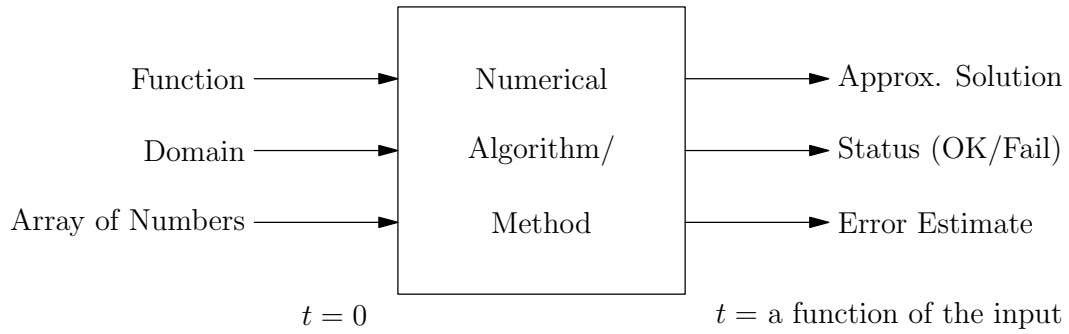
According to Hamming [1986], the *purpose of (numerical) computing is **insight**, not numbers!* Thus computing should be bound to the source of the problem and the use that is going to be made of the answers — it is not a step to be taken in isolation from reality. So we should know our problem and model and use the above framework to help us solve the problem. If more advanced numerical methods are required, we can consult the *Guide to Available Mathematical Software* (<http://gams.nist.gov/>).

40.2 What is a Numerical Algorithm?

Numerical methods are the procedures to find numerical solution of a given mathematical physics problem. There are many aspects for a numerical method — *data structures, algorithms, software implementation, numerical stability, convergence, time complexity* and *space complexity*. An algorithm is the core of a numerical method and Knuth [1998a] pointed out the five important characteristics of (any) “algorithm” are

1. *Finiteness*. An algorithm must always stop after a finite number of steps.
2. *Definiteness*. Each step of an algorithm must be precisely defined; the actions to be carried out must be rigorously and unambiguously specified for each case.
3. *Input*. An algorithm has zero or more inputs: quantities that are given to it initially before the algorithm begins, or dynamically as the algorithm runs.
4. *Output*. An algorithm has one or more outputs: quantities that have a specified relation to the inputs.
5. *Effectiveness*. An algorithm is generally expected to be *effective*, i.e. its operations must all be sufficiently basic that they can in principle be done exactly and in a finite length of time by someone using pencil and paper.

We can regard an algorithm to be the “internal” of the block:



So after the input “data” have been passed to a numerical algorithm at time $t = 0$, the algorithm will process the data and produce an approximation of solutions and hopefully an error estimation after some time.

The notion of numerical stability is related to the effect of error on a numerical algorithm.

Definition 40.1. A numerical algorithm is said to be

- *unstable* if a small initial error is magnified and propagated through the iteration process and degrading the overall accuracy of the calculation.
- *stable* if a small initial error remains small through the iteration process.
- *conditionally stable* if it is only stable for a certain set of initial condition, but unstable for all other initial conditions.

We must make sure that our numerical method is stable (or at least conditionally stable) so that we can be sure that the approximation we obtained is not too far from the true solution of the mathematical model.

40.3 Numeric Data Structures

How can we store *real numbers* such as 1000, -1.1 , $\sqrt{2}$, $\frac{80}{7}$, π in a computer? We can store them as “symbols”, but there is *no way to perform arithmetic on symbols*. By using the positional notation [Knuth, 1998b], a real number x can be expressed in decimal representation as

$$x = (-1)^s \cdot [\cdots + d_3 \cdot 10^3 + d_2 \cdot 10^2 + d_1 \cdot 10^1 + d_0 + d_{-1} \cdot 10^{-1} + d_{-2} \cdot 10^{-2} + \cdots] = (\cdots d_3 d_2 d_1 d_0 . d_{-1} d_{-2} \cdots)_{10} \quad (40.1)$$

where $d_i \in \{0, 1, \cdots, 9\}$ and s is either 0 or 1 depending on the sign of x . In particular,

$$1000 = 1000.0, \quad -1.1, \quad \frac{80}{7} = 11.428571, \quad \sqrt{2} = 1.41421356 \cdots, \quad \pi = 3.14159265 \cdots$$

However, digital computers can only *binary* data and we need a *binary representation* (instead of decimal representation) for the real number x :

$$x = (-1)^s \cdot [\cdots + a_3 \cdot 2^3 + a_2 \cdot 2^2 + a_1 \cdot 2^1 + a_0 + a_{-1} \cdot 2^{-1} + a_{-2} \cdot 2^{-2} + \cdots] = (\cdots a_3 a_2 a_1 a_0 . a_{-1} a_{-2} \cdots)_2 \quad (40.2)$$

where a_i are either 0 or 1. The ‘.’ is called the *radix point*.

The following examples demonstrate how to write an integer and a general real number in binary representation.

Example 40.1 (SPM Revision). Find the binary representation of the real number 1000.

Solution: Note that $2^{10} = 1024$, so a_i must be zero when $i \geq 10$ in (40.2) and there is no decimal places. Therefore,

$$1000 = a_9 2^9 + \cdots + a_2 2^2 + a_1 2 + a_0.$$

Using the knowledge of algebra from secondary school (or the knowledge of modular arithmetic from UECM1303 Discrete Mathematics with Applications), we will find that

$$a_9 = 1, \quad a_8 = 1, \quad a_7 = 1, \quad a_6 = 1, \quad a_5 = 1, \quad a_4 = 0, \quad a_3 = 1, \quad a_2 = 0, \quad a_1 = 0, \quad a_0 = 0.$$

Hence, the binary representation of 1000 is $\boxed{1111101000.0_2}$.

eg:bin-rep:-1.1

Example 40.2. Find the binary representation of the real number -1.1 .

Solution: Note that the integral part is 1, so $a_0 = 1$ and $a_i = 0$. Therefore, (40.2) becomes

$$1.1 = 1 + a_{-1} \cdot 2^{-1} + a_{-2} \cdot 2^{-2} + \dots$$

which reduces to the following problem

$$0.1 = \frac{a_{-1}}{2} + \frac{a_{-2}}{4} + \frac{a_{-3}}{8} + \frac{a_{-4}}{16} + \frac{a_{-5}}{32} + \dots$$

By multiplying both sides by 2, we obtain

$$0.2 = a_{-1} + \frac{a_{-2}}{2} + \frac{a_{-3}}{4} + \frac{a_{-4}}{8} + \frac{a_{-5}}{16} + \dots$$

We know that a_{-2} must be 0 because the integral part of the left hand side is 0. We can repeatedly multiply the equation by 2 to determine $a_{-2}, a_{-3}, \dots$:

decimal	integral
$0.2 \times 2 = 0.4$	0
$0.4 \times 2 = 0.8$	0
$0.8 \times 2 = 0.6$	1
$0.6 \times 2 = 0.2$	1
$0.2 \times 2 = 0.4$	0
$0.4 \times 2 = 0.8$	0 (repeating second statement)
$\vdots$	$\vdots$

A repetitive pattern is recognised and we can write down the binary representation as an infinite repetition:

$$\boxed{-1.1 = -1.0\overline{0011}_2 = -1.0001100110011\dots_2}$$

eg:bin-rep:0.1

Example 40.3. Use the binary representation in Example 40.2 to write down the binary representation of 0.1 and 0.2.

Solution. $-1.1 = -1.0\overline{0011}_2 = -1.0001100110011\dots_2$,

$$0.1 = 0.0\overline{001100110011\dots}_2 = \sum_{i=0}^{\infty} (2^{-4-4i} + 2^{-5-4i})$$

$$0.2 = 2 \times 0.1 = \sum_{i=0}^{\infty} (2^{-3-4i} + 2^{-4-4i}) = 0.001100110011\dots_2$$

□

sec:ieee-float

40.3.1 IEEE Floating Point Representation

To store every bit of -1.1 in a computer, we need infinite memory according to Example 40.2. However, this is impossible and so we have to truncate to round the binary representation for the real number x into a *floating point representation* as follows:

$$fl(x) = (-1)^s \cdot 2^e \cdot (1 + m)$$

where s denotes the sign bit, e denotes the exponent and m the mantissa of fixed finite bits.

In the early days when there was no standard representation, different companies used different representation and this made the life of the engineers that designed numerical algorithms very difficult. A new standard was proposed by IEEE to make sure that any numerical method running on a machine complying with IEEE standard produces the same result. Nearly all modern 32-bit and 64-bit CPUs comply with the following IEEE 754 standards:

- IEEE single precision (32-bit) floating point numbers: $s = 0$ for positive sign, $s = 1$ for negative sign, e is stored as an **8-bit exponent in excess 127 format**, m the **23-bit mantissa**;

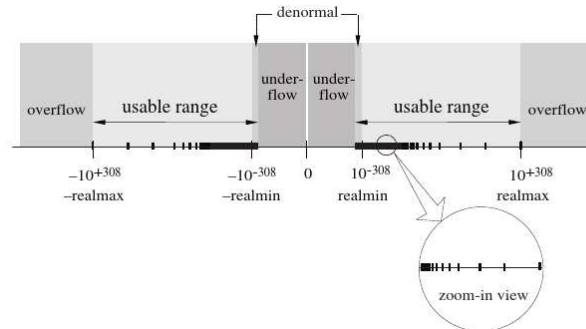
s	8-bit $e + 127$	23-bit m

- IEEE double precision (64-bit) floating point numbers: $s = 0$ for positive sign, $s = 1$ for negative sign, e is stored as an **11-bit exponent in excess 1023 format**, m the **52-bit mantissa**.

s	11-bit $e + 1023$	52-bit m

The latest standard that governs binary floating point arithmetic for most computers is the IEEE 754-2008 standard. It specifies number formats (float and double which have machine precision up to 2^{-24} and 2^{-53} respectively), basic operations, conversions, and exceptional conditions.

Floating point number system is very different from the real number system $\mathbb{R}$ used by mathematicians where the precision is infinite. Floating point numbers are finite, not equally spaced and do not occur with equal frequency. In particular, the IEEE floating point number system has the following structure:



The special floating point numbers are listed in the following table:

Special Values	Single Precision	Double Precision
$+\infty$	0x7F800000	0x7FF0000000000000
$-\infty$	0xFF800000	0xFFF0000000000000
$+0$	0x00000000	0x0000000000000000
-0	0x80000000	0x8000000000000000
Positive Quiet NaN	0x7FC00000 ~ 0x7FFFFFFF	0x7FF8000000000000 ~ 0x7FFFFFFFFFFFFFFF
Negative Quiet NaN	0xFFC00000 ~ 0xFFFFFFFF	0xFFF8000000000000 ~ 0xFFFFFFFFFFFFFFF
Positive Signaling NaN	0x7F800001 ~ 0x7FBFFFFF	0x7FF0000000000001 ~ 0x7FF7FFFFFFFF
Negative Signaling NaN	0xFF800001 ~ 0xFFBFFFFF	0xFFF0000000000001 ~ 0xFFF7FFFFFFFF

Let us look at some examples of representing real numbers in floating point format.

Example 40.4. Convert 1000 to IEEE floating point single precision format (in hexadecimal form, which will always be the case either in test or final exam).

Solution:

- 1000 is positive, Sign, $s = 0$.
- From Example 40.1, we know that $1000 = 111101000.0_2 = 1.11101000_2 \times 2^9$.
Hence, 23-bit Mantissa, $m = 1111010 \dots 0$
- 8-bit exponent in excess 127, $e + 127 = 9 + 127 = 10001000_2$

Therefore, 1000 in IEEE floating point single precision form is

0 100 0 100 0111 1010 0 0000 0000 0000 0000 000

In hexadecimal, it would be 0x447A0000.

Example 40.5. Convert 1000 to IEEE floating point double precision format.

Solution:

- Sign is positive, therefore $s = 0$.
 - From Example 40.1, we know that $1000 = 111101000.0_2 = 1.11101000_2 \times 2^9$.
Hence, 52-bit Mantissa, $m = 1111010 \dots 0$
 - 11-bit exponent in excess 1023, $e + 1023 = 9 + 1023 = 10000001000_2$
- Therefore, 1000 in IEEE floating point double precision form is

0 100 0000 1000 1111 0100 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000

In hexadecimal, it would be 0x408F400000000000.

Example 40.6 (April 2018 Final Exam Question 1(a)). Convert the real number 4321.375 to an IEEE single-precision floating-point number. Write your final result in hexadecimal representation. (12 marks)

Solution:

Sign is positive, $s = 0$ [1 mark]

With the use scientific calculator, the integral part could be converted to hexadecimal representation and then to binary representation by hand $4321 = 1'0000'1110'0001_2$ [1 mark]

Manual calculation is normally required for the decimal part:

$0.375 \times 2 = 0.75$ [1 mark]

$0.75 \times 2 = 1.5$ [1 mark]

$0.5 \times 2 = 1.0$ [1 mark]

Therefore, $0.375 = 0.0110$ [1 mark]

Expressing the number in a normalised form $4321.375 = 1.0000'1110'00010110_2 \times 2^{12}$ [2 marks]

Exponent = $12 + 127 = 1000'1011_2$ [2 marks]

The IEEE single-precision floating-point number (using chopping) is

$0100010110000'1110'000101100000000 = 45870B00_{16}$ [2 marks]

Example 40.7 (April 2014 Final Exam Question 1(a)). Convert the real number -13.6875 to 64-bit IEEE floating point number. Write your final result in hexadecimal representation. (8 marks)

Solution. Sign, $s = 1$ [1 mark]

$13.6875 = 1101.1011_2$ [2 marks]

$13.6875 = 1.1011011_2 \times 2^3$ [1 mark]

Exponent in excess 1023 = $3 + 1023 = 10000000010_2$ [2 marks]

64 bit IEEE double precision representation is

$110000000010101101100 \dots 0 = C02B600000000000_{16}$ [2 marks]

□

eg:tut1q1

Example 40.8 (Tutorial 1 Q1). Convert the decimal number $-\frac{2}{13}$ to IEEE 754 single precision floating-point format. (Always write your answer in hexadecimal form) 0xBE1D89D9

Solution. $s = 1$.

decimal $\frac{2}{13}$ $\frac{4}{13}$ $\frac{8}{13}$ $\frac{3}{13}$ $\frac{6}{13}$ $\frac{12}{13}$ $\frac{11}{13}$ $\frac{9}{13}$ $\frac{5}{13}$ $\frac{10}{13}$ $\frac{7}{13}$ $\frac{1}{13}$ $\frac{2}{13}$ repeat

$\times 2$ $\frac{4}{13}$ $\frac{8}{13}$ $\frac{16}{13}$ $\frac{6}{13}$ $\frac{12}{13}$ $\frac{24}{13}$ $\frac{22}{13}$ $\frac{18}{13}$ $\frac{10}{13}$ $\frac{20}{13}$ $\frac{14}{13}$ $\frac{2}{13}$ repeat

integral 0 0 1 0 0 1 1 1 0 1 1 0 repeat

$$\frac{2}{13} = \frac{a_{-1}}{2} + \frac{a_{-2}}{2^2} + \frac{a_{-3}}{2^3} + \dots \Rightarrow a_{-i} = \left\lfloor \frac{2 \times 2^i}{13} \right\rfloor \bmod 2$$

$$= 0.00100111011000_2$$

$$= (\text{normalisation}) 1.00111011000100111011000 \dots \times 2^{-3}$$

8-bit exponent: $e + 127 = 124 = 0111_21100_2$

Single-precision representation is $10111110000111011000100111011000 = 0xBE1D89D8_{16}$ □

Example 40.9 (Tutorial 1 Q2). Convert the decimal number 73.171875 to IEEE 754 double precision floating-point format. 0x40524B0000000000

Solution. • $73 = 1001001_2$ (binary form of the integral part)

• $0.171875 = 0.001011_2$ (binary form of the decimal places)

• $73.171875 = 1001001.001011_2 = 1.001001001011_2 \times 2^{+6}$ (normalisation)

• 11-bit exponent, $e + 1023 = 1029 = 100_20000_20101_2$

Double-precision representation is

$$01000000010100100100101100 \dots 0_2 = 40524b0000000000_{16}.$$

□

We will now look at examples of recovering real numbers from the floating point format representations.

Example 40.10. Convert the single-precision floating point number $C0C40000_{16}$ to a decimal number.

Solution:

$$C0C40000_{16} = 110000001100010000000000000000_2$$

Sign, $s = 1$ (negative)

$$\text{Exponent (excess 127)}, e + 127 = 10000001_2 = 129 \Rightarrow e = 129 - 127 = 2$$

Significant = $1.\underline{10001}_2$

$$\text{Result} = -1.10001_2 \times 2^2 = -110.001_2 = -(2^2 + 2^1 + 2^{-3}) = \boxed{-6.125}.$$

Example 40.11. Convert the double-precision floating point number $C0C4000000000000_{16}$ to a decimal number.

Solution:

$$C0C40000...0_{16} = 110000001100010000000000\cdots0000_2$$

Sign, $s = 1$ (negative)

$$\text{Exponent (excess 1023)}, e + 1023 = 10000001100_2 = 1036 \Rightarrow e = 1036 - 1023 = 13$$

Significant = $1.\underline{01}_2$

$$\text{Result} = -1.01_2 \times 2^{13} = -(2^{13} + 2^{11}) = \boxed{-10240.0}$$

Example 40.12 (April 2014 Final Exam Question 5(a)). Convert the single-precision floating point number $C2168000_{16}$ to a decimal number. (6 marks)

Solution: $C2168000_{16} = 1100\ 0010\ 0001\ 0110\ 1000\ 0000\ 0000\ 0000_2$

Negative sign $\Rightarrow s = 1$ [1 mark]

$$e + 127 = 10000100_2 = 128 + 4 \Rightarrow e = 5$$
 [2 marks]

The decimal number is [1+1+1=3 marks]

$$-1.00101101_2 \times 2^5 = -100101.101_2 = -(32 + 4 + 1 + 2^{-1} + 2^{-3}) = -37.625$$

Example 40.13 (Tutorial 1 Q3). Convert the IEEE 32-bit floating-point number $C1280000_{16}$ to decimal.

$$\dots\dots\dots \boxed{-10.5}$$

Solution.

$$C1280000_{16} = \underline{110000010010100000000000000000}_2$$

• $s = 1$, negative;

• $e + 127 = 10000010_2 \Rightarrow e = 128 + 2 - 127 = 3$

• $-1.0101000000000000000000_2 \times 2^{10000010_2 - 127} = -1.0101_2 \times 2^3 = -1010.1_2 = -10.5$ $\square$

Example 40.14 (April 2017 Final Exam Question 1(b)). Convert the single-precision floating point format $0xC0450000$ to a decimal number. (8 marks)

Solution. $0xC0450000 = 1100'0000'0100'0101'0000'0000'0000'0000_2$

Sign, $s = 1$ (negative) [1 mark]

$$e + 127 = 10000000_2 = 128 \Rightarrow e = 1$$
 [2+1=3 marks]

The decimal number is

$$-1.100'0101_2 \times 2^1 = -11.000101 = -(2 + 1 + 2^{-4} + 2^{-6}) = -3.078125$$
 [1+2+1=4 marks]

$\square$

40.3.2 Floating Point Arithmetic

In Calculus, we learn that addition and multiplication are associative. However, floating point arithmetic [Isaacson and Keller, 1966, Chapter 1] are not associative. Higham [2002a, Chapter 1 & 2] and Knuth [1998b, Chapter 4] provide detail accounts of the analysis of errors for floating point numbers. Parhami [2000], Ercegovac and Lang [2004] and Kulisch [2008] explain how floating point arithmetic work at machine level. In this section, we will explain the difference between floating point arithmetic and the arithmetic over the set of real numbers $\mathbb{R}$.

First, let us note that there are two popular ways — *copping* and *rounding*, in which the floating point number

$$fl(x) = (-1)^s \cdot 2^e \cdot (1 + 0.m_1m_2\cdots m_t), \quad -N \leq e \leq M$$

is obtained from

$$x = (-1)^s \cdot 2^e \cdot (1 + 0.b_1b_2 \cdots b_tb_{t+1} \cdots), \quad (\text{Base 2})$$

where t , N and M are some integers. In single-precision, $t = 23$, $N = 126$, $M = 127$; in double-precision, $t = 52$, $N = 1022$, $M = 1023$.

In the chopping representation,

$$m_i = b_i \quad \text{for } i = 1, \dots, t.$$

In the rounding representation, we normally have

$$m_1m_2 \cdots m_t = \begin{cases} b_1b_2 \cdots b_t, & b_{t+1} = 0, \\ b_1b_2 \cdots b_t + 1, & b_{t+1} = 1. \end{cases}$$

Lemma 40.1. *The error in t -bit floating point representation of a number $x \neq 0$ is bounded by*

$$|x - fl(x)| \leq 2^e |0.b_1 \cdots b_tb_{t+1} \cdots - 0.m_1 \cdots m_t| \leq \begin{cases} 2^e \cdot 2^{-t}, & \text{chopping,} \\ 2^e \cdot 2^{-t-1}, & \text{rounding.} \end{cases}$$

The proof can be found in Knuth [1998b, p. 232].

Therefore the floating point arithmetic has the following error estimate:

Theorem 40.15. *Let $\diamond$ denote the basic arithmetic operations $+$, $-$, $\times$, $/$. Then*

$$fl(x \diamond y) = (x \diamond y)(1 + \delta), \quad |\delta| \leq u = \begin{cases} 2^{-t}, & \text{chopping,} \\ 2^{-t-1}, & \text{rounding.} \end{cases}$$

It is common for computers to use rounding and in this case, the value u is called the *unit round off* or *theoretical machine precision*. In the single-precision case, $t = 23 \Rightarrow u \approx 5.96 \times 10^{-8}$; in the double-precision case, $t = 52 \Rightarrow u \approx 1.11 \times 10^{-16}$.

However, in some books, machine precision for single and double machine precision are defined to be approximately 1.19×10^{-7} and 2.22×10^{-16} respectively. Why is this the case? This is because “computer” machine precision is defined to be the smallest number $\epsilon_{\min}$ that adds to the floating point number 1.0 such that $1.0 + \epsilon_{\min} \neq 1.0$.

In order to see the difference between real arithmetic and floating arithmetic, we will investigate the following example.

Example 40.16. Let $a = 2234/9999$ and $b = 9999/2234$. If $fl(\cdot)$ denotes the double precision floating point number, what is the value of $a \times b$, $fl(a)$, $fl(b)$ and $fl(a \times b)$? Do we have $a \times b = fl(a \times b)$?

Solution: $a \times b = 1$
 $fl(a) = 0.22342234223422341$
 $fl(b) = 4.475828111011638$
 $fl(a \times b) = fl(a) \times fl(b) = 0.9999999999999999 \neq 1$
 Therefore $a \times b \neq fl(a \times b)$.

Example 40.17 (Tutorial 1 Q4). Computer addition $\oplus$ may be commutative but (this may not be always true since the number stored in the CPU may have higher accuracy than the number stored in memory), but it is not associative, study an example of a computer with *5-digit floating point arithmetic with rounding* to demonstrate this. For instance, 1.8345, 30.123, and 350.33.

Solution:

$$\begin{aligned} (1.8345 \oplus 30.123) \oplus 350.33 &= fl(fl(30.9575) + 350.33) = fl(30.958 + 350.33) \\ &= fl(381.288) = 382.29, \\ 1.8345 \oplus (30.123 \oplus 350.33) &= fl(1.8345 + fl(380.453)) \\ &= fl(1.8345 + 380.45) = fl(382.2845) = 382.28 \end{aligned}$$

If one types $(+ 0.1 0.2)$ in Scheme, we get 0.30000000000000004 due to rounding error, i.e.

$$fl(0.1) = fl(0.00011_2) = 3 \times \sum_{i=0}^{12} 2^{-5-4i} + 2^{-55}$$

$$fl(0.2) = fl(0.00110_2) = 3 \times \sum_{i=0}^{12} 2^{-4-4i}$$

leading to $fl(0.1) + fl(0.2) = fl(0.3) + 2^{-55}$ (a rounding error 2^{-55} is produced).

Due to the limitation of finite precision arithmetic in computer hardware, rounding error is unavoidable. However, it is not the only error that we encounter in numerical methods. The other two main source of errors in numerical methods are:

1. *truncation error* or *discretisation error*: continuous functions are approximated by polynomial functions (see Section [40.3.3](#));
2. *data uncertainty error* or *measurement error*:
 - from errors of measurement or estimation such as engineering, economics (data in economics can hardly be very precise);
 - errors in storing the data on computer (rounding errors, usually very tiny);
 - from the results of errors in an earlier computation if the data is itself the solution to another problem.

Two useful measures of error which are often used in the stopping of numerical algorithm in later topics are defined below.

Definition 40.2. Let $\hat{x}$ be a data value with error and x be the theoretically perfect data value. The *absolute error* of $\hat{x}$ is

$$E_{abs}(fl(x)) = |x - fl(x)|,$$

and the *relative error* of $\hat{x}$ is

$$E_{rel}(fl(x)) = \frac{|x - fl(x)|}{|x|}, \quad (\text{undefined when } x = 0).$$

In numerical methods, relative error is more of interest because its is theoretically scale independent, i.e. $x \rightarrow \alpha x$ and $fl(x) \rightarrow \alpha fl(x)$ leaves $E_{rel}(fl(x))$ almost unchanged.

Note that the operations of floating point multiplication and division do not magnify the relative error very much but floating point subtraction of nearly equal quantities can greatly increase the relative error. Therefore, large loss of accuracy is expected from additions and subtractions and seldom from multiplications and divisions [Knuth, 1998b, p. 229].

Suppose that an approximation $\hat{y} = fl(f(x))$ to $y = f(x)$ is computed in an arithmetic of precision u , where $f: \mathbb{R} \rightarrow \mathbb{R}$. How should we measure the “quantity” of $\hat{y}$? There are two ways:

- *Forward error analysis*: Study the forward error $f(x) - \hat{y}$
- *Backward error analysis*: Find the smallest value Δx , called the *backward error*, such that $\hat{y} = f(x + \Delta x)$.

Using the backward and forward errors, mathematicians study the numerical stability for a numerical algorithm. Brent’s Multiprecision Package from Netlib provides ways to improve precision in numerical computation.

40.3.3 Floating-Point Mathematical Functions

In calculus, we learned the following elementary functions:

- power, square roots and n th roots: x^n , $\sqrt{x}$, $\sqrt[n]{x}$.
- trigonometry functions: Sine, cosine, tangent, cotangent, secant and cosecant;
- inverse trig functions: Arcsine, arccosine, arctan, etc;
- exponents and logarithms: $\exp$, $\log$ (in numerical methods, $\log$ always means natural log);

- hyperbolic functions: $\sinh x = \frac{e^x - e^{-x}}{2}$, $\cosh x = \frac{e^x + e^{-x}}{2}$, $\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

How can we implement these functions if we are only allow to calculate them using addition, subtraction, multiplication and division over IEEE floating point numbers? We will have to approximate them using either polynomial (sometimes using Taylor series approximation) or the ratio of polynomials. Let us investigate just exponential function example below, for other functions, the references Cody, Jr and Waite [1980], Hart [1978] and Muller [1997] provide detail explanations of many elementary functions and special functions.

The implementation of an exponential function e^x over IEEE floating numbers begins with the following argument reduction of x : There is a k such that

$$2^{k-0.5} \leq e^x < 2^{k+0.5} \Rightarrow k \ln 2 - 0.5 \ln 2 \leq x < k \ln 2 + 0.5 \ln 2 \Rightarrow |x - k \ln 2| \leq 0.5 \ln 2.$$

Hence, given x , find r and integer k such that

$$x = k \ln 2 + r, \quad |r| \leq 0.5 \ln 2.$$

Therefore, the calculation is reduced to

$$\exp(x) = e^x = 2^k \exp(r)$$

and we just need to estimate $\exp(r)$ properly. By using calculator, it is not difficult to find

$$|r| \leq 0.5 \ln 2 < 0.34658.$$

We need to approximate $\exp(r)$ by a special rational function on the interval $[0, 0.34658]$. By using Remez algorithm (http://en.wikipedia.org/wiki/Remez_algorithm), a technique to generate the rational function approximation for

$$\exp(r) = 1 + r + \frac{r R_1(r)}{2 - R_1(r)}$$

where

$$R_1(r) = r - (P_1 r^2 + P_2 r^4 + P_3 r^6 + P_4 r^8 + P_5 r^{10}).$$

By taking the special IEEE floating point values in the table on Page [915](#) into account:

$$\exp(+\infty) = +\infty, \quad \exp(\text{NaN}) = \text{NaN}, \quad \exp(-\infty) = 0.$$

Sun Microsystems (which was acquired by Oracle) gave a C implementation below:

```

1 /* @(#)netlib/fdlibm/e_exp.c 1.6 04/04/22 */
2 /* (C) 2004 by Sun Microsystems, Inc. All rights reserved.
3  * Permission to use, copy, modify, and distribute this software
4  * is freely granted, provided that this notice is preserved.
5  */
6
7 /* fdlibm_exp(x): Returns the exponential of x.
8  *
9  * Accuracy:
10 * according to an error analysis, the error is always less than
11 * 1 ulp (unit in the last place).
12 *
13 * Misc. info.
14 * For IEEE double
15 * if x > 7.09782712893383973096e+02 then exp(x) overflow
16 * if x < -7.45133219101941108420e+02 then exp(x) underflow
17 *
18 * Constants:
19 * The hexadecimal values are the intended ones for the following
20 * constants. The decimal values may be used, provided that the
21 * compiler will convert from decimal to binary accurately enough
22 * to produce the hexadecimal values shown.
23 */
24
25 #define __HI(x) *(1+(int*)&x) // Only works on Little Endian CPU (e.g. Intel)
26 #define __LO(x) *(int*)&x
27
```

```

28 static const double
29 one      = 1.0,
30 halF[2]  = {0.5,-0.5},
31 huge     = 1.0e+300,
32 twom1000= 9.33263618503218878990e-302, /* 2**-1000=0x01700000,0*/
33 o_threshold= 7.09782712893383973096e+02, /* 0x40862E42, 0xFEFA39EF */
34 u_threshold= -7.45133219101941108420e+02, /* 0xc0874910, 0xD52D3051 */
35 ln2HI[2]   = { 6.93147180369123816490e-01, /* 0x3fe62e42, 0xfef00000 */
36               -6.93147180369123816490e-01}, /* 0xbfe62e42, 0xfef00000 */
37 ln2LO[2]   = { 1.90821492927058770002e-10, /* 0x3dea39ef, 0x35793c76 */
38               -1.90821492927058770002e-10}, /* 0xbdea39ef, 0x35793c76 */
39 invln2     = 1.44269504088896338700e+00, /* 0x3ff71547, 0x652b82fe */
40 P1         = 1.666666666666666019037e-01, /* 0x3FC55555, 0x5555553E */
41 P2         = -2.77777777770155933842e-03, /* 0xBF66C16C, 0x16BEBD93 */
42 P3         = 6.61375632143793436117e-05, /* 0x3F11566A, 0xAF25DE2C */
43 P4         = -1.65339022054652515390e-06, /* 0xBEBBBD41, 0xC5D26BF1 */
44 P5         = 4.13813679705723846039e-08; /* 0x3E663769, 0x72BEA4D0 */
45
46
47 double fdlibm_exp(double x) {
48     double y,hi,lo,c,t;
49     int k,xsb;
50     unsigned hx;
51
52     hx = __HI(x); /* high word of x */
53     xsb = (hx>>31)&1; /* sign bit of x */
54     hx &= 0x7fffffff; /* high word of |x| */
55
56     /* filter out non-finite argument */
57     if(hx >= 0x40862E42) { /* if |x|>=709.78... */
58         if(hx>=0x7ff00000) {
59             if(((hx&0xfffff)|__LO(x))!=0)
60                 return x+x; /* NaN */
61             else return (xsb==0)? x:0.0; /* exp(+/-inf)={inf,0} */
62         }
63         if(x > o_threshold) return huge*huge; /* overflow */
64         if(x < u_threshold) return twom1000*twom1000; /* underflow */
65     }
66
67     /* argument reduction */
68     if(hx > 0x3fd62e42) { /* if |x| > 0.5 ln2 */
69         if(hx < 0x3FF0A2B2) { /* and |x| < 1.5 ln2 */
70             hi = x-ln2HI[xsb]; lo=ln2LO[xsb]; k = 1-xsb-xsb;
71         } else {
72             k = (int)(invln2*x+halF[xsb]);
73             t = k;
74             hi = x - t*ln2HI[0]; /* t*ln2HI is exact here */
75             lo = t*ln2LO[0];
76         }
77         x = hi - lo;
78     } else if(hx < 0x3e300000) { /* when |x|<2**-28 */
79         if(huge+x>one) return one+x; /* trigger inexact */
80     } else k = 0;
81
82     /* x is now in primary range */
83     t = x*x;
84     c = x - t*(P1+t*(P2+t*(P3+t*(P4+t*P5))));
85     if(k==0) return one-((x*c)/(c-2.0)-x);
86     else y = one-((lo-(x*c)/(2.0-c))-hi);
87     if(k >= -1021) {
88         __HI(y) += (k<<20); /* add k to y's exponent */
89         return y;
90     } else {
91         __HI(y) += ((k+1000)<<20); /* add k to y's exponent */
92         return y*twom1000;
93     }
94 }

```

40.4 Programming Environment

In 1960s–1980s, computers were very expensive (see the picture of an old PDP-11 computer below) and software were implemented in *assembly language* (Section 40.4.1) and later on with *compiled programming language* (such as Fortran in Section 41.1).



<http://cm.bell-labs.com/who/dmr/picture.html>

Programs written in assembly languages were very difficult to read and debug. Compiled programming languages such as Pascal, Fortran and C were developed to allow programmers to write programs that are readable. A compiled programming language translates a program into assembly language and then calls an assembler to assemble it into either object or executable files which are binary files. During the period from 1960s to 1980s, many numerical methods have been implemented in the popular programming language such as pascal, C and Fortran 77 [Press et al., 1992a, 1992b, 1992c, 1992d, 1992e, 1992f, 1992g, 1992h, 1992i, 1992j, 1992k, 1992l, 1992m, 1992n, 1992o, 1992p, 1992q, 1992r, 1992s, 1992t, 1992u, 1992v, 1992w, 1992x, 1992y, 1992z, 1993a, 1993b, 1993c, 1993d, 1993e, 1993f, 1993g, 1993h, 1993i, 1993j, 1993k, 1993l, 1993m, 1993n, 1993o, 1993p, 1993q, 1993r, 1993s, 1993t, 1993u, 1993v, 1993w, 1993x, 1993y, 1993z, 1994a, 1994b, 1994c, 1994d, 1994e, 1994f, 1994g, 1994h, 1994i, 1994j, 1994k, 1994l, 1994m, 1994n, 1994o, 1994p, 1994q, 1994r, 1994s, 1994t, 1994u, 1994v, 1994w, 1994x, 1994y, 1994z, 1995a, 1995b, 1995c, 1995d, 1995e, 1995f, 1995g, 1995h, 1995i, 1995j, 1995k, 1995l, 1995m, 1995n, 1995o, 1995p, 1995q, 1995r, 1995s, 1995t, 1995u, 1995v, 1995w, 1995x, 1995y, 1995z, 1996a, 1996b, 1996c, 1996d, 1996e, 1996f, 1996g, 1996h, 1996i, 1996j, 1996k, 1996l, 1996m, 1996n, 1996o, 1996p, 1996q, 1996r, 1996s, 1996t, 1996u, 1996v, 1996w, 1996x, 1996y, 1996z, 1997a, 1997b, 1997c, 1997d, 1997e, 1997f, 1997g, 1997h, 1997i, 1997j, 1997k, 1997l, 1997m, 1997n, 1997o, 1997p, 1997q, 1997r, 1997s, 1997t, 1997u, 1997v, 1997w, 1997x, 1997y, 1997z, 1998a, 1998b, 1998c, 1998d, 1998e, 1998f, 1998g, 1998h, 1998i, 1998j, 1998k, 1998l, 1998m, 1998n, 1998o, 1998p, 1998q, 1998r, 1998s, 1998t, 1998u, 1998v, 1998w, 1998x, 1998y, 1998z, 1999a, 1999b, 1999c, 1999d, 1999e, 1999f, 1999g, 1999h, 1999i, 1999j, 1999k, 1999l, 1999m, 1999n, 1999o, 1999p, 1999q, 1999r, 1999s, 1999t, 1999u, 1999v, 1999w, 1999x, 1999y, 1999z, 2000a, 2000b, 2000c, 2000d, 2000e, 2000f, 2000g, 2000h, 2000i, 2000j, 2000k, 2000l, 2000m, 2000n, 2000o, 2000p, 2000q, 2000r, 2000s, 2000t, 2000u, 2000v, 2000w, 2000x, 2000y, 2000z, 2001a, 2001b, 2001c, 2001d, 2001e, 2001f, 2001g, 2001h, 2001i, 2001j, 2001k, 2001l, 2001m, 2001n, 2001o, 2001p, 2001q, 2001r, 2001s, 2001t, 2001u, 2001v, 2001w, 2001x, 2001y, 2001z, 2002a, 2002b, 2002c, 2002d, 2002e, 2002f, 2002g, 2002h, 2002i, 2002j, 2002k, 2002l, 2002m, 2002n, 2002o, 2002p, 2002q, 2002r, 2002s, 2002t, 2002u, 2002v, 2002w, 2002x, 2002y, 2002z, 2003a, 2003b, 2003c, 2003d, 2003e, 2003f, 2003g, 2003h, 2003i, 2003j, 2003k, 2003l, 2003m, 2003n, 2003o, 2003p, 2003q, 2003r, 2003s, 2003t, 2003u, 2003v, 2003w, 2003x, 2003y, 2003z, 2004a, 2004b, 2004c, 2004d, 2004e, 2004f, 2004g, 2004h, 2004i, 2004j, 2004k, 2004l, 2004m, 2004n, 2004o, 2004p, 2004q, 2004r, 2004s, 2004t, 2004u, 2004v, 2004w, 2004x, 2004y, 2004z, 2005a, 2005b, 2005c, 2005d, 2005e, 2005f, 2005g, 2005h, 2005i, 2005j, 2005k, 2005l, 2005m, 2005n, 2005o, 2005p, 2005q, 2005r, 2005s, 2005t, 2005u, 2005v, 2005w, 2005x, 2005y, 2005z, 2006a, 2006b, 2006c, 2006d, 2006e, 2006f, 2006g, 2006h, 2006i, 2006j, 2006k, 2006l, 2006m, 2006n, 2006o, 2006p, 2006q, 2006r, 2006s, 2006t, 2006u, 2006v, 2006w, 2006x, 2006y, 2006z, 2007a, 2007b, 2007c, 2007d, 2007e, 2007f, 2007g, 2007h, 2007i, 2007j, 2007k, 2007l, 2007m, 2007n, 2007o, 2007p, 2007q, 2007r, 2007s, 2007t, 2007u, 2007v, 2007w, 2007x, 2007y, 2007z, 2008a, 2008b, 2008c, 2008d, 2008e, 2008f, 2008g, 2008h, 2008i, 2008j, 2008k, 2008l, 2008m, 2008n, 2008o, 2008p, 2008q, 2008r, 2008s, 2008t, 2008u, 2008v, 2008w, 2008x, 2008y, 2008z, 2009a, 2009b, 2009c, 2009d, 2009e, 2009f, 2009g, 2009h, 2009i, 2009j, 2009k, 2009l, 2009m, 2009n, 2009o, 2009p, 2009q, 2009r, 2009s, 2009t, 2009u, 2009v, 2009w, 2009x, 2009y, 2009z, 2010a, 2010b, 2010c, 2010d, 2010e, 2010f, 2010g, 2010h, 2010i, 2010j, 2010k, 2010l, 2010m, 2010n, 2010o, 2010p, 2010q, 2010r, 2010s, 2010t, 2010u, 2010v, 2010w, 2010x, 2010y, 2010z, 2011a, 2011b, 2011c, 2011d, 2011e, 2011f, 2011g, 2011h, 2011i, 2011j, 2011k, 2011l, 2011m, 2011n, 2011o, 2011p, 2011q, 2011r, 2011s, 2011t, 2011u, 2011v, 2011w, 2011x, 2011y, 2011z, 2012a, 2012b, 2012c, 2012d, 2012e, 2012f, 2012g, 2012h, 2012i, 2012j, 2012k, 2012l, 2012m, 2012n, 2012o, 2012p, 2012q, 2012r, 2012s, 2012t, 2012u, 2012v, 2012w, 2012x, 2012y, 2012z, 2013a, 2013b, 2013c, 2013d, 2013e, 2013f, 2013g, 2013h, 2013i, 2013j, 2013k, 2013l, 2013m, 2013n, 2013o, 2013p, 2013q, 2013r, 2013s, 2013t, 2013u, 2013v, 2013w, 2013x, 2013y, 2013z, 2014a, 2014b, 2014c, 2014d, 2014e, 2014f, 2014g, 2014h, 2014i, 2014j, 2014k, 2014l, 2014m, 2014n, 2014o, 2014p, 2014q, 2014r, 2014s, 2014t, 2014u, 2014v, 2014w, 2014x, 2014y, 2014z, 2015a, 2015b, 2015c, 2015d, 2015e, 2015f, 2015g, 2015h, 2015i, 2015j, 2015k, 2015l, 2015m, 2015n, 2015o, 2015p, 2015q, 2015r, 2015s, 2015t, 2015u, 2015v, 2015w, 2015x, 2015y, 2015z, 2016a, 2016b, 2016c, 2016d, 2016e, 2016f, 2016g, 2016h, 2016i, 2016j, 2016k, 2016l, 2016m, 2016n, 2016o, 2016p, 2016q, 2016r, 2016s, 2016t, 2016u, 2016v, 2016w, 2016x, 2016y, 2016z, 2017a, 2017b, 2017c, 2017d, 2017e, 2017f, 2017g, 2017h, 2017i, 2017j, 2017k, 2017l, 2017m, 2017n, 2017o, 2017p, 2017q, 2017r, 2017s, 2017t, 2017u, 2017v, 2017w, 2017x, 2017y, 2017z, 2018a, 2018b, 2018c, 2018d, 2018e, 2018f, 2018g, 2018h, 2018i, 2018j, 2018k, 2018l, 2018m, 2018n, 2018o, 2018p, 2018q, 2018r, 2018s, 2018t, 2018u, 2018v, 2018w, 2018x, 2018y, 2018z, 2019a, 2019b, 2019c, 2019d, 2019e, 2019f, 2019g, 2019h, 2019i, 2019j, 2019k, 2019l, 2019m, 2019n, 2019o, 2019p, 2019q, 2019r, 2019s, 2019t, 2019u, 2019v, 2019w, 2019x, 2019y, 2019z, 2020a, 2020b, 2020c, 2020d, 2020e, 2020f, 2020g, 2020h, 2020i, 2020j, 2020k, 2020l, 2020m, 2020n, 2020o, 2020p, 2020q, 2020r, 2020s, 2020t, 2020u, 2020v, 2020w, 2020x, 2020y, 2020z, 2021a, 2021b, 2021c, 2021d, 2021e, 2021f, 2021g, 2021h, 2021i, 2021j, 2021k, 2021l, 2021m, 2021n, 2021o, 2021p, 2021q, 2021r, 2021s, 2021t, 2021u, 2021v, 2021w, 2021x, 2021y, 2021z, 2022a, 2022b, 2022c, 2022d, 2022e, 2022f, 2022g, 2022h, 2022i, 2022j, 2022k, 2022l, 2022m, 2022n, 2022o, 2022p, 2022q, 2022r, 2022s, 2022t, 2022u, 2022v, 2022w, 2022x, 2022y, 2022z, 2023a, 2023b, 2023c, 2023d, 2023e, 2023f, 2023g, 2023h, 2023i, 2023j, 2023k, 2023l, 2023m, 2023n, 2023o, 2023p, 2023q, 2023r, 2023s, 2023t, 2023u, 2023v, 2023w, 2023x, 2023y, 2023z, 2024a, 2024b, 2024c, 2024d, 2024e, 2024f, 2024g, 2024h, 2024i, 2024j, 2024k, 2024l, 2024m, 2024n, 2024o, 2024p, 2024q, 2024r, 2024s, 2024t, 2024u, 2024v, 2024w, 2024x, 2024y, 2024z, 2025a, 2025b, 2025c, 2025d, 2025e, 2025f, 2025g, 2025h, 2025i, 2025j, 2025k, 2025l, 2025m, 2025n, 2025o, 2025p, 2025q, 2025r, 2025s, 2025t, 2025u, 2025v, 2025w, 2025x, 2025y, 2025z, 2026a, 2026b, 2026c, 2026d, 2026e, 2026f, 2026g, 2026h, 2026i, 2026j, 2026k, 2026l, 2026m, 2026n, 2026o, 2026p, 2026q, 2026r, 2026s, 2026t, 2026u, 2026v, 2026w, 2026x, 2026y, 2026z, 2027a, 2027b, 2027c, 2027d, 2027e, 2027f, 2027g, 2027h, 2027i, 2027j, 2027k, 2027l, 2027m, 2027n, 2027o, 2027p, 2027q, 2027r, 2027s, 2027t, 2027u, 2027v, 2027w, 2027x, 2027y, 2027z, 2028a, 2028b, 2028c, 2028d, 2028e, 2028f, 2028g, 2028h, 2028i, 2028j, 2028k, 2028l, 2028m, 2028n, 2028o, 2028p, 2028q, 2028r, 2028s, 2028t, 2028u, 2028v, 2028w, 2028x, 2028y, 2028z, 2029a, 2029b, 2029c, 2029d, 2029e, 2029f, 2029g, 2029h, 2029i, 2029j, 2029k, 2029l, 2029m, 2029n, 2029o, 2029p, 2029q, 2029r, 2029s, 2029t, 2029u, 2029v, 2029w, 2029x, 2029y, 2029z, 2030a, 2030b, 2030c, 2030d, 2030e, 2030f, 2030g, 2030h, 2030i, 2030j, 2030k, 2030l, 2030m, 2030n, 2030o, 2030p, 2030q, 2030r, 2030s, 2030t, 2030u, 2030v, 2030w, 2030x, 2030y, 2030z, 2031a, 2031b, 2031c, 2031d, 2031e, 2031f, 2031g, 2031h, 2031i, 2031j, 2031k, 2031l, 2031m, 2031n, 2031o, 2031p, 2031q, 2031r, 2031s, 2031t, 2031u, 2031v, 2031w, 2031x, 2031y, 2031z, 2032a, 2032b, 2032c, 2032d, 2032e, 2032f, 2032g, 2032h, 2032i, 2032j, 2032k, 2032l, 2032m, 2032n, 2032o, 2032p, 2032q, 2032r, 2032s, 2032t, 2032u, 2032v, 2032w, 2032x, 2032y, 2032z, 2033a, 2033b, 2033c, 2033d, 2033e, 2033f, 2033g, 2033h, 2033i, 2033j, 2033k, 2033l, 2033m, 2033n, 2033o, 2033p, 2033q, 2033r, 2033s, 2033t, 2033u, 2033v, 2033w, 2033x, 2033y, 2033z, 2034a, 2034b, 2034c, 2034d, 2034e, 2034f, 2034g, 2034h, 2034i, 2034j, 2034k, 2034l, 2034m, 2034n, 2034o, 2034p, 2034q, 2034r, 2034s, 2034t, 2034u, 2034v, 2034w, 2034x, 2034y, 2034z, 2035a, 2035b, 2035c, 2035d, 2035e, 2035f, 2035g, 2035h, 2035i, 2035j, 2035k, 2035l, 2035m, 2035n, 2035o, 2035p, 2035q, 2035r, 2035s, 2035t, 2035u, 2035v, 2035w, 2035x, 2035y, 2035z, 2036a, 2036b, 2036c, 2036d, 2036e, 2036f, 2036g, 2036h, 2036i, 2036j, 2036k, 2036l, 2036m, 2036n, 2036o, 2036p, 2036q, 2036r, 2036s, 2036t, 2036u, 2036v, 2036w, 2036x, 2036y, 2036z, 2037a, 2037b, 2037c, 2037d, 2037e, 2037f, 2037g, 2037h, 2037i, 2037j, 2037k, 2037l, 2037m, 2037n, 2037o, 2037p, 2037q, 2037r, 2037s, 2037t, 2037u, 2037v, 2037w, 2037x, 2037y, 2037z, 2038a, 2038b, 2038c, 2038d, 2038e, 2038f, 2038g, 2038h, 2038i, 2038j, 2038k, 2038l, 2038m, 2038n, 2038o, 2038p, 2038q, 2038r, 2038s, 2038t, 2038u, 2038v, 2038w, 2038x, 2038y, 2038z, 2039a, 2039b, 2039c, 2039d, 2039e, 2039f, 2039g, 2039h, 2039i, 2039j, 2039k, 2039l, 2039m, 2039n, 2039o, 2039p, 2039q, 2039r, 2039s, 2039t, 2039u, 2039v, 2039w, 2039x, 2039y, 2039z, 2040a, 2040b, 2040c, 2040d, 2040e, 2040f, 2040g, 2040h, 2040i, 2040j, 2040k, 2040l, 2040m, 2040n, 2040o, 2040p, 2040q, 2040r, 2040s, 2040t, 2040u, 2040v, 2040w, 2040x, 2040y, 2040z, 2041a, 2041b, 2041c, 2041d, 2041e, 2041f, 2041g, 2041h, 2041i, 2041j, 2041k, 2041l, 2041m, 2041n, 2041o, 2041p, 2041q, 2041r, 2041s, 2041t, 2041u, 2041v, 2041w, 2041x, 2041y, 2041z, 2042a, 2042b, 2042c, 2042d, 2042e, 2042f, 2042g, 2042h, 2042i, 2042j, 2042k, 2042l, 2042m, 2042n, 2042o, 2042p, 2042q, 2042r, 2042s, 2042t, 2042u, 2042v, 2042w, 2042x, 2042y, 2042z, 2043a, 2043b, 2043c, 2043d, 2043e, 2043f, 2043g, 2043h, 2043i, 2043j, 2043k, 2043l, 2043m, 2043n, 2043o, 2043p, 2043q, 2043r, 2043s, 2043t, 2043u, 2043v, 2043w, 2043x, 2043y, 2043z, 2044a, 2044b, 2044c, 2044d, 2044e, 2044f, 2044g, 2044h, 2044i, 2044j, 2044k, 2044l, 2044m, 2044n, 2044o, 2044p, 2044q, 2044r, 2044s, 2044t, 2044u, 2044v, 2044w, 2044x, 2044y, 2044z, 2045a, 2045b, 2045c, 2045d, 2045e, 2045f, 2045g, 2045h, 2045i, 2045j, 2045k, 2045l, 2045m, 2045n, 2045o, 2045p, 2045q, 2045r, 2045s, 2045t, 2045u, 2045v, 2045w, 2045x, 2045y, 2045z, 2046a, 2046b, 2046c, 2046d, 2046e, 2046f, 2046g, 2046h, 2046i, 2046j, 2046k, 2046l, 2046m, 2046n, 2046o, 2046p, 2046q, 2046r, 2046s, 2046t, 2046u, 2046v, 2046w, 2046x, 2046y, 2046z, 2047a, 2047b, 2047c, 2047d, 2047e, 2047f, 2047g, 2047h, 2047i, 2047j, 2047k, 2047l, 2047m, 2047n, 2047o, 2047p, 2047q, 2047r, 2047s, 2047t, 2047u, 2047v, 2047w, 2047x, 2047y, 2047z, 2048a, 2048b, 2048c, 2048d, 2048e, 2048f, 2048g, 2048h, 2048i, 2048j, 2048k, 2048l, 2048m, 2048n, 2048o, 2048p, 2048q, 2048r, 2048s, 2048t, 2048u, 2048v, 2048w, 2048x, 2048y, 2048z, 2049a, 2049b, 2049c, 2049d, 2049e, 2049f, 2049g, 2049h, 2049i, 2049j, 2049k, 2049l, 2049m, 2049n, 2049o, 2049p, 2049q, 2049r, 2049s, 2049t, 2049u, 2049v, 2049w, 2049x, 2049y, 2049z, 2050a, 2050b, 2050c, 2050d, 2050e, 2050f, 2050g, 2050h, 2050i, 2050j, 2050k, 2050l, 2050m, 2050n, 2050o, 2050p, 2050q, 2050r, 2050s, 2050t, 2050u, 2050v, 2050w, 2050x, 2050y, 2050z, 2051a, 2051b, 2051c, 2051d, 2051e, 2051f, 2051g, 2051h, 2051i, 2051j, 2051k, 2051l, 2051m, 2051n, 2051o, 2051p, 2051q, 2051r, 2051s, 2051t, 2051u, 2051v, 2051w, 2051x, 2051y, 2051z, 2052a, 2052b, 2052c, 2052d, 2052e, 2052f, 2052g, 2052h, 2052i, 2052j, 2052k, 2052l, 2052m, 2052n, 2052o, 2052p, 2052q, 2052r, 2052s, 2052t, 2052u, 2052v, 2052w, 2052x, 2052y, 2052z, 2053a, 2053b, 2053c, 2053d, 2053e, 2053f, 2053g, 2053h, 2053i, 2053j, 2053k, 2053l, 2053m, 2053n, 2053o, 2053p, 2053q, 2053r, 2053s, 2053t, 2053u, 2053v, 2053w, 2053x, 2053y, 2053z, 2054a, 2054b, 2054c, 2054d, 2054e, 2054f, 2054g, 2054h, 2054i, 2054j, 2054k, 2054l, 2054m, 2054n, 2054o, 2054p, 2054q, 2054r, 2054s, 2054t, 2054u, 2054v, 2054w, 2054x, 2054y, 2054z, 2055a, 2055b, 2055c, 2055d, 2055e, 2055f, 2055g, 2055h, 2055i, 2055j, 2055k, 2055l, 2055m, 2055n, 2055o, 2055p, 2055q, 2055r, 2055s, 2055t, 2055u, 2055v, 2055w, 2055x, 2055y, 2055z, 2056a, 2056b, 2056c, 2056d, 2056e, 2056f, 2056g, 2056h, 2056i, 2056j, 2056k, 2056l, 2056m, 2056n, 2056o, 2056p, 2056q, 2056r, 2056s, 2056t, 2056u, 2056v, 2056w, 2056x, 2056y, 2056z, 2057a, 2057b, 2057c, 2057d, 2057e, 2057f, 2057g, 2057h, 2057i, 2057j, 2057k, 2057l, 2057m, 2057n, 2057o

Solution: Using “gcc -O1 -S -fno-asynchronous-unwind-tables floatadd.f90” we can obtain the following AT&T style assembly language program

```

1  .file      "floatadd.f90"
2  .text
3  .section   .rodata.str1.1,"aMS",@progbits,1
4  .LC0:
5  .string   "floatadd.f90"
6  .section   .rodata
7  .align    8
8  .LC1:
9  .ascii    "(1f4.1, \" + \", 1f4.1, \" = \", 1f4.1)"
10 .text
11 .type     MAIN__, @function
12 MAIN__:
13 pushq    %rbp
14 pushq    %rbx
15 subq     $568, %rsp
16 movl     $40, %ebp
17 movq     %fs:0(%rbp), %rax
18 movq     %rax, 552(%rsp)
19 xorl     %eax, %eax
20 leaq     .LC0(%rip), %rax
21 movq     %rax, 24(%rsp)
22 movl     $3, 32(%rsp)
23 leaq     .LC1(%rip), %rax
24 movq     %rax, 96(%rsp)
25 movq     $35, 104(%rsp)
26 movl     $4096, 16(%rsp)
27 movl     $6, 20(%rsp)
28 leaq     16(%rsp), %rbx
29 movq     %rbx, %rdi
30 call     _gfortran_st_write@PLT
31 movl     $8, %edx
32 leaq     a.3770(%rip), %rsi
33 movq     %rbx, %rdi
34 call     _gfortran_transfer_real_write@PLT
35 movl     $8, %edx
36 leaq     b.3771(%rip), %rsi
37 movq     %rbx, %rdi
38 call     _gfortran_transfer_real_write@PLT
39 movsd    a.3770(%rip), %xmm0
40 addsd    b.3771(%rip), %xmm0
41 movsd    %xmm0, 8(%rsp)
42 leaq     8(%rsp), %rsi
43 movl     $8, %edx
44 movq     %rbx, %rdi
45 call     _gfortran_transfer_real_write@PLT
46 movq     %rbx, %rdi
47 call     _gfortran_st_write_done@PLT
48 movq     552(%rsp), %rax
49 xorq     %fs:0(%rbp), %rax
50 jne     .L4
51 addq     $568, %rsp
52 popq     %rbx
53 popq     %rbp
54 ret
55 .L4:
56 call     __stack_chk_fail@PLT
57 .size    MAIN__, .-MAIN__
58 .globl   main
59 .type    main, @function
60 main:
61 subq     $8, %rsp
62 call     _gfortran_set_args@PLT
63 leaq     options.1.3778(%rip), %rsi

```

```

64     movl    $7, %edi
65     call    _gfortran_set_options@PLT
66     call    MAIN__
67     movl    $0, %eax
68     addq    $8, %rsp
69     ret
70     .size   main, .-main
71     .data
72     .align  8
73     .type   b.3771, @object
74     .size   b.3771, 8
75 b.3771:
76     .long   2576980378
77     .long   1074108825
78     .align  8
79     .type   a.3770, @object
80     .size   a.3770, 8
81 a.3770:
82     .long   858993459
83     .long   1074475827
84     .section .rodata
85     .align  16
86     .type   options.1.3778, @object
87     .size   options.1.3778, 28
88 options.1.3778:
89     .long   68
90     .long   8191
91     .long   0
92     .long   1
93     .long   1
94     .long   0
95     .long   31
96     .ident  "GCC: (GNU) 8.2.1 20180831"
97     .section .note.GNU-stack,"",@progbits

```

Commands that begin with ‘.’ signals the assembler that they are directives, meaning a command to the assembler rather than something the assembler will translate into a CPU instruction.

The `.text` directive declares the start of running program and the *entry point* is called `main` (line 60). `main` is being called by the operating system when this program is run, so line 61 actually stores the address for `main` to return (line 69).

Many of the commands are rather cryptic. However, `addsd b.3771(%rip), %xmm0` on line 40 seems to be adding two floating point numbers. Indeed, `addsd` is an SSE instruction (http://en.wikipedia.org/wiki/Streaming_SIMD_Extensions) which adds two floating point numbers and stores the result in the XMM register.

Example 40.19. Study the assembly language program in the solution of Example [40.18](#). leg:assembly:add

1. Convert 3.4 to a 64-bit floating point number.

Solution: $s = 0$.

From

$$0.4 \times 2 = 0.8 + 0, \quad 0.8 \times 2 = 0.6 + 1, \quad 0.6 \times 2 = 0.2 + 1, \quad 0.2 \times 2 = 0.4 + 0,$$

we have

$$3.4 = 11.011001100110 \dots_2 = 1.1011001100110 \dots_2 \times 2^1.$$

Hence, $e + 1023 = 1024$ and the binary representation of the 64-bit floating number is

$$010000000000 \underbrace{1011001100110 \dots 011}_{52 \text{ bits}} = 0x400b333333333333.$$

2. Convert 2.7 to 64-bit IEEE floating point number.

Solution. From

$$0.7 \times 2 = 0.4 + 1, \quad 0.4 \times 2 = 0.8 + \boxed{0}, \quad 0.8 \times 2 = 0.6 + \boxed{1}, \quad 0.6 \times 2 = 0.2 + \boxed{1}, \quad 0.2 \times 2 = 0.4 + \boxed{0}.$$

we have

$$2.7 = 10.1011001100110 \cdots_2 = 1.01011001100110 \cdots_2 \times 2^1.$$

Hence, $e + 1023 = 1024$ and the binary representation of the 64-bit floating number is

$$010000000000 \underbrace{01011001100110 \cdots 011001}_{52 \text{ bits}} = 0x4005999999999999.$$

□

3. Explain where do the values 2576980378, 1074108825, 858993459, 1074475827 in lines 59, 60, 65, 66 come from?

Solution: By using Python, we found that they are the high and low part of the IEEE floating point numbers 3.4 and 2.7:

```
hex(2576980378) => 0x9999999a; hex(1074108825) => 0x40059999
hex(858993459)  => 0x33333333; hex(1074475827) => 0x400b3333
```

The reason for us to know about assembly language has to do with a famous Intel division bug, see http://en.wikipedia.org/wiki/Pentium_FDIV_bug. Because of the bug, the CPU can return incorrect decimal results. With the knowledge of assembly language, it is possible for us to know how computation speed optimisation works and very rarely, identify bugs in CPU affecting numerical precision. However, assembly language is rarely used in numerical methods design except for researchers aiming for best performance in scientific computing.

40.4.2 C11 and C++11

C11 has surpassed the older K&R C, ANSI C and becoming the modern standard C language. C++11 has surpassed C++98 and is the modern standard C++ language. They have incorporated the IEEE 754 floating point standards into the language specifications and compiler implementations. By default, decimal numbers are double, no postfix is required. A postfix 'f' is used for single precision floating point number.

These languages not only provide floating point data structures but also logical, character and array data types which are essential to the design of numerical methods. The data types are listed in the following table.

Definition	C11	C++11	Initial Value
Include/Import	<code><stdbool.h></code> or <code><stdint.h></code> or <code><complex.h></code>	<code><cstdint></code> or <code><cstdlib></code> or <code><wchar></code> or <code><complex></code>	-
Boolean type	<code>bool</code>	<code>bool</code>	false
signed 8 bits	<code>int8_t</code>	<code>int8_t</code>	0
unsigned 8 bits	<code>uint8_t</code>	<code>uint8_t</code>	0
signed 16 bits	<code>int16_t</code>	<code>int16_t</code>	0
unsigned 16 bits	<code>uint16_t</code>	<code>uint16_t</code>	0
signed 32 bits	<code>int32_t</code>	<code>int32_t</code>	0
unsigned 32 bits	<code>uint32_t</code>	<code>uint32_t</code>	0
signed 64 bits	<code>int64_t</code>	<code>int64_t</code>	0L
unsigned 64 bits	<code>uint64_t</code>	<code>uint64_t</code>	0L
32-bit floating point	<code>float, 1.0f</code>	<code>float</code>	<code>float.nan</code>
64-bit floating point	<code>double, 1.0</code>	<code>double</code>	<code>double.nan</code>
complex number type made of two doubles	<code>double complex</code>	<code>complex<double></code>	<code>double.nan + double.nan * 1.0i</code>
complex number type made of two reals	<code>complex</code>	-	<code>real.nan + real.nan * 1.0i</code>
UTF-8 code unit	-	<code>char</code>	0xFF
UTF-16 code unit	-	<code>char16_t</code>	0xFFFF
UTF-32 code unit/Unicode code point	-	<code>char32_t</code>	0x0000FFFF

We now investigate how to check our “IEEE floating point type” with C language in the examples below.

Example 40.20 (Tutorial 1 Q6). What is the output of the following C program?

```
#include <stdio.h>
#include <stdint.h>

int main() {
    uint32_t n1 = 0xbd800000U;
    uint64_t n2 = 0x40b0e11000000000UL;
    printf("%x = %f\n", n1, *(float*)&n1);
    printf("%lx = %lf\n", n2, *(double*)&n2);
    return 0;
}
```

Solution. (a) BD800000 = 1011 1101 1000 0000 ... 0000
 Sign = -ve
 $e + 127 = 01111011_2 = 123 \Rightarrow e = -4$
 $m = 0.0 \Rightarrow n_1 = -1.0 \times 2^{-4} = -0.0625$
 (b) 40B0E11000000000 = 0100 0000 1011 0000 1110 0001 0001 0000 ... 0000
 Sign = +ve
 $e + 1023 = 10000001011_2 = 1035 \Rightarrow e = 12$
 $m = 0.0000111000010001_2$
 $\Rightarrow n_2 = 1.0000111000010001_2 \times 2^{12} = 1000011100001.0001_2 = 4321.0625$
 Therefore the program output is
 bd800000 = -0.062500
 40b0e11000000000 = 4321.062500

□

Example 40.21. By using your scientific calculator and not computer, analyse the output of the following C program?

```
#include <stdio.h>
// https://gist.github.com/khajavi/5667960
// See also : http://class.ece.iastate.edu/arun/CprE281_F05/ieee754/ie5.html
```



```

union IEEE754single {
    struct {
        unsigned int mantissa : 23;
        unsigned int exponent : 8;
        unsigned short int sign : 1;
    } raw;
    float f;
} number;

int main() {
    number.f = -6.8;
    printf("%X, %X, %X\n", number.raw.mantissa, number.raw.exponent, number.raw.sign );
    return 0;
}

```

Solution. $s = 1$

- $6_{10} = 110_2$
- $0.8_{10} = 0.\overline{1100}_2$

$$\begin{aligned}
 6.8_{10} &= 110.\underbrace{110011001100110011001100}_\text{24 bits} \\
 &= 2^2 \times 1.10110011001100110011001_2 \quad \text{truncate} \\
 &= 2^2 \times 1.10110011001100110011010_2 \quad \text{round}
 \end{aligned}$$

Note that number is single precision, $2 + 127 = 1000\,0001_2$

$$1\,1000\,0001\,\underbrace{101\,1001\,1001\,1001\,1001\,1010}_\text{59999A}$$

printf's %x will print the hex form, i.e.

59999A, 81, 1

□

40.5 Numerical Functions in Software

It was illustrated in Section [40.3.3](#) ^{sec:approx-fn} that the implementation of elementary functions can be complicated with delicate approximation with piecewise polynomials. Even the calculation of x^n where n is some positive integer can be very complicated. According to the experiment carry out in <https://baptiste-wicht.com/posts/2017/09/cpp11-performance-tip-when-to-use-std-pow.html>, calculating x^n as the product of n number of x is only faster when $n < 110$ and using $\exp(n \times \ln x)$ is faster when $n \geq 110$.

In the compiled languages such as Fortran, C, C++ and the numerical computing environments such as R, Octave and Python, elementary functions below are implemented by default.

sqrt	exp	log									
sin	cos	tan	cot	sec	csc	asin	acos	atan	acot	asec	acsc
sinh	cosh	tanh	coth	sech	csch	asinh	acosh	atanh	acoth	asech	acsch

Note that Intel processor also uses polynomial-based approximations (the older Intel 8087 Math Co-processor uses CORDIC-based approximation).

In dealing with scientific, statistical and engineering problems, there are many other special functions associated with certain integrals and the analytic solutions of Partial Differential Equations (Chapter [34](#)) ^{ch:spec:intro} only available in additional libraries or modules (such as Scipy, IMSL Numerical Libraries, Cernlib (Fortran 77), GSL (GNU C Scientific Library), ROOT (C++), etc.

- Bessel J functions of various orders: `scipy.special.jv, jn`
- Bessel Y functions of various orders: `scipy.special.yv, yn`
- Bessel I functions of various orders: `scipy.special.iv`
- Bessel K functions of various orders: `scipy.special.kv, kn`
- Bessel j functions of various orders: `scipy.special.spherical_jn`
- Bessel y functions of various orders: `scipy.special.spherical_yn`

- Clausen integral function
- Wave functions for bound states and scattering solutions
- Dawson integral `scipy.special.dawsn`
- Dilogarithms for real arguments
- Dilogarithms for complex arguments `scipy.special.spence`
- Legendre form of the various types of complete Elliptic integrals: `scipy.special.ellipj`, `ellipk`, `ellipkm1`, `ellipe`
- Carlson and Legendre form of uncomplete Elliptic integrals: `scipy.special.ellipkinc`, `ellipeinc`
- Error function (ERFC) and complementary: `scipy.special.erf`, `erfc`
- Various type of exponential integrals: `scipy.special.expn`, `expl`, `expi`
- Sinh and Cosh integrals: `scipy.special.sici`
- Trigonometric integrals
- Computation of Arctangent integral
- Complete and incomplete Fermi-Dirac functions
- Gamma and related functions: `scipy.special.gamma`, `beta`, ...
- Confluent hypergeometric function: `scipy.special.hyp2f1`, `hyp1f1`, `hyperu`, `hyp0f1`, `hyp2f0`, `hyp1f2`, `hyp3f0`
- Gauss hypergeometric function
- Laguerre polynomials of various orders
- Associate Laguerre polynomials of various orders
- Legendre polynomials, functions and spherical harmonics
- Associate Legendre polynomials
- Digamma and polygamma functions: `scipy.special.psi`, `digamma`, `polygamma`, `multigamman`
- Riemann Zeta function: `scipy.special.zeta`, `zetac`
- Struve functions of order Zero and One: `scipy.special.struve`

The details of these functions can be found in “NIST Digital Library of Mathematical Functions” (<http://dlmf.nist.gov/>).

In the previous list, the functions are from the Python Scipy module. R and `scipy.stats` provide various characterisations of random variables such as probability distribution functions:

- One and 2 variable gaussian random numbers, pdf and cumulative distributions
- One dim exponential random numbers, pdf and cumulative distributions
- Cauchy (Lorentz) random numbers, pdf and cumulative distributions
- Landau random numbers, pdf and cumulative distributions
- Gamma random numbers, pdf and cumulative distributions
- Flat (uniform) random numbers, pdf and cumulative distributions
- Lognormal random numbers, pdf and cumulative distributions
- χ^2 random numbers, pdf and cumulative distributions
- F random numbers, pdf and cumulative distributions
- Student-t random numbers, pdf and cumulative distributions
- Poisson random numbers, pdf and cumulative distributions
- Binomial random numbers, pdf and cumulative distributions
- Multinomial random numbers, pdf and cumulative distributions
- Kolgomorov distribution
- Vavilov pdf and its inverse

For students who has taken the subject UECM1703, the following example illustrates how to use the Scipy statistical functions to do simple simulation and plot the probability distribution function.

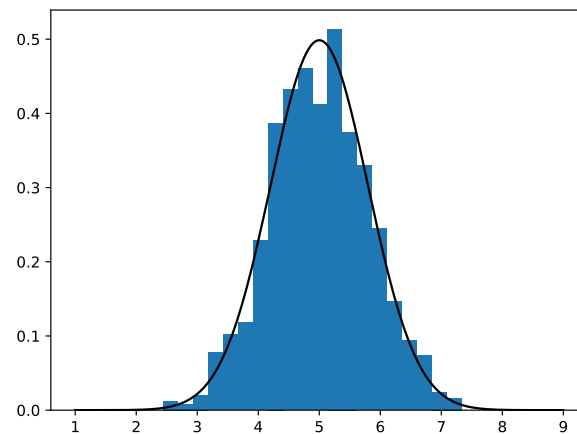
Example 40.22. Simulate 1000 data which follows normal distribution with a mean 5 and a standard deviation of 0.8. Plot the histogram for the 1000 data and try to compare it to the theoretical probability distribution.

Solution: By writing a Python script using the right Scipy and plotting routines, we can plot the normal probability distribution function $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$:

```
1 from scipy import stats
2
3 mean = 5.0; sigm = 0.8
```

```
4 normrv = stats.norm(mean, sigm)
5 simdata = normrv.rvs(size=1000, random_state=2019)
6
7 import numpy as np
8 import matplotlib.pyplot as plt
9 fig = plt.figure()
10 ax = fig.add_subplot(1,1,1)
11 ax.hist(simdata,bins=20,density=True)
12 x = np.linspace(mean-5*sigm, mean+5*sigm, 200)
13 ax.plot(x, normrv.pdf(x),'k')
14 plt.savefig("pdf_demo.pdf") #plt.show()
```

we obtain



Chapter 41

Numerical Linear Algebra

In armadillo C++, matrices and the numerical methods for matrix decompositions, matrix inverses and matrix equation solvers are encoded as follows.

```
arma::cx_mat  = arma::Mat<cx_double>   A(m, n);  // 64-bit
arma::cx_fmat = arma::Mat<cx_float>     A(m, n);  // 32-bit
```

```
arma::lu           lower-upper decomposition
arma::chol         Cholesky decomposition
arma::inv          inverse of general square matrix
arma::inv_sympd    inverse of symmetric positive definite matrix
arma::null         orthonormal basis of null space
arma::orth         orthonormal basis of range space
arma::solve        solve systems of linear equations
arma::pinv         pseudo-inverse / generalised inverse
```

```
arma::eig_sym      eigen decomposition of dense symmetric/hermitian matrix
arma::eig_gen      eigen decomposition of dense general square matrix
arma::eig_pair     eigen decomposition for pair of general dense square matrices
arma::hess         upper Hessenberg decomposition
arma::qr           QR decomposition
arma::qr_econ      economical QR decomposition
arma::qz           generalised Schur decomposition
arma::schur        Schur decomposition
arma::svd          singular value decomposition
arma::svd_econ     economical singular value decomposition
arma::sylvesterm   Sylvester equation solver
```

The numerical methods are based on the textbook Burden, R. L. and Faires, J. D., 2015. Numerical analysis. 10th ed. Boston, MA: Brooks/Cole Cengage Learning.

Need to add complex matrices and sparse matrices!!!

Trefethen, Bau III, Numerical Linear Algebra uses complex spaces or NOT?

Survey on GraphBLAS's applications ...

To replace everything with OpenBLAS, C++, etc.

To add randomised algorithms such as interpolative decomposition and randomised SVD.

In this topic, we will discuss basic numerical methods for handling three major numerical linear algebra problems: (a) the *linear system*:

$$\mathbf{Ax} = \mathbf{b}, \quad \mathbf{A} \text{ is } n \times n \text{ matrix}; \quad (41.1) \quad \text{e:linsys}$$

(b) the *least square problem*:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{Ax} - \mathbf{b}\|_2, \quad \mathbf{A} \text{ is } m \times n \text{ matrix, } m \geq n; \quad (41.2) \quad \text{e:lstsq}$$

and (c) the *eigenvalue problem*:

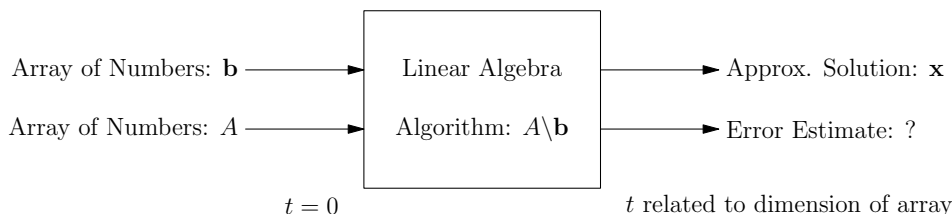
$$\mathbf{Ax} = \lambda \mathbf{x}, \quad \mathbf{A} \text{ is } n \times n \text{ matrix.} \quad (41.3) \quad \text{e:eigen}$$

For the above problems, $\mathbf{x}$ and $\mathbf{b}$ are $n \times 1$ column vectors [Stuart and Voss, 2009].

Equation (41.1) arises naturally in Lagrange interpolation (Topic 4) and linear boundary value problems (Topics 7 and 8). Problem (41.2) is closely related to linear regression problems (UECM2263 Applied

Statistical Model) and Equation (41.3) can either occur in principle component analysis or in stochastic process. More general linear equations appear in control engineering and can be found in the Fortran library SLICOT (<http://slicot.org/>, a part of the NICONET project, see <http://www.icm.tu-bs.de/NICONET/>).

Numerical methods for solving the linear system (41.1) and least square problem (41.2) can be categorised into *direct methods* (Section 41.4), which is closely related to the theory of *matrix factorisation* (Section 41.5), and *iteration methods* (Section 41.6). These numerical methods can be represented by the following block diagram:



Numerical methods for solving the eigenvalue problem (41.3) are of very different nature from problems (41.1) and (41.2). They are iteration methods and the convergence to the eigenvalue and eigenvector may be slow.

41.1 C++ Armadillo Library and OpenBLAS

Armadillo is a high quality linear algebra library (matrix maths) for the C++ language, aiming towards a good balance between speed and ease of use

OpenBLAS is an optimized Basic Linear Algebra Subprograms (BLAS) library based on GotoBLAS2 1.13 BSD version which also includes the Linear Algebra Package LAPACK.

```

1 // Monte Carlo integration with Armadillo vectors
2 call random_seed() ! initialize random number generator used in MC eval
3
4 subroutine MC_integration(n,end_val)
5   integral = 0.0
6   f = 0.0d0
7   f2 = 0.0d0
8   do i=1,n
9     call random_number(x) ! random_number only returns uniformly
10    x = x*end_val ! distributed from [0.0, 1.0]
11    f = f+integrand(x)
12    f2 = f2+ (integrand(x)**2)
13  end do
14  f = f/n
15  f2 = f2/n
16
17  integral = (end_val-0.0)*f
18  integral_err = (end_val-0.0)*sqrt((f2 - f**2)/n)
19  write (*,*) "# MC integration = ",integral,"+/-",integral_err
20 end subroutine MC_integration
21
22 function integrand(x) result (val)
23   real(c_double) :: x, val
24
25   if (x .lt. 0.00001) x = 0.00001
26   val = (x**4)*exp(x)/((exp(x)-1.0)**2)
27 end function integrand
  
```

41.2 Array and I/O

In Topic 1, we learned that real numbers are stored as the “IEEE floating point” data type. Other basic data types, such as Boolean type (called logical in Fortran), integer and characters, though heavily used in business software, are less used in numerical computation.

When we have a “rectangular” array of real numbers such as vectors and matrices, we need the **array** (of floating point values), which is a kind of *extended data type*, to store them. An *n-dimensional* (*n-D* for

short) array is a *homogeneous* collection of “items” (usually floating point numbers or integers) indexed using a tuple of n integers. In this topic, we see that a vector is stored as a 1-D array and a matrix is stored as a 2-D array. The following is an example of how a 2-D array is handled in a few programming languages.

eg:arr2

Example 41.1. Construct a 3×4 matrix with floating point numbers 1 to 12 in C, C++, R and Python 3.

Solution: In C,

```
1 #include <stdio.h>
2 int main() {
3     double arr[3][4] = {1,2,3,4,5,6,7,8,9,10,11,12};
4     for(int i=0; i<3; i++){
5         for(int j=0; j<4; j++){
6             printf("%8.2lf", arr[i][j]); // index start from 0
7         }
8         printf("\n");
9     }
10 }
```

In C++ (using C I/O instead of iostream I/O),

```
1 #include <cstdio>
2 int main() {
3     double arr[3][4] {1,2,3,4,5,6,7,8,9,10,11,12};
4     for(int i=0; i<3; i++){
5         for(int j=0; j<4; j++){
6             printf("%8.2lf", arr[i][j]); // index start from 0
7         }
8         printf("\n");
9     }
10 }
```

In R,

```
1 arr = matrix(c(1:12),3,4)
2 for(i in c(1:3)){
3     for(j in c(1:4)){
4         cat(sprintf("%8.2f", arr[i,j]))
5     }
6     cat("\n")
7 }
```

In Python 3,

```
1 import numpy as np
2 if __name__=="__main__":
3     #np.set_printoptions(precision=2)
4     np.set_printoptions(formatter={'float': '{: 6.2f}'.format})
5     arr = np.array([[1,2,3,4],[5,6,7,8],[9,10,11,12]], dtype='d')
6     print(arr)
```

eg:arr2dat

Example 41.2. Study the programs in Example 41.1, explain which will produce the right matrix below and which will produce the left matrix below:

1.00	2.00	3.00	4.00
5.00	6.00	7.00	8.00
9.00	10.00	11.00	12.00

1.00	4.00	7.00	10.00
2.00	5.00	8.00	11.00
3.00	6.00	9.00	12.00

Solution. C, C++, Python 3 will print the left matrix; R will print the right matrix. □

In the above examples, we learn to represent array in a programming language. But to store it to a file? The simplest way is to store the output of a program as text file. So how to we read the text file?

Example 41.3. Write a C program and a Fortran program to read the left matrix in Example 41.2.

eg:arr2dat

Solution: For C, it is relatively easy to read the left matrix in Example 41.2 using `fscanf`:

```
1 #include <stdio.h>
2 int main() {
3     double arr[3][4];
4     FILE *fp = fopen("eg_arr2.outc", "rt");
5     for(int i=0; i<3; i++)
6         for(int j=0; j<4; j++)
7             fscanf(fp,"%lf", &arr[i][j]);
8     // Use the printf loop in E.g. 2.1.1 to check that the data was read.
9 }
```

For Fortran, it has good array handling function `read` (remember, Fortran was developed for Scientific Computing), but we would be surprised that the data it read in is “wrong”, i.e. it reads the “left” matrix data as the right matrix in Example 41.2.

```
1 program eg_arr2read
2     use iso_c_binding, only : c_double
3     integer :: i
4     real(c_double) :: arr(3,4)
5     open(unit=10, file="eg_arr2.outc")
6     read(10,*) arr
7     do i=1,3 ! index start from 1
8         write(*,"(4f8.2)") arr(i,:)
9     end do
10 end program eg_arr2read
```

Real world scientific data are so large (from a few Gig to a few Tera) that storing as text file above is not viable. As observed above, the data may be wrongly interpreted when reading it. Special format such as HDF5 and MATLAB `.mat` were developed to overcome this issue. Note that MATLAB MAT v7.3 formatted files are just HDF5 files with a different extension and some extra meta-data (see <https://pythonhosted.org/hdf5storage/information.html>). NetCDF is another important storage format based on HDF5 [Collette, 2014]. In the following example, we will learn how to store the left matrix in Example 41.2 in a HDF5 file `hdf5eg.h5` and then reading `hdf5eg.h5`.

eg:hdf5

Example 41.4. Explain how to use HDF5 in Fortran to store the left matrix in Example 41.2. Then use C to read the data file `hdf5eg.h5`.

Solution: According to https://support.hdfgroup.org/HDF5/doc/UG/HDF5.Users.Guide-Responsive%20HTML5/index.html#t=HDF5_Users_Guide%2FDataspace%2FHDF5_Dataspace_and_Partial_I_O.htm%23TOC_7_3_2_Dataspacebc-5&rhtocid=7.1.0.2, the programming model for creating and working with HDF5 dataspace can be summarised as follows:

1. Create a dataspace `H5Screate` (append `_f` for Fortran). `H5Screate_simple` could be
2. Use the dataspace to create a dataset in the file or to describe a data array in memory
3. Modify the dataspace to define dataset elements that will participate in I/O operations
4. Use the modified dataspace while reading/writing dataset raw data or to create a region reference
5. Close the dataspace when no longer needed

HDF5 uses C storage conventions, assuming that the last listed dimension is the fastest-changing dimension and the first-listed dimension is the slowest changing. The HDF5 file format storage layout specification adheres to the C convention and the HDF5 library adheres to the same convention when storing dataspace dimensions in the file. This affects how C programs and tools interpret data written from Fortran programs and vice versa.

When a Fortran application describes a dataspace to store an array as `A(3,4)`, it specifies the value of the first dimension to be 3 and the second to be 4. Since Fortran stores data by columns, the first-listed dimension with the value 3 is the fastest-changing dimension and the last-listed dimension with the value 4 is the slowest-changing. In order to adhere to the HDF5 storage convention, the HDF5 Fortran wrapper transposes dimensions, so the first dimension becomes the last. The dataspace dimensions stored in the file will be 4,3 instead of 3,4 in order to correctly describe the Fortran data that is stored in 4 columns, each containing 3 elements.

Here is the Fortran program to generate the left matrix in Example [4I.2](#) ^{eg:arr2dat}

```
1  ! https://support.hdfgroup.org/HDF5/examples/api-fortran.html
2  ! gfortran -W -Wall -std=f2003 -I/usr/include/hdf5/serial hdf5eg.f03
3  !   -L/usr/lib/x86_64-linux-gnu/hdf5/serial -lhdf5_fortran -lhdf5
4
5  program hdf5eg
6      use hdf5
7      use iso_c_binding, only : c_double
8      implicit none
9      character(len=*), parameter :: filename = "hdf5eg.h5"
10     character(len=*), parameter :: dataset1 = "arr2"
11     integer      , parameter :: dimr      = 3
12     integer      , parameter :: dimc      = 4
13     integer      , parameter :: rank      = 2
14
15     integer :: spacestatus, i, j, hdferr
16     integer(hid_t) :: file_id, dspace_id, dset1 ! handles
17     ! HDF5 defines a data type called hsize_t which is used for sizes
18     ! of datasets and is, by default, defined as unsigned long long.
19     integer(HSIZE_T) :: storagesize, dims(2)
20     real(c_double) :: wdata(dimr,dimc) ! write buffer
21
22     dims(1:2) = (/dimr, dimc/) ! size write buffer
23     ! A 3x4 matrix of double floating point numbers
24     do i = 1, dimr
25         do j = 1, dimc
26             wdata(i,j) = real((i-1)*dimc+j)
27         end do
28     end do
29
30     ! initialize fortran interface.
31     call H5open_f(hdferr)
32     ! Create a new file using the default properties.
33     ! H5F_ACC_TRUNC_F is the file access flag to create a new file
34     ! and overwrite an existing file with the same name
35     call H5Fcreate_f(filename, H5F_ACC_TRUNC_F, file_id, hdferr)
36
37     ! Create dataspace.
38     call H5Screate_simple_f(rank, dims, dspace_id, hdferr)
39
40     ! create the dataset using the dataset creation property list.
41     call H5Dcreate_f(file_id, dataset1, H5T_IEEE_F64LE, &
42         dspace_id, dset1, hdferr)
43
44     ! retrieve and print space status and storage size for dset1.
45     storagesize = 0
46     call H5Dget_space_status_f(dset1, spacestatus, hdferr)
47     call H5Dget_storage_size_f(dset1, storagesize, hdferr)
48     if(spacestatus.eq.H5D_SPACE_STS_ALLOCATED_F)then
49         write(*,101) dataset1
50     else
51         write(*,102) dataset1
52     end if
53     write(*,100) dataset1, storagesize
54
55     write(*,'(/"writing data...",/)' )
56
57     ! write the data to the datasets.
58     call H5Dwrite_f(dset1, H5T_IEEE_F64LE, wdata, dims, hdferr)
59
60     ! retrieve and print space status and storage size for dset1.
61     call H5Dget_space_status_f(dset1, spacestatus, hdferr)
62     call H5Dget_storage_size_f(dset1, storagesize, hdferr)
63     if(spacestatus.eq.H5D_SPACE_STS_ALLOCATED_F)then
64         write(*,101) dataset1
```



```

65     else
66         write(*,102) dataset1
67     end if
68     write(*,100) dataset1, storagesize
69
70     ! close and release resources.
71     call H5Dclose_f(dset1, hdferr)
72     call H5Sclose_f(dspace_id, hdferr)
73     call H5Fclose_f(file_id, hdferr)
74
75     100 format("storage size for ",a," is: ",i3," bytes.")
76     101 format("space for ",a," has been allocated.")
77     102 format("space for ",a," has not been allocated.")
78 end program hdf5eg

```

When a Fortran application reads the data back, the HDF5 Fortran wrapper transposes the dimensions once more, returning the first dimension to be 3 and the second to be 4, describing correctly the sizes of the array that should be used to read data in the Fortran array $A(3,4)$.

When a C application reads data back, the dimensions will come out as 4 and 3, correctly describing the size of the array to read data into, since the data was written as 4 records of 3 elements each. Therefore C tools such as `h5dump` and `h5ls` always display transposed dimensions and values for the data written by a Fortran application.

```

1  #include "hdf5.h"
2
3  int main() {
4      char filename[] = "hdf5eg.h5";
5      hid_t file_id, dataset_id;
6      herr_t hdferr;
7      const int nr=3, nc=4;
8      double wdatatr[nc][nr], wdata[nr][nc];
9
10     // open & read the datasets by reading the dataset
11     file_id = H5Fopen(filename, H5F_ACC_RDONLY, H5P_DEFAULT);
12     dataset_id = H5Dopen2(file_id, "/arr2", H5P_DEFAULT);
13     // For Intel PC, H5T_NATIVE_DOUBLE=H5T_IEEE_F64LE
14     hdferr = H5Dread(dataset_id, H5T_NATIVE_DOUBLE, H5S_ALL, H5S_ALL,
15                     H5P_DEFAULT, wdatatr);
16     // Fortran Array in HDF5 is transposed.
17     for(int i=0; i<nr; i++)
18         for(int j=0; j<nc; j++)
19             wdata[i][j] = wdatatr[j][i]; // A=(A^T)^T
20     for(int i=0; i<nr; i++){
21         for(int j=0; j<nc; j++)
22             printf("%8.2lf", wdata[i][j]);
23         printf("\n");
24     }
25
26     // close and release resources.
27     hdferr = H5Dclose(dataset_id);
28     hdferr = H5Fclose(file_id);
29 }

```

41.3 Array Arithmetic and Norms

ec:array-arith

To work with arrays, we need to generalise the addition, subtraction, multiplication and division for floating point numbers to “point-wise” arithmetic operations for arrays. The programming languages C, C++, D and Python do not support point-wise arithmetic operations by default. They provide the operations through special numerical libraries such as Eigen3, Numpy and Scipy [Blanco-Silva, 2013].

eg:arr3

Example 41.5. Study the following C++ program and determine which line(s) in the program use(s) point-wise operations?

```

1  #include <armadillo>
2  #include <iostream>

```

```

3  using namespace arma;
4
5  int main() {
6      std::cout << "Armadillo version: " << arma_version::as_string() << "\n";
7      const int dim1{4}, dim2{5}, dim3{6};
8      cube A(dim1, dim2, dim3, fill::zeros);
9      cube B(dim1, dim2, dim3);
10
11      for(int i=0; i<dim1; i++)
12          for(int j=0; j<dim2; j++)
13              for(int k=0; k<dim3; k++) {
14                  A(i,j,k) = i*dim2*dim3 + j*dim3 + k+1;
15                  B(i,j,k) = A(i,j,k) + 2;
16              }
17
18      // Check if C with be a reference to A or a copy of A
19      cube C = A;
20      C(0,0,0) = 8.0;
21
22      A.print("\n ***** A *****");
23      B.print("\n ***** B *****");
24      // Elementwise multiplication is % rather than *
25      C = 2*A + (A%B - B/A);
26      C.print("\n ***** C *****");
27 }

```

Solution:

Line ??? copies each elements from array A to array C, this is “point-wise” operation (but no arithmetic is performed).

In Line ???, “point-wise” arithmetic operations on the right-hand-side and the results are copied “point-wise” into array C.

Finite dimensional vectors and matrices can be stored as 1-D and 2-D arrays of floating point numbers respectively. However, vectors and matrices has an additional arithmetic associated with them called *matrix multiplication*. High level languages such as Python (with NumPy module, read <http://www.labri.fr/perso/nrougier/from-python-to-numpy/>), R and Julia have matrix multiplication but C, C++ and D do not have matrix multiplication. Fortran provides the `matmul` function to perform matrix multiplication as illustrated below.

Example 41.6. Write a Fortran 2003 program to calculate $b^T A^T$ where

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} -1 \\ 2 \\ -3 \\ 4 \end{bmatrix}.$$

Solution: The Fortran function `transpose` allow us to transpose a two-dimensional array, i.e. a matrix. A possible implementation is given below:

```

1  program eg_matrix
2      use iso_fortran_env, only : real64
3      implicit none
4      real(real64) :: A(3,4), b(4), v(3)
5      integer i, j
6      b = (/ -1d0, 2d0, -3d0, 4d0 /)
7      do i=1,3
8          do j=1,4
9              A(i,j) = (i-1)*4 + j
10         end do
11     end do
12     print *, "***** A *****"
13     do i=1,3
14         write(*,'(10f8.4)'), (A(i,j), j=1,4)
15     end do
16     print *, "***** b *****"

```

```

17     write(*,'(10f8.4)'), (b(j), j=1,4)
18     v = matmul(b,transpose(A))
19     print *, "***** v = A*b *****"
20     write(*,'(10f8.4)'), v
21 end program eg_matrix

```

41.3.1 Vector Norms

In Topic 1, we learned the concept of absolute error for floating point numbers. To characterise the absolute error for an array, we need to measure the “length” of the array using the notion of vector norm.

Definition 41.1. A (vector) norm $\|\cdot\|$ is a real-valued function on the vector space V that satisfies the following axioms:

1. $\|\mathbf{x}\| > 0$ when $\mathbf{x} \neq \mathbf{0}$ and $\|\mathbf{x}\| = 0$ iff $\mathbf{x} = \mathbf{0}$;
2. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$;
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$

where $\mathbf{x}, \mathbf{y}$ are any vectors in the vector space V and α is any real (or complex) number.

In linear algebra, the commonly used vector norms are the p -norm (or l_p -norm)

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}, \quad p \geq 1.$$

When $p = 1, p = 2, p = \infty$, we have the most commonly used vector norms below:

- 1-norm: $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$;
- 2-norm: $\|\mathbf{x}\|_2 = (\sum_{i=1}^n |x_i|^2)^{1/2}$;
- ∞ -norm: $\|\mathbf{x}\|_\infty = \lim_{p \rightarrow \infty} \|\mathbf{x}\|_p = \max_{0 \leq i \leq n} |x_i|$.

Let's look at some examples on the computation of vector norm.

Example 41.7. Given a vector $\mathbf{x} = (1, 2, -3, -4, 1) \in \mathbb{R}^5$. Find $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$ and $\|\mathbf{x}\|_\infty$.

Solution: $\|\mathbf{x}\|_1 = 1 + 2 + 3 + 4 + 1 = 11$
 $\|\mathbf{x}\|_2 = \sqrt{1 + 2^2 + 3^2 + 4^2 + 1^2} \approx 5.56776436$
 $\|\mathbf{x}\|_\infty = 4$.

Example 41.8 (Tutorial 2, Q1). Find $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$ and $\|\mathbf{x}\|_\infty$ for the following vectors.

1. $\mathbf{x} = (3, -4, 0, 3/2)^T$;

Solution. $\|\mathbf{x}\|_1 = 3 + 4 + 3/2 = 8.5$, $\|\mathbf{x}\|_2 = \sqrt{27.25}$, $\|\mathbf{x}\|_\infty = 4$. □

2. $\mathbf{x} = (\sin k, \cos k, 2^k)^T$ for a fixed positive integer k ;

Solution. $\|\mathbf{x}\|_1 = |\sin k| + |\cos k| + 2^k$, $\|\mathbf{x}\|_2 = \sqrt{1 + 2^{2k}}$, $\|\mathbf{x}\|_\infty = 2^k$. □

Using the concept of norm, it is possible for us to measure the distance between two vectors.

Definition 41.2. Given the (vector) norm $\|\cdot\|$ of a given vector space. The *distance* of vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\text{dist}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|.$$

The fitting of data problem can be formulated based on the 2-norm distance.

Example 41.9. Use $y = mx + c$ to fit the points $(x_i, y_i) = (1, 1), (2, 3)$ and $(3, 2)$ such that $(mx_1 + c - y_1)^2 + (mx_2 + c - y_2)^2 + (mx_3 + c - y_3)^2$ is minimum. [Remark: In SPM, approximations for m and c were found using some formulae and then being plotted on a graph paper.]

1. Form a linear algebra problem (41.1) for this line-fitting problem.

$$\text{Solution.} \quad \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix} = \begin{bmatrix} m+c \\ 2m+c \\ 3m+c \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} m \\ c \end{bmatrix} \quad \square$$

2. Show that the linear algebra problem has no solution.

Solution. Suppose a solution (m^*, c^*) exists, then from the first two rows, we have $2 = m^*$ and $c^* = -1$. The third row gives us $2 = 3m^* + c^* = 6 - 1 = 5$. A contradiction. $\square$

3. Form a least square problem (41.2) for this line-fitting problem.

Solution:

$$\min_{(m,c) \in \mathbb{R}^2} \left\| \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} m \\ c \end{bmatrix} - \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix} \right\|_2 = \min_{(m,c) \in \mathbb{R}^2} \sqrt{(m+c-1)^2 + \cdots + (3m+c-2)^2}$$

4. Find the least square solution by applying the technique from UECM1303 Fundamentals of Linear Algebra.

$$\text{Solution.} \quad \begin{bmatrix} \bar{m} \\ \bar{c} \end{bmatrix} = (A^T A)^{-1} A^T \mathbf{b} = \begin{bmatrix} 14 & 6 \\ 6 & 3 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}. \quad \square$$

41.3.2 Matrix Norms

To study the numerical stability of matrix operations and convergence of iteration processes (Section 41.6.1) related to a matrix, we need ways to measure the “absolute value” of a matrix. Consider a simple problem $y = ax$. When x is mixed with an error (possibility numerical error or observation error) to become $x + \Delta x$, then the absolute error in y becomes

$$y + \Delta y = a(x + \Delta x) \Rightarrow |\Delta y| = |a||\Delta x|.$$

The error in x is magnified by $|a|$. What if x is a vector, then a becomes a matrix and the absolute value generalises to the notion of matrix norm.

Let $M_{m,n}$ denote the set of (real or complex) $m \times n$ matrices $M_{m,n}$. For a set of $n \times n$ matrix, we will normally denote it by M_n instead of $M_{n,n}$.

Definition 41.3. A matrix norm $\|\cdot\|$ is a function on $M_{m,n}$ that satisfies the following axioms:

1. $\|A\| > 0$ when $A \neq 0$ and $\|A\| = 0$ iff $A = 0$;
2. $\|\alpha A\| = |\alpha| \|A\|$;
3. $\|A + B\| \leq \|A\| + \|B\|$;
4. $\|AB\| \leq \|A\| \|B\|$

for all matrices $A \in M_{m,n}$, $B \in M_{n,k}$ and all real or complex number α .

The matrix norm is a special kind of vector norm with an extra condition 4. mentioned above. Sometimes we will use $\|\cdot\|_v$ to denote a vector norm and use $\|\cdot\|_M$ to denote a matrix norm just make a distinction.

To work with matrix-vector products, we introduce the following definition.

Definition 41.4. Let $\|\cdot\|_v$ be a vector norm of $\mathbb{R}^n$ and $\|\cdot\|_M$ be a matrix norm on M_n , then we say that the vector norm and the matrix norm are *consistent* (or *compatible*) if

$$\|A\mathbf{x}\|_v \leq \|A\|_M \|\mathbf{x}\|_v$$

for all $\mathbf{x} \in \mathbb{R}^n$ and $A \in M_n$. When a_{ij} are complex numbers, $|a_{ij}|$ is the modulus of a_{ij} .

Let A^H denote the complex conjugate transpose of A . If A is real, A^H is the same as the transpose A^T .

Definition 41.5. The *Frobenius norm* $\|\cdot\|_F$ on the set of square matrices M_n define below

$$\|A\|_F = \sqrt{\text{tr}(A^H A)} = \sqrt{\sum_{i,j=1}^n |a_{ij}|^2}, \quad A \in M_n.$$

em:cauchy-ineq

Lemma 41.1 (Cauchy-Schwarz Inequality). Let $a_i, b_i, i = 1, \dots, n$ be complex numbers.

$$\left| \sum_{i=1}^n a_i \bar{b}_i \right|^2 \leq \left(\sum_{i=1}^n |a_i|^2 \right) \left(\sum_{i=1}^n |b_i|^2 \right). \quad (41.4)$$

e:cauchy-ineq

Using the vector notation and the 2-norm notation and let $\mathbf{a} = (a_1, \dots, a_n)^T$ and $\mathbf{b} = (b_1, \dots, b_n)$, the inequality (41.4) can be written as

$$|\mathbf{a} \cdot \mathbf{b}|^2 = |\mathbf{b}^H \mathbf{a}|^2 \leq \|\mathbf{a}\|_2^2 \|\mathbf{b}\|_2^2.$$

eg:frobl

Example 41.10. Show that the 2-norm $\|\cdot\|_2$ and the Frobenius norm $\|\cdot\|_F$ are compatible for real vector $\mathbf{x}$ and real matrix A .

Proof: Let $\mathbf{x} \in \mathbb{R}^n$ be a 1×1 matrix containing one column vector and the matrix $A \in M_n$ be an

$n \times 1$ matrix containing n rows $\begin{bmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \vdots \\ \mathbf{a}_n^T \end{bmatrix}$. Then $A\mathbf{x} = \begin{bmatrix} \mathbf{a}_1^T \mathbf{x} \\ \mathbf{a}_2^T \mathbf{x} \\ \vdots \\ \mathbf{a}_n^T \mathbf{x} \end{bmatrix}$. By Cauchy-Schwarz inequality (41.4) (see Lemma 41.1), we have

$$\|A\mathbf{x}\|_2^2 = \sum_{i=1}^n |\mathbf{a}_i^T \mathbf{x}|^2 \leq \sum_{i=1}^n \|\mathbf{a}_i\|_2^2 \|\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2 \sum_{i=1}^n \|\mathbf{a}_i\|_2^2 = \|\mathbf{x}\|_2^2 \|A\|_F^2.$$

Therefore $\|A\mathbf{x}\|_2 \leq \|A\|_F \|\mathbf{x}\|_2$ and by definition, the 2-norm and the Frobenius norm are compatible.

In general it is very complicated to show that a function over a set of matrices is a matrix norm as demonstrated below.

eg:frob2

Example 41.11. Show that the Frobenius norm $\|\cdot\|_F$ on M_n is a matrix norm.

Proof: To show that the Frobenius norm is a matrix norm, we will need to show that all the four axiom Definition 41.3 are satisfied. Let $A = (a_{ij}), B = (b_{ij}) \in M_n$ and $\alpha \in \mathbb{R}$ (or $\mathbb{C}$). Then

1. If $\|A\|_F = \sqrt{\sum_{i,j=1}^n |a_{ij}|^2} = 0$, then $|a_{ij}| = 0$, i.e. $a_{ij} = 0$ for all i, j from 1 to n . If A is a zero matrix, then $\|A\|_F = 0$.

If A is not a zero matrix, there is a pair (i, j) such that $a_{ij} \neq 0$, then $|a_{ij}| \neq 0$ and hence, $\sum_{i,j=1}^n |a_{ij}|^2 > 0$. After taking square root, we have $\|A\|_F > 0$.

2. $\|\alpha A\|_F = \sqrt{\sum_{i,j=1}^n |\alpha a_{ij}|^2} = \sqrt{\sum_{i,j=1}^n |\alpha|^2 |a_{ij}|^2} = \sqrt{|\alpha|^2 \sum_{i,j=1}^n |a_{ij}|^2} = |\alpha| \sqrt{\sum_{i,j=1}^n |a_{ij}|^2} = |\alpha| \|A\|_F$

3. With a_i and b_i replaced by $|a_{ij}|$ and $|b_{ij}|$ in the Cauchy inequality (41.4), we have

$$\begin{aligned} \|A + B\|_F &= \sqrt{\sum_{i,j=1}^n |a_{ij} + b_{ij}|^2} \leq \sqrt{\sum_{i,j=1}^n (|a_{ij}|^2 + 2|a_{ij}||b_{ij}| + |b_{ij}|^2)} \\ &= \sqrt{\sum_{i,j=1}^n |a_{ij}|^2 + 2 \sum_{i,j=1}^n |a_{ij}||b_{ij}| + \sum_{i,j=1}^n |b_{ij}|^2} \\ &\leq \sqrt{\sum_{i,j=1}^n |a_{ij}|^2 + 2 \sqrt{\sum_{i,j=1}^n |a_{ij}|^2} \sqrt{\sum_{i,j=1}^n |b_{ij}|^2} + \sum_{i,j=1}^n |b_{ij}|^2} \\ &= \sqrt{\|A\|_F^2 + 2\|A\|_F \|B\|_F + \|B\|_F^2} = \|A\|_F + \|B\|_F \end{aligned}$$

4. Denoting A as a 1×1 matrix contain an $n \times n$ matrix and B as an $1 \times n$ matrix of $n \times 1$ column vectors $[\mathbf{b}_1 \cdots \mathbf{b}_n]$, and applying Example 41.10, we have

$$\|AB\|_F^2 = \|[A\mathbf{b}_1 \cdots A\mathbf{b}_n]\|_F^2 = \sum_{i=1}^n \|A\mathbf{b}_i\|_2^2 \leq \sum_{i=1}^n \|A\|_F^2 \|\mathbf{b}_i\|_2^2 = \|A\|_F^2 \left(\sum_{i=1}^n \|\mathbf{b}_i\|_2^2 \right).$$

Taking square root to both sides, we have $\|AB\|_F \leq \|A\|_F \sqrt{\sum_{i=1}^n \|\mathbf{b}_i\|_2^2} = \|A\|_F \|B\|_F$.

Are there any other matrix norms? Mathematicians found a theorem below which allows us to obtain more matrix norms from vector norms.

Theorem 41.12. Let $\|\cdot\|^{(m)}$ and $\|\cdot\|^{(n)}$ be any vector norms on the vector spaces $\mathbb{C}^m$ and $\mathbb{C}^n$ respectively. The induced norm $\|\cdot\|^{(m,n)}$ on $M_{m,n}$ defined by

$$\|A\|^{(m,n)} := \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|A\mathbf{x}\|^{(m)}}{\|\mathbf{x}\|^{(n)}} = \max_{\|\mathbf{x}\|^{(n)}=1} \|A\mathbf{x}\|^{(m)}, \quad \text{for all } A \in M_{m,n}$$

is a matrix norm. When $\|\cdot\|^{(m)}$ and $\|\cdot\|^{(n)}$ are p -norms, the induced norm $\|\cdot\|^{(m,n)}$ is called a matrix p -norm.

When $m = n$ and $p = 1, 2$ or ∞ , we have the following theorem.

Theorem 41.13. Let A be a complex $n \times n$ matrix, then

$$\|A\|_1 = \max_j \sum_{i=1}^n |a_{ij}|, \quad \|A\|_2 = \sigma_1, \quad \|A\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|$$

where σ_1 is the maximum singular value of A , i.e. the square root of the maximum eigenvalue of $A^H A$.

41.3.3 BLAS

Despite being a “scientific” programming language, Fortran does not provide some functions to handle vectors and matrices such as functions to calculate vector norm. The *Basic Linear Algebra Subprograms* (BLAS) is a specification for software library containing a set of low-level subroutines that perform common linear algebra operations such as *copying*, *vector scaling*, *vector dot products*, *linear combinations*, and *matrix multiplication* (see Wikipedia). BLAS was first published as a Fortran library in 1979 and has C, C++ implementations (e.g. CBLAS) at a later time. It served as a foundation for the standard linear algebra package LAPACK (Section 41.4.1) for handling dense matrix linear algebra problems.

BLAS functionality is divided into three levels.

Level 1 contains **vector operations** on strided arrays: generalised vector addition, Givens transform, dot products and vector norms.

$\mathbf{y} \leftarrow \mathbf{x}$	dcopy(n,dx,incx,dy,incy)
$\mathbf{x} \leftrightarrow \mathbf{y}$	dswap(n,dx,incx,dy,incy)
$\mathbf{x} \leftarrow \alpha \mathbf{x}$	dscal(n,da,dx,incx)
$\mathbf{y} \leftarrow \alpha \mathbf{x} + \mathbf{y}$	daxpy(n,da,dx,incx,dy,incy)
$\mathbf{x} \cdot \mathbf{y}$	ddot(n,dx,incx,dy,incy)
$\ \mathbf{x}\ _2$	dnorm2(n,dx,incx)
$\ \mathbf{x}\ _1$	dasum(n,dx,incx)

Example 41.14 (Level 1 BLAS Example). Given a vector $\mathbf{x} = (1, 2, -3, -4, 1) \in \mathbb{R}^5$. Write a Fortran program to calculate $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$ and $\|\mathbf{x}\|_\infty$. Compare your results to Examples 41.7.

Solution: Fortran 2003 has an abstract interface for us to interface with old Fortran 77 (as well as C) library.

```
1 #include <cblas.h>
2 #include <iostream>
```

```

3  #include <iomanip>
4  #include <cmath>
5
6  int main() {
7      const int n = 5;
8      double x[5] = {1.0, 2.0, -3.0, -4.0, 1.0};
9
10     std::cout << std::fixed << std::setprecision(16);
11     double res = cblas_dasum(n, x, 1);
12     std::cout << "The 1-norm = " << res << "\n";
13     res = cblas_dnrm2(n, x, 1);
14     std::cout << "The 2-norm = " << res << "\n";
15     res = abs(x[cblas_idamax(n, x, 1)]);
16     std::cout << "The oo-norm = " << res << "\n";
17 }

```

The output: 1-norm = 11.0, 2-norm = 5.5677643628300215, ∞ -norm = 4.0 match with our result in Examples [41.7](#).

Level 2 contains **matrix-vector operations** including a generalised matrix-vector multiplication, rank-1 and rank-2 update as well as a solver for $T\mathbf{x} = \mathbf{y}$ for $\mathbf{x}$ with T being triangular and non-singular.

$\mathbf{y} \leftarrow \alpha \tilde{A}\mathbf{x} + \beta \mathbf{y}$	<code>dgemv(trans, m, n, alpha, A, lda, x, incx, beta, y, incy)</code>
$\mathbf{x} \rightarrow \tilde{T}^{-1}\mathbf{x}$	<code>dtrsv(uplo, trans, diag, n, T, ldT, x, incx)</code>

where trans takes the value ‘N’, ‘T’, ‘C’, which means $\tilde{A}$ is A , A^T and A^H respectively. The Level 2 subroutines are especially intended to improve performance of programs using BLAS on vector processors, where Level 1 BLAS are suboptimal “because they hide the matrix-vector nature of the operations from the compiler.”

Example 41.15 (Level 2 BLAS Example). Use BLAS Level 2 subroutine to compute

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - \frac{1}{2} \begin{bmatrix} 4 \\ 3 \\ 2 \\ 1 \end{bmatrix} \quad (*)$$

Solution:

```

1  #include <cblas.h>
2  #include <iostream>
3  #include <iomanip>
4
5  int main() {
6      const int d1=3, d2=4, lda=d1;
7      const double alph=1.0, beta=-0.5;
8
9      double x[d1], y[d2], A[d1*d2];
10
11     for(int i=0; i<d1*d2; i++) A[i] = (double)(i+1);
12     for(int i=0; i<d1; i++) x[i] = 1.0;
13     for(int i=0; i<d2; i++) y[i] = double(d2-i);
14     // Fortran style array: column major, transpose is needed
15     //cblas_dgemv(CblasColMajor, CblasTrans, d1, d2, alph, A, lda, x, 1, beta, y, 1);
16     cblas_dgemv(CblasRowMajor, CblasNoTrans, d2, d1, alph, A, lda, x, 1, beta, y, 1);
17     std::cout << std::fixed << std::setprecision(2);
18     for(int i=0; i<d2; i++) {
19         std::cout << std::setw(8) << y[i];
20     }
21     std::cout << "\n";
22 }

```

Level 3, formally published in 1990, contains a **general matrix-matrix multiplication** (GEMM), rank- k and rank- $2k$ updates of a symmetric matrix and routines for solving triangular systems of equations.

$C \leftarrow \alpha \tilde{A} \tilde{B} + \beta C$	<code>dgemm(transa, transb, m, n, k, α, A, LDA, B, LDB, β, C, LDC)</code>
$B \leftarrow \alpha T^{-1} B$	<code>dtrsm(side, uplo, transa, diag, m, n, α, A, LDA, B, LDB)</code>

where T is triangular; A and B can optionally be transposed inside the routine.

GEMM is the prime target for BLAS specification implementer to optimise based on the underlying hardware (such as multilevel cache and memory structure, multicore CPU and GPU). ATLAS, Goto-BLAS, OpenBLAS (and D language Mir library?) are optimised implementations of BLAS specification.

Example 41.16 (Level 3 BLAS Example). Calculate AB using the GEMM subroutine where

$$A = \begin{bmatrix} 0 & 1 & 2 \\ 3 & 4 & 5 \\ 6 & 7 & 8 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}.$$

Solution:

```

1  #include <cblas.h>
2  #include <iostream>
3  #include <iomanip>
4
5  //based on Brock Palen (brockp@umich.edu)'s example from
6  //https://github.com/brockpalen/caen-hpc-docs/blob/master/blas/blas-code/dgemm.F90
7
8  int main() {
9      const int M=3, N=M, K=M, LDA=M, LDB=M, LDC=M;
10     // On entry, LDA specifies the first dimension of A as declared
11     // in the calling (sub) program. When TRANSA = 'N' or 'n' then
12     // LDA must be at least max( 1, m ), otherwise LDA must be at
13     // least max( 1, k ).
14     // On entry, LDB specifies the first dimension of B as declared
15     // in the calling (sub) program. When TRANSB = 'N' or 'n' then
16     // LDB must be at least max( 1, k ), otherwise LDB must be at
17     // least max( 1, n ).
18     // On entry, LDC specifies the first dimension of C as declared
19     // in the calling (sub) program. LDC must be at least
20     // max( 1, m ).
21     double alpha = 1.0, beta = 0.0;
22     double *A, *B, *C;
23     A = (double*)malloc(sizeof(double)*LDA*K);
24     B = (double*)malloc(sizeof(double)*LDB*N);
25     C = (double*)malloc(sizeof(double)*LDC*N);
26     for(int i=0; i<K*LDA; i++) A[i] = double(i);
27     for(int i=0; i<N*LDB; i++) B[i] = double(i+1);
28     cblas_dgemm(CblasRowMajor, CblasNoTrans, CblasNoTrans, M, N, K, alpha, A, LDA, B, LDB, beta, C, LDC);
29     std::cout << std::fixed << std::setprecision(16);
30     for(int i=0; i<LDC*N; i++) {
31         std::cout << std::setw(22) << C[i] << " ";
32         if((i+1)%N==0) std::cout << "\n";
33     }
34 }
```

BLAS is the lowest level linear algebra software which interact with current computer architecture illustrated below.

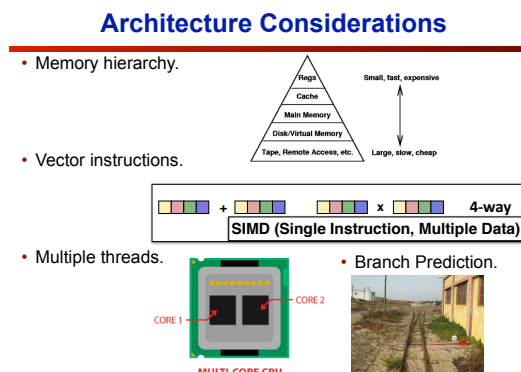


Figure 1: Carnegie Mellon 16-623 Designing Computer Vision Apps Lecture 3 Slides by Simon Lucey

41.4 Direct Solution of Linear Systems

Virtually all direct methods for solving linear systems are developed based on the idea of reducing the matrix associated with the linear system to triangular form. On serial computers triangular matrix is easy to solve by the standard back and forward substitution algorithms.

The most widely use direct method for solving linear systems is the Gaussian elimination with partial pivoting. It is a good algorithm for small to medium-sized problems. The size depends on the sparsity structure of the matrices. For dense matrices with no special property, Gaussian elimination is quite good. But when a matrix has a special structure (e.g. near upper triangular, near lower triangular, tridiagonal, upper bidiagonal, lower bidiagonal, upper Hessenberg, lower Hessenberg, symmetry), Gaussian elimination may not be sufficient and more advanced methods are developed. QR factorisation is another direct method which will be discussed in Section 41.5.4.

41.4.1 LAPACK

LAPACK (an acronym for Linear Algebra PACKage) is a library of Fortran 77 routines intended for “solving systems of simultaneous linear equations (41.1), least-squares solutions of linear systems of equations (41.2), eigenvalue problems (41.3), and singular value problems” [Anderson et al., 1999].

LAPACK was designed based on BLAS, to be efficient on a wide range of modern high-performance computers involving dense and band (single and double precision real and complex) matrices but not general sparse matrices. It is an extension to LINPACK and EISPACK and is designed to preserve the accuracy of floating point operations and to handle all the operations related to solving linear algebra problems such as matrix factorisation (e.g. LU, Cholesky, QR, SVD, Schur, generalised Schur) [Golub and Loan, 1996]. LAPACK is the standard library for various numerical computing environment such as Python/NumPy, R (using older library LINPACK), Julia, GNU Octave, MATLAB, etc. It is still being developed (e.g. the SLATE project, <https://bitbucket.org/icl/slate>) with focus on high performance computing.

The subroutines in LAPACK are classified into three groups:

- *driver routines*, each of which solves a complete problem, for example solving a system of linear equations (xGESV, to be studied later), or computing the eigenvalues of a real symmetric matrix. Users are recommended to use a driver routine if there is one that meets their requirements.
- *computational routines*, each of which performs a distinct computational task, for example an LU factorisation, or the reduction of a real symmetric matrix to tridiagonal form. Each driver routine calls a sequence of computational routines. Users (especially software developers) may need to call computational routines directly to perform tasks, or sequences of tasks, that cannot conveniently be performed by the driver routines.
- *auxiliary routines*, which in turn can be classified as follows:
 - routines that perform subtasks of block algorithms — in particular, routines that implement unblocked versions of the algorithms;
 - routines that perform some commonly required low-level computations, for example scaling a matrix, computing a matrix-norm, or generating an elementary Householder matrix; some of these may be of interest to numerical analysts or software developers and could be considered for future additions to the BLAS;
 - a few extensions to the BLAS, such as routines for applying complex plane rotations, or matrix-vector operations involving complex symmetric matrices (the BLAS themselves are not strictly speaking part of LAPACK).

41.4.2 Naming Scheme

The name of each LAPACK routine is a coded specification of its function (*within the very tight limits of standard Fortran 77 6-character names*).

All driver and computational routines have names of the form XYYZZZ, where for some driver routines the 6th character is blank.

The first letter, X, indicates the data type as follows:

S	REAL (32-bit)
D	DOUBLE PRECISION (64-bit)
C	COMPLEX (32+32-bit)
Z	COMPLEX*16 or DOUBLE COMPLEX (64+64-bit)

When we wish to refer to an LAPACK routine generically, regardless of data type, we replace the first letter by “x”. Thus xGESV refers to any or all of the routines SGESV, CGESV, DGESV and ZGESV.

The next two letters, YY, indicate the type of matrix (or of the most significant matrix). Most of these two-letter codes apply to both real and complex matrices; a few apply specifically to one or the other, as indicated in the following Table 2.1 from Anderson et al. [1999].

Table 2.1: Matrix types in the LAPACK naming scheme

BD	bidiagonal
DI	diagonal
GB	general band
GE	general (i.e., unsymmetric, in some cases rectangular)
GG	general matrices, generalized problem (i.e., a pair of general matrices)
GT	general tridiagonal
HB	(complex) Hermitian band
HE	(complex) Hermitian
HG	upper Hessenberg matrix, generalized problem (i.e a Hessenberg & a triangular matrix)
HP	(complex) Hermitian, packed storage
HS	upper Hessenberg
OP	(real) orthogonal, packed storage
OR	(real) orthogonal
PB	symmetric or Hermitian positive definite band
PO	symmetric or Hermitian positive definite
PP	symmetric or Hermitian positive definite, packed storage
PT	symmetric or Hermitian positive definite tridiagonal
SB	(real) symmetric band
SP	symmetric, packed storage
ST	(real) symmetric tridiagonal
SY	symmetric
TB	triangular band
TG	triangular matrices, generalized problem (i.e., a pair of triangular matrices)
TP	triangular, packed storage
TR	triangular (or in some cases quasi-triangular)
TZ	trapezoidal
UN	(complex) unitary
UP	(complex) unitary, packed storage

When we wish to refer to a class of routines that performs the same function on different types of matrices, we replace the first three letters by “xyy”. Thus xyySVX refers to all the expert driver routines for systems of linear equations that are listed in Table 2.2 from Anderson et al. [1999].

Table 2.2: Driver routines for linear equations

Type of matrix and storage scheme	Operation	Single precision		Double precision	
		real	complex	real	complex
general	simple driver	SGESV	CGESV	DGESV	ZGESV
	expert driver	SGESVX	CGESVX	DGESVX	ZGESVX
general band	simple driver	SGBSV	CGBSV	DGBSV	ZGBSV
	expert driver	SGBSVX	CGBSVX	DGBSVX	ZGBSVX
general tridiagonal	simple driver	SGTSV	CGTSV	DGTSV	ZGTSV
	expert driver	SGTSVX	CGTSVX	DGTSVX	ZGTSVX

Table 2.2: Driver routines for linear equations (continue)

Type of matrix and storage scheme	Operation	Single precision		Double precision	
		real	complex	real	complex
symmetric/Hermitian	simple driver	SPOSV	CPOSV	DPOSV	ZPOSV
positive definite	expert driver	SPOSVX	CPOSVX	DPOSVX	ZPOSVX
symmetric/Hermitian	simple driver	SPPSV	CPPSV	DPSPV	ZPPSV
positive definite	expert driver	SPPSVX	CPPSVX	DPSPVX	ZPPSVX
(packed storage)					
symmetric/Hermitian	simple driver	SPBSV	CPBSV	DPBSV	ZPBSV
positive definite	expert driver	SPBSVX	CPBSVX	DPBSVX	ZPBSVX
band					
symmetric/Hermitian	simple driver	SPTSV	CPTSV	DPTSV	ZPTSV
positive definite	expert driver	SPTSVX	CPTSVX	DPTSVX	ZPTSVX
tridiagonal					
symmetric/Hermitian	simple driver	SSYSV	CHESV	DSYSV	ZHESV
indefinite	expert driver	SSYSVX	CHESVX	DSYSVX	ZHESVX
complex symmetric	simple driver		CSYSV		ZSYSV
	expert driver		CSYSVX		ZSYSVX
symmetric/Hermitian	simple driver	SSPSV	CHPSV	DSPSV	ZHPSV
indefinite (packed storage)	expert driver	SSPSVX	CHPSVX	DSPSVX	ZHPSVX
complex symmetric	simple driver		CSPSV		ZSPSV
(packed storage)	expert driver		CSPSVX		ZSPSVX

The last three letters ZZZ indicate the computation performed. For example, SGBRD is a single precision routine that performs a *bidiagonal reduction* (BRD) of a real general matrix.

The names of auxiliary routines follow a similar scheme except that the 2nd and 3rd characters YY are usually LA (for example, SLASCL or CLARFG). There are two kinds of exception. Auxiliary routines that implement an unblocked version of a block algorithm have similar names to the routines that perform the block algorithm, with the sixth character being “2” (for example, SGETF2 is the unblocked version of SGETRF). A few routines that may be regarded as extensions to the BLAS are named according to the BLAS naming schemes (for example, CROT, CSYR).

LAPACK provides two types of driver routines for solving systems of linear equations (41.1) as demonstrated in Table 2.2.

- a simple driver (name ending -SV), which solves the generalised linear system

$$AX = B$$

by factorising A and overwriting B with the solution X ; and

- an expert driver (name ending -SVX), which can also perform the following functions (some of them optionally):
 - solve $A^T X = B$ or $A^H X = B$ (unless A is symmetric or Hermitian);
 - estimate the condition number of A , check for near-singularity, and check for pivot growth;
 - refine the solution and compute forward and backward error bounds;
 - equilibrate the system if A is poorly scaled.

The expert driver requires roughly twice as much storage as the simple driver in order to perform these extra functions.

These driver routines cover all the functionality of the computational routines for linear systems, except matrix inversion. It is seldom necessary to compute the inverse of a matrix explicitly, and it is certainly not recommended as a means of solving linear systems.

Example 41.17 (Tutorial 2, Q2). Determine which of the following matrices are (i) symmetric, (ii) strictly diagonally dominant (SDD), (iii) positive definite; (iv) triangular.

$$(a) \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} \quad (b) \begin{bmatrix} 2 & 1 & 0 \\ 0 & 3 & 0 \\ 1 & 0 & 4 \end{bmatrix} \quad (c) \begin{bmatrix} 4 & 2 & 6 \\ 3 & 0 & 7 \\ -2 & -1 & -3 \end{bmatrix} \quad (d) \begin{bmatrix} 4 & 0 & 0 & 0 \\ 6 & 7 & 0 & 0 \\ 9 & 11 & 1 & 0 \\ 5 & 4 & 1 & 1 \end{bmatrix}$$

Solution. (a) SY, SDD, PO

(b) SDD, PO since $\lambda_i = 2, 3, 4 > 0$

(c) None. $\mathbf{x}^T A \mathbf{x}$ can be negative when $\mathbf{x} = [0, 0, 1]^T$.

(d) PO since $\lambda_i > 0$, TR

□

Example 41.18 (Tutorial 2, Q3). LAPACK is a linear algebra package from Netlib, investigate the matrix factorisation, solution of linear equation and the corresponding computational and driver subroutines for the following types of matrix:

1. General matrix;
parbox5cm $A = PLU$. xGETRF. xGESV
2. (Symmetric) positive definite matrix; $A = LL^T$. xPOTRF. xPOSV
3. Symmetric (indefinite) matrix; $A = LDL^T$. xSYTRF. xSYSV
4. Tridiagonal matrix $A = PLU$. xGTTRF. xGTSV

The left figure below illustrates the use of LAPACK in various platforms and the right figure illustrates development framework for iOS system.

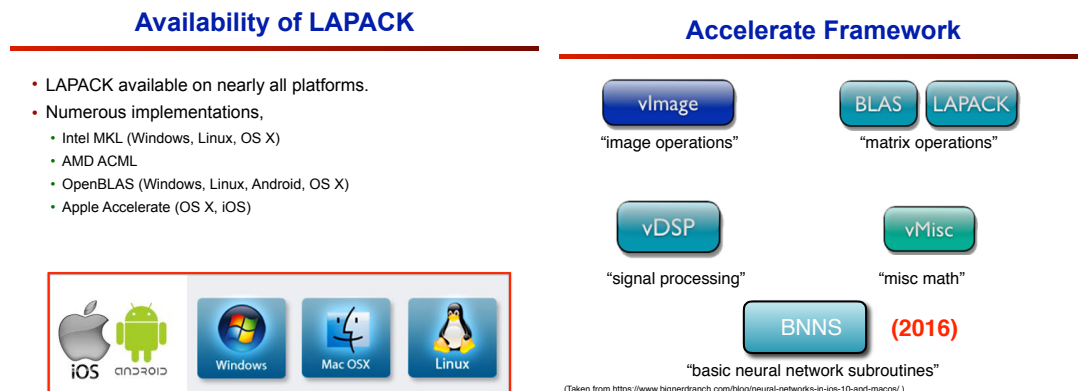


Figure 2: Carnegie Mellon 16-623 Designing Computer Vision Apps Lecture 19, 25 Slides

41.4.3 Naïve Gaussian Elimination

Naïve Gaussian elimination method is the simplest methods for solving (41.1). It may fail when one of the pivot a_{ii} (for some i) is zero. To make sure it works, it requires A to meet some criterion as follows.

Theorem 41.19. If $A(1 : k, 1 : k)$ for all $k = 1, \dots, n$ has a non-zero determinant, then naïve Gaussian elimination works.

To analyse the time complexity of naïve Gaussian elimination, we first list a simple Fortran implementation.

```

1  subroutine gausselim(n,A,b)
2  ! Warning: matrix A will be changed and b=solution vector
3      use iso_fortran_env, only : real64
4      implicit none
5      integer, intent(in) :: n
6      real(real64) :: A(n,n), b(n), c, pvt
7      integer :: i, j, k
8      ! Forward elimination
9      do i=1,n-1
10         pvt=A(i,i) ! Fixed the pivot (leading entry)
11         do j=i+1,n ! Make entries below pivot zero.
12             c=A(j,i)/pvt
13             do k=i+1,n
14                 A(j,k) = A(j,k) - A(i,k)*c
15             end do
16             b(j) = b(j) - b(i)*c
17         end do
18     end do
19     ! Backward substitution
20     do i=n,1,-1
21         do j=n,i+1,-1
22             b(i) = b(i) - A(i,j)*b(j)
23         end do
24         b(i)=b(i)/A(i,i)
25     end do
26 end subroutine gausselim

```

eg:ge:flops

Theorem 41.20. *The naïve Gaussian elimination performs approximately $\frac{2n^3}{3}$ numbers of floating point operations (flops).*

Proof: Let us investigate the lines that consist of floating point multiplication or division.

- line 12: Over the two loops (line 9 to 18), there should be

$$(n-1) + \cdots + 1 = \frac{(n-1)n}{2} \quad \text{flops (division).} \quad (41.5)$$

e:11

- line 14: Over the three loops (line 9 to 18), there should be

$$2[(n-1)^2 + \cdots + 1^2] = \frac{2(n-1)n(2n-1)}{6} \quad \text{flops (multiplication and subtraction).} \quad (41.6)$$

e:12

- line 16: Over the two loops (line 9 to 18), there should be

$$2[(n-1) + \cdots + 1] = \frac{2(n-1)n}{2} \quad \text{flops (multiplication and subtraction).} \quad (41.7)$$

e:13

- line 22: Over the two loops (line 20 to 25), there should be

$$2[1 + \cdots + (n-2)] = \frac{2(n-2)(n-1)}{2} \quad \text{flops (multiplication and subtraction).} \quad (41.8)$$

e:14

- line 24: Over a single loop (line 20 to 25), there should be

$$n \quad \text{flops (division).} \quad (41.9)$$

e:15

Adding (41.5) to (41.9) and expanding it, we obtain $\frac{2n^3}{3} + \frac{3n^2}{2} - \frac{19n}{6} + 2$ flops. When n is large, only the first term is dominant.

The following example will help us recall from the linear algebra how the above algorithm works.

eg:gel

Example 41.21 (UECM1303). Solve the following system of linear equations by applying the Gaussian Elimination:

$$\begin{bmatrix} 2 & 2 & 3 & 2 \\ 0 & 2 & 0 & 1 \\ 4 & -3 & 0 & 1 \\ 6 & 1 & -6 & -5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -2 \\ 0 \\ -7 \\ 6 \end{bmatrix}.$$

Solution: First, let us write down the augmented matrix, then we perform elementary row operations on the augmented matrix to reduce it to r.e.f.

$$\begin{aligned} & \left[\begin{array}{cccc|c} 2 & 2 & 3 & 2 & -2 \\ 0 & 2 & 0 & 1 & 0 \\ 4 & -3 & 0 & 1 & -7 \\ 6 & 1 & -6 & -5 & 6 \end{array} \right] \rightarrow \left[\begin{array}{cccc|c} 2 & 2 & 3 & 2 & -2 \\ 0 & 2 & 0 & 1 & 0 \\ 0 & -7 & -6 & -3 & -3 \\ 0 & -5 & -15 & -11 & 12 \end{array} \right] \\ & \rightarrow \left[\begin{array}{cccc|c} 2 & 2 & 3 & 2 & -2 \\ 0 & 2 & 0 & 1 & 0 \\ 0 & 0 & -6 & \frac{1}{2} & -3 \\ 0 & 0 & -15 & -\frac{17}{2} & 12 \end{array} \right] \rightarrow \left[\begin{array}{cccc|c} 2 & 2 & 3 & 2 & -2 \\ 0 & 2 & 0 & 1 & 0 \\ 0 & 0 & -6 & \frac{1}{2} & -14 \\ 0 & 0 & 0 & -\frac{39}{4} & \frac{39}{2} \end{array} \right]. \end{aligned}$$

Solving it by back substitution gives

$$x_4 = -2, \quad x_3 = \frac{1}{3}, \quad x_2 = 1, \quad x_1 = -\frac{1}{2}.$$

41.4.4 Pivoting Strategies

A serious problem with naïve Gaussian elimination is the “pivot” may be zero even if the matrix is full rank and it is numerically unstable [Higham, 2002b]. Let investigate the following example.

Example 41.22 (Taken from Strang [1988]). You are given a computer which computes with four-digit arithmetic with rounding. For more than operations involving more than 4 digits, rounding will happen. Use the Gaussian elimination to solve the following linear equation on this computer:

$$\begin{bmatrix} 0.0001 & 1.000 \\ 1.000 & 1.000 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1.000 \\ 2.000 \end{bmatrix}$$

Solution: The above linear equation can be reduced to r.e.f. as follow:

$$\begin{bmatrix} 0.0001 & 1.000 \\ 0.000 & -9999 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1.000 \\ -9998 \end{bmatrix}.$$

Solving the equation gives us $x_2 = 0.9999$ and $x_1 = (1.000 - 0.9999)/0.0001 = 0.000$, which is wrong. BEWARE: On this computer, $1.000 - 0.9999 = 0.000$ because the computer can only handle 4-digit, so 0.9999 must be rounded to 1.000 before actual calculation on the computer.

From the above example, it seems that if we have a higher precision computer, we will be able to obtain the correct answer $x_1 = 1.0001$ and $x_2 = 0.9999$. However, all computers have limited “number of bits” (instead of digits). Nowadays most personal computers are 64-bit. But who knows if 64-bit is sufficient?

To what extent should the “number of bits” be sufficient? Mathematicians have studied why the Gaussian elimination performs terribly for this example. The reason is the pivot a_{11} is too small. To avoid having a small number as pivot, pivoting strategies are developed. *Gaussian Elimination with Partial Pivoting* (GEPP) requires a maximum pivot (in absolute value) be chosen during the elimination process. A pseudo-code is given in Gerald and Wheatley [2004] and a C++ program is given in Press et al. [2007] and it is implemented as subroutines xGESV in LAPACK (Section 41.4.1). The following example explains how to obtain the solution by calling the GEPP subroutine from the LAPACK library.

eg:gepp1

Example 41.23. Solve the following system of linear equations

$$\begin{bmatrix} 0 & 2 & 0 & 1 \\ 2 & 2 & 3 & 2 \\ 4 & -3 & 0 & 1 \\ 6 & 1 & -6 & -5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ -2 \\ -7 \\ 6 \end{bmatrix} \quad (41.10)$$

e:gepp1

by using LAPACK subroutine(s) in (a) Fortran 2003; and (b) C using double precision arithmetic.

Solution: (a) We will use the standard LAPACK dgesv driver subroutine.

```

1 program gepp_eg ! Gaussian Elimination with Partial Pivoting Example
2   implicit none
3   integer, parameter :: N=4
4   double precision :: A(N,N), b(N)
5   integer :: i, info, P(N) ! Permutation for Partial Pivoting
6   A = reshape((/0,2,4,6, 2,2,-3,1, 0,3,0,-6, 1,2,1,-5/), (/N,N/))
7   b = (/0,-2,-7,6/)
8   ! LAPACK's DGESV( N, NRHS, A, LDA, IPIV, B, LDB, INFO ) solves A * X = B,
9   ! A is an N-by-N matrix and X and B are N-by-NRHS matrices.
10  call dgesv(N,1,A,N,P,b,N,info)
11  if (info .eq. 0) then
12    write(*,'(a, 4f10.6, a)') "x = (", b, ")"
13    print *, "U = "
14    do i=1,N
15      write(*,'(4f10.6)') A(i,:)
16    end do
17  else
18    print *, "Error in Gaussian Elimination with Partial Pivoting"
19  end if
20 end program gepp_eg

```

(b) One needs to beware that to call Fortran 77 subroutine DGESV in C, we need add an underscore _ to the end.

```

1 #include <stdio.h>
2
3 // LAPACK's DGESV( N, NRHS, A, LDA, IPIV, B, LDB, INFO ) solves A * X = B,
4 void dgesv_(const int *N, const int *nrhs, double *A, const int *lda,
5             int *ipiv, double *b, const int *ldb, int *info);
6
7 #define N 4
8 int main() {
9   int i,j, n=N, nrhs=1, P[N], info;
10  // Define matrix (in Fortran format), vector & suppliment variables
11  double A[N*N]={0,2,4,6,2,2,-3,1,0,3,0,-6,1,2,1,-5}, b[N] = {0,-2,-7,6};
12  dgesv_(&n,&nrhs,A,&n,P,b,&n,&info);
13  if (info == 0) {
14    printf("x = ( ");

```

```

15     for(j=0; j<N; j++) printf("%10.6f ", b[j]);
16     printf("\nU = \n");
17     for(i=0; i<N; i++) {
18         for(j=0; j<N; j++) printf("%10.6f ", A[j*N+i]); //Fortran format
19         printf("\n");
20     }
21 } else
22     printf("Error in Gaussian Elimination with Partial Pivoting");
23 }

```

Example [41.23](#) demonstrates that using the standard LAPACK library for solving linear algebra problem can be quite tedious. Numerical computing environments such as MATLAB, R, Julia were developed to simplify the tedious Fortran or C programming process.

eg:gepp2

Example 41.24 (Introduction to Scientific Computing). Write down the Julia, R and Python scripts to solve equation [\(41.10\)](#).

Solution: A sample script in Python:

```

1 import numpy as np
2 A = np.matrix("0 2 0 1; 2 2 3 2; 4 -3 0 1; 6 1 -6 -5")
3 b = np.array([0, -2, -7, 6])
4 x = np.linalg.solve(A, b)
5 print("x = {x}".format(x=x))

```

A sample script in R:

```

1 A = matrix(c(0,2,0,1,2,2,3,2,4,-3,0,1,6,1,-6,-5),4,4,byrow=T)
2 b = c(0,-2,-7,6)
3 x = solve(A,b)
4 cat("x =", x, "\n")

```

The output (with slight differences) is

```
x = [-0.5, 1.0000000000000002, 0.3333333333333337, -2.0000000000000004]
```

Despite the convenience of numerical computing environments, compiled languages such as Fortran and C++ are still much faster. With the development of advanced C++ linear algebra libraries such as Eigen3 or Armadillo, solving linear algebra problems is no longer complicated with compiled languages as demonstrated below.

eg:gepp1b

Example 41.25. Solve the linear system in Example [41.23](#) by using Armadillo C++ linear algebra library [Sanderson and Curtin, 2016] (having naming convention similar to MATLAB).

Solution: By default, Armadillo provides a wrapper for LAPACK, so one needs to link against Armadillo library only. Un-comment the first line to link against more efficient LAPACK library. Note that Armadillo makes extensive use of template meta-programming, so optimisation must be turned on to prevent slow computation.

```

1 // #define ARMA_DONT_USE_WRAPPER // http://arma.sourceforge.net/faq.html
2 #include <armadillo>
3
4 int main() {
5     arma::mat A{ {0,2,0,1}, {2,2,3,2}, {4,-3,0,1}, {6,1,-6,-5} };
6     arma::colvec b{0,-2,-7,6};
7
8     // solve Ax=b using GEPP
9     solve(A,b).print("x = ");
10    arma::mat L,U,P;
11    lu(L,U,P,A); // GEPP is closely associated with LU factorisation
12    U.print("U = ");
13 }

```


Example 41.26 (April 2014 Final Exam Q2(a)). Given $A = \begin{bmatrix} -3 & 2 & 3 \\ 2 & 2 & -1 \\ 4 & 2 & 1 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 32 \\ 4 \\ 0 \end{bmatrix}$.

1. Write down the Fortran 90 command to store the given data to the matrix A . (2 marks)

Solution: `A = reshape((-3,2,4,2,2,2,3,-1,1/), (/3,3/))` [2 marks]

2. What are the outputs of the matrix A and the vector b in the following code after the subroutine `dgesv` is called?

```
b = (/32,4,0/)
! DGESV( N, NRHS, A, LDA, IPIV, B, LDB, INFO )
call dgesv(3,1,A,3,P,b,3,info)
```

Explain your answer by writing down the Gaussian elimination with partial pivoting for solving $A\mathbf{x} = \mathbf{b}$. (7 marks)

Solution: First, let us set up the augmented matrix and reduce it to r.e.f. using GEPP:

$$\left[\begin{array}{ccc|c} -3 & 2 & 3 & 32 \\ 2 & 2 & -1 & 4 \\ 4 & 2 & 1 & 0 \end{array} \right] \rightarrow \left[\begin{array}{ccc|c} 4 & 2 & 1 & 0 \\ 2 & 2 & -1 & 4 \\ -3 & 2 & 3 & 32 \end{array} \right] \rightarrow \left[\begin{array}{ccc|c} 4 & 2 & 1 & 0 \\ 0 & 1 & -\frac{3}{2} & 4 \\ 0 & \frac{7}{2} & \frac{15}{4} & 32 \end{array} \right] \quad [1+2=3 \text{ marks}]$$

$$\rightarrow \left[\begin{array}{ccc|c} 4 & 2 & 1 & 0 \\ 0 & \frac{7}{2} & \frac{15}{4} & 32 \\ 0 & 1 & -\frac{3}{2} & 4 \end{array} \right] \rightarrow \left[\begin{array}{ccc|c} 4 & 2 & 1 & 0 \\ 0 & \frac{7}{2} & \frac{15}{4} & 32 \\ 0 & 0 & -\frac{18}{7} & -\frac{36}{7} \end{array} \right] \quad [1+1=2 \text{ marks}]$$

Solving the corresponding equation, we have the output of b :

$$\mathbf{x} = \mathbf{b} = \begin{bmatrix} -4 \\ 7 \\ 2 \end{bmatrix} \quad [1 \text{ mark}]$$

and the output of

$$A = \begin{bmatrix} 4 & 2 & 1 \\ -\frac{3}{4} & \frac{7}{2} & \frac{15}{4} \\ \frac{1}{2} & \frac{2}{7} & -\frac{18}{7} \end{bmatrix} \quad [1 \text{ mark}]$$

Example 41.27 (April 2018 Final Exam Q2(a)). You are given $A = \begin{bmatrix} 3 & 5 & -2 \\ 2 & -2 & 4 \\ 4 & 3 & 5 \end{bmatrix}$ and $\mathbf{b} =$

$$\begin{bmatrix} -15 \\ 22 \\ 16 \end{bmatrix}.$$

- (i) Your team decided to solve the linear algebra problem $A\mathbf{x} = \mathbf{b}$. You are asked to store the matrix using Fortran. Write down the Fortran statement to store the matrix properly. (2 marks)

Solution. The Fortran statement to store the matrix is [2 marks]

1 `A = reshape (/3,2,4,5,-2,3,-2,4,5/), (/3,3/))`

□

- (ii) For the following Fortran program:

```

1      ! LAPACK: DGESV( N, NRHS, A, LDA, IPIV, B, LDB, INFO )
2      call dgesv(N,1,A,N,P,b,N,info)
3      do i=1,N
4          write(*,'(4f10.6)'), (A(i,j), j=1,N)
5      end do
6      write(*,'(//,"x =",3f10.6)'), b

```

Suppose the values of the variables **A** and **b** in the program are properly set to the given matrix **A** and vector **b** respectively, write down the output of program. (8 marks)

Solution. Line 2 uses LAPACK subroutine to solve $A\mathbf{x} = \mathbf{b}$ which reduces the augmented matrix to r.e.f. using GEPP:

$$\left[\begin{array}{ccc|c} 3 & 5 & -2 & -15 \\ 2 & -2 & 4 & 22 \\ 4 & 3 & 5 & 16 \end{array} \right] \longrightarrow \left[\begin{array}{ccc|c} 4 & 3 & 5 & 16 \\ 2 & -2 & 4 & 22 \\ 3 & 5 & -2 & -15 \end{array} \right] \quad [1 \text{ mark}]$$

$$\xrightarrow{\begin{matrix} -\frac{1}{2} \\ -\frac{3}{4} \end{matrix}} \left[\begin{array}{ccc|c} 4 & 3 & 5 & 16 \\ 0 & -\frac{7}{2} & \frac{3}{2} & 14 \\ 0 & \frac{11}{4} & -\frac{23}{4} & -27 \end{array} \right] \quad [2 \text{ marks}]$$

$$\xrightarrow{\frac{22}{28}} \left[\begin{array}{ccc|c} 4 & 3 & 5 & 16 \\ 0 & -\frac{7}{2} & \frac{3}{2} & 14 \\ 0 & 0 & -\frac{32}{7} & -16 \end{array} \right] \quad [2 \text{ marks}]$$

Until line 5, the output of the Fortran code is [1 mark]

```

4.000000  3.000000  5.000000
0.500000 -3.500000  1.500000
0.750000 -0.785714 -4.571429

```

The line 6 in Fortran prints out the solution of $A\mathbf{x} = \mathbf{b}$:

$$\mathbf{b} = \mathbf{x} = \begin{bmatrix} \frac{1}{4}(16 - 3 \times (-\frac{5}{2}) - 5(\frac{7}{2})) \\ -\frac{2}{7}(14 - \frac{3}{2} \times \frac{7}{2}) \\ \frac{7}{2} \end{bmatrix} = \begin{bmatrix} 1.5 \\ -2.5 \\ 3.5 \end{bmatrix} \quad [1+1=2 \text{ marks}]$$

that is

```

x =  1.500000 -2.500000  3.500000

```

□

a-faires:6.2q3c

Example 41.28 (Tutorial 2, Q4). Use the Gaussian elimination with partial pivoting to solve the following linear system

$$\begin{aligned} 2x_1 - 3x_2 + 2x_3 &= 5 \\ -4x_1 + 2x_2 - 6x_3 &= 14. \\ 2x_1 + 2x_2 + 4x_3 &= 8 \end{aligned}$$

Solution.

$$\begin{aligned} \left[\begin{array}{cccc} 2 & -3 & 2 & 5 \\ -4 & 2 & -6 & 14 \\ 2 & 2 & 4 & 8 \end{array} \right] &\rightarrow \left[\begin{array}{cccc} -4 & 2 & -6 & 14 \\ 0 & -2 & -1 & 12 \\ 0 & 3 & 1 & 15 \end{array} \right] \rightarrow \left[\begin{array}{cccc} -4 & 2 & -6 & 14 \\ 0 & 3 & 1 & 15 \\ 0 & -2 & -1 & 12 \end{array} \right] \\ &\rightarrow \left[\begin{array}{cccc} -4 & 2 & -6 & 14 \\ 0 & 3 & 1 & 15 \\ 0 & 0 & -\frac{1}{3} & 22 \end{array} \right] \Rightarrow \begin{cases} x_3 = -66 \\ x_2 = \frac{15 - (-66)}{3} = 27 \\ x_1 = \frac{14 - 2(27) + 6(-66)}{-4} = 109 \end{cases} \end{aligned}$$

□

eg:gepp3

le:gepp1

Example 41.29. Solve the linear equation (41.10) by applying the GEPP. Show your steps in detail.

Solution.

$$\begin{array}{l}
 \begin{array}{c}
 \xrightarrow{R_1 \leftrightarrow R_4} \\
 \xrightarrow{R_2 \leftarrow R_2 - \frac{1}{3}R_1} \\
 \xrightarrow{R_3 \leftrightarrow R_2} \\
 \xrightarrow{R_4 \leftarrow R_4 - \frac{6}{11}R_2}
 \end{array}
 \begin{bmatrix}
 0 & 2 & 0 & 1 & 0 \\
 2 & 2 & 3 & 2 & -2 \\
 4 & -3 & 0 & 1 & -7 \\
 6 & 1 & -6 & -5 & 6 \\
 6 & 1 & -6 & -5 & 6 \\
 0 & \frac{5}{3} & 5 & \frac{11}{3} & -4 \\
 4 & -3 & 0 & 1 & -7 \\
 0 & 2 & 0 & 1 & 0
 \end{bmatrix}
 \xrightarrow{\begin{array}{c} R_3 \leftarrow R_3 - \frac{2}{3}R_1 \\ R_4 \leftarrow R_4 - \frac{24}{75}R_3 \end{array}}
 \begin{bmatrix}
 6 & 1 & -6 & -5 & 6 \\
 2 & 2 & 3 & 2 & -2 \\
 4 & -3 & 0 & 1 & -7 \\
 0 & 2 & 0 & 1 & 0 \\
 6 & 1 & -6 & -5 & 6 \\
 0 & \frac{5}{3} & 5 & \frac{11}{3} & -4 \\
 0 & -\frac{11}{3} & 4 & \frac{13}{3} & -11 \\
 0 & 2 & 0 & 1 & 0
 \end{bmatrix}
 \xrightarrow{\begin{array}{c} R_3 \leftrightarrow R_3 - \frac{5}{11}R_2 \\ R_4 \leftrightarrow R_4 - \frac{24}{75}R_3 \end{array}}
 \begin{bmatrix}
 6 & 1 & -6 & -5 & 6 \\
 0 & -\frac{11}{3} & 4 & \frac{13}{3} & -11 \\
 0 & \frac{5}{3} & 5 & \frac{11}{3} & -4 \\
 0 & 2 & 0 & 1 & 0 \\
 6 & 1 & -6 & -5 & 6 \\
 0 & -\frac{11}{3} & 4 & \frac{13}{3} & -11 \\
 0 & 0 & \frac{75}{11} & \frac{62}{11} & -9 \\
 0 & 0 & \frac{24}{11} & \frac{37}{11} & -6
 \end{bmatrix}
 \end{array}$$

Now we perform back substitution:

$$\begin{aligned}
 x_4 &= (-\frac{78}{25}) / (\frac{39}{25}) = -2 \\
 x_3 &= (-9 - \frac{62}{11}(-2)) / (\frac{75}{11}) = \frac{1}{3} \\
 x_2 &= (-11 - \frac{13}{3}(-2) - 4(\frac{1}{3})) / (-\frac{11}{3}) = 1 \\
 x_1 &= (6 + 5(-2) + 6(\frac{1}{3}) - 1(1)) / 6 = -\frac{1}{2}.
 \end{aligned}$$

□

Remark 41.30. In the Gaussian elimination with partial pivoting, we only find the maximum value from the first non-zero entry of each row. The *complete pivoting* (or *maximal pivoting*) searches not just down the column but also to the right through the remaining columns and it uses a combination of row interchanges (corresponding to renumbering equations) and column interchanges (corresponding to renaming variables) to bring the pivot position the largest entry in the southeast part of the matrix. However, complete pivoting is expensive time-wise.

Although it is possible to construct matrices for which partial pivoting gives arbitrarily bad results, using arbitrarily large matrices, but experience shows that these matrices simply do not occur in practice. The rapid growth of error that can happen in principle simply does not occur in practice. One reason might be the cancellation of errors that occurs during the computation process. So partial pivoting is sufficient in general (but not naïve Gaussian elimination).

Example 41.31 (Tutorial 2, Q5). Investigate the following Julia 0.6 script

```

# Forster (1994), SIAM J. Matrix Anal. Appl. 14, pp1354--1362
T = 40.0
for n=2:65
    h = T/(n-1)
    # collect() is faster and the recommended method of converting ranges to
    # arrays. Using range() requires splat operator `...`, i.e. b=[-(0:n-1)*h...]
    # Ref: https://en.wikibooks.org/wiki/Introducing_Julia/Arrays_and_tuples
    b = -collect(0:n-1)*h # convert range to array
    A = -h*tril(ones(n,n),-1)
    A = A + (1-h/2)*diagm(ones(n))
    A[:,1] = -h/2*ones(n)

```

```

A[1,1] = 1
A[:,n] = -ones(n)
A[n,n] = -h/2
xapprox = A\b
exact = norm(A*ones(n)-b)
appro = norm(A*xapprox-b)
fprintf "n=%2d: |A(exact)-b|_2=%8.4f; " n exact
fprintf " but |A(xapprox)-b|_2=%8.4f\n" appro
end

```

Explain what does the script do? What can we learn from running this script?

Solution. It is a script to demonstrate that “Gaussian Elimination with Partial Pivoting Can Fail in Practice”. In this example, when $n = 2 : 60$, the error $\|Ax^{\text{gepp}} - \mathbf{b}\|_2 \approx 0$ but the error will be large starting from 61:

```

n=61: |A(exact)-b|_2= 0.0000; but |A(xapprox)-b|_2=151.8298
n=62: |A(exact)-b|_2= 0.0000; but |A(xapprox)-b|_2=187.1947
n=63: |A(exact)-b|_2= 0.0000; but |A(xapprox)-b|_2= 6.4159
n=64: |A(exact)-b|_2= 0.0000; but |A(xapprox)-b|_2= 26.3758
n=65: |A(exact)-b|_2= 0.0000; but |A(xapprox)-b|_2= 53.4079

```

□

41.4.5 Matrix Inversion

There are situations in which a matrix inverse must be computed. For example, in statistics, the inverse can convey important statistical information, in certain matrix iterations arising in eigenvalue-related problems, etc. [Higham, 1996]

Matrix inversion can be easily carried out by solving the following linear equation using Gaussian elimination with partial pivoting:

$$AX = I_n,$$

where I_n is the $n \times n$ identity matrix. The whole process is similar to the *Gauss-Jordan elimination* which carries out elementary row operations on $[A|I]$ and transform it to $[I|A^{-1}]$ form.

eg:gepp4

Example 41.32 (Introduction to Scientific Computing). Write a script to find the inverse of the coefficient matrix in (41.10).

eg:gepp1

Solution: A sample Python script is listed below.

```

1 import numpy as np
2 import scipy.linalg as linalg
3 A = np.array([[0,2,0,1],[2,2,3,2],[4,-3,0,1],[6,1,-6,-5]], dtype=np.double)
4 inverseOfA = linalg.inv(A)
5 print(f"{A}^(-1) =\n{inverseOfA}")

```

Remark 41.33. Since the solution of $A\mathbf{x} = \mathbf{b}$ is written as $\mathbf{x} = A^{-1}\mathbf{b}$. Many people are misled to use `inv(A)*b` to find $\mathbf{x}$. However, according to Matlab help <http://www.mathworks.com/help/matlab/ref/inv.html>, it takes almost 2.5 times as long to compute the solution with $\mathbf{x} = \text{inv}(A)*\mathbf{b}$ as with $\mathbf{x} = A \setminus \mathbf{b}$. Both produce computed solutions with about the same error, 10^{-6} , reflecting the condition number of the matrix. When A is badly conditioned, $\mathbf{x} = A \setminus \mathbf{b}$ gives a more accurate answer.

41.4.6 Application in Statistics

In statistics, a linear regression problem has a least square solution satisfying the normal equation which can be solved using GEPP. Other applications of linear algebra in statistics are given in Holmes [2016, Chapter 9].

eg:lr-simple

Example 41.34. Investigate the data `headbrain.csv` from <https://github.com/mubaris/potential-enigma> with a linear regression model using (a) statistical formula; (b) normal equation.

Solution: A linear regression has the form:

$$Y = a + bX \quad (41.11)$$

e:lr-simple

(a) According to <https://mubaris.com/2017-09-28/linear-regression-from-scratch>, the statistical formula is

$$b = \frac{\sum_{i=1}^m (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^m (x_i - \bar{x})^2}; \quad a = \bar{y} - \beta_1 \bar{x} \quad (41.12)$$

e:lr-simple:ab

Using the Array Arithmetic (Section [41.3](#)), the Python code for (41.12) is

```
1 import pandas as pd # depends on numpy
2 df = pd.read_csv("headbrain.csv")
3 head_size = 'Head Size(cm^3)'; brain_wgt = 'Brain Weight(grams)'
4 X = df[head_size].values; Y = df[brain_wgt].values
5 import numpy as np
6 xbar = np.mean(X); ybar = np.mean(Y)
7 numer = sum((X-xbar)*(Y-ybar)) # Point-wise multiplication
8 denom = sum((X-xbar)**2) # Point-wise 'square'
9 b = numer/denom; a = ybar - b*xbar # Equation (2.11)
10 import matplotlib.pyplot as pl
11 pl.plot([np.min(X), np.max(X)], [a+b*np.min(X), a+b*np.max(X)],
12         color='#58b970', label='Regression Line')
13 pl.scatter(X, Y, c='#ef5423', label='Scatter Plot')
14 pl.xlabel('Head Size in cm3'); pl.ylabel('Brain Weight in grams')
15 pl.legend(); pl.show()
```

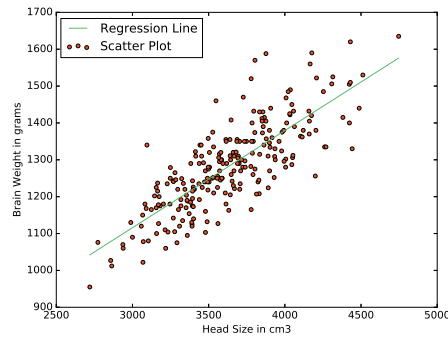
(b) The samplings of the model ([41.11](#)) can be written in the linear algebra form below:

$$\begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

which is associated with a least-square problem ([41.2](#)) and the least-square solution can be found from the normal equation using GEPP. The Python code is shown below.

```
1 import pandas as pd # depends on numpy
2 df = pd.read_csv("headbrain.csv")
3 head_size = 'Head Size(cm^3)'; brain_wgt = 'Brain Weight(grams)'
4 X = df[head_size].values; Y = df[brain_wgt].values
5 import numpy as np
6 A = np.ones((X.size, 2)); A[:,1] = X
7 coeff = np.linalg.solve(A.T @ A, A.T @ Y)
8 b = coeff[1]; a = coeff[0]
9 import matplotlib.pyplot as pl
10 pl.plot([np.min(X), np.max(X)], [a+b*np.min(X), a+b*np.max(X)],
11         color='#58b970', label='Regression Line')
12 pl.scatter(X, Y, c='#ef5423', label='Scatter Plot')
13 pl.xlabel('Head Size in cm3'); pl.ylabel('Brain Weight in grams')
14 pl.legend(); pl.show()
```

The simple linear regression result of (a) and (b) are shown below:



Note: linear regression is a basic tool in machine learning. The Python code above can be replaced by Python's machine learning library scikit-learn as follows.

```

1 import pandas as pd # depends on numpy
2 df = pd.read_csv("headbrain.csv")
3 head_size = 'Head Size(cm^3)'; brain_wgt = 'Brain Weight(grams)'
4 X = df[head_size].values; Y = df[brain_wgt].values
5 from sklearn.linear_model import LinearRegression
6 reg = LinearRegression()
7 X = X.reshape((len(X),1))
8 reg = reg.fit(X, Y)
9 Y_pred = reg.predict(X)
10 import matplotlib.pyplot as plt
11 plt.plot(X, Y_pred, color='#58b970', label='Regression Line')
12 plt.scatter(X, Y, c='#ef5423', label='Scatter Plot')
13 plt.xlabel('Head Size in cm3'); plt.ylabel('Brain Weight in grams')
14 plt.legend(bbox_to_anchor=(0.4, 1)); plt.show()

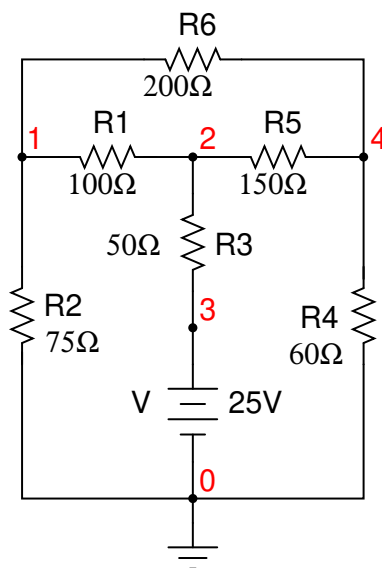
```

41.4.7 Applications in Electrical Engineering

Many linear electrical and electronic circuit problems can be transformed into linear algebra problem with the appropriate use of graph theory. We will illustrate this point with an example below.

eg:spice1

Example 41.35. Assume that the ground voltage is 0, find the voltages V_i at the nodes 1 to 4 for the following circuit. Note that by using the Kirchoff current (and voltage) law, we are able to obtain the linear equation on the right.



$$\begin{aligned}
 \frac{V_1}{75} + \frac{V_1 - V_4}{200} + \frac{V_1 - V_2}{100} &= 0 \\
 \frac{V_2 - V_1}{100} + \frac{V_2 - V_3}{50} + \frac{V_2 - V_4}{150} &= 0 \\
 V_3 &= 25 \\
 \frac{V_4 - V_1}{200} + \frac{V_4 - V_2}{150} + \frac{V_4}{60} &= 0
 \end{aligned}$$

Solution: Rewriting the above equation in matrix form:

$$\begin{bmatrix} \frac{1}{75} + \frac{1}{100} + \frac{1}{200} & -\frac{1}{100} & 0 & -\frac{1}{200} \\ -\frac{1}{100} & \frac{1}{100} + \frac{1}{50} + \frac{1}{150} & -\frac{1}{50} & -\frac{1}{150} \\ 0 & 0 & 1 & 0 \\ -\frac{1}{200} & -\frac{1}{150} & 0 & \frac{1}{150} + \frac{1}{60} + \frac{1}{200} \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 25 \\ 0 \end{bmatrix}$$

By either using hand calculation, Python, Octave or R, we obtain

$$V_1 = 6.66406859, \quad V_2 = 16.36788776, \quad V_3 = 25, \quad V_4 = 5.02727981$$

Remark 41.36. In engineering, it is inefficient and error prone to key in the values into a matrix and solve it. A practical method is to use computer aided design software or simulate it using a SPICE language below.

```
Bridge Circuit for Use with Basic Circuit Laws
* John Keown, OrCAD PSpice and Circuit Analysis 4e, p31
* To Run: ngspice -b dc-eg.cir
V 3 0 25V
R1 2 1 100
R2 1 0 75
R3 3 2 50
R4 4 0 60
R5 4 2 150
R6 4 1 200

.print DC V(1) V(2) V(3) V(4)
.dc V 25 25 1

* Print all operating conditions
* .op
.end
```

Running using the SPICE simulator, a solution to the voltages could be obtained.

Doing analysis at TEMP = 27.000000 and TNOM = 27.000000

No. of Data Rows : 1

bridge circuit for use with basic circuit laws
DC transfer characteristic Mon Nov 25 15:03:44 2013

Index	v-sweep	v(1)	v(2)	v(3)	v(4)
0	2.500000e+01	6.664069e+00	1.636789e+01	2.500000e+01	5.027280e+00

41.5 Matrix Factorisation

By applying the right elementary row operations to the augmented matrix associated with the linear system (41.1) ^{e:lin sys} transform it to a triangulated form. It is easy to obtain a solution of the triangulated form. But why is the solution obtained from the triangulated form is the same as the original linear system (41.1)? ^{e:lin sys} By investigating the structure of matrix and the corresponding matrix factorisation, mathematicians proved that the two forms are equivalent and so the solutions must be the same. LU and PLU decompositions (or factorisations, see

Sections [41.5.1](#) and [41.5.2](#)) are discovered as a consequence of this research. A rather special case when A is symmetric positive definite, we have a special LU decomposition called the Cholesky decomposition. Finally, based on the Gram-Schmidt orthogonalisation, it is possible to obtain a QR decomposition (Section [41.5.4](#)).

Gentle [2009] stated two facts that all statisticians should be aware of:

Computer numbers are not the same as real numbers, and the arithmetic operations on computer numbers are not exactly the same as those of ordinary arithmetic.

and

The form of a mathematical expression and the way the expression should be evaluated in actual practice may be quite different.

The first fact has been discussed in Topic 1. As for the second fact, a prime example is the use of the normal equations in linear regression, $X^T X b = X^T y$. The solution is often written as $\mathbf{b} = (X^T X)^{-1} X^T \mathbf{y}$. That does not mean that we ever actually compute $X^T X$ or its inverse. Instead, we will learn in Section [41.5.4](#), the linear regression is solved using QR factorisation.

41.5.1 LU Decomposition

Definition 41.6 (Fundamentals of Linear Algebra). An *LU decomposition* of a matrix A is the representation of A as the product $A = LU$ of a (square) lower triangular matrix L and an upper triangular matrix U , which is the same size as A .

Theorem 41.37. If a matrix A can be reduced to row echelon form without row interchanges (by using naïve Gaussian elimination), then A has an LU decomposition. The lower triangular matrix L can be obtained as follows:

$$L = L_1^{-1} \cdots L_{n-1}^{-1} = \begin{bmatrix} 1 & & & & \\ l_{21} & 1 & & & \\ l_{31} & l_{32} & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \cdots & l_{n,n-1} & 1 \end{bmatrix} \quad \text{where} \quad L_k = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -l_{k+1,k} & 1 & \\ & & \vdots & & \ddots \\ & & -l_{nk} & & & 1 \end{bmatrix}$$

where $l_{jk} = \frac{a'_{jk}}{a'_{kk}}$, $k < j \leq n$ and a'_{kk} are the pivots during the Gaussian elimination process.

Example 41.38. Solve the following linear system below

$$\begin{aligned} x_1 + 2x_2 + 3x_3 &= 0 \\ 4x_1 - 6x_2 - 6x_3 &= 1 \\ -7x_1 - 7x_2 + 9x_3 &= 0 \end{aligned}$$

given the LU factorisation of the coefficient matrix A :

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & -6 & -6 \\ -7 & -7 & 9 \end{bmatrix} = \begin{bmatrix} 1 & & \\ 4 & 1 & \\ -7 & -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 0 & -14 & -18 \\ 0 & 0 & 21 \end{bmatrix}.$$

Solution: First let $\mathbf{y} = U\mathbf{x}$. Then

$$A\mathbf{x} = \mathbf{b} \Rightarrow L\mathbf{y} = \mathbf{b} \Rightarrow \mathbf{y} = \begin{bmatrix} 0 \\ 1 \\ \frac{1}{2} \end{bmatrix}.$$

We can now solve

$$U\mathbf{x} = \mathbf{y} \Rightarrow \begin{cases} x_3 = \frac{1}{42}, \\ x_2 = (1 + 18(\frac{1}{42})) / (-14) = -\frac{5}{49}, \\ x_1 = -2(-\frac{5}{49}) - 3(\frac{1}{42}) = \frac{13}{98}. \end{cases}$$

41.5.2 PLU Decomposition

sec:plu

Some matrices do not have LU decomposition, $\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \neq \begin{bmatrix} 1 & 0 \\ m & 1 \end{bmatrix} \begin{bmatrix} a & b \\ 0 & c \end{bmatrix}$ for example. However, after permuting a matrix suitably, we will always have a *PLU decomposition* or an *LU factorisation (associated) with (the Gaussian elimination with partial) pivoting*, i.e.

$$A = PLU.$$

Here P is a permutation matrix which satisfies $P^T P = I_n$. This is the theoretical foundation for GEPP.

eg:plu

Example 41.39 (UECM1313). Given the matrix $A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & -1 & -1 & 4 \\ -2 & 3 & 2 & 1 \\ -6 & 4 & 2 & -8 \end{bmatrix}$.

1. Find the PLU factorisation of A .

Solution:

$$\begin{aligned} A &= \underbrace{\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}}_{P_1} \begin{bmatrix} -6 & 4 & 2 & -8 \\ 0 & -1 & -1 & 4 \\ -2 & 3 & 2 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} = P_1 \underbrace{\begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ \frac{1}{3} & & 1 & \\ 0 & & & 1 \end{bmatrix}}_{L_1} \underbrace{\begin{bmatrix} -6 & 4 & 2 & -8 \\ 0 & -1 & -1 & 4 \\ 0 & \frac{5}{3} & \frac{4}{3} & \frac{11}{3} \\ 0 & 1 & 1 & 0 \end{bmatrix}}_{U_1} \\ &= P_1 P_2 P_2 L_1 P_2 U_1 = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{P_2} U_1 = \underbrace{\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}}_{P_1 P_2} \underbrace{\begin{bmatrix} 1 & & & \\ \frac{1}{3} & 1 & & \\ 0 & & 1 & \\ 0 & & & 1 \end{bmatrix}}_{L_2 = P_2 L_1 P_2} \underbrace{\begin{bmatrix} -6 & 4 & 2 & -8 \\ 0 & \frac{5}{3} & \frac{4}{3} & \frac{11}{3} \\ 0 & -1 & -1 & 4 \\ 0 & 1 & 1 & 0 \end{bmatrix}}_{U_2 = P_2 U_1} \\ &= \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & & & \\ \frac{1}{3} & 1 & & \\ 0 & -\frac{3}{5} & 1 & \\ 0 & \frac{3}{5} & & 1 \end{bmatrix} \begin{bmatrix} -6 & 4 & 2 & -8 \\ 0 & \frac{5}{3} & \frac{4}{3} & \frac{11}{3} \\ 0 & 0 & -\frac{1}{5} & \frac{31}{5} \\ 0 & 0 & \frac{1}{5} & -\frac{11}{5} \end{bmatrix} \end{aligned}$$

$$= \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & & & \\ \frac{1}{3} & 1 & & \\ 0 & -\frac{3}{5} & 1 & \\ 0 & \frac{3}{5} & -1 & 1 \end{bmatrix} \begin{bmatrix} -6 & 4 & 2 & -8 \\ 0 & \frac{5}{3} & \frac{4}{3} & \frac{11}{3} \\ 0 & 0 & -\frac{1}{5} & \frac{31}{5} \\ 0 & 0 & 0 & 4 \end{bmatrix} = PLU$$

2. Based on what you learned in UECM1703 Introduction to Scientific Computing, write either Octave/MATLAB or Python instructions to compute the PLU factorisation of A .

Solution: The Python instructions are

```
1 import numpy as np, scipy.linalg as linalg
2 A = np.array([[0,1,1,0],[0,-1,-1,4],[-2,3,2,1],[-6,4,2,-8]],dtype=np.double)
3 L,U,P = linalg.lu(A)
4 print(f"P={P}\nL={L}\nU={U}")
```

and the Octave/MATLAB instructions are

```
1 A = [0 1 1 0; 0 -1 -1 4; -2 3 2 1; -6 4 2 -8]
2 [L,U,P] = lu(A)
```

h-faires:6.5q9c

Example 41.40 (Tutorial 2, Q6). Obtain the PLU factorisation for the matrix $A = \begin{bmatrix} 1 & -2 & 3 & 0 \\ 3 & -6 & 9 & 3 \\ 2 & 1 & 4 & 1 \\ 1 & -2 & 2 & -2 \end{bmatrix}$.

Solution. Permutation matrices P_i satisfy $P_i^{-1} = P_i^T = P_i$. In general, permutation matrix only satisfies $P^T = P^{-1}$.

$$\begin{aligned} A &= \underbrace{\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{P_1} \underbrace{\begin{bmatrix} 3 & -6 & 9 & 3 \\ 1 & -2 & 3 & 0 \\ 2 & 1 & 4 & 1 \\ 1 & -2 & 2 & -2 \end{bmatrix}}_{P_1 A} = \underbrace{P_1}_{L_1} \underbrace{\begin{bmatrix} 1 & & & \\ \frac{1}{3} & 1 & & \\ \frac{2}{3} & & 1 & \\ \frac{1}{3} & & & 1 \end{bmatrix}}_{L_1} \underbrace{\begin{bmatrix} 3 & -6 & 9 & 3 \\ 0 & 0 & 0 & -1 \\ 0 & 5 & -2 & -1 \\ 0 & 0 & -1 & -3 \end{bmatrix}}_{U_1} \\ &= \underbrace{P_1 P_2 P_2}_{L_1} \underbrace{\begin{bmatrix} 1 & & & \\ \frac{1}{3} & 1 & & \\ \frac{2}{3} & & 1 & \\ \frac{1}{3} & & & 1 \end{bmatrix}}_{L_1} \underbrace{\begin{bmatrix} 3 & -6 & 9 & 3 \\ 0 & 5 & -2 & -1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & -3 \end{bmatrix}}_{U_2 = P_2 U_1} = \underbrace{\begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{P_1 P_2} \underbrace{\begin{bmatrix} 1 & & & \\ \frac{2}{3} & 1 & & \\ \frac{1}{3} & & 1 & \\ \frac{1}{3} & & & 1 \end{bmatrix}}_{L_2 = P_2 L_1 P_2} \underbrace{\begin{bmatrix} 3 & -6 & 9 & 3 \\ 0 & 5 & -2 & -1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & -3 \end{bmatrix}}_{U_2 = P_2 U_1} \\ &= \underbrace{\begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{P = P_1 P_2 P_3} \underbrace{\begin{bmatrix} 1 & & & \\ \frac{2}{3} & 1 & & \\ \frac{1}{3} & & 1 & \\ \frac{1}{3} & & & 1 \end{bmatrix}}_{P_3 L_2 P_3} \underbrace{\begin{bmatrix} 3 & -6 & 9 & 3 \\ 0 & 5 & -2 & -1 \\ 0 & 0 & -1 & -3 \\ 0 & 0 & 0 & -1 \end{bmatrix}}_{U_3 = P_3 U_2} \end{aligned}$$

□

Example 41.41 (Tutorial 2, Q7). Obtain the PLU factorisation for the matrix $A = \begin{bmatrix} -3 & 1 & 2 \\ 1 & 3 & 2 \\ 7 & 4 & -3 \end{bmatrix}$.

$$\begin{aligned}
\text{Solution } \mathbf{A} &= \underbrace{\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}}_{P_1} \underbrace{\begin{bmatrix} 7 & 4 & -3 \\ 1 & 3 & 2 \\ -3 & 1 & 2 \end{bmatrix}}_{P_1 A} = P_1 \underbrace{\begin{bmatrix} 1 & & \\ \frac{1}{7} & 1 & \\ -\frac{3}{7} & & 1 \end{bmatrix}}_{L_1} \underbrace{\begin{bmatrix} 7 & 4 & -3 \\ 0 & \frac{17}{7} & \frac{17}{7} \\ 0 & \frac{19}{7} & \frac{5}{7} \end{bmatrix}}_{U_1} = P_1 L_1 P_2 \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}}_{P_2} U_1 \\
&= P_1 P_2 P_2 \underbrace{\begin{bmatrix} 1 & & \\ \frac{1}{7} & & 1 \\ -\frac{3}{7} & 1 & \end{bmatrix}}_{L_1 P_2} \underbrace{\begin{bmatrix} 7 & 4 & -3 \\ 0 & \frac{19}{7} & \frac{5}{7} \\ 0 & \frac{17}{7} & \frac{17}{7} \end{bmatrix}}_{U_2 = P_2 U_1} = \underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}}_{P_1 P_2} \underbrace{\begin{bmatrix} 1 & & \\ -\frac{3}{7} & 1 & \\ \frac{1}{7} & & 1 \end{bmatrix}}_{L_2 = P_2 L_1 P_2} U_2 \\
&= \underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}}_{P = P_1 P_2} \underbrace{\begin{bmatrix} 1 & & \\ -\frac{3}{7} & 1 & \\ \frac{1}{7} & \frac{17}{19} & 1 \end{bmatrix}}_L \underbrace{\begin{bmatrix} 7 & 4 & -3 \\ 0 & \frac{19}{7} & \frac{5}{7} \\ 0 & 0 & \frac{34}{19} \end{bmatrix}}_U
\end{aligned}$$

□

eg:gepp6

Example 41.42. Find the PLU factorisation of the augmented matrix in the linear equation (41.10) which is associated with the Gaussian elimination with partial pivoting.

$$\begin{aligned}
\text{Solution. The augmented matrix } \tilde{A} &= \begin{bmatrix} 0 & 2 & 0 & 1 & 0 \\ 2 & 2 & 3 & 2 & -2 \\ 4 & -3 & 0 & 1 & -7 \\ 6 & 1 & -6 & -5 & 6 \end{bmatrix} \\
\tilde{A} &= \underbrace{\begin{bmatrix} & & & 1 \\ & 1 & & \\ & & 1 & \\ 1 & & & \end{bmatrix}}_{P_1} \underbrace{\begin{bmatrix} 6 & 1 & -6 & -5 & 6 \\ 2 & 2 & 3 & 2 & -2 \\ 4 & -3 & 0 & 1 & -7 \\ 0 & 2 & 0 & 1 & 0 \end{bmatrix}}_{P_1 A} = P_1 \underbrace{\begin{bmatrix} 1 & & & \\ \frac{1}{3} & 1 & & \\ \frac{2}{3} & & 1 & \\ 0 & & & 1 \end{bmatrix}}_{L_1} \underbrace{\begin{bmatrix} 6 & 1 & -6 & -5 & 6 \\ 0 & \frac{5}{3} & 5 & \frac{11}{3} & -4 \\ 0 & -\frac{11}{3} & 4 & \frac{13}{3} & -11 \\ 0 & 2 & 0 & 1 & 0 \end{bmatrix}}_{U_1} \\
&= \underbrace{\begin{bmatrix} & & & 1 \\ & 1 & & \\ & & 1 & \\ 1 & & & \end{bmatrix}}_{P_1 P_2} \underbrace{\begin{bmatrix} 1 & & & \\ \frac{2}{3} & 1 & & \\ \frac{1}{3} & & 1 & \\ 0 & & & 1 \end{bmatrix}}_{L_2 = P_2 L_1 P_2} \underbrace{\begin{bmatrix} 6 & 1 & -6 & -5 & 6 \\ 0 & -\frac{11}{3} & 4 & \frac{13}{3} & -11 \\ 0 & \frac{5}{3} & 5 & \frac{11}{3} & -4 \\ 0 & 2 & 0 & 1 & 0 \end{bmatrix}}_{U_2 = P_2 U_1} \\
&= \underbrace{\begin{bmatrix} & & & 1 \\ & 1 & & \\ & & 1 & \\ 1 & & & \end{bmatrix}}_{P = P_1 P_2} \underbrace{\begin{bmatrix} 1 & & & \\ \frac{2}{3} & 1 & & \\ \frac{1}{3} & -\frac{5}{11} & 1 & \\ -\frac{6}{11} & & & 1 \end{bmatrix}}_{L_3} \underbrace{\begin{bmatrix} 6 & 1 & -6 & -5 & 6 \\ 0 & -\frac{11}{3} & 4 & \frac{13}{3} & -11 \\ 0 & 0 & \frac{75}{11} & \frac{62}{11} & -9 \\ 0 & 0 & \frac{24}{11} & \frac{37}{11} & -6 \end{bmatrix}}_{U_3} \\
&= \underbrace{\begin{bmatrix} & & & 1 \\ & 1 & & \\ & & 1 & \\ 1 & & & \end{bmatrix}}_{P = P_1 P_2} \underbrace{\begin{bmatrix} 1 & & & \\ \frac{2}{3} & 1 & & \\ \frac{1}{3} & -\frac{5}{11} & 1 & \\ -\frac{6}{11} & \frac{24}{75} & 1 & \end{bmatrix}}_L \underbrace{\begin{bmatrix} 6 & 1 & -6 & -5 & 6 \\ 0 & -\frac{11}{3} & 4 & \frac{13}{3} & -11 \\ 0 & 0 & \frac{75}{11} & \frac{62}{11} & -9 \\ 0 & 0 & 0 & \frac{39}{25} & -\frac{78}{25} \end{bmatrix}}_U
\end{aligned}$$

□

Example 41.43 (April 2018 Final Exam Q2(d)). A numerical method to find the eigenvalue of the matrix A closest to a given number α is implemented as a Fortran subroutine below:

```

1  subroutine some_numerical_method(n,A,x0,alpha)
2      integer, intent(in) :: n
3      real(c_double), intent(in) :: A(n,n), x0(n), alpha

```

```

4      real(c_double) :: m, x(n), y(n), B(n,n)
5      integer, external :: idamax ! BLAS Level 1
6      integer :: i, j, info, P(n)
7      m = x0(idamax(n,x0,1))
8      x = x0/m
9      write(*,'(a)',advance="no") "Step    0: "
10     call show(m,x)
11     do i=1,nstep
12         B = A
13         do j=1,n
14             B(j,j) = A(j,j) - alpha
15         end do
16         write(*,'("Step ",i3," : ")',advance="no") i
17         y = x
18         call dgesv(n,1,B,n,P,y,n,info)
19         m = y(idamax(n,y,1))
20         x = y/m
21         call show(m,x,y)
22     end do
23 end subroutine

```

For a large $n \times n$ matrix A , this Fortran subroutine performs poorly. Explain why this is the case and suggest an improvement to increase the speed of the subroutine. Justify your answer. (3 marks)

Solution. For large matrix A , because (i) A is repeatedly copy to B , (ii) $dgesv$ is solving in $O(n^3)$, therefore the subroutine performs poorly. [1+1=2 marks]
 The speed can be increased by performing PLU factorisation before the do loop and solving $PLUx = b$ takes $O(n^2)$ steps and is more efficient. [1 mark] □

41.5.3 Cholesky Decomposition

When A is a **symmetric positive definite matrix**, i.e. for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}^T A \mathbf{x} > 0$, it is possible to have an LU factorisation in which $L = U^T$ or $U = L^T$, called the Cholesky decomposition:

$$A = U^T U = LL^T.$$

Symmetric matrices often appear in nonlinear optimisation, least squares curve-fitting problems and numerical solutions of PDEs. The Cholesky decomposition is the method of choice because of its superior efficiency and numerical stability. Compared to the LU decomposition, it is roughly twice as efficient.

Mathematicians have developed an algorithm specifically for the Cholesky decomposition, a pseudo-algorithm can be found in Burden and Faires [2015]. However, it is easier for us to just perform hand calculation based on the definition.

eg:chol1

Example 41.44. Given a symmetric matrix $A = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 1 & 4 \end{bmatrix}$.

1. Find the Cholesky factorisation of A .

Solution: By definition $U = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$. So

$$\begin{bmatrix} 4 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 1 & 4 \end{bmatrix} = U^T U = \begin{bmatrix} u_{11}^2 & u_{11}u_{12} & u_{11}u_{13} \\ . & u_{12}^2 + u_{22}^2 & u_{12}u_{13} + u_{22}u_{23} \\ . & . & u_{13}^2 + u_{23}^2 + u_{33}^2 \end{bmatrix}$$

Solving the above equation, we have

$$\begin{aligned}
u_{11} &= \sqrt{4} = 2 & u_{12} &= \frac{2}{2} = 1 & u_{13} &= \frac{1}{2} \\
u_{22} &= \sqrt{3 - 1^2} = \sqrt{2} & u_{23} &= (1 - 1 \times \frac{1}{2})/\sqrt{2} = \frac{1}{\sqrt{8}} \\
u_{33} &= \sqrt{4 - \frac{1}{4} - \frac{1}{8}} = \sqrt{\frac{29}{8}}
\end{aligned}$$

Therefore, the Cholesky factorisation of A is

$$A = \begin{bmatrix} 2 & & \\ 1 & \sqrt{2} & \\ \frac{1}{2} & \frac{1}{\sqrt{8}} & \sqrt{\frac{29}{8}} \end{bmatrix} \begin{bmatrix} 2 & 1 & \frac{1}{2} \\ & \sqrt{2} & \frac{1}{\sqrt{8}} \\ & & \sqrt{\frac{29}{8}} \end{bmatrix}$$

2. Write down a single-line Octave/MATLAB or Julia command to find the Cholesky factorisation of A .

Solution: `chol([4 2 1; 2 3 1; 1 1 4])`

eg:chol2

Example 41.45. What LAPACK subroutine should we use to perform the Cholesky factorisation for double precision? Write a Fortran program to calculate the Cholesky factorisation for

the symmetric matrix $A = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 1 & 4 \end{bmatrix}$. What is the output for U ?

Solution: By using the LAPACK subroutine DPOTRF, the following program

```

1 program potrf_eg1 ! Cholesky Decomposition Example
2   use iso_fortran_env, only : real64
3   implicit none
4   integer, parameter :: N=3
5   real(real64) :: A(N,N)
6   integer :: i, j, info
7   A = reshape((/4, 2, 1, 2, 3, 1, 1, 1, 4/), (/N,N/))
8   ! DPOTRF( UPLO, N, A, LDA, INFO )
9   ! A = U**T * U, if UPLO = 'U', or A = L * L**T, if UPLO = 'L',
10  call dpotrf('U', N, A, N, info)
11  if (info .eq. 0) then
12    print *, "U = "
13    do i=1,N
14      write(*,'(4f10.6)'), (A(i,j), j=1,N)
15    end do
16  else
17    print *, "Error in Cholesky Decomposition"
18  end if
19 end program potrf_eg1

```

produce the result

$$U = \begin{bmatrix} 2.000000 & 1.000000 & 0.500000 \\ 2.000000 & 1.414214 & 0.353553 \\ 1.000000 & 1.000000 & 1.903943 \end{bmatrix}$$

eg:chol3

Example 41.46. The supervisor asked an engineer to write down the output of the following program without running the program. What is the output?

```

1 #include <armadillo>
2
3 int main() {
4   arma::mat L, A {
5     {2,-1,0}, {-1,2,-1}, {0,-1,2}
6   };
7   arma::chol(L,A,"lower"); // by default, it is upper
8   L.print("L = ");
9 }

```

Solution.

$$\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} = \begin{bmatrix} l_{11} & & \\ l_{21} & l_{22} & \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ & l_{22} & l_{32} \\ & & l_{33} \end{bmatrix} = \begin{bmatrix} l_{11}^2 & l_{11}l_{21} & l_{11}l_{31} \\ \cdot & l_{21}^2 + l_{22}^2 & l_{21}l_{31} + l_{22}l_{32} \\ \cdot & \cdot & l_{31}^2 + l_{32}^2 + l_{33}^2 \end{bmatrix}$$

$$l_{11} = \sqrt{2} \qquad l_{21} = \frac{-1}{\sqrt{2}} \qquad l_{31} = 0$$

$$l_{22} = \sqrt{2 - \frac{1}{2}} = \sqrt{\frac{3}{2}} \qquad l_{32} = -\sqrt{\frac{2}{3}}$$

$$l_{33} = \sqrt{2 - \frac{2}{3}} = \frac{2}{\sqrt{3}}$$

The output is

$$L = \begin{bmatrix} 1.41421 & 0 & 0 \\ -0.707107 & 1.22474 & 0 \\ 0 & -0.816497 & 1.1547 \end{bmatrix}$$

□

pku:nla:p23

Example 41.47 (Tutorial 2, Q8). Use Cholesky decomposition to show that $A = \begin{bmatrix} 16 & 4 & 8 & 4 \\ 4 & 10 & 8 & 4 \\ 8 & 8 & 12 & 10 \\ 4 & 4 & 10 & 12 \end{bmatrix}$ is positive definite and solve the equation $A\mathbf{x} = [36, 36, 46, 34]^T$.

Solution.

$$\begin{bmatrix} 16 & 4 & 8 & 4 \\ 4 & 10 & 8 & 4 \\ 8 & 8 & 12 & 10 \\ 4 & 4 & 10 & 12 \end{bmatrix} = \underbrace{\begin{bmatrix} l_{11} & & & \\ l_{21} & l_{22} & & \\ l_{31} & l_{32} & l_{33} & \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix}}_L \underbrace{\begin{bmatrix} l_{11} & l_{21} & l_{31} & l_{41} \\ & l_{22} & l_{32} & l_{42} \\ & & l_{33} & l_{43} \\ & & & l_{44} \end{bmatrix}}_{L^T}$$

Hence, $l_{11}^2 = 16 \Rightarrow l_{11} = 4$

$l_{21}l_{11} = 4 \Rightarrow l_{21} = 4/4 = 1, l_{31} = 8/4 = 2, l_{41} = 4/4 = 1$

$l_{21}^2 + l_{22}^2 = 10 \Rightarrow l_{22}^2 = 10 - 1^2 = 9 \Rightarrow l_{22} = 3,$

$l_{21}l_{31} + l_{22}l_{32} = 8 \Rightarrow l_{32} = \frac{8-1(2)}{3} = 2$

$l_{21}l_{41} + l_{22}l_{42} = 4 \Rightarrow l_{42} = \frac{4-1(1)}{3} = 1$

$l_{31}^2 + l_{32}^2 + l_{33}^2 = 12 \Rightarrow l_{33}^2 = 12 - 2^2 - 2^2 = 4 \Rightarrow l_{33} = 2,$

$l_{31}l_{41} + l_{32}l_{42} + l_{33}l_{43} = 10 \Rightarrow l_{43} = \frac{10-2(1)-2(1)}{2} = 3$

$l_{41}^2 + l_{42}^2 + l_{43}^2 + l_{44}^2 = 12 \Rightarrow l_{44}^2 = 12 - 1^2 - 1^2 - 3^2 = 1 \Rightarrow l_{44} = 1$

$$L = \begin{bmatrix} 4 & & & \\ 1 & 3 & & \\ 2 & 2 & 2 & \\ 1 & 1 & 3 & 1 \end{bmatrix}$$

$$\text{Let } \mathbf{y} = L^T \mathbf{x}. \text{ Then } \left\{ \begin{array}{l} A\mathbf{x} = L\mathbf{y} = \begin{bmatrix} 4 \\ 1 & 3 \\ 2 & 2 & 2 \\ 1 & 1 & 3 & 1 \end{bmatrix} \mathbf{y} = \begin{bmatrix} 36 \\ 36 \\ 46 \\ 34 \end{bmatrix} \Rightarrow \mathbf{y} = \begin{bmatrix} 9 \\ 9 \\ 5 \\ 1 \end{bmatrix} \\ L^T \mathbf{x} = \begin{bmatrix} 4 & 1 & 2 & 1 \\ & 3 & 2 & 1 \\ & & 2 & 3 \\ & & & 1 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 9 \\ 9 \\ 5 \\ 1 \end{bmatrix} \Rightarrow \mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 1 \\ 1 \end{bmatrix} \end{array} \right. \quad \square$$

41.5.4 QR Decomposition

sec:qr

The QR decomposition is a versatile computational tool that finds use in linear system problem (41.1), least square problem (41.2) and eigenvalue problem (41.3). It can be computed in three main ways. The Gram-Schmidt orthogonalisation process, which sequentially orthogonalises the columns of A , though numerically unstable, is the oldest method and is most convenient for hand calculation. *Givens transformation*, which is based on the *Givens rotation*

$$G(i, j, \theta) = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & c & \cdots & -s & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & s & \cdots & c & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix},$$

are preferred when A has a special sparsity structure, such as band or Hessenberg structure. Householder triangularisation, which uses Householder transforms $H(\mathbf{w}) = I_n - 2\mathbf{w}\mathbf{w}^T$, provides the most generally useful way to compute the QR decomposition [Higham, 1996].

Solving a least square problem (41.2) using QR decomposition would be

$$\min_{\mathbf{x} \in \mathbb{C}} \|A\mathbf{x} - \mathbf{b}\|_2 \Rightarrow \boxed{R\mathbf{x} = Q^T \mathbf{b}}. \quad (41.13)$$

e:linsys2

In Julia and Octave/MATLAB, the `\` (backslash command) uses Gaussian elimination to solve (41.1) when A is square. But when A is $m \times n$ and $m \neq n$, the backslash command uses QR decomposition to find the least-square solution for (41.13).

In the following example, we will explore how to find QR factorisation by using Gram-Schmidt orthogonalisation.

201405final:q4b

Example 41.48 (April 2014 Final Exam Question 4(b)). Given a 4×2 matrix $A = \begin{bmatrix} 1 & 1 \\ 1 & 4 \\ 1 & 3 \\ 1 & -1 \end{bmatrix}$.

- (i) Find a QR factorisation for A (7 marks)

Solution: $\mathbf{f}_1 = \mathbf{a}_1$

$$\mathbf{f}_2 = \mathbf{a}_2 - \mathbf{f}_1(\mathbf{f}_1 \cdot \mathbf{f}_1)^{-1} \mathbf{f}_1 \cdot \mathbf{a}_2 = \begin{bmatrix} 1 \\ 4 \\ 3 \\ -1 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \frac{1 + 4 + 3 - 1}{4} = \frac{1}{4} \begin{bmatrix} -3 \\ 9 \\ 5 \\ -11 \end{bmatrix}$$

..... [3 marks]

$$R = Q^T A = \begin{bmatrix} 2.0 & 3.5 \\ 0 & 3.84057 \end{bmatrix}. \quad [2 \text{ marks}]$$

Thus the QR factorisation of A is

$$A = \begin{bmatrix} 0.5 & -0.195283 \\ 0.5 & 0.585850 \\ 0.5 & 0.325472 \\ 0.5 & -0.716039 \end{bmatrix} \begin{bmatrix} 2.0 & 3.5 \\ 0 & 3.84057 \end{bmatrix}. \quad [2 \text{ marks}]$$

(ii) Find the least-square solution of $A\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$ (Hint: Use the result in part (i) above).

(5 marks)

Solution: There are two ways to solve this problem: (a) Use projection method since $A^T A$ is a 2×2 matrix. (b) Use QR factorisation of A from (i):

$$A\mathbf{x} = \mathbf{b} \Rightarrow R\mathbf{x} = Q^T \mathbf{b} = \begin{bmatrix} 5 \\ -0.91132 \end{bmatrix} \quad [3 \text{ marks}]$$

$$\begin{aligned} x_2 &= \frac{-0.91132}{3.84057} = -0.23729 \\ x_1 &= \frac{5 - 3.5(-0.23729)}{2.0} = 2.9153 \end{aligned} \quad [2 \text{ marks}]$$

eg:qr

Example 41.49 (Wikipedia, Gram-Schmidt reduction). Find a QR factorisation for the matrix

$$A = \begin{bmatrix} 12 & -51 & 4 \\ 6 & 167 & -68 \\ -4 & 24 & -41 \end{bmatrix}.$$

Solution: Let $A = [\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]$. Then

$$\mathbf{f}_1 = \mathbf{a}_1 = \begin{bmatrix} 12 \\ 6 \\ -4 \end{bmatrix}, \quad \mathbf{f}_2 = \mathbf{a}_2 - \mathbf{f}_1(\mathbf{f}_1 \cdot \mathbf{f}_1)^{-1} \mathbf{f}_1 \cdot \mathbf{a}_2 = \begin{bmatrix} -51 \\ 167 \\ 24 \end{bmatrix} - \frac{294}{196} \begin{bmatrix} 12 \\ 6 \\ -4 \end{bmatrix} = \begin{bmatrix} -69 \\ 158 \\ 30 \end{bmatrix}$$

$$\begin{aligned} \mathbf{f}_3 &= \mathbf{a}_3 - \mathbf{f}_1(\mathbf{f}_1 \cdot \mathbf{f}_1)^{-1} \mathbf{f}_1 \cdot \mathbf{a}_3 - \mathbf{f}_2(\mathbf{f}_2 \cdot \mathbf{f}_2)^{-1} \mathbf{f}_2 \cdot \mathbf{a}_3 \\ &= \begin{bmatrix} 4 \\ -68 \\ -41 \end{bmatrix} - \frac{-196}{196} \begin{bmatrix} 12 \\ 6 \\ -4 \end{bmatrix} - \frac{-12250}{30625} \begin{bmatrix} -69 \\ 158 \\ 30 \end{bmatrix} = \begin{bmatrix} -58/5 \\ 6/5 \\ -33 \end{bmatrix}. \end{aligned}$$

Hence,

$$Q = \left[\frac{\mathbf{f}_1}{\|\mathbf{f}_1\|}, \frac{\mathbf{f}_2}{\|\mathbf{f}_2\|}, \frac{\mathbf{f}_3}{\|\mathbf{f}_3\|} \right] = \begin{bmatrix} \frac{12}{14} & \frac{-69}{175} & \frac{-58/5}{35} \\ \frac{6}{14} & \frac{158}{175} & \frac{6/5}{35} \\ \frac{-4}{14} & \frac{30}{175} & \frac{-33}{35} \end{bmatrix}.$$

Since $Q^T Q = I_3$ and $A = QR$, therefore

$$R = Q^T A = \begin{bmatrix} 14 & 21 & -14 \\ 0 & 175 & -70 \\ 0 & 0 & 35 \end{bmatrix}$$

Computers normally use Householder transformation (based on the idea of “mirror reflection”) to find the QR factorisation instead of the manual Gram-Schmidt reduction in Example 41.49 because Gram-Schmidt reduction is subject to numerical stability leading to inaccurate and non-orthogonal vectors.

Definition 41.7. Let $\mathbf{w} \in \mathbb{R}^n$ with $\|\mathbf{w}\|_2 = 1$. The $n \times n$ matrix

$$H(\mathbf{w}) = I_n - 2\mathbf{w}\mathbf{w}^T$$

is called a *Householder transformation*.

The following example will probably convince us why calculating using Householder transformation by hand is a bad idea.

Example 41.50. Apply the Householder transform $H(\mathbf{w}_1)$ to make values below the leading entry of the first column of $A = \begin{bmatrix} 12 & -51 & 4 \\ 6 & 167 & -68 \\ -4 & 24 & -41 \end{bmatrix}$ become zeros. What is $\mathbf{w}_1$? Write down $H(\mathbf{w}_1)A$.

Solution. First, we find the length of the first column and form the following vector:

$$\pm \mathbf{a} = \pm \begin{bmatrix} \sqrt{12^2 + 6^2 + 4^2} \\ 0 \\ 0 \end{bmatrix} = \pm \begin{bmatrix} 14 \\ 0 \\ 0 \end{bmatrix}$$

We can either choose + or – but for numerical stability, we need to choose the vector so that the sign is the same as the value of the first column and calculate the sum and obtain $\mathbf{w}_1$:

$$\mathbf{c}_1 + \mathbf{a} = \begin{bmatrix} 26 \\ 6 \\ -4 \end{bmatrix} \Rightarrow \mathbf{w}_1 = \frac{1}{\sqrt{26^2 + 6^2 + 4^2}} \begin{bmatrix} 26 \\ 6 \\ -4 \end{bmatrix} = \frac{1}{\sqrt{728}} \begin{bmatrix} 26 \\ 6 \\ -4 \end{bmatrix}$$

$$\begin{aligned} \text{Hence, } H(\mathbf{w}_1)A &= A - 2\mathbf{w}_1\mathbf{w}_1^T A = \begin{bmatrix} 12 & -51 & 4 \\ 6 & 167 & -68 \\ -4 & 24 & -41 \end{bmatrix} - \frac{2}{728} \begin{bmatrix} 26 \\ 6 \\ -4 \end{bmatrix} \begin{bmatrix} 26 & 6 & -4 \end{bmatrix} \\ &= \begin{bmatrix} 12 & -51 & 4 \\ 6 & 167 & -68 \\ -4 & 24 & -41 \end{bmatrix} - \begin{bmatrix} 26 & -30 & -10 \\ 6 & -\frac{90}{13} & -\frac{30}{13} \\ -4 & \frac{60}{13} & \frac{20}{13} \end{bmatrix} = \begin{bmatrix} -14 & -21 & 14 \\ 0 & \frac{2261}{13} & -\frac{854}{13} \\ 0 & \frac{252}{13} & -\frac{553}{13} \end{bmatrix} \end{aligned}$$

□

Example 41.51. You are working in a company which has a program written in Fortran 95 below using the LAPACK subroutine dgeqrf:

```
1 program geqrf_eg
2   implicit none
3   integer, parameter :: N=4
4   real(kind=8) :: A(N,N), Q(N,N), H(N,N), v(N), tau(N), work(N)
5   integer :: info, i, j
6   A = reshape((/1d0,2d0,2d0,1d0, 4d0,3d0,5d0,1d0, 3d0,2d0,3d0,5d0, &
```

```

7      & 4d0,2d0,1d0,6d0/), (/N,N/))
8      ! DGEQRF( M, N, A, LDA, TAU, WORK, LWORK, INFO )
9      call dgeqrf(N,N,A,N,tau,work,N,info)
10     if (info .eq. 0) then
11         do i=1,N
12             Q(i,i)=1.0d0
13         end do
14         do i=1,N
15             if (tau(i) .ne. 0.0) then
16                 v = (/ (0.0d0,j=1,i-1), 1.0d0, A(i+1:N,i) /)
17                 do j=1,N
18                     H(j,1:N) = -tau(i)*v(j)*v(1:N)
19                     H(j,j) = 1.0d0 + H(j,j)
20                 end do
21                 Q = matmul(H, Q)
22             end if
23         end do
24         Q = transpose(Q)
25         print *, "Q = "
26         do i=1,N
27             write(*,'(5f12.6)') Q(i,:)
28         end do
29         print *, "R = "
30         do i=1,N
31             write(*,'(5f12.6)') A(i,:)
32         end do
33     else
34         print *, "Error in QR Factorisation"
35     end if
36 end program geqrf_eg

```

1. Explain what is the function of this program.

Solution. The program finds the QR factorisation of $A = \begin{bmatrix} 1 & 4 & 3 & 4 \\ 2 & 3 & 2 & 2 \\ 2 & 5 & 3 & 1 \\ 1 & 1 & 5 & 6 \end{bmatrix}$. □

2. Write a Octave/MATLAB or Julia command which produces output similar to the Fortran 95 program above.

```

Q =
-0.316228    0.723318    0.299701   -0.535715
-0.632456   -0.456832   -0.409717   -0.472690
-0.632456    0.304555   -0.163128    0.693279
-0.316228   -0.418763    0.845991    0.094538

R =
-3.162278   -6.640783   -5.692100   -5.059644
 0.480506    2.626785    0.076139   -0.228416
 0.480506    0.023833    3.820236    5.292184
 0.240253    0.328389   -0.738693   -1.827735

```

Solution.
`[Q,R] = qr([1 2 2 1;4 3 5 1;3 2 3 5;4 2 1 6]')` □

eg:qr:givens

Example 41.52 (Wikipedia). Write down the matrix after applying the Givens transform to

make values below the leading entry of the first column of $A = \begin{bmatrix} 6 & 5 & 0 \\ 5 & 1 & 4 \\ 0 & 4 & 3 \end{bmatrix}$ become zeros.

Solution: We are trying to find $c = \cos \theta$ and $s = \sin \theta$ such that

$$\begin{bmatrix} c & -s & 0 \\ s & c & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 6 & 5 & 0 \\ 5 & 1 & 4 \\ 0 & 4 & 3 \end{bmatrix} = \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & 4 & 3 \end{bmatrix}$$

By using trigonometry, it is possible for us to find

$$c = \frac{6}{\sqrt{6^2 + 5^2}} = \frac{6}{\sqrt{61}}, \quad s = -\frac{5}{\sqrt{6^2 + 5^2}} = -\frac{5}{\sqrt{61}}$$

The matrix after applying the Givens transform is

$$\begin{bmatrix} \sqrt{61} & \frac{35}{\sqrt{61}} & \frac{20}{\sqrt{61}} \\ 0 & -\frac{19}{\sqrt{61}} & \frac{24}{\sqrt{61}} \\ 0 & 4 & 3 \end{bmatrix}$$

Example 41.53 (Tutorial 2, Q9). Compute the QR decomposition of $A = \begin{bmatrix} 1 & -1 & 4 \\ 1 & 4 & -2 \\ 1 & 4 & 2 \\ 1 & -1 & 0 \end{bmatrix}$ using

1. Gram-Schmidt method;

Solution. $\mathbf{u}_1 = \mathbf{a}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$; $\mathbf{u}_2 = \mathbf{a}_2 - \mathbf{u}_1(\mathbf{u}_1 \cdot \mathbf{u}_1)^{-1} \mathbf{u}_1 \cdot \mathbf{a}_2 = \begin{bmatrix} -1 \\ 4 \\ 4 \\ -1 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \frac{-1 + 4 + 4 - 1}{4} =$

$$\frac{5}{2} \begin{bmatrix} -1 \\ 1 \\ 1 \\ -1 \end{bmatrix};$$

$$\mathbf{u}_3 = \mathbf{a}_3 - \mathbf{u}_1 \frac{\mathbf{u}_1 \cdot \mathbf{a}_3}{\mathbf{u}_1 \cdot \mathbf{u}_1} - \mathbf{u}_2 \frac{\mathbf{u}_2 \cdot \mathbf{a}_3}{\mathbf{u}_2 \cdot \mathbf{u}_2} = \begin{bmatrix} 4 \\ -2 \\ 2 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \frac{4}{4} - \begin{bmatrix} -1 \\ 1 \\ 1 \\ -1 \end{bmatrix} \frac{-4}{4} = 2 \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}$$

$$Q = \frac{1}{2} \begin{bmatrix} 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \\ 1 & -1 & -1 \end{bmatrix}, \quad R = Q^T A = \begin{bmatrix} 2 & 3 & 2 \\ & 5 & -2 \\ & & 4 \end{bmatrix}, \quad A = QR.$$

QR factorisation is $\begin{bmatrix} 1/2 & -1/2 & 1/2 \\ 1/2 & 1/2 & -1/2 \\ 1/2 & 1/2 & 1/2 \\ 1/2 & -1/2 & -1/2 \end{bmatrix} \begin{bmatrix} 2 & 3 & 2 \\ & 5 & -2 \\ & & 4 \end{bmatrix}$

□

2. Householder transform;

Solution. Let $A^{(0)} = A$.

(a) Compute the reflector $\mathbf{w}_1 = \mathbf{a}_1 + \text{sgn}(a_{11}) \|\mathbf{a}_1\| \mathbf{e}_1 = [3 \ 1 \ 1 \ 1]^T$.

$$\begin{aligned}
\text{(b) Compute } A^{(1)} &= H(\mathbf{w}_1)A^{(0)} = (I_4 - \frac{2}{\mathbf{w}_1 \cdot \mathbf{w}_1} \mathbf{w}_1 \mathbf{w}_1^T)A^{(0)} = \begin{bmatrix} -2 & -3 & -2 \\ 0 & 10/3 & -4 \\ 0 & 10/3 & 0 \\ 0 & -5/3 & -2 \end{bmatrix} \\
\text{(c) Compute the reflector } \mathbf{w}_2 &= \tilde{\mathbf{a}}_2 + \text{sgn}(a_{22})\|\tilde{\mathbf{a}}_2\|\mathbf{e}_2 = \frac{5}{3} [0 \ 5 \ 2 \ -1]^T. \\
\text{(d) Compute } A^{(2)} &= H(\mathbf{w}_2)A^{(1)} = (I_4 - \frac{2}{\mathbf{w}_2 \cdot \mathbf{w}_2} \mathbf{w}_2 \mathbf{w}_2^T)A^{(1)} = \begin{bmatrix} -2 & -3 & -2 \\ 0 & -5 & 2 \\ 0 & 0 & \frac{12}{5} \\ 0 & 0 & -\frac{16}{5} \end{bmatrix} \\
\text{(e) Compute the reflector } \mathbf{w}_3 &= \tilde{\mathbf{a}}_3 + \text{sgn}(a_{33})\|\tilde{\mathbf{a}}_3\|\mathbf{e}_3 = \frac{4}{5} [0 \ 0 \ 3+5 \ -4]^T. \\
\text{(f) Compute } A^{(3)} &= H(\mathbf{w}_3)A^{(2)} = (I_4 - \frac{2}{\mathbf{w}_3 \cdot \mathbf{w}_3} \mathbf{w}_3 \mathbf{w}_3^T)A^{(2)} = \begin{bmatrix} -2 & -3 & -2 \\ 0 & -5 & 2 \\ 0 & 0 & -4 \\ 0 & 0 & 0 \end{bmatrix} = R
\end{aligned}$$

Hence, $A^{(3)} = H(\mathbf{w}_3)H(\mathbf{w}_2)H(\mathbf{w}_1)A = R \Rightarrow A = H(\mathbf{w}_1)H(\mathbf{w}_2)H(\mathbf{w}_3)R$.

$$\text{where } H(\mathbf{w}_1)H(\mathbf{w}_2)H(\mathbf{w}_3) = \begin{bmatrix} -0.5 & 0.5 & -0.5 & -0.5 \\ -0.5 & -0.5 & 0.5 & -0.5 \\ -0.5 & -0.5 & -0.5 & 0.5 \\ -0.5 & 0.5 & 0.5 & 0.5 \end{bmatrix}.$$

$$\text{QR factorisation is } \begin{bmatrix} -0.5 & 0.5 & -0.5 & -0.5 \\ -0.5 & -0.5 & 0.5 & -0.5 \\ -0.5 & -0.5 & -0.5 & 0.5 \\ -0.5 & 0.5 & 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} -2 & -3 & -2 \\ 0 & -5 & 2 \\ 0 & 0 & -4 \\ 0 & 0 & 0 \end{bmatrix}$$

□

In Example 41.48, we learn that QR factorisation could be used to solve the least square problem. Another examples is given below.

eg:lls-qr

Example 41.54 (Fundamental of Linear Algebra, Topic 6). Use QR factorisation to find the least square solution for the linear system below.

$$\begin{cases} x_1 + x_2 = 2 \\ x_1 - x_2 = 0 \\ 2x_1 + x_2 = -4. \end{cases} \quad (41.14) \quad \text{e:6.1}$$

Solution. Equation (41.14) in matrix form: e:6.1

$$A\mathbf{x} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ -4 \end{bmatrix} = \mathbf{b} \quad (41.15) \quad \text{e:6.2}$$

Let $A = [\mathbf{v}_1, \mathbf{v}_2]$. Then

$$\mathbf{f}_1 = \mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$$

$$\mathbf{f}_2 = \mathbf{v}_2 - \mathbf{f}_1(\mathbf{f}_1 \cdot \mathbf{f}_1)^{-1} \mathbf{f}_1 \cdot \mathbf{v}_2 = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} - \frac{2}{6} \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} \frac{2}{3} \\ -\frac{4}{3} \\ \frac{1}{3} \end{bmatrix}$$

Let $\mathbf{e}_1 = \mathbf{f}_1/\|\mathbf{f}_1\|$ and $\mathbf{e}_2 = \mathbf{f}_2/\|\mathbf{f}_2\|$, where $\|\mathbf{f}_1\| = \sqrt{6}$, $\|\mathbf{f}_2\| = \sqrt{7/3}$. Then

$$\mathbf{v}_1 = \mathbf{f}_1 = \sqrt{6}\mathbf{e}_1$$

$$\mathbf{v}_2 = \mathbf{f}_2 + \frac{1}{3}\mathbf{f}_1 = \frac{\sqrt{6}}{3}\mathbf{e}_1 + \sqrt{\frac{7}{3}}\mathbf{e}_2$$

$$\text{Therefore } A = [\mathbf{v}_1, \mathbf{v}_2] = [\mathbf{e}_1, \mathbf{e}_2] \begin{bmatrix} \sqrt{6} & \sqrt{6}/3 \\ 0 & \sqrt{7/3} \end{bmatrix} = \begin{bmatrix} 1/\sqrt{6} & 2/\sqrt{21} \\ 1/\sqrt{6} & -4/\sqrt{21} \\ 2/\sqrt{6} & 1/\sqrt{21} \end{bmatrix} \begin{bmatrix} \sqrt{6} & \sqrt{6}/3 \\ 0 & \sqrt{7/3} \end{bmatrix}.$$

Since $Q^T Q = I$, Equation (41.15) becomes

$$\begin{bmatrix} \sqrt{6} & \sqrt{6}/3 \\ 0 & \sqrt{7/3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{6} & 1/\sqrt{6} & 2/\sqrt{6} \\ 2/\sqrt{21} & -4/\sqrt{21} & 1/\sqrt{21} \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ -4 \end{bmatrix}$$

Simplifying it by multiplying first row with $\sqrt{6}$ and second row with $\sqrt{21}$, we have

$$\begin{bmatrix} 6 & 2 \\ 0 & 7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 2 \\ 2 & -4 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ -4 \end{bmatrix}$$

The least square solution is

$$x_2 = \frac{0}{7}, \quad x_1 = \frac{-6}{6} = -1.$$

□

Remark 41.55. The LAPACK driver routine xGELS solves the least square problem (41.2) and the general least square problem

$$\min\{\|B - AX\|_F : X \in M_{n,k}\}, \quad A \in M_{m,n}, \quad B \in M_{m,k}$$

by Householder QR factorisation on the assumption that $\text{rank}(A) = \min\{m, n\}$, i.e., A has *full rank*. So xGELS find a least squares solution of an overdetermined system when $m > n$ and a minimum norm solution of an undetermined system when $m < n$. xGELS uses a QR or LQ factorisation of A .

If $m < n$, the factorisation takes the form

$$A = Q[R_1 \ R_2], \quad R_1 \in M_{m,m}.$$

The routine xORGQR (or xUNGQR in the complex case) is provided to form all or part of Q .

An LQ factorisation is computed by xGELQF. When $A \in M_{m,n}$ and $m \leq n$. It takes the form $A = [L \ O]Q$. It is essentially the QR factorisation of A^T . It can be used to find the minimum 2-norm solution to an underdetermined system.

In general, we may have $\text{rank}(A) < \min\{m, n\}$, i.e. A may be *rank-deficient*, we seek the *minimum norm least squares solution* $\mathbf{x}$ which minimises both $\|\mathbf{x}\|_2$ and $\|A\mathbf{x} - \mathbf{b}\|_2$. The

LAPACK driver routines xGELSX, xGELSY, xGELSS and xGELSD handle the rank-deficient case. xGELSX and xGELSY use a *complete orthogonal factorisation* of A , while xGELSS and xGELSD uses the singular value decomposition.

Example 41.56. Write a Fortran program using DGELS to solve the least-square problem in Examples 41.54.

Solution:

```

1  program leastsq_eg
2      implicit none
3      integer, parameter :: M=3, N=2, NRHS=1, LDA=M, LDB=M, lwork=2*M
4      real(kind=8) :: A(M,N), b(M), work(lwork)
5      integer :: info, i, j
6      A = reshape((/1d0,1d0,2d0, 1d0,-1d0,1d0/), (/M,N/))
7      b = (/2d0, 0d0, -4d0/)
8      call dgels('N',M,N,NRHS,A,LDA,b,LDB,work,lwork,info)
9      if (info .eq. 0) then
10         print *, "A = "
11         do i=1,M
12             write(*,'(5f12.6)'), (A(i,j), j=1,N)
13         end do
14         print *, "b = "
15         write(*,'(5f12.6)'), (b(i), i=1,M)
16     else
17         print *, "Error in Linear Least Square Problem"
18     end if
19 end program leastsq_eg

```

Remark 41.57. Apart from the important matrix factorisations above, the other important matrix factorisation is the singular value decomposition (SVD). It is a very powerful technique for dealing with (41.1) when A is either singular or numerically very close to singular. Therefore, SVD can be applied in noise reduction and data compression. LAPACK provides the driver routines xGESVD, xGESDD and xBDSCR for solving (41.1) using SVD.

For dense matrix, MATLAB relies on the matrix factorisation functions in LAPACK [Davis, 2013].

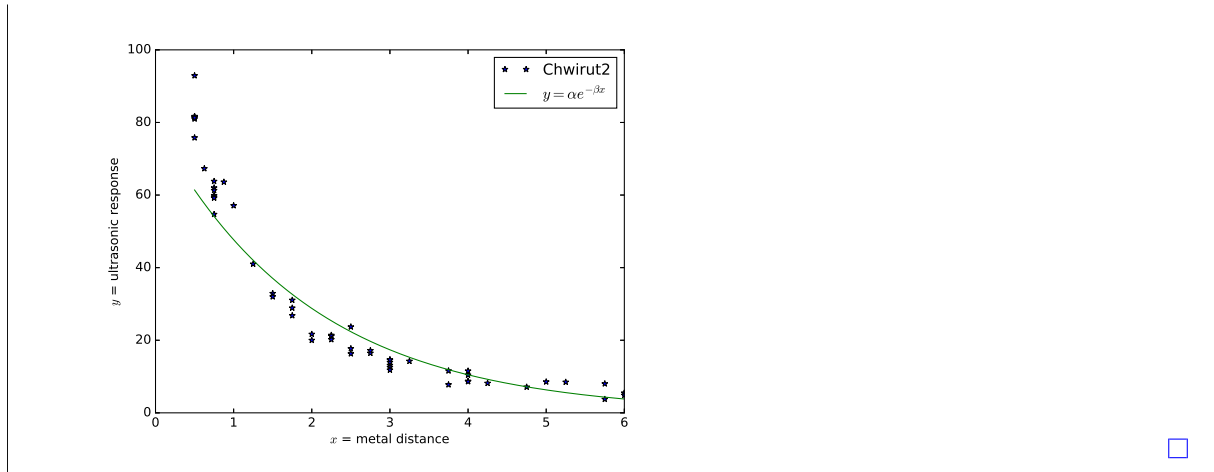
Sometimes, it is necessary for us to use the matrix factorisation, particularly in iterative refinement. After obtaining an approximation $\mathbf{x}^0$ by solving (41.1) in $O(n^3)$ floating point operations, we can reuse the LU factorisation to refine the approximation $\mathbf{x}^0$ by solving

$$A\delta\mathbf{x} = A\mathbf{x}^0 - \mathbf{b}$$

in $O(n^2)$ operations and a better floating point approximation should be $\mathbf{x} = \mathbf{x}^0 - \delta\mathbf{x}$ since $A\mathbf{x} = \mathbf{b}$.

Example 41.58 (Tutorial 2, Q10). Get the data from <http://www.itl.nist.gov/div898/strd/nls/data/chwirut2.shtml> and use the model $y = \alpha e^{-\beta x}$ to fit the data. Estimate the values of α and β .

Solution. $\ln y = -\beta x + \ln \alpha \Rightarrow \begin{bmatrix} -x_i & 1 \end{bmatrix} \begin{bmatrix} \beta \\ \ln \alpha \end{bmatrix} = [\ln y_i]$
 $\alpha = 79.03053442773087; \beta = 0.5051374207790204$



41.5.5 SVD Decomposition

sec:svd

Singular value decomposition (SVD) can be viewed as an improved technique compare to QR factorisation. In QR factorisation, only the column space basis are taken into consideration. However, SVD deals with the row and column space in a symmetric fashion.

p:5.3.4

Theorem 41.59 ((Reduced) Singular Value Decomposition, SVD). *A matrix A of order $m \times n$ with rank k can be factorised as*

$$A = U\Sigma V^* \quad (41.16)$$

e:5.3.5

where U is of order $m \times k$ s.t. $U^*U = I_k$, V is of order $n \times k$ s.t. $V^*V = I_k$ and Σ is a diagonal matrix of order $k \times k$ with all positive entries.

First thing to notice is that SVD is not unique. We have the following example.

engineers-2004

Example 41.60.

$$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos \theta & \sin \theta \\ \sin \theta & -\cos \theta \end{bmatrix}$$

is not unique since θ is arbitrary.

However, transforming SVD to the eigenproblem of $A^T A$, we are able to obtain the SVD for any matrix A .

eg:svd

Example 41.61. Compute the full SVD for the following matrix

$$A = \begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix}.$$

Solution.

1. Compute $A^T A$.

$$A^T A = \begin{bmatrix} 25 & -15 \\ -15 & 25 \end{bmatrix}.$$

Rationale: $A^* A = V\Sigma U^* U\Sigma V^* = V\Sigma^2 V^*$. In the case of real matrix, $A^* = A^T$.

2. Find the eigenvalues and unit eigenvectors of $A^T A$ and singular values of A .

$$\begin{vmatrix} 25 - \lambda & -15 \\ -15 & 25 - \lambda \end{vmatrix} = 25^2 - 50\lambda + \lambda^2 - 15^2 = \lambda^2 - 50\lambda + 400 = 0$$

Solving it gives $\lambda = 40, 10$.

When $\lambda = 40$,

$$\begin{bmatrix} -15 & -15 \\ -15 & -15 \end{bmatrix} \mathbf{v}_1 = 0 \Rightarrow \mathbf{v}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

When $\lambda = 10$,

$$\begin{bmatrix} 15 & -15 \\ -15 & 15 \end{bmatrix} \mathbf{v}_2 = 0 \Rightarrow \mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

The relation of singular value of A and eigenvalues of $A^T A$ are $\sigma = \sqrt{\lambda}$, therefore

$$\sigma_1 = \sqrt{40} \approx 6.3245, \quad \sigma_2 = \sqrt{10} \approx 3.1622.$$

$$3. \Sigma = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} = \begin{bmatrix} \sqrt{40} & 0 \\ 0 & \sqrt{10} \end{bmatrix}$$

$$4. V = [\frac{\mathbf{v}_1}{\|\mathbf{v}_1\|}, \frac{\mathbf{v}_2}{\|\mathbf{v}_2\|}] = \begin{bmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

5. Compute $U = AV\Sigma^{-1}$.

$$U = \begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix} \begin{bmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{40}} & 0 \\ 0 & \frac{1}{\sqrt{10}} \end{bmatrix} = \begin{bmatrix} -1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & -1/\sqrt{5} \end{bmatrix} = \begin{bmatrix} -0.4472 & 0.8944 \\ -0.8944 & -0.4472 \end{bmatrix}$$

$$6. A = U\Sigma V^T = \begin{bmatrix} -1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & -1/\sqrt{5} \end{bmatrix} \begin{bmatrix} \sqrt{40} & 0 \\ 0 & \sqrt{10} \end{bmatrix} \begin{bmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

□

Remark 41.62. Computing an SVD by working with the eigenproblem for $A^T A$ is numerically poor in finite-precision arithmetic. In practice, SVD is calculated via a sequence of orthogonal transformations.

Example 41.63. Solve the problem in Example [41.61](#) ^{leg:svd} using LAPACK routines for SVD `xGESVD`. Here, `x` can be `s, d, c, z` which stands for single real, double real, single complex and double complex precision.

Solution:

```
1 ! http://software.intel.com/sites/products/documentation/doclib/mkl_sa/11/mkl_lapack_examples/d
2 program svd_eg
3     use iso_c_binding, only : c_double
4     implicit none
5     integer, parameter :: m=2, n=2, lwmax=1000, lda=m, ldu=m, ldvt=n
6     real(c_double) :: A(lda,n), U(ldu,m), Vt(ldvt,n), S(n), work(lwmax)
7     integer :: i, j, info, lwork
8     data A /4,3,0,-5/ ! In fortran column is filled followed by row.
9
10    write(*,*) 'A = '
11    do i = 1, lda
12        write(*,9998) ( A( i, j ), j = 1, n )
13    end do
14    ! Query the optimal workspace.
15    lwork = -1
16    call dgesvd('all', 'all', m, n, A, lda, S, U, ldu, Vt, ldvt, &
17        work, lwork, info)
18    lwork = min(lwmax, int(work(1)))
19    call dgesvd('all', 'all', n, n, A, lda, S, U, ldu, Vt, ldvt, &
```



```

20      work, lwork, info)
21      ! Check for convergence.
22      if( info.gt.0 ) then
23          write(*,*) 'The algorithm computing SVD failed to converge.'
24          stop
25      end if
26      write(*,*) 'U = '
27      do i = 1, ldu
28          write(*,9998) ( U( i, j ), j = 1, n )
29      end do
30      write(*,*) 'S = '
31      do i = 1, n
32          write(*,9998) ( S(i) )
33      end do
34      write(*,*) 'Vt = '
35      do i = 1, ldvt
36          write(*,9998) ( Vt( i, j ), j = 1, n )
37      end do
38      9998 format( 11(:,1x,f10.6) )
39  end program svd_eg

```

Consider the reduced SVD of A , i.e. $A = U\Sigma V^T$ where $m \geq n$. Since U is orthogonal, we have

$$A = U\Sigma(U^T U)V^T = (U\Sigma U^T)(UV^T) = PZ$$

where $P = U\Sigma U^T$ and $Z = UV^T$. Z has orthonormal columns and P is symmetric positive definite since $x^T P x = (U^T x)^T \Sigma (U^T x) \geq 0$. With this we obtain the following theorem.

Theorem 41.64 (Polar Decomposition). *Any A with rank k can be factorised as $A = PZ$, where $P \in \text{Mat}_n(\mathbb{C})$ is n.n.d. with rank k and $Z \in \text{Mat}_n(\mathbb{C})$ is unitary. If A is nonsingular, then A can be factorised as $A = GQ$, where $QQ' = I$ and $G = G'$.*

Proof: This can be obtained from the SVD of A :

$$A = U\Sigma V^* = (U\Sigma U^*)(UV^*) = PZ.$$

Since $(UV^*)^*(UV^*) = VU^*UV^* = I$, $Z = UV^*$ is unitary.

The decomposition is called the *polar decomposition* because it is comparable to the decomposition of a complex number z into the modulus and angular part $z = |z|e^{i \arg z}$.

When A is symmetric positive semi-definite, A has a Cholesky decomposition $A = GG^T$. From the SVD of the matrix $G = U\Sigma V^T$, we obtain the *matrix square root decomposition* of A :

$$A = GG^T = (U\Sigma V^T)(U\Sigma V^T)^T = (U\Sigma V^T)(V\Sigma U^T) = (U\Sigma^2 U^T) = (U\tilde{\Sigma} U^T U\tilde{\Sigma} U^T) = X^2$$

where $X = U\Sigma U^T$ is called the *square root* of A .

Because of the theoretical beauty of SVD, SVD is routinely used in statistics for principle component analysis (PCA) and also used in the simulations of large LTI models using order reduction. An example of PCA is given below. See <https://stats.stackexchange.com/questions/134282/relationship-between-svd-and-pca-how-to-use-svd-to-perform-pca> for details.

Example 41.65. Modern machine learning libraries (mostly written in C++, Python, etc.) will provide PCA as part of them. Investigate how to implement PCA to handle a CSV file using C++ or Python.

Solution: Using one popular C++ library mlpack, we can code it as

```

1 // http://mlpack.org/docs/mlpack-git/doxygen.php?doc=sample.html
2 #include <mlpack/core.hpp>
3 // pca -v -i input.data -o modified.data

```

```

4 using namespace mlpack;
5 int main()
6 {
7     // First, load the data.
8     arma::mat data;
9     // Use data::Load() which transposes the matrix.
10    data::Load("data.csv", data, true);
11    // Now compute the covariance. We assume that the data is already centered.
12    // Remember, because the matrix is column-major, the covariance operation is
13    // transposed.
14    arma::mat cov = data * trans(data) / data.n_cols;
15    // Save the output.
16    data::Save("cov.csv", cov, true);
17 }

```

There are a lot of researches on the application of SVD to image compression. A simple illustration is given below.

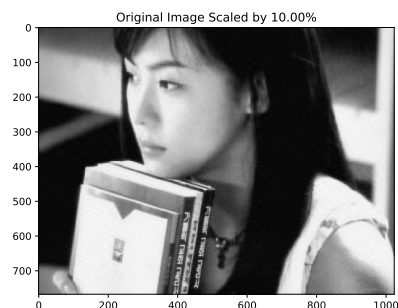
Example 41.66. Write a Python program to compress a colour image into grey image using SVD. Explain how to achieve the compression? Is it possible to compress colour image?

Solution: An example Python script is given below:

```

1 import numpy as np, matplotlib.pyplot as pl, matplotlib.image as im
2
3 def rgb2gray(rgb): return np.dot(rgb[...,:3], [0.299, 0.587, 0.114])
4
5 img_file_name = "svdimg-2009528232045197425484.jpg"
6 orig_img = im.imread(img_file_name)
7 gray_img = rgb2gray(orig_img)
8
9 U, S, V = np.linalg.svd(gray_img)
10
11 percentage = 0.1
12 percentage_size = int(percentage*min(gray_img.shape))
13 Sapprox = np.zeros(gray_img.shape)
14 print(percentage_size)
15 for i in range(percentage_size):
16     Sapprox[i,i] = S[i]
17 appr_img = U @ Sapprox @ V
18
19 #pl.imshow(gray_img, cmap=pl.get_cmap("gray"))
20 pl.imshow(appr_img, cmap=pl.get_cmap("gray"))
21 pl.title("Original Image Scaled by {p:.2%}".format(p=percentage))
22 pl.savefig("svdimg2.pdf")
23 pl.show()

```



To show how the compression is achieved, we will list the memory consumption in a table (assuming 1Mb=1000,000b, b=bytes):

	Original Image	Reduced to 10%
Characteristic	1024 × 768 8-bit RGB jpeg	1024 × 768 double
Theoretical Memory	1024 × 768 × 3 = 2.36Mb	1024 × 768 × 8 = 6.29Mb
Real Memory	2.36Mb	(1024 × 77 + 77 × 768) × 8 = 1.10Mb

So we have a 50% reduction in memory storage from a colour image to a black-and-white image.

There are disadvantages of SVD compare to the most-commonly used discrete cosine transform (DCT) method (used in JPEG image): SVD is not a good block-based transformation technique and using SVD on the whole image requires lots of memory. SVD is very computation intensive, whereas DCT has fast implementations (i.e. FFT).

41.6 Iteration Methods

Gaussian elimination and its variants in early section are *direct methods*, which are based on the theory of matrix factorisation discussed in the previous section. LU decomposition of $n \times n$ matrix requires approximately $\frac{2}{3}n^3$ flops according to Theorem 41.20. A modern Intel i7 CPU has approximately 60 GigaFLOPS ability, which means i7 can compute 6×10^{-10} floating point operations per second. If $n = 1000$, Gaussian elimination takes about $\frac{2}{3} \times 1000^3 \times 6 \times 10^{-10} = \frac{2}{5}$ seconds on a Intel i7 CPU. But when $n = 10,000$, it will take $\frac{2}{5} \times 10^3 = 400$ seconds or 6.67 minutes. In real world applications, n can be much more than 10,000, in which the computations can take a few hours to a few days.

To speed up the computation, we must make full use of the sparse structure of the large matrix and sometimes use iteration methods. The term “iterative method” refers to a wide range of techniques that use successive approximations to obtain more accurate solutions to a linear system at each step. In this section we will cover two types of iterative methods. *Stationary methods*, which include Jacobi method, Gauss-Seidel method and SOR, are older, simpler to understand and implement, but usually not as effective. *Nonstationary methods*, which include the Conjugate Gradient (CG) method and Generalised Minimal Residue (GMRES) method, are a relatively new, theoretically related to the properties of orthogonal vectors, can be highly effective [Barrett et al., 1994].

41.6.1 Iteration Process and Convergence Rate

Majority of the solution of numerical problems need to be translated into some kind of *iteration processes*. The most important class of iteration process is the *fixed point iteration process*

$$\mathbf{x}_{k+1} = T(\mathbf{x}_k), \quad k = 0, 1, 2, 3, \dots; \quad \mathbf{x}_0 = \text{initial guess vector} \quad (41.17)$$

where $T : D \rightarrow \mathbb{R}^n$ is a mapping of a domain in n -dimensional space. Examples from this topic are (41.20), (41.25), (41.28). However, one has to be careful, even if the numerical method can be written in the form (43.26), the form may not be the best way to perform calculation, (41.25) is calculated using (41.24).

The following table shows the history of various iterative methods.

1845	Jacobi	Jacobi method
1874	Seidel	Gauss-Seidel method
≈1939	Temple	Method of steepest descent
1940s	Various	Successive overrelaxation (SOR) method
1952	Hestenes and Stiefel	Conjugate gradient method

In an iteration process (43.26), a sequence of vectors $\{\mathbf{x}_k\}_{k=1}^{\infty}$ will be generated. The sequence is said to be *converge* to a fixed vector $\mathbf{x}^*$ if for every $\varepsilon > 0$, there is an integer N such that for all $k > N$, $\|\mathbf{x}_k - \mathbf{x}^*\| < \varepsilon$, where $\|\cdot\|$ is a vector norm. We call $\mathbf{x}^*$ a *limit* of the sequence $\{\mathbf{x}_k\}$:

$$\lim_{k \rightarrow \infty} \mathbf{x}_k = \mathbf{x}^*. \quad (41.18)$$

How fast does the sequence $\{\mathbf{x}_k\}$ converges to $\mathbf{x}^*$? The speed of convergence is characterised by the notion of *order of convergence*.

def:conv-rate

Definition 41.8. Suppose a sequence $\{\mathbf{x}_k\}_{k=1}^{\infty}$ converges to $\mathbf{x}^*$. The convergence rate is said to be at least

- *linear* if there is a constant $c < 1$ and an integer N such that for all $k > N$, $\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leq c\|\mathbf{x}_k - \mathbf{x}^*\|$.
- *of order α* if there is a constant $c < 1$ and an integer N such that for all $k > N$, $\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leq c\|\mathbf{x}_k - \mathbf{x}^*\|^\alpha$. When $\alpha = 1$, it is linear convergence; When $\alpha = 2$, it is called *quadratic* convergence.
- *superlinear* if there is another $\{\epsilon_k\}_{k=1}^{\infty}$ converging to 0 and an integer N such that for all $k > N$, $\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leq \epsilon_k\|\mathbf{x}_k - \mathbf{x}^*\|$.

The successful convergence of iterative methods may be dependent on the positive definite property of the coefficient matrix A in (41.1), but checking this property is as costly as solving it. However, some conditions affecting convergence, such as positive diagonal elements, diagonal dominance, and symmetry are relatively easy to verify. Therefore, the convergence theorems are developed around these ideas. Rapid convergence cannot be guaranteed unless the matrix of the system is symmetric and positive definite. For mildly nonsymmetric systems, iterative methods may still be successful. However, iterative methods may not converge when applied to systems with coefficient matrices which are completely general with no special properties.

41.6.2 Convergence Acceleration when $n = 1$

Convergence acceleration algorithms are used to obtain improved estimates of the limit $\mathbf{x}^*$ of a sequence $\{\mathbf{x}_k\}$ in (41.18) when:

- the members of the sequence can be computed to high precision;
- the behaviour of $\mathbf{x}_k$ as a function of k is regular enough.

When $n = 1$ in (43.26), a technique called *Aitken's Δ^2 process* can be used to speed up convergence of a sequence $\{s_k\}$ that is linearly convergent, based on the following theorem.

Theorem 41.67 (Aitken's Acceleration). Assume that the sequence $\{s_k\}_{k=0}^{\infty}$ converges linearly to the limit p and that $s_k \neq p$ for all $k \geq 0$. If there exists a real number A with $|A| \leq 1$ such that

$$\lim_{k \rightarrow \infty} \frac{s_{k+1} - p}{s_k - p} = A$$

then the sequence $\{q_n\}_{n=0}^{\infty}$ defined by

$$q_k = s_k - \frac{(s_{k+1} - s_k)^2}{s_{k+2} - 2s_{k+1} + s_k} \quad (41.19)$$

converges to p faster than $\{s_k\}_{k=0}^{\infty}$, in the sense that $\lim_{k \rightarrow \infty} \left| \frac{q_k - p}{s_k - p} \right| = 0$.

Introducing the Aitken's process in the fixed point iteration (43.26) with $n = 1$ leads to the *Steffensen's (acceleration) method*. A Fortran 2003 implementation is given below.

```

1  procedure(func) :: T
2  double precision, intent(in) :: x0
3  double precision, intent(out) :: root
4  ! info is optional, if unchecked, can cause SEGV
5  integer, intent(out), optional :: info, nsteps
6  double precision :: v0, v1, v2, approx, d2, relerr
7  integer :: iter
8  v0 = x0
9  do iter=1,maxIter
10     v1 = T(v0)
11     v2 = T(v1)
12     d2 = v2 - 2*v1 + v0
13     if (d2 .eq. 0d0) then                                ! Ignore Fortran Warning
14         root = v2
15         if (present(info)) info = 0
16         if (present(nsteps)) nsteps=iter
17         return
18     else
19         approx = v0 - (v1-v0)**2/d2 ! Aitken's acceleration
20     end if
21     if (v0 .eq. 0d0) then                                ! Ignore Fortran Warning
22         relerr = approx
23     else
24         relerr = (approx-v0)/v0
25     end if
26     if (abs(relerr) < accu) then
27         root = approx
28         if (present(info)) info = 0
29         if (present(nsteps)) nsteps=iter
30         return
31     end if
32     v0 = approx
33 end do
34 if (present(info)) info = 1
35 end subroutine fixedpt
36
37 subroutine bisect(f, a, b, root, info, ndiv)

```

Example 41.68. Study the following Fortran program and explain what it does? In how many iterations will the subroutine fixedpt stop?

```

1  program fixedptac
2      use textbook_root1
3      implicit none
4      real(c_double) :: x0 = 1.0d0, fp
5      real(c_double), external :: T
6      integer :: info, N
7      call setaccu(1d-8)
8      call fixedpt(T, x0, fp, info, N)
9      print *, "Fixed point with prec 1e-8 = ", fp, "after ", N, "steps"
10 end program fixedptac
11 function T(x)
12     use iso_c_binding, only : c_double
13     real(c_double) :: x, T
14     T = 20d0/(x**2+2.*x+10d0)
15 end function T

```

Solution. The program computes iteration process (43.26) with $T(x) = \frac{20}{x^2+2x+10}$ and $n = 1$ such that

$$x_{k+1} = \frac{20}{x_k^2 + 2x_k + 10}, \quad k = 0, 1, 2, \dots; x_0 = 1$$

using *Steffensen's (acceleration) method* as follows.

- $x_0^{(1)} = x_0 = 1$
 $x_1^{(1)} = T(x_0) = \frac{20}{1^2 + 2 + 10} = 1.5384615384615385;$
 $x_2^{(1)} = T(x_1^{(1)}) = \frac{20}{x_1^2 + 2x_1 + 10} = 1.2950191570881224;$
 $x_3^{(1)} = x_0 - \frac{(x_1^{(1)} - x_0)^2}{x_2^{(1)} - 2x_1^{(1)} + x_0} = 1.3708138826872336$
- $x_0^{(2)} = x_3^{(1)} = 1.3708138826872336;$
 $x_1^{(2)} = T(x_0^{(2)}) = 1.3679180902988501;$
 $x_2^{(2)} = T(x_1^{(2)}) = 1.369203162587276;$
 $x_3^{(2)} = x_0^{(2)} - \frac{(x_1^{(2)} - x_0^{(2)})^2}{x_2^{(2)} - 2x_1^{(2)} + x_0^{(2)}} = 1.3688081699448114$
- $x_0^{(3)} = 1.3688081699448114$
 $x_1^{(3)} = 1.3688080802492308;$
 $x_2^{(3)} = 1.3688081200586704;$
 $x_3^{(3)} = 1.3688081078213727$
- $x_0^{(4)} = 1.3688081078213727$
 $x_1^{(4)} = 1.3688081078213725;$
 $x_2^{(4)} = 1.3688081078213725;$
 $x_3^{(4)} = 1.3688081078213725$

Fixed point 1.3688081078213725 with a precision of 10^{-8} is obtained after 4 iterations. $\square$

Example 41.69 (Tutorial 3, Q1). Some older scientific calculator has the ability to calculate cosine straight away when the button is pressed. For any number x_0 , continue pressing the cosine many times converge to a value. The process can be modelled by a sequence $\{x_i\}_{i=0}^{\infty}$ where $x_{k+1} = \cos x_k$, $k \geq 0$.

Show that this is a contraction mapping on $[-1, 1]$ and use Steffensen's acceleration method (based on Aitken's Process) to improve the convergence of $\{x_i\}_{i=0}^{\infty}$ with $x_0 = 0$ to an accuracy of 10^{-8} . Write out the first two steps by hand and the rest can be computed using software.

Solution. Since the derivative $-\sin x$ has an absolute value of less than 1 in the interval $[-1, 1]$. So $\cos x$ is a contraction mapping.

According to the Steffensen's acceleration method, we need 3 values:

$$x_0 = 0, x_1 = \cos 0 = 1, x_2 = \cos 1 = 0.54030231$$

$$x_3 = x_0 - \frac{(x_1 - x_0)^2}{x_2 - 2x_1 + x_0} = 0 - \frac{(1 - 0)^2}{0.54030231 - 2(1) + 0} = 0.68507336$$

$$x_0^{(2)} = 0.68507336, x_1^{(2)} = 0.77437263, x_2^{(2)} = 0.71485987$$

$$x_3^{(2)} = x_0^{(2)} - \frac{(x_1^{(2)} - x_0^{(2)})^2}{x_2^{(2)} - 2x_1^{(2)} + x_0^{(2)}} = 0.73866015$$

Using a Fortran or Python, we can obtain a fixed point 0.73908513 of 8 decimal places after 4 function iterations. $\square$

Remark 41.70. An important question which is left out in the above discussion is the stability of the acceleration. In numerical computation, there are two main sources of error: *truncation error*, which arises because only a finite number of terms of an infinite sequence are taken into account; and *roundoff error*, which arises because the computer cannot perform calculations exactly. Usually roundoff error is only an important issue when the original sequence is monotonic.

In the case of acceleration methods, we typically start from a sequence in which truncation error is much larger than roundoff error. The effect of the extrapolation is to reduce the

truncation error, but to increase the roundoff error. Therefore, we need to be careful with the acceleration technique because when roundoff error dominates, the result will not be useful.

41.6.3 Sparse Matrix

LAPACK only handles dense matrices and banded matrices and is not suitable for handling sparse matrices. In engineering and science, very large sparse matrices are very common, <http://math.nist.gov/MatrixMarket/data/Harwell-Boeing/> has many examples.

Sparse matrix linear algebra problems can be solved by direct methods mentioned above but require more processing steps than the dense matrix situation we dealt with. Direct methods for solving sparse linear systems depend crucially on the pattern of the matrix and typically split into three phases [Davis, 2006]:

1. ordering and symbolic analysis, typically based solely on the nonzero pattern of A ;
2. numeric factorisation (which preserves the sparsity of the matrix);
3. uses the factorisation to solve a linear system (41.1) or a least-square problem (41.2).

Netlib has a few libraries to solve sparse matrix problem, for example Y12M [Zlatev et al., 1981]. The current standard library which is used in many open source numerical computing environment is the *SuiteSparse*, which includes GraphBLAS, Mongoose (multilevel graph partitioning), UMFPACK (Unsymmetric pattern Multi-Frontal factorisation), etc.

For sparse matrix, MATLAB uses SparseLU (UMFPACK), Cholesky (CHOLMOD), QR factorisations (SuiteSparseQR) from SuiteSparse [Davis, 2013] to solve (41.1) and (41.3). Armadillo's `spsolve` uses SuperLU to solve (41.1). There are other sparse linear algebra libraries such as MA57 and MUMPS with more restrictive license.

Octave/MATLAB and Julia relieve one from the complexity of handling sparse matrix, the following is an example in Julia and Fortran.

Example 41.71. Solving the linear algebra problem
$$\begin{bmatrix} 2 & 3 & 0 & 0 & 0 \\ 3 & 0 & 4 & 0 & 6 \\ 0 & -1 & -3 & 2 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 4 & 2 & 0 & 1 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 8 \\ 45 \\ -3 \\ 3 \\ 19 \end{bmatrix}$$
 using

1. Fortran and UMFPACK.

Solution: UMFPACK requires Compressed sparse column (CSC) format:

```
1 #include <stdio.h>
2 #include "umfpack.h"
3
4 int main(void) {
5     int i, n = 5;
6     double x[5];
7     void *Symbolic, *Numeric;
8
9     /* cumulative count of entries, as matrix is scanned columnwise */
10    int Ap[] = { 0, 2, 5, 9, 10, 12 };
11    /* row indices of entries, as matrix is scanned columnwise */
12    int Ai[] = { 0, 1, 0, 2, 4, 1, 2, 3, 4, 2, 1, 4 };
13    /* matrix entries */
14    double Ax[] = { 2, 3, 3, -1, 4, 4, -3, 1, 2, 2, 6, 1 };
15
16    double b[] = { 8, 45, -3, 3, 19 }; /* the right hand side */
17
18    /* phase 1: symbolic analysis */
19    umfpack_di_symbolic(n, Ap, Ai, Ax, &Symbolic, NULL, NULL);
20    /* phase 2: LU factorization */
21    umfpack_di_numeric(Ap, Ai, Ax, Symbolic, &Numeric, NULL, NULL);
22    umfpack_di_free_symbolic(&Symbolic);
```

```

23      /* phase 3: solve system */
24      umfpack_di_solve(UMFPACK_A, Ap, Ai, Ax, x, b, Numeric, NULL, NULL);
25      umfpack_di_free_numeric(&Numeric);
26
27      /* print results and exit */
28      for (i = 0; i < n; i++) printf("x[%d] = %g\n", i, x[i]);
29  }

```

Example 41.72 (Tutorial 2, Q11). Study the following Julia script, explain what it does and determine the output of the script:

```

1  # vec: convert a 2D array to column vector
2  Arow = vec([1 2 2 2 3 4 4 4 5 5 5 5 6 6 6])
3  Acol = vec([1 2 3 4 3 1 4 5 1 4 5 6 1 2 6])
4  Av = vec([10 12 -3 -1 15 -2 20 -2 -1 -5 1 -1 -1 -2 6])
5  b = vec(float([10 11 45 68 -22 31]))
6  A = sparse(Arow, Acol, Av)
7  x = float(A\b)
8  println("x = $x")

```

Solution. The script solves the linear algebra problem

$$\begin{bmatrix} 10 & 0 & 0 & 0 & 0 & 0 \\ 0 & 12 & -3 & -1 & 0 & 0 \\ 0 & 0 & 15 & 0 & 0 & 0 \\ -2 & 0 & 0 & 20 & -2 & 0 \\ -1 & 0 & 0 & -5 & 1 & -1 \\ -1 & -2 & 0 & 0 & 0 & 6 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 10 \\ 11 \\ 45 \\ 33 \\ -22 \\ 31 \end{bmatrix} \Rightarrow \mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{bmatrix}$$

□

41.6.4 Jacobi Method

The Jacobi method is based on solving for every variable locally with respect to the other variables; one iteration of the method corresponds to solving for every variable once. The resulting method is easy to understand and implement, but convergence is slow (so it is not useful in practice).

Let us rewrite (41.1) into a fixed point iteration form $\mathbf{x} = T(\mathbf{x})$. To do this, A is split into

$$A = U + L + D$$

where U and L are strict upper and lower triangle matrices respectively. Then (41.1) can be rearranged as

$$D\mathbf{x} = \mathbf{b} - (U + L)\mathbf{x} \xrightarrow[\text{index } k]{\text{tagging with}} D\mathbf{x}_{k+1} = \mathbf{b} - (U + L)\mathbf{x}_k.$$

Now it is easy to obtain the Jacobi iteration

$$\mathbf{x}_{k+1} = -D^{-1}(U + L)\mathbf{x}_k + D^{-1}\mathbf{b} \quad (41.20)$$

in which the component form is

$$x_i^{(k+1)} = \frac{b_i}{a_{ii}} - \frac{1}{a_{ii}} \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \frac{1}{a_{ii}} \sum_{j=i+1}^n a_{ij}x_j^{(k)}, \quad i = 1, \dots, n \quad (41.21)$$

where $x_i^{(k)} = [\mathbf{x}_k]_i$.

It is not difficult to transform (41.21) into a Fortran subroutine below.

```

1  subroutine jacobi(n,A,b,x0,x,cnt,info)
2      integer, intent(in) :: n
3      real(c_double), intent(in) :: x0(n)

```



```

4      real(c_double), intent(inout) :: A(n,n), b(n)
5      real(c_double), intent(out)  :: x(n)
6      integer, intent(out), optional :: info, cnt
7      real(c_double) :: xprev(n), error
8      integer :: i, j, iter
9
10     xprev = x0                      ! or BLAS' DCOPY(N,DX,INCX,DY,INCY)
11     do i=1,n                        ! Scale b, A according to eqn (2.15)
12         b(i) = b(i)/A(i,i)
13         A(i,1:i-1) = A(i,1:i-1) / A(i,i)    ! Warning: A is changed!
14         A(i,i+1:n) = A(i,i+1:n) / A(i,i)
15     end do
16     do iter=1,maxIter
17         do i=1,n
18             x(i) = b(i)
19             do j=1,i-1
20                 x(i) = x(i) - A(i,j)*xprev(j)
21             end do
22             do j=i+1,n
23                 x(i) = x(i) - A(i,j)*xprev(j)
24             end do
25         end do
26         error = dasum(n,x-xprev,1)      ! Using 1-norm
27         if (error<accu) then
28             if (present(cnt)) cnt = iter
29             if (present(info)) info = 0
30             return
31         end if
32         xprev = x
33     end do
34     if (present(info)) info=1 ! Max. number of iterations exceeded
35 end subroutine jacobi

```

An important task is to know when the Jacobi method is applicable. Mathematicians have shown that it is applicable to diagonally dominant matrices.

Definition 41.9. An $n \times n$ matrix A is *diagonally dominant* if $|a_{ii}| \geq \sum_{j=1, j \neq i}^n |a_{ij}|$, for $i = 1, \dots, n$. A is called *strictly diagonally dominant* if $|a_{ii}| > \sum_{j=1, j \neq i}^n |a_{ij}|$, for $i = 1, \dots, n$.

Using the eigenvalue analysis of A , mathematicians proved the following theorem using Gershgorin theorem [Beers, 2006, p. 126].

Theorem 41.73. If A is strictly diagonally dominant or diagonally dominant and irreducible, then Jacobi iteration (41.20) converges. (e:jacobi)

Assuming that $G = -D^{-1}(U + L)$, $\mathbf{f} = D^{-1}\mathbf{b}$ in (41.20). (e:jacobi) Then (41.20) becomes (e:jacobi)

$$\mathbf{x}_{k+1} = G\mathbf{x}_k + \mathbf{f}.$$

Suppose $\mathbf{x}^*$ is a fixed point, i.e. $\mathbf{x}^* = G\mathbf{x}^* + \mathbf{f}$, then Jacobi converges linearly (see Definition 41.8) (def:conv-rate) according to

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| = \|(G\mathbf{x}_k + \mathbf{f}) - (G\mathbf{x}^* + \mathbf{f})\| = \|G(\mathbf{x}_k - \mathbf{x}^*)\| \leq \|G\| \|\mathbf{x}_k - \mathbf{x}^*\|.$$

Theorem 41.74. If $\mathbf{x}^*$ is a fixed point of (41.20) (e:jacobi) (which is the same as the solution of (41.1)) (e:lin sys) and the matrix norm $\|G\| < 1$, then

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq \frac{\|G\|^k}{1 - \|G\|} \|\mathbf{x}_1 - \mathbf{x}_0\| \quad (41.22)$$

where $\mathbf{x}_0$ is the initial vector.

(e:ineq:lininter)

Outline of Proof:

$$\|\mathbf{x}_k - \mathbf{x}^*\| = \|G(\mathbf{x}_{k-1} - \mathbf{x}^*)\| = \|G^2(\mathbf{x}_{k-2} - \mathbf{x}^*)\| = \cdots = \|G^k(\mathbf{x}_0 - \mathbf{x}^*)\| \leq \|G^k\| \|\mathbf{x}_0 - \mathbf{x}^*\|.$$

Now, we need to estimate the link between $\|\mathbf{x}_0 - \mathbf{x}^*\|$ and $\|\mathbf{x}_1 - \mathbf{x}_0\|$. Since $\|G\| < 1$, $(I - G)^{-1}$ exists and

$$\begin{aligned} \|\mathbf{x}_0 - \mathbf{x}^*\| &= \|\mathbf{x}_0 - (I - G)^{-1}\mathbf{f}\| = \|(I - G)^{-1}((I - G)\mathbf{x}_0 - \mathbf{f})\| \\ &= \|(I - G)^{-1}(\mathbf{x}_0 - G\mathbf{x}_0 - \mathbf{f})\| = \|(I - G)^{-1}(\mathbf{x}_0 - \mathbf{x}_1)\| \\ &\leq \|(I - G)^{-1}\| \|\mathbf{x}_0 - \mathbf{x}_1\| \leq \frac{1}{\|I - G\|} \|\mathbf{x}_0 - \mathbf{x}_1\| \\ &\leq \frac{1}{\|I\| - \|G\|} \|\mathbf{x}_0 - \mathbf{x}_1\| = \frac{1}{1 - \|G\|} \|\mathbf{x}_0 - \mathbf{x}_1\|. \end{aligned}$$

Combining the above two inequalities, we obtain [e:ineq:lininter](#) (41.22).

[eg:jacobi-eg](#)

Example 41.75 (Non-Symmetric Matrix). Use Jacobi iteration technique to solve

$$\begin{aligned} x_1 - 10x_2 + x_3 &= 13 \\ 20x_1 + x_2 - x_3 &= 17 \\ -x_1 + x_2 + 10x_3 &= 18 \end{aligned}$$

Solution: Assuming ∞ -norm $\|\cdot\|_\infty$.

1. To ensure the convergence of this method, we rearrange the equations to obtain a strictly diagonally dominant system:
$$\begin{cases} 20x_1 + x_2 - x_3 = 17 \\ x_1 - 10x_2 + x_3 = 13 \\ -x_1 + x_2 + 10x_3 = 18 \end{cases}$$
2. Make each diagonal element the subject of each equation:

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{20}(17 - x_2^{(k)} + x_3^{(k)}) \\ x_2^{(k+1)} &= \frac{1}{10}(-13 + x_1^{(k)} + x_3^{(k)}) \\ x_3^{(k+1)} &= \frac{1}{10}(18 + x_1^{(k)} - x_2^{(k)}) \end{aligned} \tag{41.23}$$

[e:jacobi-eg](#)

3. Start with a reasonable initial guess to the solution, e.g. $x_1^{(0)} = x_2^{(0)} = x_3^{(0)} = 0$.
4. Substitute this initial approximation into the RHS of [e:jacobi-eg](#) (41.23), and calculate the new approximation

$$\begin{aligned} x_1^{(1)} &= \frac{1}{20}(17 - x_2^{(0)} + x_3^{(0)}) = 0.850 \\ x_2^{(1)} &= \frac{1}{10}(-13 + x_1^{(0)} + x_3^{(0)}) = -1.3 \\ x_3^{(1)} &= \frac{1}{10}(18 + x_1^{(0)} - x_2^{(0)}) = 1.8 \end{aligned}$$

5. To improve the approximation, we can repeat the substitution process. The next approximation is

$$\begin{aligned} x_1^{(2)} &= \frac{1}{20}(17 - (-1.3) + 1.8) = 1.005 \\ x_2^{(2)} &= \frac{1}{10}(-13 + 0.85 + 1.8) = -1.035 \\ x_3^{(2)} &= \frac{1}{10}(18 + 0.85 - (-1.3)) = 2.015 \end{aligned}$$

6. Iterate until $\max_i |x_i^{(k+1)} - x_i^{(k)}| < \epsilon$, say 0.0002. As $|x_i^{(6)} - x_i^{(5)}| < 0.0002$ for all i , we stop the computation. The results obtained are summarised in the following table:

k	0	1	2	3	4	5	6
$x_1^{(k)}$	0	0.850	1.005	1.0025	1.0001	0.99997	1.0000
$x_2^{(k)}$	0	-1.3	-1.035	-0.9980	-0.99935	-0.99999	-1.0000
$x_3^{(k)}$	0	1.8	2.015	2.004	2.0000	1.9999	2.0000

Example 41.76 (Symmetric Matrix). With an initial guess of zero vector, use the Fortran subroutine to solve the following linear equation and determine how many Jacobi iterations are there to find the solution to an accuracy of $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_1 < 10^{-6}$?

$$\begin{aligned} 3x_1 + x_2 &= 0 \\ x_1 + 3x_2 + x_3 &= -2 \\ x_2 + 3x_3 + x_4 &= -7 \\ x_3 + 3x_4 &= 6 \end{aligned}$$

Solution: Solution $\mathbf{x} = (-0.200000, 0.600000, -3.600000, 3.200000)$ is found in 26 iterations.

Example 41.77 (Tutorial 2, Q12). By using $\mathbf{x}^{(0)} = \mathbf{0}$, find the first two iterations of the Jacobi method for the linear system:

$$\begin{cases} 10x_1 + 5x_2 = 6 \\ 5x_1 + 10x_2 - 4x_3 = 25 \\ -4x_2 + 8x_3 - x_4 = -11 \\ -x_3 + 5x_4 = -11 \end{cases}$$

Solution.

$$\begin{aligned} \begin{cases} x_1^{(1)} = \frac{6 - 5x_2^{(0)}}{10} = 0.6 \\ x_2^{(1)} = \frac{25 - 5x_1^{(0)} + 4x_3^{(0)}}{10} = 2.5 \\ x_3^{(1)} = \frac{-11 + 4x_2^{(0)} + x_4^{(0)}}{8} = -1.375 \\ x_4^{(1)} = \frac{-11 + x_3^{(0)}}{5} = -2.2 \end{cases} &\Rightarrow \mathbf{x}^{(1)} = \begin{bmatrix} 0.6 \\ 2.5 \\ -1.375 \\ -2.2 \end{bmatrix} \\ \begin{cases} x_1^{(2)} = \frac{6 - 5x_2^{(1)}}{10} = -0.65 \\ x_2^{(2)} = \frac{25 - 5x_1^{(1)} + 4x_3^{(1)}}{10} = 1.65 \\ x_3^{(2)} = \frac{-11 + 4x_2^{(1)} + x_4^{(1)}}{8} = -0.4 \\ x_4^{(2)} = \frac{-11 + x_3^{(1)}}{5} = -2.475 \end{cases} &\Rightarrow \mathbf{x}^{(2)} = \begin{bmatrix} -0.65 \\ 1.65 \\ -0.4 \\ -2.475 \end{bmatrix} \end{aligned}$$

□

Example 41.78 (Tutorial 2, Q13). Given the linear system $\begin{cases} x_1 - x_3 = 0.2, \\ -\frac{1}{2}x_1 + x_2 - \frac{1}{4}x_3 = -1.425, \\ x_1 - \frac{1}{2}x_2 + x_3 = 2 \end{cases}$

with the solution $[0.9, -0.8, 0.7]^T$.

1. Is the coefficient matrix strictly diagonally dominant?

Solution. No since $1 \not\geq 1 + \frac{1}{2}$. □

2. At the 589 step, $\mathbf{x} = [0.89999992, -0.80000000, 0.69999998]$, will the Jacobi iteration stop at the 590 step if the ∞ -norm is used with an accuracy less than 10^{-7} ?

Solution.

$$\begin{aligned}x_1^{(590)} &= 0.2 + x_3^{(589)} = 0.89999998, \\x_2^{(590)} &= -1.425 + \frac{1}{2}x_1^{(589)} + \frac{1}{4}x_3^{(589)} = -1.425 + 0.89999992 + 0.69999998 \\&= -0.800000045, \\x_3^{(590)} &= 2 - x_1^{(589)} + \frac{1}{2}x_2^{(589)} = 0.700000008\end{aligned}$$

Hence, $\|\mathbf{x}^{(590)} - \mathbf{x}^{(589)}\|_\infty = \max\{0.00000006, 0.000000045, 0.00000001\} = 10^{-7} \not< 10^{-7}$. So one more step is necessary to stop. □

41.6.5 Gauss-Seidel Method

The Gauss-Seidel method is like the Jacobi method (41.21), ^{e:jacobi2} except that it uses updated values of x_j as soon as they are available, i.e.

$$x_i^{(k+1)} = \frac{b_i}{a_{ii}} - \frac{1}{a_{ii}} \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \frac{1}{a_{ii}} \sum_{j=i+1}^n a_{ij}x_j^{(k)}, \quad i = 1, \dots, n. \quad (41.24) \quad \text{e:gs}$$

This formula is related to the linear algebra problem ^{e:linsys} (41.1) by

$$(L + D)\mathbf{x} = -U\mathbf{x} + \mathbf{b} \Rightarrow \mathbf{x} = -(L + D)^{-1}U\mathbf{x} + (L + D)^{-1}\mathbf{b}$$

and the Gauss-Seidel iteration:

$$\mathbf{x}_{k+1} = -(L + D)^{-1}U\mathbf{x}_k + (L + D)^{-1}\mathbf{b}. \quad (41.25) \quad \text{e:gauss-seidel}$$

In component form, Gauss-Seidel iteration is Equation ^{e:gs} (41.24) given above.

Similar to Jacobi method ^{e:jacobi2} (41.21), it is easy to transform ^{e:gs} (41.24) into a Fortran subroutine below.

e:gauss-seidel

```
1 subroutine gaussseidel(n, A, b, x0, x, cnt, info)
2   integer, intent(in) :: n
3   integer, intent(out), optional :: info, cnt
4   real(c_double), intent(in) :: A(n,n), b(n), x0(n)
5   real(c_double), intent(out) :: x(n)
6   real(c_double) :: error, xprev(n)
7   integer :: i, j, iter
8   x = x0
9   do iter=1,maxIter
10    xprev = x
11    do i=1,n ! Updating x
12      x(i)=b(i)
13      do j=1,i-1
14        x(i) = x(i) - A(i,j)*x(j)
15      end do
16      do j=i+1,n
17        x(i) = x(i) - A(i,j)*x(j)
18      end do
19      x(i) = x(i) / A(i,i)
20    end do
21    error = dasum(n,x-xprev,1)
22    if (error < accu) then
23      if (present(cnt)) cnt = iter
24      if (present(info)) info = 0
```

```

25         return
26     end if
27 end do
28 if (present(info)) info=1 ! Max. number of iterations exceeded
29 end subroutine gaussseidel

```

When A is diagonal dominant, both Jacobi and Gauss-Seidel iterations will converge but without diagonal dominance, neither is sure to converge. In general, if the Jacobi method converges, the Gauss-Seidel method will converge faster than the Jacobi method, though still relatively slowly. It is applicable to strictly diagonally dominant, or symmetric positive definite matrices. It is normally considered as a special case of SOR method ($\omega = 1$) in the next subsection.

gauss-seidel-eg **Example 41.79** (Non-Symmetric Matrix). Use Gauss-Seidel method to solve

$$\begin{aligned}
 x_1 - 10x_2 + x_3 &= 13 \\
 20x_1 + x_2 - x_3 &= 17 \\
 -x_1 + x_2 + 10x_3 &= 18
 \end{aligned}$$

Solution:

1. To ensure the convergence of this method, we rearrange the equations to obtain a strictly diagonally dominant system:

$$\begin{aligned}
 20x_1 + x_2 - x_3 &= 17 \\
 x_1 - 10x_2 + x_3 &= 13 \\
 -x_1 + x_2 + 10x_3 &= 18
 \end{aligned}$$

2. Rewrite the above system into an iteration form:

$$\begin{aligned}
 x_1^{(k+1)} &= \frac{1}{20}(17 - x_2^{(k)} + x_3^{(k)}) \\
 x_2^{(k+1)} &= \frac{1}{10}(-13 + x_1^{(k+1)} + x_3^{(k)}) \\
 x_3^{(k+1)} &= \frac{1}{10}(18 + x_1^{(k+1)} - x_2^{(k+1)})
 \end{aligned} \tag{41.26}$$

3. Start with a reasonable initial guess to the solution, e.g. $x_1^{(0)} = x_2^{(0)} = x_3^{(0)} = 0$.
4. Substitute this initial approximation into the RHS of (41.26), and calculate the new approximation

$$\begin{aligned}
 x_1^{(1)} &= \frac{1}{20}(17 - x_2^{(0)} + x_3^{(0)}) = 0.850 \\
 x_2^{(1)} &= \frac{1}{10}(-13 + x_1^{(1)} + x_3^{(0)}) = -1.215 \\
 x_3^{(1)} &= \frac{1}{10}(18 + x_1^{(1)} - x_2^{(1)}) = 2.0065
 \end{aligned}$$

5. To improve the approximation, we can repeat the substitution process. The next approximation is

$$\begin{aligned}
 x_1^{(2)} &= \frac{1}{20}(17 - (-1.215) + 2.0065) = 1.0111 \\
 x_2^{(2)} &= \frac{1}{10}(-13 + 1.0111 + 2.0065) = -0.99824 \\
 x_3^{(2)} &= \frac{1}{10}(18 + 1.0111 - (-0.99824)) = 2.0009
 \end{aligned}$$

The results obtained are summarised in the following table:

k	0	1	2	3	4
$x_1^{(k)}$	0	0.850	1.0111	0.99995	1.0000
$x_2^{(k)}$	0	-1.215	-0.99824	-0.99992	-1.0000
$x_3^{(k)}$	0	2.0065	2.0009	2.0000	2.0000

gauss-seidel-eg2 **Example 41.80** (Symmetric Matrix). With an initial guess of zero vector, use the Fortran subroutine to solve the following linear equation and determine how many Gauss-Seidel iterations

are there to find the solution to an accuracy of $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_1 < 10^{-6}$?

$$\begin{aligned} 3x_1 + x_2 &= 0 \\ x_1 + 3x_2 + x_3 &= -2 \\ x_2 + 3x_3 + x_4 &= -7 \\ x_3 + 3x_4 &= 6 \end{aligned}$$

Solution: $\hat{\mathbf{x}} = (-0.200000, 0.600000, -3.600000, 3.200000)$ is found in 15 iterations.

-fares:7.3q3c

Example 41.81 (Tutorial 2, Q14). Find the first two iterations of the the Gauss-Siedel method for the linear system

$$\begin{cases} 10x_1 + 5x_2 = 6 \\ 5x_1 + 10x_2 - 4x_3 = 25 \\ -4x_2 + 8x_3 - x_4 = -11 \\ -x_3 + 5x_4 = -11 \end{cases}$$

using $\mathbf{x}^{(0)} = \mathbf{0}$.

Solution.

$$\begin{aligned} x_1^{(1)} &= \frac{6 - 5x_2^{(0)}}{10} = 0.6 \\ x_2^{(1)} &= \frac{25 - 5x_1^{(1)} + 4x_3^{(0)}}{10} = \boxed{2.2} \\ x_3^{(1)} &= \frac{-11 + 4x_2^{(1)} + x_4^{(0)}}{8} = \boxed{-0.275} \\ x_4^{(1)} &= \frac{-11 + x_3^{(1)}}{5} = \boxed{-2.255} \\ x_1^{(2)} &= \frac{6 - 5x_2^{(1)}}{10} = \boxed{-0.5} \\ x_2^{(2)} &= \frac{25 - 5x_1^{(2)} + 4x_3^{(1)}}{10} = \boxed{2.64} \\ x_3^{(2)} &= \frac{-11 + 4x_2^{(2)} + x_4^{(1)}}{8} = \boxed{-0.33687} \\ x_4^{(2)} &= \frac{-11 + x_3^{(2)}}{5} = \boxed{-2.2674} \end{aligned}$$

$$\begin{aligned} \mathbf{x}^{(1)} &= [0.6, 2.2, -0.275, -2.255]^T, \\ \mathbf{x}^{(2)} &= [-0.5, 2.64, -0.33687, -2.2674]^T \end{aligned}$$

□

41.6.6 Successive Over-Relaxation (SOR) Method

Overrelaxation can be applied to Gauss-Seidel method (^{e:gs}41.24) by introducing an extrapolation parameter ω to the right-hand-side of (^{e:gs}41.24):

$$x_i^{(k+1)} = (1 - \omega)x_i^{(k)} + \omega \left(\frac{b_i}{a_{ii}} - \frac{1}{a_{ii}} \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \frac{1}{a_{ii}} \sum_{j=i+1}^n a_{ij}x_j^{(k)} \right), \quad i = 1, \dots, n. \quad (41.27) \quad \text{e:sor}$$

This is called the *successive overrelaxation (SOR) method*, the parameter ω is called the *over-relaxation factor* and *under-relaxation factor* when $1 < \omega$ and $0 < \omega < 1$ respectively.

To study the convergence of SOR method, we need to write down the matrix form of (^{e:sor}41.27):

$$\mathbf{x}_{k+1} = (1 - \omega)\mathbf{x}_k + \omega D^{-1}(\mathbf{b} - L\mathbf{x}_{k+1} - U\mathbf{x}_k)$$

Rearranging it so that $\mathbf{x}_{k+1}$ is on the left-hand-side and $\mathbf{x}_k$ is on the right-hand-side:

$$(D + \omega L)\mathbf{x}_{k+1} = \omega\mathbf{b} + [(1 - \omega)D - \omega U]\mathbf{x}_k$$

leads to a fixed point iteration form (43.26):

$$\mathbf{x}_{k+1} = \underbrace{(D + \omega L)^{-1}[(1 - \omega)D - \omega U]}_{G_\omega} \mathbf{x}_k + \underbrace{\omega(D + \omega L)^{-1}\mathbf{b}}_{\mathbf{f}_\omega}. \quad (41.28)$$

When $\omega = 1$, we obtain the Gauss-Seidel method. Therefore, SOR can be seen as the acceleration of the convergence of Gauss-Seidel method. So, it is essential to identify when will SOR converge. The following theorem characterises the type of matrices where SOR converges.

Theorem 41.82. *If SOR method (41.27) converges, then $0 < \omega < 2$. If A is strictly diagonally dominant and $0 < \omega < 1$, then SOR method converges. If A is real symmetric and positively definite and $0 < \omega < 2$, then SOR method converges.*

The speed of convergence depends critically on ω and the optimal value for ω may be estimated from the spectral radius of the Jacobi iteration matrix under certain conditions. The proof of Theorem 41.82 involves complicated eigenvalue estimation of A and will be ignored here. We will use the following example to demonstrate the influence of ω on the convergence of SOR.

Similar to the Gauss-Seidel method, a Fortran 2003 subroutine for (41.27) can be naively implemented.

```

1  subroutine sor(n, A, b, x0, w, x, cnt, info)
2      integer, intent(in) :: n
3      integer, intent(out), optional :: info, cnt
4      real(c_double), intent(in) :: A(n,n), b(n), x0(n), w
5      real(c_double), intent(out) :: x(n)
6      real(c_double) :: error, xprev(n), res
7      integer :: i, j, iter
8      x = x0
9      do iter=1,maxIter
10         xprev = x
11         do i=1,n
12             x(i) = b(i)
13             do j=1,n ! Similar to Gauss-Seidel but using comparison instead
14                 if (j .ne. i) x(i) = x(i) - A(i,j)*x(j)
15             end do
16             x(i) = (1-w)*xprev(i) + w*x(i)/A(i,i) ! Linear Extrapolation
17         end do
18         error = dasum(n,x-xprev,1)
19         res = dnorm2(n,matmul(A,x)-b,1) ! When w=0, error=0 because xprev=0
20         if (error < accu .and. res < accu) then
21             if (present(cnt)) cnt = iter
22             if (present(info)) info = 0
23             return
24         end if
25     end do
26     if (present(info)) info=1 ! Max. number of iterations exceeded
27 end subroutine sor

```

eg:sor-eg2

Example 41.83 (Symmetric Matrix). With an initial guess of zero vector, use the Fortran subroutine to solve the following linear equation and determine how many SOR iterations with $\omega = 1.1$ are there to find the solution to an accuracy of $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_1 < 10^{-6}$? What about the case $\omega = 0.9, 1.3, 1.5$?

$$\begin{cases} 3x_1 + x_2 & = 0 \\ x_1 + 3x_2 + x_3 & = -2 \\ x_2 + 3x_3 + x_4 & = -7 \\ x_3 + 3x_4 & = 6 \end{cases} \quad (41.29)$$

Solution: The solution vector to an absolute accuracy of 6 decimal places is $\hat{\mathbf{x}} = (-0.200000, 0.600000, -3.600000, 3.200000)$.

$\omega = 1.1$, the solution $\hat{\mathbf{x}}$ is found in 9 iterations.

$\omega = 0.9$, the solution $\hat{\mathbf{x}}$ is found in 20 iterations.

$\omega = 1.3$, the solution $\hat{\mathbf{x}}$ is found in 15 iterations.

$\omega = 1.5$, the solution $\hat{\mathbf{x}}$ is found in 25 iterations.

Remark: $\omega = 1$, $\hat{\mathbf{x}}$ is found in 15 iterations according to Example [41.80](#).

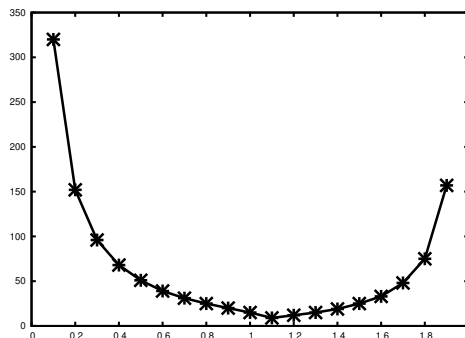
To determine the number of iterations against ω , we run the Fortran program:

```

1 program sor_eg2    ! Symmetric matrix example
2   use linearalg
3   integer, parameter :: N=4
4   real(c_double) :: b(N) = (/0,-2,-7,6/), zero(N) = (/0,0,0,0/)
5   real(c_double) :: A(N,N), x(N), w
6   integer :: cnt, info, i
7   call setmter(1000)
8   A = reshape((/3,1,0,0, 1,3,1,0, 0,1,3,1, 0,0,1,3/), (/N,N/))
9   do i=1,19
10    w = i/10d0
11    call sor(N,A,b,zero,w,x,cnt,info)
12    if (info .eq. 0) then
13      write(*,100) "w=", w, ", x = (", (x(j), j=1,N), ")"
14      write(*,'(a, i4, 1x, a)') "is found in", cnt, "iterations."
15    else
16      print *, "Error: SOR Method not converging with w=", w
17    end if
18  end do
19  100 format (a, f8.4, a, 4f10.6, a)
20 end program sor_eg2

```

and obtain the plot:



Example 41.84 (Tutorial 2, Q15). Find the first two iterations of the SOR method with $\omega = 1.1$ for the linear system

$$\begin{cases} 10x_1 + 5x_2 = 6, \\ 5x_1 + 10x_2 - 4x_3 = 25, \\ -4x_2 + 8x_3 - x_4 = -11, \\ -x_3 + 5x_4 = -11, \end{cases}$$

using $\mathbf{x}^{(0)} = \mathbf{0}$.

Solution. The best way to remember SOR method is to view it as an extrapolation between previous value and Gauss-Seidel approximation:

$$x_i^{(k+1)} = (1 - \omega)x_i^{(k)} + \frac{\omega}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} + \sum_{j=i+1}^n a_{ij}x_j^{(k)} \right]$$

$$\begin{aligned}
x_1^{(1)} &= -0.1x_1^{(0)} + 1.1 \left[\frac{6 - 5x_2^{(0)}}{10} \right] = 0.66 \\
x_2^{(1)} &= -0.1x_2^{(0)} + 1.1 \left[\frac{25 - 5x_1^{(1)} + 4x_3^{(0)}}{10} \right] = 2.387 \\
x_3^{(1)} &= -0.1x_3^{(0)} + 1.1 \left[\frac{-11 + 4x_2^{(1)} + x_4^{(0)}}{8} \right] = -0.19965 \\
x_4^{(1)} &= -0.1x_3^{(0)} + 1.1 \left[\frac{-11 + x_3^{(1)}}{5} \right] = -2.463923 \\
x_1^{(2)} &= -0.1x_1^{(1)} + 1.1 \left[\frac{6 - 5x_2^{(1)}}{10} \right] = -0.71885 \\
x_2^{(2)} &= -0.1x_2^{(1)} + 1.1 \left[\frac{25 - 5x_1^{(2)} + 4x_3^{(1)}}{10} \right] = 2.8188215 \\
x_3^{(2)} &= -0.1x_3^{(1)} + 1.1 \left[\frac{-11 + 4x_2^{(2)} + x_4^{(1)}}{8} \right] = -0.2809725875 \\
x_4^{(2)} &= -0.1x_4^{(1)} + 1.1 \left[\frac{-11 + x_3^{(2)}}{5} \right] = -2.23542166925 \\
\mathbf{x}^{(1)} &= [0.66, 2.387, -0.19965, -2.463923]^T, \\
\mathbf{x}^{(2)} &= [-0.71885, 2.8188215, -0.2809725875, -2.23542166925]^T
\end{aligned}$$

□

ir.gohyk:tut3q2

Example 41.85 (Tutorial 2, Q16). Let $A_2 = \begin{bmatrix} 164 & 47 & 5 & 18 & 26 \\ 47 & 140 & 13 & 26 & 34 \\ 5 & 13 & 156 & 39 & 47 \\ 18 & 26 & 39 & 172 & 5 \\ 26 & 34 & 47 & 5 & 148 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}$.

1. Determine if A_2 is (i) symmetric, (ii) strictly diagonally dominant, (iii) positive definite.

Solution. A_2 is SY, SDD, and PO.

□

2. What is the output of the following C++ program? What numerical method has been applied to obtain it?

```

#include <iostream>
#include <iomanip>
#include <eigen3/Eigen/Dense>

using namespace std;
const unsigned int N=5;
int main() {
    Eigen::MatrixXd A2(N,N);
    Eigen::VectorXd b(N), x(N);

    A2 << 164, 47, 5, 18, 26, 47, 140, 13, 26, 34, 5, 13, 156,
        39, 47, 18, 26, 39, 172, 5, 26, 34, 47, 5, 148;
    b << 1, 2, 3, 4, 5;
    x = A2.partialPivLu().solve(b);
    cout << "x = " << endl << x << endl;
}

```

Solution. $\mathbf{x} = [-0.00221179, 0.00308221, 0.00430663, 0.0211326, 0.0313827]$.

□

3. Explain if it is a good idea to use GEPP to solve the linear system $A_2\mathbf{x} = \mathbf{b}$ or a more specialised algorithm should be used?

Solution. Cholesky is better. GEPP can solve the problem but A_2 is symmetric positive definite, therefore, using Cholesky decomposition based algorithm will be more storage and time efficient. □

4. Explain what does the C++ program below do?

```
#include <iostream>
#include <eigen3/Eigen/Dense>

using namespace std;
const unsigned int N=5;
int main() {
    Eigen::MatrixX<double> A2(N,N);
    Eigen::VectorX<double> b(N), x(N);
    A2 << 164, 47, 5, 18, 26,
          47, 140, 13, 26, 34,
          5, 13, 156, 39, 47,
          18, 26, 39, 172, 5,
          26, 34, 47, 5, 148;
    b << 1, 2, 3, 4, 5;
    x = A2.llt().solve(b);
    cout << "Solution, x = " << endl << x << endl;
}
```

Solution. The $A_2\mathbf{x} = \mathbf{b}$ is solved by using Cholesky LL^T factorisation. □

5. Explain if iterative methods, such as Jacobi or Gauss-Seidel methods, will converge to the solution $\mathbf{x}$?

Solution. Yes. A_2 is SDD □

6. Given the implementation of Gauss-Seidel method on page [986](#). Assume that `maxIter` is 100 and $\mathbf{x}_0$ is a zero vector, if `gaussseidel(5,A2,b,x0,x)` is called in a Fortran program, explain what is the value of $\mathbf{x}$ at line 21 when `iter` has a value of 2?

Solution. $\mathbf{x} = \begin{bmatrix} -0.0035341948370575318 \\ 0.0049657219798669898 \\ 0.007.6443829726899094 \\ 0.020455200146435763 \\ 0.03.0145219716212257 \end{bmatrix}$ □

Remark 41.86. The Jacobi, Gauss-Seidel and SOR methods were used to be widely employed for solving large sparse systems [Leader, 2004]. However, the modern approach is different, the technique called *iterative refinement* or *iterative improvement* of the form $A\mathbf{e} = \mathbf{r}$, $\mathbf{e} = \mathbf{x} - \mathbf{x}_0$ is used. The initial guess $\mathbf{x}_0$ that is suitably close to the true solution by performing several iterations of Gauss-Seidel iteration to generate $\mathbf{x}_0$. Then $\mathbf{x}_k$ is calculated using Krylov space method (where GMRES, Arnoldi process are involved).

Iterative improvement is one way in which one can improve iteration method. The other way that is standard now a days is *preconditioning*, where $A\mathbf{x} = \mathbf{b}$ is transformed to an equivalent form $M_1AM_2\mathbf{y} = M_1\mathbf{b}$ such that the condition number $K(M_1AM_2) \ll K(A)$ where $K(A) = \|A\|\|A^{-1}\|$.

41.6.7 Conjugate Gradient (CG) Methods

sec:cg

The *conjugate gradient (CG) method* is a popular iterative method for solving sparse system. The name is derived from the fact that it generates a sequence of “conjugate”-orthogonal vectors. These vectors are the residuals of the iterates. They are also the gradients of a quadratic functional, the minimisation of which is equivalent to solving the linear system. CG is an extremely effective method when the coefficient matrix is *symmetric positive definite*, since storage for only a limited number of vectors is required. The array operations can be efficient for

a properly stored sparse matrix. The speed of convergence depends on the condition number of the coefficient matrix A ; if extremal eigenvalues are well-separated then superlinear convergence behaviour can result.

Assume that A is a **symmetric positive definite** matrix, i.e. $\mathbf{x}^T A^T \mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$.

Definition 41.10. The A -inner product of $\mathbf{u}$ and $\mathbf{v}$ is defined as

$$\langle \mathbf{u}, \mathbf{v} \rangle_A = (A\mathbf{u}) \cdot \mathbf{v} = \mathbf{u}^T A^T \mathbf{v}.$$

The vectors $\mathbf{u}$ and $\mathbf{v}$ are said to be A -conjugate if $\langle \mathbf{u}, \mathbf{v} \rangle_A = 0$.

Consider a convex quadratic function

$$F(\mathbf{x}) = \frac{1}{2} \langle \mathbf{x}, \mathbf{x} \rangle_A - \langle \mathbf{b}, \mathbf{x} \rangle. \quad (41.30)$$

Its gradient is $\nabla F(\mathbf{x}) = \frac{1}{2}(A^T \mathbf{x} + A\mathbf{x}) - \mathbf{b} = A\mathbf{x} - \mathbf{b}$, since A is symmetric. Therefore the linear system (41.1) is a critical point for $F(\mathbf{x})$. Thus, if A is a positive definite and symmetric, then by minimising $F(\mathbf{x})$, we can obtain a solution for $A\mathbf{x} = \mathbf{b}$.

Suppose $\mathbf{x}^* = c_1 \mathbf{p}^{(1)} + \dots + c_n \mathbf{p}^{(n)}$ is the true solution for (41.1), where $\{\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(n)}\}$ forms a mutually A -conjugate basis of $\mathbb{R}^n$. So, by definition, we can derive the coefficient

$$c_k = \frac{\langle \mathbf{p}^{(k)}, \mathbf{b} \rangle}{\langle \mathbf{p}^{(k)}, \mathbf{p}^{(k)} \rangle_A} \text{ by multiplying } (\mathbf{p}^{(k)})^T A \text{ to } \mathbf{x}^* \text{ and simplifying it.}$$

So, our task now is to find $\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(n)}$. With an inspiration Multi-variable Calculus, for an initial guess $\mathbf{x}^{(0)}$, we should choose the $\mathbf{p}^{(1)}$ to be the negative gradient of $F(\mathbf{x})$:

$$\mathbf{p}^{(1)} = -\nabla F(\mathbf{x}_0) = \mathbf{b} - A\mathbf{x}^{(0)} =: \mathbf{r}^{(0)}.$$

Using the idea from “steepest decent”, let $\mathbf{x} = \mathbf{x}^{(0)} + \alpha \mathbf{p}^{(1)}$, we find α so that (41.30) is minimum.

$$\frac{d}{d\alpha} F(\mathbf{x}^{(0)} + \alpha \mathbf{p}^{(1)}) = \frac{d}{d\alpha} \left[\frac{1}{2} ((\mathbf{x}^{(0)})^T A \mathbf{x}^{(0)} + 2\alpha (\mathbf{x}^{(0)})^T A \mathbf{p}^{(1)} + \alpha^2 (\mathbf{p}^{(1)})^T A \mathbf{p}^{(1)}) - \mathbf{b}^T \mathbf{x}^{(0)} - \alpha \mathbf{b}^T \mathbf{p}^{(1)} \right] = 0.$$

Simplifying the above gives us the first step ($k = 1$):

$$\alpha^{(1)} = \frac{\mathbf{b}^T \mathbf{p}^{(1)} - (\mathbf{x}^{(0)})^T A \mathbf{p}^{(1)}}{(\mathbf{p}^{(1)})^T A \mathbf{p}^{(1)}} = \frac{(\mathbf{r}^{(0)})^T \mathbf{p}^{(1)}}{(\mathbf{p}^{(1)})^T A \mathbf{p}^{(1)}} = \frac{(\mathbf{r}^{(0)})^T \mathbf{r}^{(0)}}{(\mathbf{p}^{(1)})^T A \mathbf{p}^{(1)}}, \quad \mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \alpha^{(1)} \mathbf{p}^{(1)}, \quad \mathbf{r}^{(1)} = \mathbf{b} - A\mathbf{x}^{(1)}.$$

If we are at the k -th ($k \geq 2$) step with $\alpha^{(k)} = \frac{(\mathbf{r}^{(k-1)})^T \mathbf{p}^{(k)}}{(\mathbf{p}^{(k)})^T A \mathbf{p}^{(k)}}$, $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \alpha^{(k)} \mathbf{p}^{(k)}$ and

$$\mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{x}^{(k)} = \mathbf{b} - A(\mathbf{x}^{(k-1)} + \alpha^{(k)} \mathbf{p}^{(k)}) \Rightarrow \boxed{\mathbf{r}^{(k)} = \mathbf{r}^{(k-1)} - \alpha^{(k)} A \mathbf{p}^{(k)}} \quad (41.31)$$

and we want to determine the $(k+1)$ -th step so that “going down the hill $F(\mathbf{x})$ ” is minimum with respect to the 2-D plane $\pi = \{\mathbf{x}^{(k)} + \xi \mathbf{r}^{(k)} + \eta \mathbf{p}^{(k)} : \xi, \eta \in \mathbb{R}\}$ at $\mathbf{x}^{(k)}$. Therefore,

$$\begin{aligned} \frac{\partial}{\partial \xi} F(\mathbf{x}^{(k)} + \eta \mathbf{p}^{(k)} + \xi \mathbf{r}^{(k)}) &= \xi (\mathbf{r}^{(k)})^T A \mathbf{r}^{(k)} + \underline{(\mathbf{x}^{(k)})^T A \mathbf{r}^{(k)}} + \eta (\mathbf{p}^{(k)})^T A \mathbf{r}^{(k)} - \underline{\mathbf{b}^T \mathbf{r}^{(k)}} \\ &= \xi (\mathbf{r}^{(k)})^T A \mathbf{r}^{(k)} + \eta (\mathbf{p}^{(k)})^T A \mathbf{r}^{(k)} - \underline{(\mathbf{r}^{(k)})^T \mathbf{r}^{(k)}} = 0 \\ \frac{\partial}{\partial \eta} F(\mathbf{x}^{(k)} + \eta \mathbf{p}^{(k)} + \xi \mathbf{r}^{(k)}) &= \eta (\mathbf{p}^{(k)})^T A \mathbf{p}^{(k)} + \underline{(\mathbf{x}^{(k)})^T A \mathbf{p}^{(k)}} + \xi (\mathbf{p}^{(k)})^T A \mathbf{r}^{(k)} - \underline{\mathbf{b}^T \mathbf{p}^{(k)}} \\ &= \eta (\mathbf{p}^{(k)})^T A \mathbf{p}^{(k)} + \xi (\mathbf{p}^{(k)})^T A \mathbf{r}^{(k)} - \underline{(\mathbf{r}^{(k)})^T \mathbf{p}^{(k)}} = 0 \quad (*) \end{aligned}$$

By calculation, we know that $(\mathbf{r}^{(1)})^T \mathbf{p}^{(1)} = 0$. We want to choose $\xi^{(k)}$ and $\eta^{(k)}$ so that $\boxed{\mathbf{r}^{(k)} \cdot \mathbf{p}^{(k)} = 0}$, $\mathbf{r}^{(k)} \cdot \mathbf{r}^{(k-1)} = 0$ and from (*) and (41.31),

$$\beta^{(k+1)} := -\frac{\eta^{(k)}}{\xi^{(k)}} = -\frac{(\mathbf{r}^{(k)})^T A \mathbf{p}^{(k)}}{(\mathbf{p}^{(k)})^T A \mathbf{p}^{(k)}} = -\frac{(\mathbf{r}^{(k)})^T \left[\frac{\mathbf{r}^{(k-1)} - \mathbf{r}^{(k)}}{\alpha^{(k)}} \right]}{(\mathbf{p}^{(k)})^T \left[\frac{(\mathbf{r}^{(k-1)} - \mathbf{r}^{(k)})}{\alpha^{(k)}} \right]} = -\frac{-(\mathbf{r}^{(k)})^T \mathbf{r}^{(k)}}{(\mathbf{p}^{(k)})^T \mathbf{r}^{(k-1)}} = \frac{(\mathbf{r}^{(k)})^T \mathbf{r}^{(k)}}{(\mathbf{r}^{(k-1)})^T \mathbf{r}^{(k-1)}} \quad (41.32)$$

$$\mathbf{p}^{(k+1)} = \mathbf{r}^{(k)} + \beta^{(k+1)} \mathbf{p}^{(k)}, \quad \alpha^{(k+1)} = \frac{(\mathbf{r}^{(k)})^T \mathbf{p}^{(k+1)}}{(\mathbf{p}^{(k+1)})^T A \mathbf{p}^{(k+1)}} = \frac{(\mathbf{r}^{(k)})^T \mathbf{r}^{(k)}}{(\mathbf{p}^{(k+1)})^T A \mathbf{p}^{(k+1)}} \quad (41.33) \quad \text{e:cg:p}$$

From the above analysis, mathematicians obtain a sequence of residuals $\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(n-1)}$ which are mutually orthogonal and a sequence of direction vectors $\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(n)}$ which are mutually A -conjugate. With these two sequences, the solution of (41.1) can be obtained and a Fortran 2003 implementation of CG method is given below.

```

1  subroutine cg(n, A, b, x0, x, cnt, info)
2      integer, intent(in) :: n
3      integer, intent(out), optional :: info, cnt
4      real(c_double), intent(in) :: A(n,n), b(n), x0(n)
5      real(c_double), intent(out) :: x(n)
6      real(c_double) :: res(n), p(n), alp, bet, re2
7      integer :: iter
8      res = b - matmul(A, x0)
9      p = res ! Steepest Descent Direction
10     do iter=1, maxIter
11         re2 = dot_product(res, res) ! To make (r_{k+1}, r_k) = 0
12         alp = re2/dot_product(matmul(A, p), p) ! Equation (2.25)
13         x = x + alp*p
14         res = res - alp*matmul(A, p) ! Equation (2.23)
15         if (dasum(n, res, 1) < accu) then
16             if (present(cnt)) cnt = iter
17             if (present(info)) info = 0
18             return
19         end if
20         bet = dot_product(res, res)/re2 ! Equation (2.24)
21         ! To make p_{k+1} conjugate A-orthogonal to p_k
22         p = res + bet*p ! Equation (2.25)
23     end do
24     if (present(info)) info = 1 ! Maximum Iteration Reach
25 end subroutine cg

```

CG method uses relatively simple matrix-vector multiplications and this makes the calculation fast. With a good theoretical convergence rate as shown in the following theorem, CG method is used in many modern large-scaled iterative methods.

Theorem 41.87 (Quarteroni and Saleri [2006]). *Let A be a symmetric positive definite matrix. The conjugate gradient method for solving (41.1) converges after at most n steps (in exact arithmetic). Moreover, the error $\mathbf{e}^{(k)}$ at the k -th iteration (with $k < n$) is orthogonal to $\mathbf{p}^{(i)}$, for $i = 1, \dots, k$ and*

$$\|\mathbf{e}^{(k)}\|_A \leq \frac{2c^k}{1+c^{2k}} \|\mathbf{e}^{(0)}\|_A, \quad c = \frac{\sqrt{K_2(A)} - 1}{\sqrt{K_2(A)} + 1} \quad (41.34) \quad \text{e:quarteroni06:5}$$

where $K_2(A) = \|A\|_2 \|A^{-1}\|_2$ is the condition number of A .

Practically, due to rounding errors, we need a proper pre-conditioning so that CG method converges quickly. However, CG method may be fast from a computer point of view, it is definitely a terrible method for hand calculation as the following example demonstrates.

Example 41.88. Use the CG method to solve $\begin{bmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 24 \\ 30 \\ -24 \end{bmatrix}$ by taking $\mathbf{x}^{(0)} = \mathbf{0}$.

Solution: $\mathbf{p}^{(0)} = \mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)} = [24, 30, -24]^T$

$$\alpha^{(1)} = \frac{(\mathbf{r}^{(0)})^T \mathbf{r}^{(0)}}{(\mathbf{p}^{(1)})^T A \mathbf{p}^{(1)}} = \frac{2052}{13968}$$

$$\begin{aligned}
\mathbf{x}^{(1)} &= \mathbf{x}^{(0)} + \alpha^{(1)} \mathbf{p}^{(1)} = \left[\frac{342}{97}, \frac{855}{194}, -\frac{342}{97} \right] \\
\mathbf{r}^{(1)} &= \mathbf{b} - A\mathbf{x}^{(1)} = \left[-\frac{645}{194}, -\frac{168}{97}, -\frac{1065}{194} \right] \\
\beta^{(2)} &= \frac{\mathbf{r}^{(1)} \mathbf{r}^{(1)}}{\mathbf{r}^{(0)} \mathbf{r}^{(0)}} = \frac{1621}{75272} \\
\mathbf{p}^{(2)} &= \mathbf{r}^{(1)} + \beta^{(2)} \mathbf{p}^{(1)} = \left[-\frac{52839}{18818}, -\frac{40869}{37636}, -\frac{113031}{18818} \right] \\
\alpha^{(2)} &= \frac{(\mathbf{r}^{(1)})^T \mathbf{r}^{(1)}}{(\mathbf{p}^{(2)})^T A \mathbf{p}^{(2)}} = \frac{314474}{1322343} \\
\mathbf{x}^{(2)} &= \mathbf{x}^{(1)} + \alpha^{(2)} \mathbf{p}^{(2)} = \left[\frac{22101}{7733}, \frac{32084}{7733}, -\frac{38311}{7733} \right] \\
\mathbf{r}^{(2)} &= \mathbf{b} - A\mathbf{x}^{(2)} = \left[-\frac{936}{7733}, -\frac{960}{7733}, -\frac{24}{703} \right] \\
\beta^{(3)} &= \frac{\mathbf{r}^{(2)} \mathbf{r}^{(2)}}{\mathbf{r}^{(1)} \mathbf{r}^{(1)}} = \frac{2408704}{3408559473} \\
\mathbf{p}^{(3)} &= \mathbf{r}^{(2)} + \beta^{(3)} \mathbf{p}^{(2)} = \left[\frac{7119432}{59799289}, -\frac{7469568}{59799289}, -\frac{2295336}{59799289} \right] \\
\alpha^{(3)} &= \frac{(\mathbf{r}^{(2)})^T \mathbf{r}^{(2)}}{(\mathbf{p}^{(3)})^T A \mathbf{p}^{(3)}} = \frac{7733}{6484} \\
\mathbf{x}^{(3)} &= \mathbf{x}^{(2)} + \alpha^{(3)} \mathbf{p}^{(3)} = [3, 4, -5]
\end{aligned}$$

With floating-point arithmetic, the $\mathbf{r}^{(3)}$ may not be exactly zero but close to zero. If the matrix is too large and the numerical error is large, then there will be more iterations but should stop eventually.

eg:cg-eg2

Example 41.89 (Symmetric Matrix). With an initial guess of zero vector, use the Fortran subroutine to solve the linear equation (41.29) in Example 41.83 and determine how many CG iterations are there to find the solution to an accuracy of $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_1 < 10^{-6}$?

Solution: The solution vector $\mathbf{x} = (-0.200000, 0.600000, -3.600000, 3.200000)$ is found in 4 iterations.

41.6.8 Fortran Module and ITPACK

Unlike the direct method, there is no widely used standard library for iterative methods. In this section, we will study how to gather numerical methods into a simple Fortran 2003 module and learn how to use a freely available Fortran 77 library ITPACK from Netlib.

Since there is no standard software for the four iterative methods mentioned above, we have implemented them as `jacobi`, `gaussseidel`, `sor` and `cg` in Fortran 2003. To gather them in a Fortran module, we need to identify the common variables and functions use by these methods:

1. The variable `maxIter` is required to indicate the maximum number of iterations allowed. It will be set to 100 since we are not solving large system. We will provide a subroutine call `setmiter` to allow the user to change the maximum iteration allowed.
2. The variable `accu` is required to determine if the difference between a vector $\mathbf{x}_k$ and its update $\mathbf{x}_{k+1}$, $\|\mathbf{x}_k - \mathbf{x}_{k+1}\|$ is small enough for us to stop. We will set `accu` to 10^{-6} to allow our solutions to have an accuracy of 6 decimal places.
3. The vector norm we used are 1-norm and 2-norm, which are available as `dasum` and `dnorm2` from BLAS respectively. We need to define an interface for them in the module.
4. We will be performing all calculations in 64-bit IEEE floating point arithmetic (refer to Topic 1) and it is defined as `c_double` in Fortran 2003 (and `real64` in Fortran 2008).

With these considerations, we can define all the variables and gather all the subroutines of the numerical methods into a Fortran module as follows.

```

module textbook_la_it
    !use iso_fortran_env, only : real64 ! Only in Fortran2008

```

```

use iso_c_binding, only : c_double
implicit none
integer, private :: maxIter=100      ! type declarations
real(c_double), private :: accu=1d-6
abstract interface                   ! type/interface definitions
    function normfn(n, dx, incx)
        use iso_c_binding, only : c_double
        integer, intent(in) :: n, incx
        real(c_double) :: dx(n), normfn
    end function normfn
end interface
procedure(normfn) :: dasum, dnorm2
contains

.....

! Naive ``Chasing`` algorithm (or Thomas algorithm) for tridiagonal
! linear algebra problem based on Crout factorisation (LU factorisation).
subroutine chasing(N,D,L,U,B,x)
    use iso_c_binding, only : c_double
    integer, intent(in) :: N
    real(c_double) :: D(N), L(N-1), U(N-1), B(N), y(N), LP(N), DP(N)
    real(c_double), intent(out) :: x(N)
    integer :: i
    ! Crout factorisation
    DP = D
    do i=2,N
        LP(i) = L(i-1)/DP(i-1)
        DP(i) = DP(i) - LP(i)*U(i-1)
    end do
    ! Forward substitution
    y(1) = B(1)
    do i=2,N
        y(i) = B(i) - LP(i)*y(i-1)
    end do
    ! Backward substitution
    x(N) = y(N)/DP(N)
    do i=N-1,1,-1
        x(i) = (y(i)-U(i)*x(i+1))/DP(i)
    end do
end subroutine chasing

subroutine setmiter(m) ! set maximum iteration allowed
    integer, intent(in) :: m
    maxIter = m
end subroutine setmiter

subroutine setaccu(a)
    integer, intent(in) :: a
    accu = a
end subroutine setaccu

end module textbook_la_it

```

The design, testing, documentation and upgrading of a software are part of a software engineering flow. For example, a more efficient Jacobi method called *Scheduled Relaxation Jacobi (SRJ)* scheme was introduced in Yang and Mittal [2014]. If SRJ were to be added into our Fortran module, then we need ways to keep track of the implementation by using version control tools and to verify our implementation by testing tools. Rouson et al. [2011] explains how to apply software engineering methodology to maintain a scientific software.

Our implementation of iterative methods are too primitive and cannot be applied in advanced engineering software. Real world iterative methods are based to sparse matrix and have much more advanced acceleration techniques and proper stopping criteria than our implementation [Hageman and Young, 1981, Barrett et al., 1994]. ITPACK is a typical real world software, which is used in the numerical solution of elliptic partial difference equation software ELLPACK.

Example 41.90 (Non-Symmetric Matrix). Store the matrix $A = \begin{bmatrix} 10 & 0 & 0 & 0 & -2 & 0 \\ 3 & 9 & 0 & 0 & 0 & 3 \\ 0 & 7 & 8 & 7 & 0 & 0 \\ 3 & 0 & 8 & 7 & 5 & 0 \\ 0 & 8 & 0 & 9 & 9 & 13 \\ 0 & 4 & 0 & 0 & 2 & -1 \end{bmatrix}$ in

CSR (Fortran) format. Notice that the non-symmetric storage is slightly different from the symmetric case.

Solution: The major difference between CSR Fortran and CSR C format is the C index begins from 0 instead of 1:

Nonzero data: $A = [10, -2, 3, 9, 3, 7, 8, 7, 3, 8, 7, 5, 8, 9, 9, 13, 4, 2, -1]$

Column indexing: $JA = [1, 5, 1, 2, 6, 2, 3, 4, 1, 3, 4, 5, 2, 4, 5, 6, 2, 5, 6]$

Position indexing: $IA = [1, 3, 6, 9, 13, 17, 20]$ $\text{size}(IA) = n + 1 = 7$

Let nz be “non-zero elements” Note: First row has $3 - 1 = 2$ nz ; second row has $6 - 3 = 3$ nz ; third row has $9 - 6 = 3$ nz ; 4th row has $13 - 9 = 4$ nz ; 5th row has $17 - 13 = 4$ nz ; 6th row has $20 - 17 = 3$ nz .

We fill up A **by row** instead of column in Compressed sparse column (CSC or CCS or Harwell-Boeing) format.

eg:itpack

Example 41.91 (Symmetric Matrix). An engineer is reading a Fortran 95 program which calls the functions from ITPACK library. By studying the user manual of ITPACK, explain what is the purpose of this program and what is the output of this program.

```

1 program mytst_sor
2   implicit none
3   integer, parameter :: dp=selected_real_kind(15, 307) ! double precision
4   integer, parameter :: N=4, IAN=5, JAN=7, NW=2*N
5   ! Matrix is symmetric & diagonally dominant
6   real(kind=dp) :: A(JAN), RHS(N), U(N), RPARM(12), WKSP(NW)
7   integer :: IA(IAN), JA(JAN), IPARM(12), IER, IWKSP(NW)
8
9   A = (/3,1, 3,1, 3,1, 3/)
10  JA = (/1,2, 2,3, 3,4, 4/)
11  IA = (/1,3,5,7,8/)
12  RHS = (/0,-2,-7,6/)
13  call vfill(N,U,0.D0)
14  call dfaulst(IPARM,RPARM)
15  write (*,"(a,e12.7)") "tol=", rparm(8)
16  iparm(6) = 0 ! ITPACK Manual: textbook methods are 'nonadaptive methods'
17  iparm(7) = 2
18  call jsr(N,IA,JA,A,RHS,U,IWKSP,NW,WKSP,IPARM,RPARM,IER)
19  if (IER .eq. 0) then
20    write(*,100) "Jacobi: x =", U, IPARM(1) ! Jacobi Semi-Iteration
21  else
22    print *, "Error in ITPACK JSI"
23  end if
24  call vfill(N,U,0.D0)
25  call dfaulst(IPARM,RPARM)
26  iparm(6) = 0
27  call sor(N,IA,JA,A,RHS,U,IWKSP,NW,WKSP,IPARM,RPARM,IER)
28  if (IER .eq. 0) then
29    write(*,100) "Gauss-Seidel: x =", U, IPARM(1) ! G-S == SOR(Omega=1)
30  else
31    print *, "Error in ITPACK SOR"
32  end if
33  call vfill(N,U,0.D0)
34  call dfaulst(IPARM,RPARM)
35  iparm(6) = 0
36  rparm(5)=1.1 ! SOR(Omega=1.1)
37  call sor(N,IA,JA,A,RHS,U,IWKSP,NW,WKSP,IPARM,RPARM,IER)

```

```

38     if (IER .eq. 0) then
39         write(*,100) "SOR(Omega=1.1): x =", U, IPARM(1)
40     else
41         print *, "Error in ITPACK SOR"
42     end if
43 100 format(a20,4f10.6," in ",i3," iterations")
44 end program mytst_sor

```

Solution: By looking at the variables A, JA and IA from line 9 to line 11, the comment at line 5 and by looking up “sparse matrix” from http://en.wikipedia.org/wiki/Sparse_matrix, we should recognise that it as the *compressed row storage (CRS)* or *Yale Sparse Matrix* format. The subroutine dfaulst initialises IPARM(5) to 0 which indicates that the matrix is symmetric. Hence, A, JA and IA represents the 4×4 matrix of 16 elements (compare to the CRS with $7+7+5 = 19$ elements. When the matrix is much larger, e.g. $n = 100$, then $100^2 > 2 \times \#(\text{non-zero elements}) + 100 + 1$, the CRS format saves a lot of memory):

$$A = \begin{bmatrix} 3 & 1 & & \\ * & 3 & 1 & \\ & * & 3 & 1 \\ & & * & 3 \end{bmatrix}$$

With an initial guess of zero vector (line 13) and $\mathbf{b} = \text{RHS}$ the program solve the following linear equation using Jacobi (jsi with special parameters), Gauss-Seidel (SOR with $\omega = 1$) and SOR($\omega = 1.1$):

$$A\mathbf{x} = \begin{bmatrix} 0 \\ -2 \\ -7 \\ 6 \end{bmatrix}.$$

Running the program generates the following output:

```

tol=.7110000E-12
      Jacobi: x = -0.199991 0.599978 -3.599985 3.199987 in 19 iterations
      Gauss-Seidel: x = -0.199997 0.599997 -3.599998 3.199999 in 11 iterations
      SOR(Omega=1.1): x = -0.200000 0.600000 -3.600000 3.200000 in 7 iterations

```

41.7 Eigenvalue Problems

Eigenvalue problem (41.3) is such a difficult problem that either one whole chapter (e.g. Isaacson and Keller [1966, Chapter 4]) or books (e.g. Wilkinson [1988], Stewart [2001]) are dedicated to it. There is no direct method for solving eigenvalue problem (41.3). We learned to solve eigenvalue problem by transforming an eigenvalue problem (41.3) to a polynomial equation called the *characteristic equation*:

$$|A - \lambda I| = 0. \quad (41.35)$$

However, mathematicians have proven that there is no general way of solving a polynomial equation by algebraic manipulation (Galois theory). Since eigenvalue problem (41.3) is equivalent to a polynomial equation (41.35), eigenvalue problem is theoretically impossible to solve with algebraic manipulation. Therefore, iterative method may be the only method to find all eigenvalues.

There are two basic iterative methods for finding eigenvalues, the power methods (Section 41.7.1) and the QR algorithm (Section 41.7.3).

41.7.1 Power Methods

sec:powerm

The power method (or its variations) is used to find the “dominant” eigenvalue of a square matrix A . Before proceeding to the description of this method, we review a little bit theories that characterise the eigenvalues of A from UECM1303 Fundamentals of Linear Algebra.

Theorem 41.92. *Let f be any polynomial. If λ is eigenvalue of A , then*

1. λ^k is an eigenvalue of A^k where k is a positive integer.
2. $f(\lambda)$ is an eigenvalue of $f(A)$.
3. If A is nonsingular, then $\frac{1}{\lambda}$ and $f(\frac{1}{\lambda})$ are the eigenvalues of A^{-1} and $f(A^{-1})$ respectively.

If matrix A has different eigenvalues, then it is possible to diagonalise A . However, A may have the same eigenvalues and this may cause the power methods to fail. Mathematicians had identified the culprit as the “multiplicity” defined below.

Definition 41.11. The *algebraic multiplicity* of an eigenvalue is the multiplicity of the corresponding root of the characteristic polynomial.

Definition 41.12. The *geometric multiplicity* of an eigenvalue λ_i is the dimension of the associated eigenspace $\{\mathbf{x} : A\mathbf{x} = \lambda_i\mathbf{x}\}$.

The “structure” theorem of square matrices below characterises the multiplicity.

thm:jordan

Theorem 41.93 (Should be thought in the AM subject Linear Algebra). *Any matrix $A \in M_n(\mathbb{R})$ can be put in Jordan canonical form (or Jordan normal form) by a similarity transformation, i.e.*

$$T^{-1}AT = J = \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_q \end{bmatrix}$$

where

$$J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix}, \quad \lambda_i \in \mathbb{C}$$

is called a Jordan block of size n_i with eigenvalue λ_i and $\sum_{i=1}^q n_i = n$.

The value n_i and the matrix J_i are related to the geometric multiplicity. This theorem says that if a matrix has eigenvalues which appears $n_i > 1$ times for some i , then the matrix A may not be similar to a diagonal matrix.

According to Theorem 41.93, it is easy to find examples which A^k as $k \rightarrow \infty$ does not provide us with information of the eigenvalue of A . However, when there is a dominant eigenvalue, defined below, A^k will gives us the information about the dominant eigenvalue of A .

Definition 41.13. Given a matrix A , the *dominant eigenvalue* of A is the eigenvalue that is largest in absolute value than all of the other eigenvalues. The corresponding eigenvalues is called the *dominant eigenvector*.

Let us look at the following example to get ourselves familiar with the definition above.

Example 41.94. Given that matrix A has eigenvalues -4 , -3 , 1 and 3 and matrix B has eigenvalues -4 , -3 , 3 and 4 . Find the dominant eigenvalue for A and B .

Solution:

1. Since $|-4|$ is the largest absolute value for A , -4 is the dominant eigenvalue.
2. Since there are two eigenvalues with the same absolute value for B , there is **no** dominant eigenvalue.

Now, let us state a theorem which tells how A^k can tell us the information about the dominant eigenvalue of A .

Theorem 41.95 (Rate of Convergence of Power Method). *Let A be a diagonalisable $n \times n$ matrix with eigenvalues λ_i such that*

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_{n-1}| \geq |\lambda_n|.$$

Assume that $\mathbf{v}_1, \dots, \mathbf{v}_n$ are unit eigenvectors of A associated with $\lambda_1, \lambda_2, \dots, \lambda_n$ respectively. Let $\mathbf{x}_0$ be a nonzero vector in $\mathbb{R}^n$ that is an initial guess for the dominant eigenvector $\mathbf{v}_1$. Then the vector $\mathbf{x}_k = \frac{1}{\max_i |(\mathbf{y}_k)_i|} \mathbf{y}_k$ is a good approximation to a dominant vector of $\mathbf{v}_1$ when k is sufficiently large. Here, $\mathbf{y}_k = A\mathbf{x}_{k-1}$. (Expanding out, we have a sequence $A\mathbf{x}_0, A^2\mathbf{x}_0, \dots, A^k\mathbf{x}_0, \dots$)

1. The dominant eigenvalue λ_1 can be approximated by the Rayleigh quotient

$$\lambda_1 \approx \frac{A\mathbf{x}_k \cdot \mathbf{x}_k}{\mathbf{x}_k \cdot \mathbf{x}_k}.$$

2. The rate of convergence depends on the ratios $\frac{\lambda_2}{\lambda_1}, \dots, \frac{\lambda_n}{\lambda_1}$. The smaller the ratios, the faster the rate of convergence.

Remark 41.96. If A is not diagonalisable, then A^k may not converge.

The value λ_1^k will be very large for large k . So we need to perform a scaling in order to apply the above theorem. An algorithm to find the dominant eigenvalue involving the power of A and scaling is called the *(scaled) power method*. The procedure is given below.

1. Let $\mathbf{x}_0$ be any initial vector in $\mathbb{R}^n$ whose largest component is 1.
2. Repeat the following steps for $k = 1, 2, \dots$:
 - (a) Compute $\mathbf{y}_k = A\mathbf{x}_{k-1}$.
 - (b) Let m_k be the component of $\mathbf{y}_k$ with the largest absolute value.
 - (c) Set $\mathbf{x}_k = \frac{1}{m_k} \mathbf{y}_k$.

For most choices of the initial vector $\mathbf{x}_0$, $m_k \rightarrow \lambda_1$ and $\mathbf{x}_k$ converges to a dominant eigenvector.

eg:3.5.2

Example 41.97. Approximate the dominant eigenvector of $A = \begin{bmatrix} 1 & 1 \\ 2 & 0 \end{bmatrix}$ using the (scaled)

power method by taking $\mathbf{x}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ as the initial vector. Perform the power method for 7 steps.

Solution:

k	0	1	2	3	4	5	6	7	8
$\mathbf{y}_k$	—	$\begin{bmatrix} 1 \\ 2 \end{bmatrix}$	$\begin{bmatrix} 1.5 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1.67 \\ 2 \end{bmatrix}$	$\begin{bmatrix} 1.83 \\ 1.67 \end{bmatrix}$	$\begin{bmatrix} 1.91 \\ 2 \end{bmatrix}$	$\begin{bmatrix} 1.95 \\ 1.91 \end{bmatrix}$	$\begin{bmatrix} 1.98 \\ 2 \end{bmatrix}$	$\begin{bmatrix} 1.99 \\ 1.98 \end{bmatrix}$
$\mathbf{x}_k$	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0.5 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0.67 \end{bmatrix}$	$\begin{bmatrix} 0.83 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0.91 \end{bmatrix}$	$\begin{bmatrix} 0.95 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0.98 \end{bmatrix}$	$\begin{bmatrix} 0.99 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0.99 \end{bmatrix}$
m_k	1	2	1.5	2	1.83	2	1.95	2	1.99

eg:powerm2

Example 41.98. Let $A = \begin{bmatrix} 4 & -2 \\ 3 & -1 \end{bmatrix}$. Use the power method to approximate the dominant eigenvalue and a corresponding eigenvector. **[Remark:** Without explicitly stating $\mathbf{x}_0$, we will always take $\mathbf{x}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.]

Solution. Let $\mathbf{x}_0 = [1 \ 0]^T$.

$$\begin{aligned} \mathbf{y}_1 = A\mathbf{x}_0 &= \begin{bmatrix} 4 & -2 \\ 3 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 4 \\ 3 \end{bmatrix}, & \mathbf{x}_1 &= \frac{1}{\max\{4, 3\}} \mathbf{y}_1 = \begin{bmatrix} 1 \\ 0.75 \end{bmatrix} \\ \mathbf{y}_2 = A\mathbf{x}_1 &= \begin{bmatrix} 2.5 \\ 2.25 \end{bmatrix}, & \mathbf{x}_2 &= \frac{1}{2.5} \mathbf{y}_2 = \begin{bmatrix} 1.0 \\ 0.9 \end{bmatrix} \\ \mathbf{y}_3 = A\mathbf{x}_2 &= \begin{bmatrix} 2.2 \\ 2.1 \end{bmatrix}, & \mathbf{x}_3 &= \frac{1}{2.2} \mathbf{y}_3 = \begin{bmatrix} 1.0 \\ 0.954545 \end{bmatrix} \\ \mathbf{y}_4 = A\mathbf{x}_3 &= \begin{bmatrix} 2.090909 \\ 2.045454 \end{bmatrix}, & \mathbf{x}_4 &= \frac{1}{2.090909} \mathbf{y}_4 = \begin{bmatrix} 1 \\ 0.978261 \end{bmatrix} \\ \mathbf{y}_5 = A\mathbf{x}_4 &= \begin{bmatrix} 2.043478 \\ 2.021739 \end{bmatrix}, & \mathbf{x}_5 &= \begin{bmatrix} 1. \\ 0.989362 \end{bmatrix} \\ \mathbf{y}_6 = A\mathbf{x}_5 &= \begin{bmatrix} 2.021277 \\ 2.010638 \end{bmatrix}, & \mathbf{x}_6 &= \begin{bmatrix} 1 \\ 0.994737 \end{bmatrix} \\ \mathbf{y}_7 = A\mathbf{x}_6 &= \begin{bmatrix} 2.010526 \\ 2.005263 \end{bmatrix}, & \mathbf{x}_7 &= \begin{bmatrix} 1 \\ 0.997382 \end{bmatrix} \end{aligned}$$

From $\mathbf{y}_7$, we can estimate that $\lambda_1 \approx 2$. □

In the following examples, we will investigate the possibilities for power method to fail.

eg:powerm4

Example 41.99. Write a script to find the dominant eigenvalue using power method. Apply it to the matrix $A = \begin{bmatrix} 2 & 1 & 6 \\ 0 & 2 & 5 \\ 0 & 0 & 2 \end{bmatrix}$ with (a) $\mathbf{x}_0 = [1, 0, 0]^T$ and (b) $\mathbf{x}_0 = [2, 1, 0]^T$. Comment on the outcome. Does power method succeed in finding the dominant eigenvalue?

Solution: A sample Python script to calculate using power method is listed below.

```

1 import numpy as np
2
3 def dominant(a,b): return (a if (abs(a)>abs(b)) else b)
4
5 def scaled_power_method(A,x0=None,maxiter=10,shift=0.0,verbose=False):
6     if x0 is None:
7         x = np.zeros(A.shape[1],dtype=np.double)
8         x[0] = m = 1.0
9     elif A.shape[1] != len(x0):
10         print("Error: Matrix A and size x0 mismatch")
11         return

```

```

12     else:
13         m = dominant(np.min(x0), np.max(x0))
14         x = np.array(x0)/m
15     if shift != 0.0: A = A - np.diag(np.full(A.shape[0], shift))
16     if verbose: print("Step 0: y=%s, m=%s, x=%s" % (None, m, x))
17     for i in range(1, maxiter+1):
18         y = A @ x
19         m = dominant(y.min(), y.max())
20         x = y.copy()/m
21         if verbose: print("Step %d: y=%s, m=%s, x=%s" % (i, y, m, x))

```

(a) When $\mathbf{x}_0 = [1, 0, 0]^T$, we will get $m_1 = m_2 = 2$. Since $\mathbf{x}_0$ is an eigenvector of A , we are not getting the dominant eigenvalue and eigenvector. Power method failed in this case.

(b) When $\mathbf{x}_0 = [2, 1, 0]^T$, we will get $m_{100} = 2.019417$ (vs $\lambda_1 = 2$) which indicates that power method sometimes converges very slowly to the dominant eigenvalue.

Example 41.100 (Tutorial 2, Q17. May 2012 Final Exam Question 3(b)). Consider the matrix

$$\begin{bmatrix} 4 & 2 & 1 \\ 0 & 3 & 2 \\ 1 & 1 & 4 \end{bmatrix}.$$

1. Apply three iterations of the power method to approximate the dominant eigenvalue of the above matrix. Use $(1, 1, 1)^T$ as the initial guess of the eigenvector.

$$\begin{aligned}
 \text{Solution. } \mathbf{x}^{(1)} &= \frac{1}{m^{(1)}} A\mathbf{x}^{(0)} = \frac{1}{7} \begin{bmatrix} 7 \\ 5 \\ 6 \end{bmatrix} = \begin{bmatrix} 1. \\ 0.71428571 \\ 0.85714286 \end{bmatrix} \\
 \mathbf{x}^{(2)} &= \frac{1}{m^{(2)}} A\mathbf{x}^{(1)} = \frac{1}{6.28571429} \begin{bmatrix} 6.28571429 \\ 3.85714286 \\ 5.14285714 \end{bmatrix} = \begin{bmatrix} 1. \\ 0.61363636 \\ 0.81818182 \end{bmatrix} \\
 \mathbf{x}^{(3)} &= \frac{1}{m^{(3)}} A\mathbf{x}^{(2)} = \frac{1}{6.04545454} \begin{bmatrix} 6.04545454 \\ 3.47727273 \\ 4.88636363 \end{bmatrix} = \begin{bmatrix} 1. \\ 0.57518797 \\ 0.80827068 \end{bmatrix}.
 \end{aligned}$$

□

2. Compute a better approximation for λ_1 with Aitken's Δ^2 method $q_0 = m^{(1)} - \frac{(m^{(2)} - m^{(1)})^2}{m^{(3)} - 2m^{(2)} + m^{(1)}}$.

$$\begin{aligned}
 \text{Solution. } q_0 &= m^{(1)} - \frac{(m^{(2)} - m^{(1)})^2}{m^{(3)} - 2m^{(2)} + m^{(1)}} = 7 - \frac{(6.28571429 - 7)^2}{6.04545454 - 2(6.28571429) + 7} \\
 &= 5.92367904
 \end{aligned}$$

Note: It is very close the dominant eigenvalue 5.91963956583942.

□

There are situations in which power method will fail regardless of the initial vector we choose.

Example 41.101. Apply power method to $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ for any initial vector $\mathbf{x}_0 \in \mathbb{R}^2$ and explain if power method works.

$$\begin{aligned}
 \text{Solution. Let } \mathbf{x}_0 &= \begin{bmatrix} a \\ b \end{bmatrix} \text{ and assume } a > b \text{ without loss of generality. Then } A\mathbf{x}_0 = \begin{bmatrix} b \\ -a \end{bmatrix} \Rightarrow \\
 m_1 &= -a, \mathbf{x}_1 = \begin{bmatrix} -b/a \\ 1 \end{bmatrix} \Rightarrow A\mathbf{x}_1 = \begin{bmatrix} 1 \\ b/a \end{bmatrix} \Rightarrow m_2 = 1, \mathbf{x}_2 = \begin{bmatrix} 1 \\ b/a \end{bmatrix} \Rightarrow A\mathbf{x}_2 = \begin{bmatrix} b/a \\ -1 \end{bmatrix}, m_3 =
 \end{aligned}$$

$$-1, \mathbf{x}_3 = \begin{bmatrix} -b/a \\ 1 \end{bmatrix}, \dots$$

In general, we will have $m_i = \begin{cases} -1, & i \text{ is odd}, i > 1 \\ 1, & i \text{ is even}, i > 1. \end{cases}$

Power method never converges and fails. $\square$

What if we are interested in finding other eigenvalues instead of the dominant eigenvalue? Then we need the *shifted power method*. It uses this property: If λ is an eigenvalue of A , then $\lambda - \alpha$ is an eigenvalue of $A - \alpha I$ for any scalar α . Therefore, if λ_1 is the dominant eigenvalue of A , the eigenvalues of $A - \lambda_1 I$ will be $0, \lambda_2 - \lambda_1, \lambda_3 - \lambda_1, \dots, \lambda_n - \lambda_1$. We can apply the power method to compute $\lambda_2 - \lambda_1$, and from this value we can find λ_2 . Repeating this process will allow to compute all the eigenvalues. However, this is impractical and in practise, QR algorithm (Section 41.7.3) is much better numerical method.

eg:3.5.2b

Example 41.102. Use the shifted power method to compute the second eigenvalue of the matrix $A = \begin{bmatrix} 1 & 1 \\ 2 & 0 \end{bmatrix}$ from Example 41.97.

Solution: From Example 41.97, we know that $\lambda_1 = 2$. We therefore calculate the dominant eigenvalue of

$$B = A - 2I = \begin{bmatrix} -1 & 1 \\ 2 & -2 \end{bmatrix}.$$

Surprisingly,

$$\begin{aligned} \mathbf{x}_1 = B\mathbf{x}_0 = \begin{bmatrix} -1 \\ 2 \end{bmatrix} &\Rightarrow m_1 = 2, \quad \mathbf{y}_1 = \frac{1}{2}\mathbf{y}_0 = \begin{bmatrix} -\frac{1}{2} \\ 1 \end{bmatrix} \\ \mathbf{x}_2 = B\mathbf{x}_1 = \begin{bmatrix} 1.5 \\ -3 \end{bmatrix} &\Rightarrow m_2 = -3, \quad \mathbf{y}_2 = \frac{1}{-3}\mathbf{y}_1 = \begin{bmatrix} -\frac{1}{2} \\ 1 \end{bmatrix}. \end{aligned}$$

Since $\mathbf{y}_2 - \mathbf{y}_1$ is a zero vector, we can stop! The eigenvalue is $\lambda_2 - 2 = -3 \Rightarrow \lambda_2 = -1$.

Let us investigate the application of power method to A^{-1} . According to Theorem 41.92, if matrix A is invertible with eigenvalues λ , then A^{-1} has eigenvalues $\frac{1}{\lambda}$. So when we apply the power method to A^{-1} , its dominant eigenvalue will be the reciprocal of the smallest (in absolute value) eigenvalue of A . Combining “inverse” and “power method”, we obtain the *inverse power method*. In practise, we would be using the *shifted inverse power method*, which is a combination of “inverse” and “shifted power method”, in which after estimating an eigenvalue α , we can use the *shifted inverse power method* to find the eigenvalue λ and the eigenvector of A that is closest to α .

If λ is an eigenvalue of A and $\alpha \neq \lambda$, then $A - \alpha I$ is invertible if α is not an eigenvalue of A and $\frac{1}{\lambda - \alpha}$ is an eigenvalue of $(A - \alpha I)^{-1}$. If α is very close to λ , then $\frac{1}{\lambda - \alpha}$ will be much bigger in absolute value than the next eigenvalue, so the convergence will be very rapid.

Example 41.103. Write a computer script for the shifted inverse power method and use it to

approximate the eigenvalue of $A = \begin{bmatrix} 4 & 1 & 1 \\ 0 & 2 & 1 \\ -2 & 0 & 9 \end{bmatrix}$ that is closest to 7.

Solution: We are using $(A - 7I)^{-1}$ in our calculation. It is not really practical to calculate by hand. Rewriting the Python script in Example 41.99:

```
1 def shift_inv_power_method(A, x0=None, maxiter=10, shift=0.0, verbose=False):
2     import scipy.linalg as linalg
3     if x0 is None:
```

```

4         x = np.zeros(A.shape[1], dtype=np.double)
5         x[0] = m = 1.0
6     elif A.shape[1] != len(x0):
7         print("Error: Matrix A and size x0 mismatch")
8         return
9     else:
10        m = dominant(np.min(x0), np.max(x0))
11        x = np.array(x0)/m
12    B = A - np.diag(np.full(A.shape[0], shift))
13    if verbose: print("Step 0: y=%s, m=%s, x=%s" % (None, m, x))
14    for i in range(1, maxiter+1):
15        y = linalg.solve(B, x)
16        m = dominant(y.min(), y.max())
17        x = y.copy()/m
18        if verbose: print("Step %d: y=%s, m=%s, x=%s" % (i, y, m, x))
19    return (m, x)

```

By running the script by trial and error, we discovered that it will take us approximately 35 steps to get to an accuracy of $< 10^{-6}$:

- Step 1: $y = [-0.55555556, -0.11111111, -0.55555556], m = -0.555556, x = [1., 0.2, 1.]$
- Step 2: $y = [-0.24444444, 0.01111111, 0.25555556], m = 0.255556, x = [-0.95652174, 0.04347826, 1.]$
- Step 3: $y = [0.85990338, 0.26328502, 1.35990338], m = 1.359903, x = [0.63232682, 0.19360568, 1.]$
- Step 4: $y = [-0.03947109, 0.05338465, 0.46052891], m = 0.460529, x = [-0.08570816, 0.11592029, 1.]$
- ...
- Step 32: $y = [0.17324167, 0.10380957, 0.67324167], m = 0.673242, x = [0.25732464, 0.15419362, 1.]$
- Step 33: $y = [0.17324258, 0.10380979, 0.67324258], m = 0.673243, x = [0.25732564, 0.15419374, 1.]$
- Step 34: $y = [0.17324201, 0.10380965, 0.67324201], m = 0.673242, x = [0.25732501, 0.15419367, 1.]$
- Step 35: $y = [0.17324236, 0.10380974, 0.67324236], m = 0.673242, x = [0.25732541, 0.15419371, 1.]$

So, $\frac{1}{\lambda - 7} = 0.673242 \Rightarrow \lambda = 8.4853$.

Google is the company that make full use of power method. Google claims to index more than 25 billion pages. In fact, the word “Google” originated from the number “googol”, which means 10^{100} . Roughly 95% of the text in web pages is composed from a mere 10000 words. So for most searches, there will be a huge number of pages containing the words in the search phrase. How can we rank the pages so that the most important pages are at the top of the list?

One idea to solve this problem is proposed by the creators of PageRank algorithm — Sergey Brin and Lawrence Page¹: the importance of a page $I(P)$ should be judged by the number of pages linking to it as well as their importance.

Suppose that page P_j has ℓ_j links. If one of those links is to page P_i , then P_j will pass on $1/\ell_j$ of its importance to P_i . The importance ranking of P_i is then the sum of all contributions made by pages linking to it. That is, if we denote the set of pages linking to P_i by B_i , then

$$I(P_i) = \sum_{P_j \in B_i} \frac{I(P_j)}{\ell_j}. \quad (41.36) \quad \text{e:pagerank1}$$

Expressing (41.36) in matrix form, we have

$$\mathbf{r} = H\mathbf{r}, \quad H = [H_{ij}], \quad H_{ij} = \begin{cases} 1/\ell_j, & \text{if } P_j \in B_i, \\ 0, & \text{otherwise.} \end{cases} \quad (41.37) \quad \text{e:pagerank2}$$

The *hyperlink matrix* H is a stochastic matrix² and $\mathbf{r} = [I(P_i)]$ is the importance ranking of all web pages.

In the modern collective intelligence terminology, a PageRank $I(P_i)$ is “vote” or “like”, by all other pages on the Web, about how important a page is. A link to a page counts as a vote of support. If there is no link, there is no support.

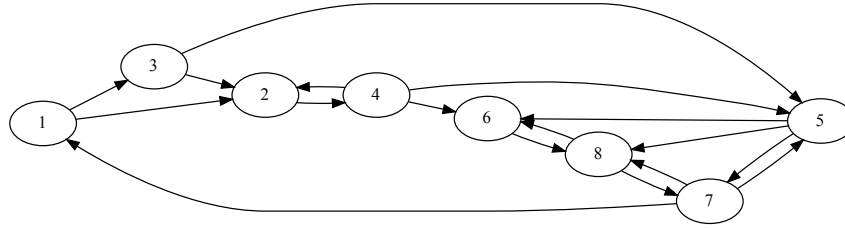
¹Their paper “The Anatomy of a Large-Scale Hypertextual Web Search Engine” can be found on the Internet

²each column of the stochastic matrix sums to 1

Equation (41.37) can be view as an eigenvalue problem with $\lambda = 1$ in (41.3). Mathematically, it is found that for a “good” stochastic matrix, power method is applicable to find the dominant “PageRank” eigenvector $\mathbf{r}$.

feature-google

Example 41.104 (AMS Feature Article: “How Google Finds Your Needle in the Web’s Haystack”). Write down the hyperlink matrix and find a PageRank for the following network:



Solution: The hyperlink matrix is

$$H = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{matrix} & \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 1 & \frac{1}{3} & 0 \end{pmatrix} \end{matrix}$$

Solving $H\mathbf{r} = \mathbf{r}$ by using (scaled) power method, at step 70, we obtain

$$\mathbf{x} = \begin{bmatrix} 0.20340011 \\ 0.22881661 \\ 0.10169419 \\ 0.22882173 \\ 0.33052252 \\ 0.68644168 \\ 0.61016973 \\ 1.00004437 \end{bmatrix}, \quad m = 1.00004437346, \quad \mathbf{y} = \begin{bmatrix} 0.20339109 \\ 0.22880645 \\ 0.10168968 \\ 0.22881157 \\ 0.33050786 \\ 0.68641122 \\ 0.61014265 \\ 1. \end{bmatrix}$$

By scaling $\mathbf{y}$, we obtain the PageRank eigenvector:

$$\mathbf{r} = \frac{1}{\sum_{i=1}^8 y_i} \mathbf{y} = \begin{bmatrix} 0.06000161 \\ 0.0674993 \\ 0.02999908 \\ 0.06750081 \\ 0.09750183 \\ 0.20249549 \\ 0.1799958 \\ 0.29500609 \end{bmatrix}.$$

More sophisticated theories are required to guaranteed the power method works in general since power method can fail with matrices which does not have dominant eigenvector as demonstrated below.

Example 41.105 (Kurt Bryan, Tanya Leise). Consider the hyperlink matrix

$$H = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.5 \\ 0 & 0 & 1 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Let $c_1, c_2 \in \mathbb{R}$, $\mathbf{x} = \begin{bmatrix} 1/2 \\ 1/2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} 0 \\ 0 \\ 1/2 \\ 1/2 \\ 0 \end{bmatrix}$. Show that $c_1\mathbf{x} + c_2\mathbf{y}$ is the eigenvector of H and

determine whether a unique PageRank exists for H .

Solution: By calculation, it is easy to show that

$$H(c_1\mathbf{x} + c_2\mathbf{y}) = c_1\mathbf{x} + c_2\mathbf{y}, \quad c_1, c_2 \in \mathbb{R}.$$

Therefore $c_1\mathbf{x} + c_2\mathbf{y}$ are all eigenvectors of H . Since $\mathbf{x}$ and $\mathbf{y}$ are eigenvectors of H , the PageRank is not unique.

Remark 41.106. In fact, Google proposed the use of the following iteration in their paper to make sure the “power method” converges: $\mathbf{r} = (0.85H + 0.15 \begin{bmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \ddots & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix})\mathbf{r}$ where n is number of web pages.

Google’s Pagerank calculation using the power method can take several days. Several enhancements of the iteration procedure have been proposed by academics:

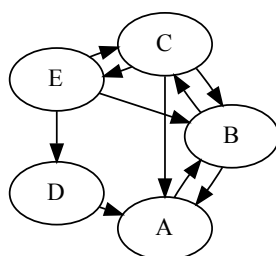
1. an adaptive method that checks the convergence of the components of the Pagerank vector and avoids performing the power iteration for those components;
2. Using the block structure of the web for speedup;
3. An acceleration method based on Aitken extrapolation (Topic 3).

41.7.2 Google PageRank and Markov Chain

We will continue to study Google’s *PageRank algorithm* (Example 41.104), one of the most famous application of eigenvalue and eigenvector, which is described in the article “The \$25,000,000,000 Eigenvector The Linear Algebra Behind Google”.

To explain the PageRank algorithm, we model the web as an oriented graph. The *vertices* are the *pages*, and the *oriented edges* are the *links between pages*. Each page corresponds to a different url. Hence, a website may contain many pages. The model makes no difference between the individual pages of a website and its front page. But, most likely, the algorithm will rank better the front page of an important website.

Let us look at the simple web with five pages named A, B, C, D, and E.



This web has few links. If we are on page A, then there is only one link to page B, while, if we are on page C, we find three links, and we can choose to move to either page A, or B, or E. Notice that there is at least one link from each page.

If we walk randomly on the oriented graph, starting from a page, at each step we choose at random a link from the page where we are, and we follow it. For instance, if we start on page B, then we can go to A or to C with probability $1/2$ for each case while, if we start on D, then we necessarily go to A with probability 1. Where will we be after n steps? This can be characterised by the left transition matrix (this is the transpose of the right transition matrix mentioned in Topic 2 and Probabilistic State Machine from Discrete Mathematics)

$$P = \begin{matrix} & \begin{matrix} A & B & C & D & E \end{matrix} \\ \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{3} & 1 & 0 \\ 1 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} & 0 & 0 \end{pmatrix} & \begin{matrix} A \\ B \\ C \\ D \\ E \end{matrix} \end{matrix} = (p_{ij}). \quad (41.38) \quad \text{e:pagerank}$$

If we consider a random variable X_n with values in the set of pages $\{A, B, C, D, E\}$, which contains $N = 5$ pages. X_n represents the page where we are after n steps of the random walk. Hence, p_{ij} , the entry of the matrix P located on the i -th row and j -th column, is the conditional probability that we are on the i -th page at step $n + 1$, given that we were on the j -th page at step n :

$$p_{ij} = \text{Prob}(X_{n+1} = i \mid X_n = j).$$

Note that the sum of entries of any column of P is equal to 1 and that all entries are greater or equal to zero. Matrices having these two properties are very special: each such matrix is the matrix of a *Markov chain process*, also called *Markov transition matrix*. It always has 1 as an eigenvalue and there exists an eigenvector with eigenvalue 1, all the components of which are less than or equal to 1 and greater than or equal to 0, with sum equal to 1.

As mentioned in Discrete Mathematics, the entry $(p_{ij}^{(m)})$ of the matrix P^m describes the probability $\text{Prob}(X_{n+m} = i \mid X_n = j)$. When strange conditions do not happen, after n (which is sufficiently large) steps, the probability of being on a page is independent from where we started! In fact the eigenvector π corresponding to the eigenvalue $\lambda = 1$ is the *stationary vector* of P :

$$P\pi = \pi.$$

The following deep theorem of Frobenius guarantees that for the matrix associated to a web graph, we will always find a stationary solution.

Theorem 41.107 (Frobenius Theorem). *Consider an $N \times N$ Markov transition matrix $P = (p_{ij})$ (i.e. $p_{ij} \in [0, 1]$ for all i, j , and the sum of entries on each column is equal to 1, namely $\sum_{i=1}^N p_{ij} = 1$). Then*

1. $\lambda = 1$ is one eigenvalue of P .
2. Any eigenvalue λ of P satisfies $|\lambda| \leq 1$. There exists an eigenvector π of the eigenvalue 1, all the coordinates of which are greater or equal than zero. Without loss of generality we can suppose that the sum of its coordinates be 1.

It is now time to admire the power of this theorem. For this purpose, we will make the simplifying hypothesis that the matrix P has a basis of eigenvectors $\mathcal{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ (we will learn the definition in Topic 4) and we suppose that $\mathbf{v}_1$ is the vector π of Frobenius theorem. For each $\mathbf{v}_i$, there exists λ_i such that $P\mathbf{v}_i = \lambda_i\mathbf{v}_i$. Let us take any nonzero vector X with

$X^T = (x_1, \dots, x_N)$, where $x_i \in [0, 1]$ and $\sum_{i=1}^N x_i = 1$. We decompose X in the basis $\mathcal{V}$:

$$X = \sum_{i=1}^N a_i \mathbf{v}_i.$$

Mathematicians prove that $a_1 = 1$. Now, let us calculate PX :

$$PX = \sum_{i=1}^N a_i P\mathbf{v}_i = \sum_{i=1}^n a_i \lambda_i \mathbf{v}_i,$$

since $\mathbf{v}_i$ is an eigenvector of the eigenvalue λ_i . And if we iterate we get

$$P^n X = \sum_{i=1}^n a_i \lambda_i^n \mathbf{v}_i.$$

If all λ_i for $i > 1$ satisfy $|\lambda| < 1$, then

$$\lim_{n \rightarrow \infty} P^n X = a_1 \mathbf{v}_1 = \pi.$$

Notice however that the theorem does not guarantee that any matrix P satisfying the hypotheses of the theorem will have this property. Let us describe the possible pathologies and the remedy.

Possible pathologies:

1. The eigenvalue 1 may be a multiple root of the characteristic polynomial $\det(\lambda I - A) = 0$.
2. The matrix P may have other eigenvalues λ than 1, with modulus equal to 1. Here the eigenvalues can be complex numbers.

What do we do in this case? First, we define what a “good” matrix with no pathology.

Definition 41.14. A Markov transition matrix is *well-behaved* if

1. The eigenvalue 1 is a simple root of the characteristic polynomial $\det(\lambda I - A) = 0$.
2. All other eigenvalues λ of P than 1 have modulus less than 1.

Mathematicians discover that most matrices P are well-behaved. So, if the Markov transition matrix of a web is not well-behaved, then the strategy is to deform it slightly to become well-behaved. The **remedy**: consider the $N \times N$ matrix $Q = (q_{ij})$, with $q_{ij} = \frac{1}{N}$ for all i, j and replace the matrix P of the web by the matrix

$$P_\beta = (1 - \beta)P + \beta Q \quad (41.39) \quad \text{e:1}$$

for some small $\beta \in [0, 1]$. The value $\beta = 0.15$ has been used by Google.

Notice that the matrix P_β still has nonnegative entries and that the sum of the entries of each column is still equal to 1, so it is still a Markov transition matrix. The following theorem guarantees that there exists some small β for which we have fixed the pathologies.

Theorem 41.108. *Given any Markov transition matrix P , there always exists a positive β , as small as we wish, such that the matrix P_β is well-behaved. Let π be the eigenvector of 1 for the matrix P_β , normalized so that the sum of its coordinates be 1. For the matrix P_β , given any nonzero vector X , where $X^T = (p_1, \dots, p_N)$ with $p_i \in [0, 1]$ and $\sum_{i=1}^N p_i = 1$, then*

$$\lim_{n \rightarrow \infty} (P_\beta)^n X = \pi. \quad (41.40) \quad \text{e:2}$$

Example 41.109. Find the stationary vector π for the Markov transition matrix (41.38). le:pagerank

Solution:

$$P\pi = \pi \Rightarrow (P - I)\pi = \begin{bmatrix} -1 & \frac{1}{2} & \frac{1}{3} & 1 & 0 \\ 1 & -1 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & -1 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & -1 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} & 0 & -1 \end{bmatrix} \pi = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Solving it gives

$$\pi = \begin{bmatrix} 4 \\ 16/3 \\ 3 \\ 1/3 \\ 1 \end{bmatrix} t.$$

Pick $t = 3/41$, we have the stationary distribution (a unit vector): $\pi = \begin{bmatrix} 12/41 \\ 16/41 \\ 9/41 \\ 1/41 \\ 3/41 \end{bmatrix}$.

Remark 41.110. Finding the stationary distribution π , i.e. an eigenvector of the eigenvalue 1 for the matrix P_β in Equation (41.39), is not at all a trivial task when the matrix has billions of lines and columns: both the time of the computation and the memory space to achieve it represent real challenges. The usual method of Gauss elimination (the centre of attention in Topic 2) is useless in this case, both because of the size of the computation, and because it requires to divide by small coefficients. A more efficient algorithm makes use of the property (41.40) (see the paper by A. N. Langville and C. D. Meyer, “A Survey of Eigenvector Methods for Web Information Retrieval”, SIAM Review, Volume 47, Issue 1, (2005), pp. 135–161).

Another important application of eigenvalues and eigenvectors is in the study of long term behaviour of Markov chains. If X_{n+1} and X_n are two consecutive state vectors of a Markov chain with transition matrix P , then $X_{n+1} = PX_n$. What makes X_n approach a limiting vector as $n \rightarrow \infty$? Mathematicians discovered that a type of Markov chain define below has a limiting vector.

Definition 41.15. A transition matrix P is called *regular* if, for some integer r , all entries of P^r are strictly positive. Beware that 0 is not strictly positive.

A Markov chain process is called *regular* if its transition matrix is regular.

Example 41.111. Show that $A = \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0 & 1 \\ 0 & 0.5 & 0 \end{bmatrix}$ is regular.

Proof: Since $A^4 = \begin{bmatrix} 0.4375 & 0.3125 & 0.5 \\ 0.3125 & 0.625 & 0.125 \\ 0.25 & 0.0625 & 0.375 \end{bmatrix}$ has all positive entries, A is regular.

Theorem 41.112 (Main Theorem in Markov Chain Theory). 1. If P is a regular transition matrix, then as n approaches infinity, $P^n \rightarrow S$ where S is a matrix of the form $[\mathbf{v}, \mathbf{v}, \dots, \mathbf{v}]$ with $\mathbf{v}$ being a constant vector.

2. If P is a regular transition matrix of a Markov chain process, and if X is any state vector, then as n approaches infinity, $P^n X \rightarrow \mathbf{p}$, where $\mathbf{p}$ is a fixed probability vector (the sum of its entries is 1), all of whose entries are positive. This $\mathbf{p}$ is called the steady-state vector of the system.

Example 41.113. People who lived in Ottawa for a while are concerned with the weather. An unofficial study of the weather in the city in early spring yields the following observations:

- It is almost impossible to have two nice days in a row.
- If we have a nice day, we just as likely to have snow or rain the next day.
- If we have snow or rain, then we have an even chance to have the same the next day.
- If there is a change from snow or rain, only half of the time is this a change to a nice day.

1. Write the (left) transition matrix to model this system.
2. If it is nice today, what is the probability of being nice after one week?
3. If there is no sudden climate change (due to global warming or pollution), find the long time behaviour of the weather.

Solution:

1. Since the weather tomorrow depends only on today, this is a Markov chain process. The (left) transition matrix of this system is

$$\begin{array}{c} N \quad R \quad S \\ \begin{array}{l} N \\ R \\ S \end{array} \left(\begin{array}{ccc} 0 & 0.25 & 0.25 \\ 0.5 & 0.5 & 0.25 \\ 0.5 & 0.25 & 0.5 \end{array} \right)$$

where the letters N , R , S represent Nice, Rain, Snow respectively.

Note that this is NOT the way we learned in Discrete Mathematics (adjacency matrix in digraph) and it is NOT the way statistician represent transition matrix. Since we are writing vectors in column, therefore the matrix is a transpose of the “adjacency matrix”.

2. If it is nice today, then the initial state-vector is $X_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$. After seven days (one week), the state-vector would be

$$X_7 = \begin{bmatrix} 0 & 0.25 & 0.25 \\ 0.5 & 0.5 & 0.25 \\ 0.5 & 0.25 & 0.5 \end{bmatrix}^7 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.19995117 \\ 0.40002441 \\ 0.40002441 \end{bmatrix}.$$

So, there is about 20% chance of being nice in one week.

3. Notice first that we are dealing with a regular Markov chain since the transition matrix is regular, so we are sure that the steady-state vector exists. To find it we solve the homogeneous system

$$(P - I)X = 0 \Rightarrow \begin{bmatrix} -1 & 0.25 & 0.25 \\ 0.5 & -0.5 & 0.25 \\ 0.5 & 0.25 & -0.5 \end{bmatrix} X = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Reducing the matrix $P - I$ to r.r.e.f. gives $\begin{bmatrix} 1 & 0 & -0.5 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix}$ and the general solution

$$X = \begin{bmatrix} 0.5t \\ t \\ t \end{bmatrix}.$$

So what solution do we choose? Remember that a steady-state vector is in particular a probability vector; that is the sum of its components is 1: $0.5t + t + t = 1$ gives $t = 0.4$. Thus, the steady-state vector is

$$X = \begin{bmatrix} 0.2 \\ 0.4 \\ 0.4 \end{bmatrix}.$$

In the long term, there is 20% chance of getting a nice day, 40% chance of having a rainy day and 40% chance of having a snowy day.

Example 41.114 (Tutorial 2, Q20). Suppose a population has three age classes $\mathbf{x} \in \mathbb{R}^3$ and the females in the second and third age classes produces 4 and 3 offspring, respectively, every iteration. Suppose further that 50% of the females in the age class live on into the second age class and 25% of the females in the second age class live on into the age class. Then the *Leslie model* for this population is

$$\mathbf{x}_{k+1} = \begin{bmatrix} 0 & 4 & 3 \\ 0.5 & 0 & 0 \\ 0 & 0.25 & 0 \end{bmatrix} \mathbf{x}_k.$$

Use numerical computing environment to find the explicit solution of this model given $\mathbf{x}_0 = [10, 10, 10]^T$.

Solution.

$$\mathbf{x}_k = \lambda_1^k \mathbf{v}_1 + \lambda_2^k \mathbf{v}_2 + \lambda_3^k \mathbf{v}_3 = 1.5^k \begin{bmatrix} 0.947368 \\ 0.315789 \\ 0.0526316 \end{bmatrix} + (-1.30902)^k \begin{bmatrix} -0.932011 \\ 0.355996 \\ -0.0679893 \end{bmatrix} + (-0.190983)^k \begin{bmatrix} 0.225884 \\ -0.591372 \\ 0.774116 \end{bmatrix}$$

□

41.7.3 QR Algorithm

QR algorithm or *QR method* was invented by J. G. F. Francis in 1961 [Lax, 2007]. Unlike the power methods which can only find one dominant eigenvalues, QR algorithm can find all the eigenvalues.

In practise, instead of reducing eigenvalue problem (41.3) to a polynomial equation (41.35), we need to use an auxiliary matrix of the polynomial and find all the eigenvalues of the matrix, which are the zeros of the polynomial.

Theorem 41.115. The (multi-)set of eigenvalues of the matrix

$$A = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & 0 & \cdots & 0 & -a_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & -a_{n-2} \\ 0 & 0 & 0 & \cdots & 1 & -a_{n-1} \end{bmatrix}$$

is the same as the (multi-)set of roots of the polynomial

$$f(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_1z + a_0.$$

This theorem is significant in electrical filter design and linear control engineering[Boukas and AL-Sunni, 2011a] because the task of finding all poles and zeros of a transfer function $H(s)$ can be transformed to finding all eigenvalues of an corresponding matrix. The matrix can be solved using the QR algorithm which will be explained after the following example.

eg:transfer-fn

Example 41.116 (Ordinary Differential Equations). Use Julia 0.6's `eig` function to find the poles of the following transfer function

$$H(s) = \frac{3s + 6}{s^3 + 3s^2 + 7s + 5}.$$

Solution: According to Theorem [41.115](#), we can use Julia to find the eigenvalues of the matrix

$$A = \begin{bmatrix} 0 & 0 & -5 \\ 1 & 0 & -7 \\ 0 & 1 & -3 \end{bmatrix}$$

corresponding to the denominator of $H(s)$ which are $\boxed{-1 \pm 2i, -1}$.

Using the command `eig([0 0 -5; 1 0 -7; 0 1 -3])`, we obtain the poles of $H(s)$, i.e. the roots of the denominator.

QR algorithm can be seen as an adaptation of power method, where, instead using a vector $\mathbf{x}_0$ and the iteration $A^k \mathbf{x}_0$, we use an orthogonal matrix and for every iteration, we re-orthogonalise the new matrix as follows:

1. $A = Q_1 R_1$
2. $A^2 = (Q_1 R_1)(Q_1 R_1) = Q_1 (R_1 Q_1) R_1 = Q_1 (Q_2 R_2) R_1 = (Q_1 Q_2)(R_2 R_1)$
3. $A^3 = A \cdot A^2 = (Q_1 R_1)(Q_1 Q_2)(R_2 R_1) = (Q_1 Q_2 R_2 Q_2)(R_2 R_1) = Q_1 Q_2 (Q_3 R_3) R_2 R_1$
4. In general, $A^k = Q_1 Q_2 \cdots Q_k R_k \cdots R_2 R_1$.

Mathematicians discovered that the R_k above will converge and summarised it in the following theorem.

Theorem 41.117. *If A is symmetric and satisfies $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0$ (and A satisfies some regularity conditions), then $Q_k R_k$ converges to some diagonal matrix Λ which consists of eigenvalues of A .*

A naïve Julia implementation of QR algorithm is shown below.

```
1 m,n = size(A)
2 B = A
3 for i=1:50
4     Q, R = qr(B)
5     B = R*Q
6     print("Step $i: RQ=")
7     for i=1:n
8         if i>1 print(" ")
9         for j=1:n
10             @printf "%10.6f " B[i,j]
11         end
12     println()
13 end
14 end
```

A more sophisticated implementation is given in Burden and Faires [2015].

Symmetric matrices are common in engineering problem and special LAPACK subroutines are designed to handle them. The behaviour of real symmetric matrices and Hermitian matrices are governed by two spectral theorems stated below.

Theorem 41.118 (Spectral Theorem). *If A is real and symmetric, then its eigenvalues are real, and eigenvectors are mutually orthogonal.*

Theorem 41.119 (Spectral Theorem). *If A is complex and Hermitian, then its eigenvalues are real, and eigenvectors are mutually orthogonal (inner product with conjugate transpose). Furthermore, if A is Hermitian and positive definite, then its eigenvalues are positive.*

LAPACK provides the following subroutines to handle symmetric matrices.

- **xORGTR**: Generates the orthogonal/unitary transformation matrix from a reduction to tridiagonal form determined by **xSYTRD**.
- **xORMTR**: Multiplies a general matrix by the orthogonal/unitary transformation matrix from a reduction to tridiagonal form determined by **xSYTRD**.
- **xSTEQR**: Computes all eigenvalues and eigenvectors of a real symmetric tridiagonal matrix, using the implicit QL or QR algorithm.
- **xPTEQR**: Computes all eigenvalues and eigenvectors of a real symmetric positive definite tridiagonal matrix, by computing the SVD of its bidiagonal Cholesky factor.

Using LAPACK subroutines would too complex and we will just use the Julia 0.3 implementation above to study the symmetric matrix from Example 41.122.

Example 41.120. Use the naïve QR algorithm to find all the eigenvalues of the 4×4 matrix

$$A = \begin{bmatrix} 4 & 1 & -2 & 2 \\ 1 & 2 & 0 & 1 \\ -2 & 0 & 3 & -2 \\ 2 & 1 & -2 & -1 \end{bmatrix}.$$

Solution: We find that even up to 50 steps, some of the eigenvalues (e.g. 2.263170, -2.192156) are still converging extremely slowly. Using the `eig` function, it is easy to find the eigenvalues -2.19751698, 1.08436446, 2.26853141, 6.84462111 for comparison.

1	Step 48: RQ=	6.844621	0.000000	0.000000	0.000000
2		-0.000000	2.262444	-0.164774	-0.000000
3		-0.000000	-0.164774	-2.191429	-0.000000
4		-0.000000	-0.000000	-0.000000	1.084364
5	Step 49: RQ=	6.844621	0.000000	0.000000	-0.000000
6		-0.000000	2.262819	0.159629	0.000000
7		0.000000	0.159629	-2.191804	-0.000000
8		0.000000	0.000000	-0.000000	1.084364
9	Step 50: RQ=	6.844621	0.000000	0.000000	0.000000
10		-0.000000	2.263170	-0.154644	-0.000000
11		-0.000000	-0.154644	-2.192156	-0.000000
12		-0.000000	-0.000000	-0.000000	1.084364

Example 41.121 (Tutorial 2, Q19). Given $C = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$. Carry out the following task by using an interactive numerical computing environment.

1. By carrying out just 10 iterations using a numerical computing environment, find the dominant eigenvalue of C by power method with $\mathbf{x}_0 = [1, 1, 1]^T$ $\lambda_1 \approx 3.414216$

	i	$\mathbf{x}^{(i)}$	$m^{(i)}$	
<i>Solution.</i>	0	$[1.000000, 1.000000, 1.000000]^T$	1.000000	□
	1	$[1.000000, 0.000000, 1.000000]^T$	1.000000	
	2	$[1.000000, -1.000000, 1.000000]^T$	2.000000	
	3	$[-0.750000, 1.000000, -0.750000]^T$	4.000000	
	4	$[-0.714286, 1.000000, -0.714286]^T$	3.500000	
	5	$[-0.708333, 1.000000, -0.708333]^T$	3.428571	
	6	$[-0.707317, 1.000000, -0.707317]^T$	3.416667	
	7	$[-0.707143, 1.000000, -0.707143]^T$	3.414634	
	8	$[-0.707113, 1.000000, -0.707113]^T$	3.414286	
	9	$[-0.707108, 1.000000, -0.707108]^T$	3.414226	
	10	$[-0.707107, 1.000000, -0.707107]^T$	<u>3.414216</u>	

2. Use Aitken's Δ^2 to improve the estimates for the dominant eigenvalue using $\lambda_1^{(8)}$, $\lambda_1^{(9)}$ and $\lambda_1^{(10)}$.

Solution.

$$q_8 = \lambda_1^{(8)} - \frac{(\lambda_1^{(9)} - \lambda_1^{(8)})^2}{\lambda_1^{(10)} - 2\lambda_1^{(9)} + \lambda_1^{(8)}} \approx 3.414214$$

□

3. Use the simple QR algorithm to find all the eigenvalues C .

Solution. 3.414214, 2, 0.585786

□

4. A matrix is said to defective if for one of the eigenvalue, the algebraic multiplicity and geometric multiplicity are different. So is C a defective matrix?

Solution. No. For each eigenvalue, the algebraic and geometric multiplicities are equal to one. Therefore, C is non-defective.

□

QR algorithm is too slow for finding the eigenvalues of *symmetric matrix* A in the real world. Suitable Householder transforms H_i are used to transform A to *tridiagonal matrix* $T = H_n \cdots H_1 A H_1^T \cdots H_n^T$ and then QR algorithm is used to find the common eigenvalues of T and A .

For *non-symmetric matrix* B , it is only possible to transform B to *upper Hessenberg matrix* Γ (see Table 2.1 HG) using suitable Householder transforms. The QR algorithm can obtain the common eigenvalues for both Γ and B although it is much slower compare to the symmetric matrix case. LAPACK provides the following subroutines to handle non-symmetric matrices.

- xORGHR: Generates the orthogonal/unitary transformation matrix from a reduction to Hessenberg form determined by xGEHRD
- xORMHR: Multiplies a general matrix by the orthogonal/unitary transformation matrix from a reduction to Hessenberg form determined by xGEHRD
- xHSEQR: Computes the eigenvalues and Schur factorization of an upper Hessenberg matrix, using the multishift QR algorithm
- xHSEIN: Computes specified right and/or left eigenvectors of an upper Hessenberg matrix by inverse iteration
- xGEBAL: Balances a general matrix in order to improve the accuracy of computed eigenvalues
- xGEBAK: Transforms eigenvectors of a balanced matrix to those of the original matrix supplied to xGEBAL

In *Fundamentals of Linear Algebra*, we learned that when A is similar to matrix B if there is a matrix P such that $B = PAP^{-1}$. When A and B are similar, they have the same eigenvalues.

Householder's method is a technique to reduce the matrix A to either a tridiagonal matrix or a Hessenberg matrix, which having the same eigenvalues as A , can be more efficiently estimated by the QR algorithm [Burden and Faires, 2015].

Householder's method begins by determining a Householder transformation $H(\mathbf{w}_1)$ with the property that $A^{(2)} = H(\mathbf{w}_1)AH(\mathbf{w}_1)$ zero's out the entries in the first column of A beginning with the third row, i.e.

$$a_{j1}^{(2)} = 0, \quad \text{for each } j = 3, 4, \dots, n.$$

If A is symmetrical, we also have $a_{j1}^{(2)} = 0$. We can choose $\mathbf{w}_2$ so that $A^{(3)} = H(\mathbf{w}_2)A^{(2)}H(\mathbf{w}_2)$ zero's out the entries in the second column beginning from the fourth row, etc. In the end, we will obtain a Hessenberg matrix (or tridiagonal matrix).

Let's just look at an example to see how this works.

ten-faires:p615

Example 41.122 (Symmetric matrix, Burden and Faires [2015]). Apply Householder transformations to the symmetric 4×4 matrix

$$A = \begin{bmatrix} 4 & 1 & -2 & 2 \\ 1 & 2 & 0 & 1 \\ -2 & 0 & 3 & -2 \\ 2 & 1 & -2 & -1 \end{bmatrix}$$

to produce a symmetric tridiagonal matrix that is similar to A . **[Remark:** Try to notice the difference of this example to Example 41.50]

Solution. First, we need to find

$$\mathbf{u}_1 = \begin{bmatrix} 0 \\ 1 \\ -2 \\ 2 \end{bmatrix} - \begin{bmatrix} 0 \\ \sqrt{1 + (-2)^2 + 2^2} \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 4 \\ -2 \\ 2 \end{bmatrix} \Rightarrow \mathbf{w}_1 = \frac{1}{\sqrt{24}} \begin{bmatrix} 0 \\ 4 \\ -2 \\ 2 \end{bmatrix}.$$

The first step reduction is

$$\begin{aligned} A^{(2)} &= (I - 2\mathbf{w}_1\mathbf{w}_1^T)A(I - 2\mathbf{w}_1\mathbf{w}_1^T) \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -\frac{1}{3} & \frac{2}{3} & -\frac{2}{3} \\ 0 & \frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ 0 & -\frac{2}{3} & \frac{1}{3} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} 4 & 1 & -2 & 2 \\ 1 & 2 & 0 & 1 \\ -2 & 0 & 3 & -2 \\ 2 & 1 & -2 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -\frac{1}{3} & \frac{2}{3} & -\frac{2}{3} \\ 0 & \frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ 0 & -\frac{2}{3} & \frac{1}{3} & \frac{2}{3} \end{bmatrix} = \begin{bmatrix} 4 & -3 & 0 & 0 \\ -3 & \frac{10}{3} & 1 & \frac{4}{3} \\ 0 & 1 & \frac{5}{3} & -\frac{4}{3} \\ 0 & \frac{4}{3} & -\frac{4}{3} & -1 \end{bmatrix} \end{aligned}$$

Let us continue to find

$$\mathbf{u}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \frac{4}{3} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \sqrt{1 + (\frac{4}{3})^2} \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \frac{8}{3} \\ \frac{4}{3} \end{bmatrix} \Rightarrow \mathbf{w}_2 = \frac{1}{\sqrt{80}} \begin{bmatrix} 0 \\ 0 \\ 8 \\ 4 \end{bmatrix}$$

The second step reduction is

$$\begin{aligned}
 A^{(3)} &= (I - 2\mathbf{w}_2\mathbf{w}_2^T)A(I - 2\mathbf{w}_2\mathbf{w}_2^T) \\
 &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{3}{5} & -\frac{4}{5} \\ 0 & 0 & -\frac{4}{5} & \frac{3}{5} \end{bmatrix} \begin{bmatrix} 4 & -3 & 0 & 0 \\ -3 & \frac{10}{3} & 1 & \frac{4}{3} \\ 0 & 1 & \frac{5}{3} & -\frac{4}{3} \\ 0 & \frac{4}{3} & -\frac{4}{3} & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{3}{5} & -\frac{4}{5} \\ 0 & 0 & -\frac{4}{5} & \frac{3}{5} \end{bmatrix} \\
 &= \begin{bmatrix} 4 & -3 & 0 & 0 \\ -3 & \frac{10}{3} & -\frac{5}{3} & 0 \\ 0 & -\frac{5}{3} & -\frac{33}{25} & \frac{68}{75} \\ 0 & 0 & \frac{68}{75} & \frac{149}{75} \end{bmatrix}
 \end{aligned}$$

□

en-faires:p615b

Example 41.123. Use LAPACK subroutine to transform the symmetric 4×4 matrix

$$A = \begin{bmatrix} 4 & 1 & -2 & 2 \\ 1 & 2 & 0 & 1 \\ -2 & 0 & 3 & -2 \\ 2 & 1 & -2 & -1 \end{bmatrix}$$

to produce a symmetric tridiagonal matrix that is similar to A .

Solution: The subroutine xSYTRD reduces a symmetric/Hermitian matrix to real symmetric tridiagonal form by an orthogonal/unitary similarity transformation.

```

1  program eigen_reduce1 ! Symmetric Matrix Example
2      integer, parameter :: N=4
3      integer :: LWORK=N
4      real(kind=8) :: A(N,N), D(N), E(N-1), TAU(N), WORK(N)
5      A = reshape((/4,1,-2,2, 1,2,0,1, -2,0,3,-2, 2,1,-2,-1/), (/N,N/))
6      ! DSYTRD(UPLO, N, A, LDA, D, E, TAU, WORK, LWORK, INFO)
7      ! 'U' (or 'L'): Upper (or Lower) triangle of A is stored
8      call dsytrd('L',N,A,N,D,E,TAU,WORK,LWORK,INFO)
9      if (INFO .eq. 0) then
10         print *, "A (after DSYTRD) = "
11         do i=1,N
12             write(*,'(4f10.6)', (A(i,j), j=1,N)
13         end do
14         print *, "B = "
15         do i=1,N
16             do j=1,N
17                 if (i .eq. j) then
18                     write (*,100,advance='no'), D(i)
19                 else if (i .eq. j+1) then
20                     write (*,100,advance='no'), E(j)
21                 else if (j .eq. i+1) then
22                     write (*,100,advance='no'), E(i)
23                 else
24                     write (*,100,advance='no'), 0.0
25                 end if
26             end do
27             print *,
28         end do
29     else
30         print *, "Error in the Householder's Method reduction"
31     end if
32     100 format(4f10.6, x)
33 end program eigen_reduce1

```

The output is shown below.

For the non-symmetric example, the calculations are similar but much more complicated.

eg:tut2q18

Example 41.124 (Tutorial 2, Q18, Non-symmetric matrix). Use Householder's method to reduce the matrix $A = \begin{bmatrix} 1 & -3 & 3 \\ 3 & -5 & 3 \\ 6 & -6 & 4 \end{bmatrix}$ to a Hessenberg matrix.

Solution. First, find

$$\mathbf{u}_1 = \begin{bmatrix} 0 \\ 3 \\ 6 \end{bmatrix} + \begin{bmatrix} 0 \\ \sqrt{3^2 + 6^2} \\ 0 \end{bmatrix}, \quad \mathbf{w}_1 = \begin{bmatrix} 0 \\ 0.85065081 \\ 0.52573111 \end{bmatrix}$$

$$\text{Then } A^{(2)} = (I - 2\mathbf{w}_1\mathbf{w}_1^T)A(I - 2\mathbf{w}_1\mathbf{w}_1^T) = \begin{bmatrix} 1.0 & -1.34164079 & 4.02492236 \\ -6.70820393 & 1.0 & -9.0 \\ 0.0 & 0.0 & -2.0 \end{bmatrix} \quad \square$$

utionEigen.pdf

Example 41.125. Use LAPACK subroutine to transform the following matrix

$$A = \begin{bmatrix} 1 & -3 & 3 \\ 3 & -5 & 3 \\ 6 & -6 & 4 \end{bmatrix}$$

to a similar Hessenberg matrix.

Solution: The subroutine xGEHRD reduces a general matrix to upper Hessenberg form by an orthogonal/unitary similarity transformation.

```

1  program eigen_reduce2 ! Non-Symmetric Matrix Example
2      integer, parameter :: N=3
3      integer :: LWORK=N
4      real(kind=8) :: A(N,N), TAU(N), WORK(N)
5      A = reshape((/1,3,6, -3,-5,-6, 3,3,4/), (/N,N/))
6      ! DGEHRD(N, ILO, IHI, A, LDA, TAU, WORK, LWORK, INFO)
7      ! It is assumed that A is already upper triangular in rows and
8      ! columns 1:ILO-1 and IHI+1:N. ILO and IHI are normally set by
9      ! a previous call to DGEBAL; otherwise they should be set
10     ! to 1 and N respectively.
11     call dgehrd(N,1,N,A,N,TAU,WORK,LWORK,INFO)
12     if (INFO .eq. 0) then
13         print *, "A (after DGEHRD) = "
14         do i=1,N
15             write(*,'(4f10.6)'), (A(i,j), j=1,N)
16         end do
17     else
18         print *, "Error in the Householder's Method reduction"
19     end if
20     100 format(4f10.6, x)
21 end program eigen_reduce2

```

eg:tut2q18

The output is shown below (which matches the results in Example 41.124 using hand calculation).

Again studying the implementation details of LAPACK subroutines are beyond the scope of this subject. We will just explore two examples, one with all the eigenvalues real, while the other example has complex eigenvalues.

Example 41.126 (Real Eigenvalues). Use QR algorithm to find the eigenvalues of

$$A = \begin{bmatrix} 0 & 0 & -6 \\ 1 & 0 & -1 \\ 0 & 1 & 4 \end{bmatrix}.$$

Solution: In approximately 38 steps, we obtain the eigenvalues of A , which are 3, 2, -1 .

```
Step 36: RQ=  3.000001  -5.196152   3.535534
             0.000000   1.999999  -1.224744
             0.000000   0.000000  -1.000000
Step 37: RQ=  3.000001  -5.196152  -3.535534
             0.000000   1.999999   1.224744
             0.000000   0.000000  -1.000000
Step 38: RQ=  3.000000  -5.196152   3.535534
             0.000000   2.000000  -1.224745
             0.000000   0.000000  -1.000000
```

Example 41.127. Write a Fortran program using the appropriate subroutine from LAPACK to find the eigenvalues in Example 41.126. eg:nev:real

Solution:

```
1  program eigen_eg  ! Non-Symmetric Matrix Example
2      integer, parameter :: N=3, LDA=N, LDVL=N, LDVR=N, LWORK=4*N
3      integer :: info, i
4      real(kind=8) :: A(N,N), WR(N), WI(N), WORK(LWORK)
5      real(kind=8) :: VL(LDVL,N), VR(N,LDVR)
6      A = reshape((/0d0,1d0,0d0, 0d0,0d0,1d0, -6d0,-1d0,4d0/), (/N,N/))
7      ! Don't compute both left and right eigenvectors
8      call dgeev('N','N',N,A,LDA,WR,WI,VL,LDVL,VR,LDVR,WORK,LWORK,info)
9      if (INFO .eq. 0) then
10         print *, "Eigenvalues = "
11         do i=1,N
12             write(*,'(f10.6, "+", f10.6, "i")'), WR(i), WI(i)
13         end do
14         print *, "A (after DGEEV) = "
15         do i=1,N
16             write(*,'(4f10.6)'), (A(i,j), j=1,N)
17         end do
18     else
19         print *, "Error in finding eigenvalues"
20     end if
21 end program eigen_eg
```

The program should generate the following output:

```
1  Eigenvalues =
2  -1.000000+ 0.000000i
3   2.000000+ 0.000000i
4   3.000000+ 0.000000i
5  A (after DGEEV) =
6  -1.000000 -1.340848  0.299718
7   0.000000  2.000000  3.887454
8   0.000000  0.000000  3.000000
```

Example 41.128 (Complex Eigenvalues). Is it possible to find all the eigenvalues

$$A = \begin{bmatrix} 0 & 0 & -5 \\ 1 & 0 & -7 \\ 0 & 1 & -3 \end{bmatrix}$$

using QR algorithm?

Solution: We know from Example 41.116 that the eigenvalues are $-1 \pm 2i$ and -1 but the QR algorithm fails to converge as follows:

```
Step 46: RQ= -1.384776  -7.532194  -4.898479
              0.550710  -0.615224  -0.819496
              0.000000  -0.000000  -1.000000
Step 47: RQ=  1.117621  -6.843061   4.248907
              1.239842  -3.117621   2.571665
              0.000000   0.000000  -1.000000
Step 48: RQ= -4.005575  -5.858002  -4.754999
              2.224902   2.005575   1.434103
              0.000000   0.000000  -1.000000
Step 49: RQ= -1.046056  -7.553034   4.853163
              0.529869  -0.953944   1.055214
              0.000000  -0.000000  -1.000000
Step 50: RQ=  1.803854  -6.156100  -3.852592
              1.926804  -3.803854  -3.134359
              0.000000   0.000000  -1.000000
```

However, LAPACK implemented much more advanced QR algorithm which allows us to solve eigenvalue problems with complex eigenvalues as illustrated below.

Example 41.129. Write a Fortran program using the appropriate subroutine from LAPACK to find the eigenvalues in Example 41.128.

Solution.

```
1  program eigen_eg  ! Non-Symmetric Matrix Example
2      integer, parameter :: N=3, LDA=N, LDVL=N, LDVR=N, LWORK=4*N
3      integer :: info, i
4      real(kind=8) :: A(N,N), WR(N), WI(N), WORK(LWORK)
5      real(kind=8) :: VL(LDVL,N), VR(N,LDVR)
6      A = reshape((/0d0,1d0,0d0, 0d0,0d0,1d0, -5d0,-7d0,-3d0/), (/N,N/))
7      ! Don't compute both left and right eigenvectors
8      call dgeev('N','N',N,A,LDA,WR,WI,VL,LDVL,VR,LDVR,WORK,LWORK,info)
9      if (INFO .eq. 0) then
10         print *, "Eigenvalues = "
11         do i=1,N
12             write(*,'(f10.6, "+", f10.6, "i")') WR(i), WI(i)
13         end do
14         print *, "A (after DGEEV) = "
15         do i=1,N
16             write(*,'(4f10.6)') (A(i,j), j=1,N)
17         end do
18     else
19         print *, "Error in finding eigenvalues"
20     end if
21 end program eigen_eg
```

The program should generate the following output:

```
1  Eigenvalues =
2  -1.000000+  0.000000i
3  -1.000000+  2.000000i
4  -1.000000+ -2.000000i
5  A (after DGEEV) =
6  -1.000000  2.062966 -3.903395
7   0.000000 -1.000000  2.302832
8   0.000000 -1.736992 -1.000000
```

Remark 41.130. For large scale eigenvalue problems, instead of using the LAPACK library, ARPACK Fortran 77 library or SLEPc are used. SLEPc (<http://slepc.upv.es/>), the Scalable Library for Eigenvalue Problem Computations, is a software library for the solution of large scale sparse eigenvalue problems on parallel computers. It is an extension of PETSc and can be used for linear eigenvalue problems in either standard or generalized form, with real or complex arithmetic. It can also be used for computing a partial SVD of a large, sparse, rectangular matrix, and to solve nonlinear eigenvalue problems (polynomial or general). Additionally, SLEPc provides solvers for the computation of the action of a matrix function on a vector. For issues related to symmetric eigenvalue problems, consult Parlett [1998]. For the complexity of non-symmetric eigenvalue problems, one can continue their investigation by reading Trefethen and Embree [2005].

Chapter 42

Interpolation & Curve Fitting

Standard techniques for the approximation of the numerical differentiation, numerical quadrature, the solution to differential equations rely on replacing intractable functions by interpolation polynomials [Acton, 1990]. There are a few ways to describe the *difference* between the “intractable” function and the approximating polynomial — (i) *interpolation / collocation*, where Lagrange polynomials is used to match data points; (ii) *osculation*, where osculating polynomials not only match data points but also the derivatives!

In engineering (e.g. medical imaging), topography, computer graphics, CAD, CAM and 3D-printer, we often want to fit a smooth curve / surface / hyper-surfaces through a set of (experimental) data points. For example, adaptive method of ODE normally generates irregular data points, interpolation is necessary to obtain the solution at some particular time. Another example is from experimental physics, where the scientists need to interpret measured data from a radioactive source by fitting data with an appropriate analytical function.

When we have “1-D” points to fit using a one-variable (or *univariate*) polynomial or a function to be approximated by a polynomial

$$P_n(x) = a_0 + a_1x + \cdots + a_{n-1}x^{n-1} + a_nx^n, \quad a_0, a_1, \dots, a_n \in \mathbb{R}, \quad (42.1) \quad \text{e:poly}$$

we have two ways to fit the data.

One way is to require a *single polynomial* to pass through all the given points and this leads to *interpolating polynomials* and *osculating polynomials*. Note that a_k , $k = 0, 1, \dots, n$ are called the *coefficient* of the polynomial (42.1), a_0 and a_n are called the *constant coefficient* and *leading coefficient* respectively. e:poly

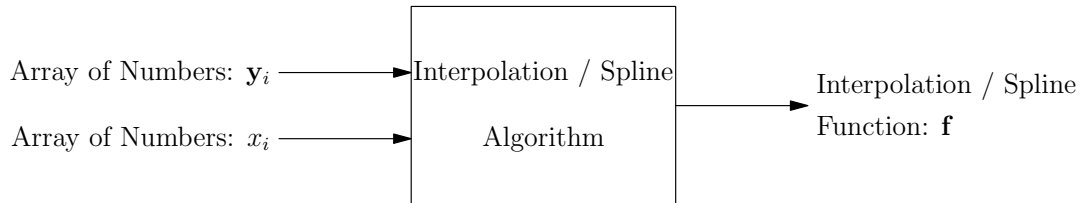
Another way is to have multiple *piecewise polynomials* to pass through the points. Two types of “smooth” piecewise cubic polynomials — *cubic splines* and *Hermite cubics* are often recommended for fitting a smooth curve through accurate data.

The requirement for approximating polynomials to pass through the all given data points is too restrictive in computer graphics design. For interactive design, such as in CG/CAD/CAM, **control points** are used instead of **data points**, “Bézier” curves (Section 42.6) and B-splines change with respect to the control points but normally do not pass through the control points [Kahaner et al., 1989]. sec:bezier

When we have m -dimensional ($m \geq 2$) points, a *multivariate* polynomial of degree n is needed to “fit” them:

$$P_n(x_1, \dots, x_d) = b_0 + \sum_{1 \leq i_1 + \dots + i_d \leq n} b_{i_1 \dots i_d} x^{i_1} \cdots x^{i_d}, \quad b_{i_1 \dots i_d} \in \mathbb{R}. \quad (42.2) \quad \text{e:poly-d}$$

The Lagrange interpolation (Section sec:interp 42.2), Hermite interpolation (Section sec:hermite 42.3) and splines (Section sec:spline 42.4) are used in the generation of curvilinear coordinate system for grid mapping in PDE [Thompson et al., 1985]. Either interpolation curve and spline curves can be regarded as a block which takes in data points and returns a function.



Sometimes, a curve is required to “try” to pass through the data points in the way to minimise errors leading to *linear regression* (Section 53.1) or *nonlinear regressions* (Section 53.5).

42.1 Working with Polynomials

In solving problems related to ODE or 1-D interpolation problems, a univariate polynomial (42.1) is sufficient. However, when dealing with 2-D or 3-D spatial PDE problems, two-variable and three-variable polynomials, i.e. (42.2) with $d = 2$ or $d = 3$, are required.

Polynomials (42.1) and (42.2) are used in the approximation of data or functions because they can be evaluated, differentiated and integrated easily.

Definition 42.1. The *order* of a polynomial $P_n(x)$ in Formula (42.1) is $n + 1$. $P_n(x)$ has a *degree* $\leq n$. When $a_n \neq 0$, $P_n(x)$ is said to have degree n but when $a_n = 0$, the degree of $P_n(x)$ is $< n$.

Most programming languages only support polynomials and their operations through special mathematical libraries such as the C++ library CGAL [The CGAL Project, 2017].

eg:cgall

Example 42.1 (Working with polynomials using CGAL library). Explain what does the following C++ code does:

```

1 #include <CGAL/config.h>
2 #include <CGAL/Polynomial.h>
3 #include <CGAL/Polynomial_traits_d.h>
4 #include <CGAL/Polynomial_type_generator.h>
5
6 int main(){
7   CGAL::set_pretty_mode(std::cout); // without this, the output is ugly
8   // Data Structure for 2D points
9   typedef CGAL::Polynomial_type_generator<int,2>::Type Poly_2;
10  typedef CGAL::Polynomial_traits_d<Poly_2> PT_2;
11  // Data Structure for 3D points
12  typedef CGAL::Polynomial_type_generator<int,3>::Type Poly_3;
13  typedef CGAL::Polynomial_traits_d<Poly_3> PT_3;
14
15  // Construction of ``symbols`` using shift
16  Poly_2 x = PT_2::Shift()(Poly_2(1),1,0); //  $x_0^1$ 
17  Poly_2 y = PT_2::Shift()(Poly_2(1),1,1); //  $x_1^1$ 
18  Poly_3 x2 = PT_3::Shift()(Poly_3(1),1,0); //  $x_0^1$ 
19  Poly_3 y2 = PT_3::Shift()(Poly_3(1),1,1); //  $x_1^1$ 
20  Poly_3 z2 = PT_3::Shift()(Poly_3(1),1,2); //  $x_2^1$ 
21
22  // Construction of 2-variable & 3-variable ``polynomials``
23  Poly_2 F = 11 * CGAL::ipower(y,4) * CGAL::ipower(x,2)
24             + 5 * CGAL::ipower(y,4) * CGAL::ipower(x,1)
25             + 7 * CGAL::ipower(y,3) * CGAL::ipower(x,2);
26  Poly_3 F2 = x2*y2*y2*z2*z2*z2;
27
28  { // Reuse the variables degree, degt, degv
29    std::cout << "The bivariate polynomial F: " << F << "\n";
30    PT_2::Degree degree;
31    PT_2::Total_degree degt;
32    PT_2::Degree_vector degv;
  
```



```

33     std::cout << "The degrees of F w.r.t. x and y are "
34         << degree(F,0) << " and " << degree(F) << " respectively.\n";
35     std::cout << "The total degree of F          : " << degt(F) << "\n";
36     std::cout << "The degree vector of F          : " << degv(F) << "\n";
37     //PT_2::Get_coefficient coeff;
38     //for(int k=0; k<degt(F); k++) {
39     //    std::cout << "Coefficient of y^" << k << " is " << coeff(F, k) << "\n";
40     //}
41 }
42 { // Reuse the variables degree, degt, degv
43     std::cout << "\nThe trivariate polynomial F2: " << F2 << "\n";
44     PT_3::Degree degree;
45     PT_3::Total_degree degt;
46     PT_3::Degree_vector degv;
47     std::vector<std::string> var{"x", "y", "z"};
48     for(int k : std::vector<int> {0,1,2}) {
49         std::cout << "Degree of F2 w.r.t. " << var[k] << ": "
50             << degree(F2, k) << "; ";
51     }
52     std::cout << "\nThe total degree of F2          : " << degt(F2) << "\n";
53     std::cout << "The degree vector of F2          : " << degv(F2) << "\n";
54 }
55 }

```

Solution: The program constructs a 2-variate polynomial $11x^2y^4 + 5xy^4 + 7x^2y^3$ and a 3-variate monomial xy^2z^3 as illustrated by the output of this program.

```

1 The bivariate polynomial F: (11*x^2 + 5*x)*y^4 + (7*x^2)*y^3
2 The degrees of F w.r.t. x and y are 2 and 4 respectively.
3 The total degree of F          : 6
4 The degree vector of F          : (2,4)
5
6 The trivariate polynomial F2: ((x)*y^2)*z^3
7 Degree of F2 w.r.t. x: 1; Degree of F2 w.r.t. y: 2; Degree of F2 w.r.t. z: 3;
8 The total degree of F2          : 6
9 The degree vector of F2          : (1,2,3)

```

Example [42.1](#) ^{eg:cgall} demonstrates that the manipulation of polynomials for computational geometry (the maths behind CAD) can be very complicated. CGAL only supports *polynomial reduction, linear and quadratic programming* for computational geometry and does not support general polynomial manipulation. One normally uses a computer algebra system such as Maxima (or Reduce) for this purpose.

Example 42.2. Investigate how to represent the polynomial $p(x) = \frac{1}{2}x^3 + 2x^2 + 3x + 4$ and $q(x) = 5x + 7$ and perform addition, multiplication and differentiation in Maxima.

Solution:

```

1 /* To run script: maxima -q < script.mac > script.out or maxima -qb script.mac */
2 p : 0.5*x^3 + 2*x^2 + 3*x + 4$
3 q : 5*x + 7$
4 print("p(x)=", p, "; q(x)=", q)$
5 print("p(x) + q(x) =", p + q)$
6 print("p(x) * q(x) =", expand(p * q))$
7 print("diff(p(x), x) =", diff(p, x))$
8 print("diff(q(x), x) =", diff(q, x))$

```

which generates the following output:

```

1 (%i1)
2 (%i2)
3 (%i3)
4          3      2
5 p(x)= 0.5 x  + 2 x  + 3 x + 4 ; q(x)= 5 x + 7
6 (%i4)

```

```

7
8 p(x) + q(x) = 0.5 x3 + 2 x2 + 8 x + 11
9 (%i5)
10
11 p(x) * q(x) = 2.5 x4 + 13.5 x3 + 29 x2 + 41 x + 28
12 (%i6)
13
14 diff(p(x), x) = 1.5 x2 + 4 x + 3
15 (%i7)
16 diff(q(x), x) = 5
17 (%i8)

```

42.2 (Lagrange) Interpolation

sec:interp

Definition 42.2. Assume that the values of a function f are only known at some discrete points $x_0 < x_1 < \cdots < x_m$. It is one of the classical problems of numerical analysis to estimate the value of f at some point $\xi \notin \{x_0, x_1, \dots, x_m\}$. If $x_0 < \xi < x_m$, this problem is called *interpolation*, and if either $\xi < x_0$ or $x_m < \xi$, this problem is called *extrapolation*.

Definition 42.3. A polynomial that passes through all the given sequence of data points $(x_0, y_0), \dots, (x_n, y_n)$ is called an *(Lagrange) interpolating polynomial*.

thm:interp:exist

Theorem 42.3 (Interpolating polynomial exist uniquely). *If $n + 1$ points $(x_0, y_0), \dots, (x_n, y_n)$ are given such that $x_0 < \cdots < x_n$ then there is a unique polynomial $P(x)$ of order $n + 1$ such that $P(x_k) = y_k, \quad k = 0, 1, \dots, n$.*

The procedure of finding the interpolating polynomial is called *Lagrange interpolation*. Even though the interpolating polynomial is unique, it can be written in different forms. The *natural form*, with $\{1, x, \dots, x^n\}$ as the basis, will be discussed in Section [42.2.1](#). The *Lagrange form*, which is useful in the analysis of numerical integration, with $\{L_0(x), \dots, L_n(x)\}$ as the basis, as well as the *modified Lagrange form*, will be discussed in Section [42.2.2](#). Section [42.2.4](#) discusses the *Newton form*, with $\{1, (x - x_0), (x - x_0)(x - x_1), \dots, (x - x_0)(x - x_1) \cdots (x - x_{n-1})\}$ as the basis, is useful in the analysis of difference method.

Assume that $n + 1$ points $\{(x_k, y_k)\} = \{(x_0, y_0), \dots, (x_n, y_n)\}$ such that $x_0 < \cdots < x_n$ are given in the following subsections.

42.2.1 Natural Form

sec:natural-f

Since the Lagrange interpolating polynomial is a polynomial of order $n + 1$, it has the form

$$P(x) = a_0 + a_1x + \cdots + a_nx^n$$

and the graph $\{(x, P(x)) : x \in \mathbb{R}\}$ passes through all the points (x_k, y_k) , i.e.

$$\begin{cases} P(x_0) = a_0 + a_1x_0 + \cdots + a_{n-1}x_0^{n-1} + a_nx_0^n = y_0, \\ P(x_1) = a_0 + a_1x_1 + \cdots + a_{n-1}x_1^{n-1} + a_nx_1^n = y_1, \\ \quad \quad \quad \dots \\ P(x_n) = a_0 + a_1x_n + \cdots + a_{n-1}x_n^{n-1} + a_nx_n^n = y_n. \end{cases}$$

The above equations can be written in matrix form

$$\underbrace{\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix}}_V \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}. \quad (42.3)$$

The coefficient matrix V is called *Vandermonde matrix*. It has a determinant $\det(V) = \prod_{1 \leq i < j \leq n} (x_j - x_i)$. So, as long as the points x_i are different, $\det(V) \neq 0$, the equation (42.3) has a unique solution and this proves Theorem 42.3.

Example 42.4. Given the data points $(0,0)$, $(\pi/2,1)$, $(\pi,0)$, $(3\pi/2,-1)$.

1. Write a program to find the coefficients of the interpolating polynomial for the data points.

Solution: A C++ program could be written based on Equation (42.3).

```
1 // eg_interp_natural.cxx
2 #include <openblas/lapack.h>
3 #include <cmath> // atan, pow
4 #include <cstdio> // printf
5
6 int main() {
7     int N=4, LDA=N, LDB=N, NRHS=1;
8     const double pi = 4.0*atan(1.0);
9     double A[LDA*N], coeffs[N];
10    double x[N] {0.0, pi/2.0, pi, 3.0*pi/2.0};
11    double y[N] {0.0, 1.0, 0.0, -1.0};
12    int ipiv, info;
13    // Constructing Vandermonde matrix in C style array
14    for (int i=0; i<N; i++) {
15        A[i] = 1.0; // first row
16        for (int j=1; j<N; j++) { // rest rows
17            A[i+j*N] = pow(x[i], j);
18        }
19        coeffs[i] = y[i];
20    }
21    LAPACK_dgesv(&N,&NRHS,A,&LDA,&ipiv,coeffs,&LDB,&info);
22    if( info > 0 ) {
23        printf( "The diagonal element of the triangular factor of A,\n" );
24        printf( "U(%i,%i) is zero, so that A is singular;\n", info, info );
25        printf( "the solution could not be computed.\n" );
26        exit( 1 );
27    }
28    printf("lagrange(x,y): y = ");
29    for (int i=0; i<N; i++) {
30        printf("%10.6f x^%d", coeffs[i], i);
31        if(i<N-1 && coeffs[i+1]>0.0) printf(" +");
32    }
33    printf("\n");
34 }
```

2. Find the interpolating polynomial using the Vandermonde matrix.

Solution. Solving

$$\begin{bmatrix} 1 & 0 & 0^2 & 0^3 \\ 1 & \pi/2 & (\pi/2)^2 & (\pi/2)^3 \\ 1 & \pi & \pi^2 & \pi^3 \\ 1 & 3\pi/2 & (3\pi/2)^2 & (3\pi/2)^3 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \end{bmatrix}$$

we obtain $a_0 = 0$, $a_1 = 1.697653$, $a_2 = -0.810569$, and $a_3 = 0.086004$ using scientific calculator and the interpolating polynomial is

$$y = 0.000000x^0 + 1.697653x^1 - 0.810569x^2 + 0.086004x^3.$$

□

p244:ex14.3.2

Example 42.5 (Tutorial 4, Q1). Find a quadratic through $(-1, a)$, $(0, b)$ and $(1, c)$. Write your answer in natural form.

Solution. We form the linear algebra problem and solve it to obtain the coefficients of the interpolating polynomial in natural form:

$$\begin{bmatrix} 1 & -1 & (-1)^2 \\ 1 & 0 & 0^2 \\ 1 & 1 & 1^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \Rightarrow y(x) = b + \frac{c-a}{2}x + \frac{c-2b+a}{2}x^2.$$

□

sec:lagrange-f

42.2.2 Lagrange Form

Another way to find the interpolating polynomial $P(x)$ for $\{(x_k, y_k)\}$ is to use the *Lagrange basis polynomials*

$$L_{n,k}(x) = \frac{(x-x_0) \cdots (x-x_{k-1})(x-x_{k+1}) \cdots (x-x_n)}{(x_k-x_0) \cdots (x_k-x_{k-1})(x_k-x_{k+1}) \cdots (x_k-x_n)} = \prod_{\substack{i=0, \\ i \neq k}}^n \frac{(x-x_i)}{(x_k-x_i)} \quad (42.4)$$

e:lagrange-bas

for $k = 0, 1, \dots, n$, such that

$$P(x) = L_{n,0}(x) \cdot y_0 + L_{n,1}(x) \cdot y_1 + \cdots + L_{n,n}(x) \cdot y_n. \quad (42.5)$$

e:lagrange-f

This was an easy consequence of a formula published in 1783 by Leonhard Euler and were discovered in 1779 by Edward Waring, and in 1795 by Joseph Louis Lagrange.

We will illustrate how to find the interpolating polynomial for a given data using (42.5) and (42.4) with the following examples.

eg:lagrange-f

Example 42.6. Given $(0, 0)$, $(\pi/2, 1)$, $(\pi, 0)$, $(3\pi/2, -1)$. Find the interpolating polynomial in the Lagrange form and determine its degree.

Solution: By applying (42.5), the interpolating polynomial in Lagrange form is obtained:

$$\begin{aligned} P(x) &= \frac{(x-\frac{\pi}{2})(x-\pi)(x-\frac{3\pi}{2})}{(0-\frac{\pi}{2})(0-\pi)(0-\frac{3\pi}{2})} \cdot 0 + \frac{(x-0)(x-\pi)(x-\frac{3\pi}{2})}{(\frac{\pi}{2}-0)(\frac{\pi}{2}-\pi)(\frac{\pi}{2}-\frac{3\pi}{2})} \cdot 1 \\ &\quad + \frac{(x-0)(x-\frac{\pi}{2})(x-\frac{3\pi}{2})}{(\pi-0)(\pi-\frac{\pi}{2})(\pi-\frac{3\pi}{2})} \cdot 0 + \frac{(x-0)(x-\frac{\pi}{2})(x-\pi)}{(\frac{3\pi}{2}-0)(\frac{3\pi}{2}-\frac{\pi}{2})(\frac{3\pi}{2}-\pi)} \cdot (-1) \\ &= \frac{4}{\pi^3}x(x-\pi)(x-\frac{3\pi}{2}) - \frac{4}{3\pi^3}x(x-\frac{\pi}{2})(x-\pi). \end{aligned}$$

Example 42.7 (April 2014 Final Exam Question 1(c)). Find the Lagrange form of the polynomial which interpolates the function $f(x) = \frac{10x+2}{x+2}$ at 0, 1 and 3. (6 marks)

Solution. At $x = 0, 1, 3$, $f(x) = 1, 4, \frac{32}{5}$ [1 mark]

$$\begin{aligned} P(x) &= \frac{(x-1)(x-3)}{(0-1)(0-3)}(1) + \frac{(x-0)(x-3)}{(1-0)(1-3)}(4) + \frac{(x-0)(x-1)}{(3-0)(3-1)}(\frac{32}{5}) \\ &= \frac{1}{3}(x-1)(x-3) - 2x(x-3) + \frac{32}{30}x(x-1) \end{aligned} \quad [5 \text{ marks}]$$

With what we learned in Topics 1 and 2 regarding Fortran, it is not difficult for us to implement (42.5) and (42.4) in Fortran as follows.

```

1 function lagrange(xv, yv, x, stat)
2   real(c_double), intent(in) :: xv(:), yv(:), x
3   real(c_double) :: lagrange, xi, lb, denom
4   integer :: i, j, stat, N
5   N = size(xv)
6   lagrange = 0.0
7   stat = 0 ! status == 0: No error
8   do i=1,N
9     lb = 1.0
10    xi = xv(i)
11    do j=1,N ! Calculate Lagrange Basis Polynomial at x
12      if (j .ne. i) then
13        denom = xi-xv(j)
14        if (abs(denom)<1d-100) then
15          stat = 1 ! Error: xv[$i] and xv[$j] are too close!
16          cycle ! ignore the current point & use the next
17        end if
18        lb = lb*(x-xv(j))/denom
19      end if
20    end do
21    lagrange = lagrange + yv(i)*lb
22  end do
23 end function lagrange

```

In the above implementation, for N points, the Lagrange basis needs two $4(N - 1)$ floating point operations at line 11 and line 20 and altogether, (42.5) requires $N(4(N - 1) + 2)$ floating point operations according to lines 8 to 22. With some algebra operations, we obtain two modified Lagrange forms.

A method developed by Eric Harold Neville, called Neville's algorithm, is based on a recursively generated Lagrange polynomial approximations.

Definition 42.4. Let f be a function defined at $x_0, x_1, \dots, x_n$, and suppose that $m_1, m_2, \dots, m_k$ are k distinct integers, with $0 \leq m_i \leq n$ for each i . The Lagrange polynomial that agrees with $f(x)$ at the k points $x_{m_1}, x_{m_2}, \dots, x_{m_k}$ is denoted as $P_{m_1, m_2, \dots, m_k}(x)$.

Theorem 42.8 (Recursively Generated Lagrange Polynomials). *Let f be defined at $x_0, x_1, \dots, x_k$ and x_j, x_i be two different numbers in this set. If*

$$P(x) = \frac{(x - x_j)P_{0,1,\dots,j-1,j+1,\dots,k}(x) - (x - x_i)P_{0,1,\dots,i-1,i+1,\dots,k}(x)}{x_i - x_j}$$

then $P(x)$ is the k th Lagrange polynomial that interpolates, or agrees with, $f(x)$ at the $k + 1$ points $x_0, x_1, \dots, x_k$.

The above theorem is the foundation for *Neville's algorithm*, which uses the *linear interpolation between successive iterations* to find the interpolating polynomial. We begin by assigning y_i (relabelling the index from 1 to $n + 1$ instead of starting from 0) which is equal to P_i , to the first column of an interpolation table $Q_{i,1}$. The recursive formula in Theorem 42.8 becomes

$$Q_{i,j} = \frac{(x - x_{i-j+1})Q_{i,j-1} - (x - x_i)Q_{i-1,j-1}}{x_i - x_{i-j+1}}$$

and a special case (when we have 5 points) of the above formula would look like

$$\begin{array}{llllll}
x_1 & y_1 = P_1 = Q_{1,1} & & & & \\
x_2 & y_2 = P_2 = Q_{2,1} & P_{1,2} = Q_{2,2} & & & \\
x_3 & y_3 = P_3 = Q_{3,1} & P_{2,3} = Q_{3,2} & P_{1,2,3} = Q_{3,3} & & \\
x_4 & y_4 = P_4 = Q_{4,1} & P_{3,4} = Q_{4,2} & P_{2,3,4} = Q_{4,3} & P_{1,2,3,4} = Q_{4,4} & \\
x_5 & y_5 = P_5 = Q_{5,1} & P_{4,5} = Q_{5,2} & P_{3,4,5} = Q_{5,3} & P_{2,3,4,5} = Q_{5,4} & P_{1,2,3,4,5} = Q_{5,5}
\end{array}$$

and a simple Fortran implementation is listed below.

```

1  ! Calc. estimated value at x using Neville's method
2  ! http://www.youtube.com/watch?v=c_m5mElHLEg
3  function neville(xv, yv, x, echo)
4      real(c_double), intent(in) :: xv(:), yv(:), x
5      logical, intent(in), optional :: echo
6      real(c_double), dimension(:,:), allocatable :: Q
7      real(c_double) :: neville, numer, denom
8      integer :: i, j, n
9      n = size(xv)
10     allocate(Q(n,n))
11     Q(:,1) = yv
12     do i=2,n
13         do j=2,i
14             numer = (x-xv(i-j+1))*Q(i,j-1)-(x-xv(i))*Q(i-1,j-1)
15             denom = xv(i)-xv(i-j+1)
16             Q(i,j) = numer / denom
17         end do
18     end do
19     if (present(echo)) call show_array(n,Q) ! defined in this module
20     neville = Q(n,n)
21 end function neville

```

In the above implementation, for $n + 1$ points, the overall floating point operations is $\frac{n^2}{2}(5 + 1 + 1) = \frac{7n^2}{2}$ floating point operations, which is half of the previous implementation.

Definition 42.5. The *modified Lagrange formula* of (42.5) is e:lagrange-f

$$P(x) = \prod_{i=0}^n (x - x_i) \left[\frac{w_0}{x - x_0} \cdot y_0 + \frac{w_1}{x - x_1} \cdot y_1 + \cdots + \frac{w_n}{x - x_n} \cdot y_n \right] \quad (42.6)$$
e:mlangrange-f

where $w_i = \prod_{i=0, i \neq k}^n \frac{1}{x_k - x_i}$. When $y_i \equiv 1$, $i = 0, 1, \dots, n$,

$$1 = \prod_{i=0}^n (x - x_i) \left[\frac{w_0}{x - x_0} + \frac{w_1}{x - x_1} + \cdots + \frac{w_n}{x - x_n} \right]. \quad (42.7)$$
e:mlangrange-1

Dividing (42.6) over (42.7), we obtain the *barycentric formula* e:mlangrange-fmlangrange-1

$$P(x) = \frac{\frac{w_0}{x - x_0} \cdot y_0 + \frac{w_1}{x - x_1} \cdot y_1 + \cdots + \frac{w_n}{x - x_n} \cdot y_n}{\frac{w_0}{x - x_0} + \frac{w_1}{x - x_1} + \cdots + \frac{w_n}{x - x_n}}. \quad (42.8)$$
e:mlangrange-f2

The modified Lagrange formula (42.6) is backward stable with respect to perturbations in the function values while the barycentric formula (42.8) is not backward stable, it is forward stable for nicely chosen interpolating points [Higham, 2004]. Unless the degree of the polynomial is low, it is usually not a good idea to use uniformly spaced interpolation points. Instead one should use Chebyshev points or other systems of points clustered at the boundary of the interval [Berrut and Trefethen, 2004].

Example 42.9. For Example 42.6, write down (a) the modified Lagrange formula; and (b) the barycentric formula. eg:lagrange-f

Solution: (a) From the Lagrange form in Example 42.6, the modified Lagrange formula is eg: lagrange-f

$$P(x) = (x-0)(x-\frac{\pi}{2})(x-\pi)(x-\frac{3\pi}{2}) \left[\frac{w_0}{x-0} \cdot 0 + \frac{w_1}{x-\frac{\pi}{2}} \cdot 1 + \frac{w_2}{x-\pi} \cdot 0 + \frac{w_3}{x-\frac{3\pi}{2}} \cdot (-1) \right]$$

where $w_0 = \frac{1}{(0-\frac{\pi}{2})(0-\pi)(0-\frac{3\pi}{2})}$, $w_1 = \frac{1}{(\frac{\pi}{2}-0)(\frac{\pi}{2}-\pi)(\frac{\pi}{2}-\frac{3\pi}{2})}$, $w_2 = \frac{1}{(\pi-0)(\pi-\frac{\pi}{2})(\pi-\frac{3\pi}{2})}$, $w_3 = \frac{1}{(\frac{3\pi}{2}-0)(\frac{3\pi}{2}-\frac{\pi}{2})(\frac{3\pi}{2}-\pi)}$.

The barycentric formula is

$$P(x) = \frac{\frac{w_0}{x-0} \cdot 0 + \frac{w_1}{x-\frac{\pi}{2}} \cdot 1 + \frac{w_2}{x-\pi} \cdot 0 + \frac{w_3}{x-\frac{3\pi}{2}} \cdot (-1)}{\frac{w_0}{x-0} + \frac{w_1}{x-\frac{\pi}{2}} + \frac{w_2}{x-\pi} + \frac{w_3}{x-\frac{3\pi}{2}}}.$$

One thing which is not nice of Lagrange form (42.5) e: lagrange-f (or its variations) is that we cannot tell the degree of the interpolating polynomial without expanding it as illustrated in the following examples.

schaum2e: 6.19 **Example 42.10** (Tutorial 4, Q2). Find a polynomial of degree two which collocates with $y(x) = \sqrt{x}$ at $x = 0, 1, 4$. Write your answer in both Lagrange form and natural form.

Solution. The interpolating polynomial “collocates” the three data points of y is

$$P_2(x) = \underbrace{0 + 1 \frac{(x-0)(x-4)}{(1-0)(1-4)}}_{\text{Lagrange form}} + 2 \underbrace{\frac{(x-0)(x-1)}{(4-0)(4-1)}}_{\text{Natural form}} = \frac{7x - x^2}{6}$$

We can see that P_2 is having a degree of 2. □

Example 42.11. Find the interpolating polynomial for the points $(-2, 21)$, $(1, 6)$, $(3, 26)$, $(2, 13)$ in the Lagrange form and determine its degree.

Solution.
$$P(x) = 21 \frac{(x-1)(x-3)(x-2)}{(-2-1)(-2-3)(-2-2)} + 6 \frac{(x+2)(x-3)(x-2)}{(1+2)(1-3)(1-2)} \\ + 26 \frac{(x+2)(x-1)(x-2)}{(3+2)(3-1)(3-2)} + 13 \frac{(x+2)(x-1)(x-3)}{(2+2)(2-1)(2-3)}$$

$$a_3 = \frac{21}{(-3)(-5)(-4)} + \frac{6}{3(-2)(-1)} + \frac{26}{5(2)(1)} + \frac{13}{4(1)(-1)} = 0$$

$$a_2 = \frac{21}{(-3)(-5)(-4)}(-1-3-2) + \frac{6}{3(-2)(-1)}(2-3-2) \\ + \frac{26}{5(2)(1)}(2-1-2) + \frac{13}{4(1)(-1)}(2-1-3) = 3 \neq 0$$

Therefore the degree of the interpolating polynomial is 2. □

42.2.3 Application of Lagrange Form

The interpolating polynomial e: lagrange-f (42.5) is normally use to approximate a function f in a range $[a, b]$ containing the collection of points $\{x_k\}$. Therefore, we are interested in estimating the difference $f(x) - P(x)$ over $[a, b]$. Without further hypotheses, there is no way we can quantify this difference. However, if f is sufficiently “smooth”, i.e. $f \in C^{n+1}$, then we can have the following *error characterisation theorem* [Henrici, 1982].

Theorem 42.12. Suppose $x_0, \dots, x_n$ are distinct numbers in the interval $[a, b]$ and $f \in C^{n+1}[a, b]$. Then, for each x in $[a, b]$, a number $\xi(x)$ (generally unknown) in (a, b) exists with

$$f(x) = P(x) + \frac{f^{(n+1)}(\xi(x))}{(n+1)!} (x-x_0)(x-x_1) \cdots (x-x_n). \quad (42.9)$$

where $P(x)$ is the interpolating polynomial defined by e: lagrange-f (42.5) and e: lagrange-basis (42.4).

e: lagrange-error

This error formula (42.9) for $f(x) - P(x)$ is an important theoretical result because Lagrange polynomials are used extensively in numerical differentiation and integration methods. For example, the error formula (42.9) could be applied to obtain general derivative formulae for numerical differentiation in Topic 3. Suppose that $x_0, \dots, x_n$ are $n+1$ distinct numbers in some interval $[a, b]$ and that $f \in C^{n+1}[a, b]$. Differentiating $f(x)$ in (42.9) leads to

$$f'(x) = \sum_{j=0}^n f(x_j) L'_{n,j}(x) + \frac{d}{dx} \left[\frac{f^{(n+1)}(\xi(x))}{(n+1)!} \right] \prod_{i=0}^n (x - x_i) + \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \frac{d}{dx} \prod_{i=0}^n (x - x_i).$$

When x is one of the numbers x_k , $k = 0, \dots, n$, the second term becomes zero, the third term can be simplified by substituting x_k into the product formula, and $f'(x)$ becomes

$$f'(x_k) = \sum_{j=0}^n f(x_j) L'_j(x_k) + \underbrace{\frac{f^{(n+1)}(\xi(x_k))}{(n+1)!} \prod_{\substack{j=0 \\ j \neq k}}^n (x_k - x_j)}_{\text{Error term}}$$

which is called an $(n+1)$ -point approximation of $f'(x_k)$.

Theoretically, using more evaluation points produces greater accuracy, however the number of functional evaluations and growth of round-off error discourages this somewhat. The most common formulae are those involving three and five evaluation points.

When $x_i = x_0 + ih$, $i = 0, 1, 2$ we have the *three-point endpoint formula*:

$$\begin{aligned} f'(x_0) &= f(x_0) \frac{d}{dx} \left[\frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} \right]_{x=x_0} + \dots + f(x_2) \frac{d}{dx} \left[\frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} \right]_{x=x_0} \\ &\quad + \frac{f'''(\xi_1)}{3!} (x_0 - x_1)(x_0 - x_2) \\ &= f(x_0) \left[\frac{2x_0 - (x_1 + x_2)}{(-h)(-2h)} \right] + \dots + f(x_2) \left[\frac{2x_0 - (x_0 + x_2)}{(h)(-h)} \right] + \frac{f'''(\xi_1)}{3!} (-h)(-2h) \\ &= f(x_0) \left[\frac{-h - 2h}{(-h)(-2h)} \right] + f(x_1) \left[\frac{-2h}{(h)(-h)} \right] + f(x_2) \left[\frac{-2h}{(2h)(h)} \right] + \frac{h^2}{3} f'''(\xi_1) \\ &= \frac{1}{2h} \left[-3f(x_0) + 4f(x_0 + h) - f(x_0 + 2h) \right] + \frac{h^2}{3} f'''(\xi_1) \end{aligned}$$

for some $\xi_1 \in (x_0, x_2)$.

When $x_i = x_0 + ih$, $i = -1, 0, 1$ we have the *three-point midpoint formula*:

$$f'(x_0) = \frac{1}{2h} \left[-f(x_0 - h) + f(x_0 + h) \right] + \frac{h^2}{6} f'''(\xi_2)$$

for some $\xi_2 \in (x_{-1}, x_1)$.

When $x_i = x_0 + ih$, $i = -2, -1, 0$ we have the *three-point endpoint formula*:

$$f'(x_0) = \frac{1}{2h} \left[f(x_0 - 2h) - 4f(x_0 - h) + 3f(x_0) \right] + \frac{h^2}{3} f'''(\xi_3)$$

for some $\xi_3 \in (x_{-2}, x_0)$.

Five-point formulae are even more complex and can be found in Burden and Faires [2015].

Remark 42.13. All differentiation formulae will face round-off error problems because division by small numbers tends to exaggerate round-off error. In the case of numerical differentiation, it is impossible to avoid the problem entirely, although the higher-order methods reduce the difficulty.

As approximations, numerical differentiation is *unstable*, since small values of h needed to reduce truncation error also cause the round-off error to grow. However, in addition to being used for computational purposes, the formulae are needed for approximating the solutions of ODEs (Chapters 45, 46) and PDEs (Chapter 47).

42.2.4 Newton Form and Newton Forward Divided-Differences

In this section, the *Newton form* of interpolating polynomial $P(x)$, which allows us to study information concerning differences of sequence of function values, is discussed.

Assume that the values y_i of $n+1$ points x_i are defined by a function f such that $y_i = f(x_i)$. Let $f[x_i] = f(x_i)$ for $i = 0, \dots, n$, the concept of *divided difference* is defined as:

$$f[x_i, x_{i+1}] = \frac{f[x_{i+1}] - f[x_i]}{x_{i+1} - x_i}, \quad i = 0, \dots, n-1,$$

and the n th *divided difference* is defined recursively as

$$f[x_0, x_1, \dots, x_n] = \frac{f[x_1, \dots, x_n] - f[x_0, \dots, x_{n-1}]}{x_n - x_0}.$$

With this notation, the interpolation polynomial for f with respect to $x_0, \dots, x_n$ can be expressed as

$$\begin{aligned} P(x) &= f[x_0] + f[x_0, x_1](x - x_0) + \dots + f[x_0, x_1, \dots, x_n](x - x_0) \dots (x - x_{n-1}) \\ &= f[x_0] + \sum_{i=1}^n f[x_0, x_1, \dots, x_i](x - x_0) \dots (x - x_{i-1}). \end{aligned} \quad (42.10)$$

e:newton-f

It is called the *Newton's (divided-difference) formula* or the *interpolating polynomial in Newton form*. A proof is given in Hoffman [1992, Chapter 4]. It is based on the verification of $P(x_i) = y_i$ for $i = 1, 2, \dots, n$ using the definition of divided difference. Another more complicated proof is given in Isaacson and Keller [1966, Chapter 6].

When $\{x_i\}$ are equally spaced, i.e. $x_i - x_{i-1} = h$ or $x_i = x_0 + ih$ for $i = 1, \dots, n$, then there is some real number s such that $x = x_0 + sh \in [x_0, x_n]$ and the Newton form (42.10) can be written as the *Newton forward divided-difference formula (NDD)*:

$$\begin{aligned} P(x) &= f[x_0] + f[x_0, x_1](sh) + \dots + f[x_0, x_1, \dots, x_n]s(s-1) \dots (s-n+1)h^n \\ &= f[x_0] + \sum_{i=1}^n f[x_0, \dots, x_i] \binom{s}{i} i! h^i = f[x_0] + \underbrace{\sum_{i=1}^n \binom{s}{i} \Delta^i f(x_0)}_{\text{NDD}} \end{aligned}$$

by using the finite difference notation from Topic 3

$$f[x_0, x_1, \dots, x_i] = \frac{\Delta^i f(x_0)}{i! h^i}$$

and the *generalised binomial-coefficient* formula:

$$\binom{s}{i} = \frac{s(s-1) \dots (s-i+1)}{i \times \dots \times 1}.$$

A sample computer program to find the coefficients of the Newton's formula (42.10) and stored in an array is shown below.

```

1 function newton_dd(xv, yv, echo)
2   real(c_double), intent(in) :: xv(:), yv(:)
3   real(c_double), dimension(:), allocatable :: newton_dd
4   logical, intent(in), optional :: echo
5   real(c_double), dimension(:,:), allocatable :: Q
6   integer :: i, j, n
7   n = size(xv)
8   allocate(Q(n,n))
9   Q(:,1) = yv
10  do i=2,n
11    do j=2,i
12      Q(i,j) = (Q(i,j-1) - Q(i-1,j-1)) / (xv(i) - xv(i-j+1))

```

e:newton-f

```

13         end do
14     end do
15     if (present(echo)) call show_array(n,Q)
16     allocate(newton_dd(n))
17     newton_dd = ((Q(i,i), i=1,n)/)
18 end function newton_dd

```

Example 42.14. Apply Newton's formula to find a polynomial of degree four or less which takes the values $f(0) = 1$, $f(1) = -1$, $f(2) = 1$, $f(3) = -1$, $f(4) = 1$. Use a computer program to verify your answer.

Solution: The finite divided-difference table is

$$\left\{ \begin{array}{cccccc} 0 & 1 & & & & \\ 1 & -1 & -2 & & & \\ 2 & 1 & 2 & 2 & & \\ 3 & -1 & -2 & -2 & -\frac{4}{3} & \\ 4 & 1 & 2 & 2 & \frac{4}{3} & \frac{2}{3} \end{array} \right.$$

and the interpolating polynomial is

$$P_4(x) = 1 - 2(x-0) + 2(x-0)(x-1) - \frac{4}{3}(x-0)(x-1)(x-2) + \frac{2}{3}(x-0)(x-1)(x-2)(x-3).$$

The following program verifies our calculation above:

```

1  #include "textbook_interp.h"    // newton_dd
2
3  int main() {
4      arma::vec xv {0., 1., 2., 3., 4.};
5      arma::vec yv {1.0, -1.0, 1.0, -1.0, 1.0};
6      arma::vec coeff = newton_dd(xv, yv, 1);
7      coeff.print("coeff=");
8  }

```

ex:1

Example 42.15. Given the points $(1.0, 0.7651977)$, $(1.3, 0.6200860)$, $(1.6, 0.4554022)$, $(1.9, 0.2818186)$, $(2.2, 0.1103623)$. Find the interpolating polynomial $P(x)$ in Newton form.

Solution: Complete the following divided-difference table

i	x_i	y_i	$\Delta f/h$	$\Delta^2 f/h^2$	$\Delta^3 f/h^3$	$\Delta^4 f/h^4$
0	1.0	0.7651977				
1	1.3	0.6200860	-0.48370567			
2	1.6	0.4554022	-0.54894600	-0.10873389		
3	1.9	0.2818186	-0.57861200	-0.04944333	0.06587840	
4	2.2	0.1103623	-0.57152100	0.01181833	0.06806852	0.00182510

From the divided-difference table, we have

$$\begin{aligned}
 P_4(x) = & 0.7651977 - 0.4837057(x-1.0) - 0.1087339(x-1.0)(x-1.3) \\
 & + 0.0658784(x-1.0)(x-1.3)(x-1.6) \\
 & + 0.0018251(x-1.0)(x-1.3)(x-1.6)(x-1.9).
 \end{aligned}$$

schaum2e:6.16

Example 42.16 (Tutorial 4, Q4). Given $y(x) = \sin(\frac{\pi x}{2})$. Is there a polynomial of degree 4 which collocates with $y(x)$ at $x = 0, 1, 2, 3, 4$? Justify your answer.

Solution. From the divided-difference table:

$$\begin{array}{ccccccc} & 0 & 0 & & & & \\ & 1 & 1 & 1 & & & \\ & 2 & 0 & -1 & -1 & & \\ & 3 & -1 & -1 & 0 & \frac{1}{3} & \\ & 4 & 0 & 1 & 1 & \frac{1}{3} & 0 \end{array}$$

we learn that the interpolating polynomial can only be degree 3. $\square$

t3:q1

Example 42.17 (Tutorial 4, Q5). For the given function $f(x) = \sqrt{1+x}$, let $x_0 = 0$, $x_1 = 0.6$, $x_2 = 0.9$. Construct interpolation polynomials of degree at most one and at most two to approximate $f(0.45)$, and find the absolute error.

Solution. Based on the three points $(0, 1)$, $(0.6, 1.26491106)$, $(0.9, 1.37840488)$, we can obtain the following difference table

0	1		
0.6	1.26491106	0.44151843	
0.9	1.37840488	0.37831273	-0.07022856

$$P_1(0.45) = 1.19868329 \quad P_2(0.45) = 1.20342372.$$

Since $f(0.45) = 1.20415946$,

$$|P_1(0.45) - f(0.45)| = 0.00547617, \quad |P_2(0.45) - f(0.45)| = 0.00073574.$$

$$P_1(x) = 1 + 0.44151843(x - 0)$$

$$P_2(x) = 1 + 0.44151843(x - 0) - 0.07022856(x - 0)(x - 0.6)$$

$\square$

Example 42.18 (Tutorial 4, Q3). Find an interpolating polynomial for the following specification:

x	-2	-1	0	1	2
$f(x)$	-1	3	1	-1	3

- Write the interpolating polynomial in Lagrange form.

Solution.

$$\begin{aligned} L(x) = & -\frac{(x+1)x(x-1)(x-2)}{(-2+1)(-2)(-2-1)(-2-2)} + 3\frac{(x+2)x(x-1)(x-2)}{(-1+2)(-1)(-1-1)(-1-2)} \\ & + \frac{(x+2)(x+1)(x-1)(x-2)}{(0+2)(0+1)(0-1)(0-2)} - \frac{(x+2)(x+1)x(x-2)}{(1+2)(1+1)(1-2)} \\ & + 3\frac{(x+2)(x+1)x(x-1)}{(2+2)(2+1)(2-0)(2-1)} \end{aligned}$$

$\square$

- Write the interpolating polynomial in Newton form.

Solution.

-2	-1				
-1	3	4			
0	1	-2	-3		
1	-1	-2	0	1	
2	3	4	3	1	0

$$N(x) = -1 + 4(x+2) - 3(x+2)(x+1) + (x+2)(x+1)(x)$$

$\square$

- Show that the cubic polynomial

$$P(x) = 3 - 2(x+1) + 0(x+1)(x) + (x+1)(x)(x-1)$$

is also an interpolating polynomial for the data.

Solution. By substituting all the value into P , we find that $P(x_i) = f(x_i)$ matches for all x_i . Therefore $P(x)$ is an interpolating polynomial by interpolation theorem. $\square$

Example 42.19 (Tutorial 4, Q7). For a function f , the forward divided differences are given by

$x_0 = 0.0$	$f[x_0]$		
		$f[x_0, x_1]$	
$x_1 = 0.4$	$f[x_1]$		$f[x_0, x_1, x_2] = \frac{50}{7}$
		$f[x_1, x_2] = 10$	
$x_2 = 0.7$	$f[x_2] = 6$		

Determine the missing entries in the table.

Solution.

$$\frac{10 - f[x_0, x_1]}{0.7 - 0.0} = \frac{50}{7} \Rightarrow f[x_0, x_1] = 5$$

$$\frac{6 - f[x_1]}{0.7 - 0.4} = 10 \Rightarrow f[x_1] = 6 - 3 = 3$$

$$\frac{3 - f[x_0]}{0.4 - 0.0} = 5 \Rightarrow f[x_0] = 1$$

$\square$

Example 42.20 (April 2014 Final Exam Question 3(a)). Use Newton's divided-difference formula to construct the interpolating polynomial of degree 3 for

$$f(8.1) = 16.94410, f(8.3) = 17.56492, f(8.6) = 18.50515, f(8.7) = 18.82091$$

and approximate the value $f(8.4)$. (10 marks)

Solution.

8.1	16.94410				
8.3	17.56492	3.10410			
8.6	18.50515	3.13410	0.06000		 [6 marks]
8.7	18.82091	3.15760	0.05875	-0.0020833	

The Newton's divided-difference formula is

$$P(x) = 16.94410 + 3.10410(x - 8.1) + 0.06(x - 8.1)(x - 8.3) - 0.0020833(x - 8.1)(x - 8.3)(x - 8.6)$$

[2 marks]

Hence, $f(8.4) \approx P(8.4) = 17.87714$ [2 marks] $\square$

Example 42.21 (April 2018 Final Exam Question 2). Given the Malaysian Government Securities (MGS) as of 26 January 2018 to be approximated by the following term structure (Table 1):

Maturity, T (year)	1	2	5	15	28
Yield rate, $y(T)$ (percent)	3.74	3.48	3.50	4.10	4.79

Table 1: Term Structure

- (a) Derive the forward Newton formula of the interpolating polynomial $P(T)$ for the above data. Perform all calculations to at least 6 decimal places. Here the variable T is the maturity period. (14 marks)

Solution. To derive the forward Newton formula, we construct the table

1	3.74				
2	3.48	-0.260000			
5	3.50	0.006667	0.066667		
15	4.10	0.060000	0.004103	-0.004469	
28	4.79	0.053077	-0.000301	-0.000169	0.000159

..... [1+4+3+2+1=11 marks]

The forward Newton formula is

$$P(T) = 3.74 - 0.260000(T - 1) + 0.066667(T - 1)(T - 2) \\ - 0.004469(T - 1)(T - 2)(T - 5) + 0.000159(T - 1)(T - 2)(T - 5)(T - 15)$$

[3 marks]
□

- (b) Use the formula $P(T)$ in part (a) to estimate $y(1.5)$ to at least 6 decimal places. (4 marks)

Solution. From part (a),

$$P(1.5) = 3.74 - 0.13 - 0.016667 - 0.003910 - 0.001878 \\ = 3.587545$$

[4 marks]
□

- (c) A mathematical model of term structure normally requires $|y'(T)| < 1$ for $T > 10$. Is the interpolating polynomial $P(T)$ in part (a) suitable for this? Justify your answer. (2 marks)

Solution. No. The interpolating polynomial is not suitable for this because after differentiating $P(T)$ with respect to T , the degree of the derivative is 3, when T is large, $P'(T)$ is large and hence $|P'(T)| > 1$ [2 marks]
□

[Total : 20 marks]

42.2.5 Newton Backward Divided Differences

If the interpolating nodes are rearranged from last to first as $x_n, \dots, x_0$, we can write the interpolating formula as

$$P(x) = f[x_n] + f[x_n, x_{n-1}](x - x_n) + \dots + f[x_n, \dots, x_0](x - x_n)(x - x_{n-1}) \dots (x - x_1).$$

If the nodes are equally spaced with $x = x_n + sh$ and $x = x_i + (s + n - i)h$, then

$$P(x) = f[x_n] + f[x_n, x_{n-1}](sh) + f[x_n, x_{n-1}, x_{n-2}]s(s+1)h^2 + \dots + f[x_n, \dots, x_0]s(s+1) \dots (s+n-1)h^n.$$

This form is called the *Newton backward divided-difference* formula.

By introducing *backward difference* notation (mentioned in Topic 3) for a sequence $\{p_k\}_{k=0}^{\infty}$ by

$$\nabla p_k = p_k - p_{k-1}, \quad k \geq 1$$

and *higher order backward difference* notation defined recursively by

$$\nabla^m p_k = \nabla(\nabla^{m-1} p_k), \quad m \geq 2, \quad k \geq m,$$

we obtain

$$f[x_n, x_{n-1}, \dots, x_{n-i}] = \frac{1}{i!h^i} \nabla^i f(x_n).$$

Using the above notation, the Newton backward divided-difference can be written as the *Newton backward-difference formula*

$$P(x) = f[x_n] + s\nabla f(x_n) + \frac{s(s+1)}{2} \nabla^2 f(x_n) + \dots + \frac{s(s+1) \dots (s+n-1)}{n!} \nabla^n f(x_n) \\ = f[x_n] + \sum_{k=1}^n (-1)^k \binom{-s}{k} \nabla^k f(x_n).$$

Example 42.22. For Example ^{ex:1}42.15, find an approximation to $f(1.1)$ by using the forward divided-difference formula and find an approximation to $f(2.0)$ by using the backward divided-difference formula.

Solution:

$$\begin{aligned} P(1.1) &= P(1.0 + \tfrac{1}{3}(0.3)) \\ &= 0.7651977 + \tfrac{1}{3}(0.3)(-0.4837057) + \tfrac{1}{3}(-\tfrac{2}{3})(0.3)^2(-0.1087339) \\ &\quad + \tfrac{1}{3}(-\tfrac{2}{3})(-\tfrac{5}{3})(0.3)^3(0.0658784) + \tfrac{1}{3}(-\tfrac{2}{3})(-\tfrac{5}{3})(-\tfrac{8}{3})(0.3)^4(0.0018251) = 0.7196460. \end{aligned}$$

To approximate a value when x is close to the end of the tabulated values, e.g. $x = 2.0$, we would like to make earlier use of data points closest to x . To do so requires using the Newton backward divided-difference formula with $x_4 = 2.2$, ..., $x_0 = 1.0$, and $s = -\frac{2}{3}$:

$$\begin{aligned} P(2.0) &= P(2.2 - \tfrac{2}{3}(0.3)) \\ &= 0.1103623 - \tfrac{2}{3}(0.3)(-0.5715210) - \tfrac{2}{3}(\tfrac{1}{3})(0.3)^2(0.0118183) \\ &\quad - \tfrac{2}{3}(\tfrac{1}{3})(\tfrac{4}{3})(0.3)^3(0.0658785) - \tfrac{2}{3}(\tfrac{1}{3})(\tfrac{4}{3})(\tfrac{7}{3})(0.3)^4(0.0018251) = 0.2238754. \end{aligned}$$

Example 42.23 (Tutorial 4, Q6). Write down the Newton forward divided-difference formula and backward divided-difference formula for the following data:

$$\begin{aligned} f(-0.75) &= -0.07181250, & f(-0.5) &= -0.02475000, \\ f(-0.25) &= 0.33493750, & f(0) &= 1.10100000. \end{aligned}$$

and approximate the specified value $f(-\frac{1}{3})$ using each of the polynomials.

Solution. The divided-difference table is

-0.75	-0.07181250			
-0.5	-0.02475000	0.188250		
-0.25	0.33493750	1.438750	2.501000	
0	1.10100000	3.064250	3.251000	1.000000

From the table, we can obtain the forward divided-difference formula

$$\begin{aligned} P_f(x) &= -0.07181250 + 0.188250(x + 0.75) + 2.501000(x + 0.75)(x + 0.5) \\ &\quad + (x + 0.75)(x + 0.5)(x + 0.25). \end{aligned}$$

and the backward divided-difference formula

$$P_b(x) = 1.10100000 + 3.064250x + 3.251000x(x + 0.25) + x(x + 0.25)(x + 0.5)$$

Therefore, the approximation, $P_f(-\frac{1}{3}) = P_b(-\frac{1}{3}) =$ 0.17451852

□

Remark 42.24. The Newton divided-difference methods, whether forward, backward or central, which can be found in Burden and Faires [2015, p132], are used when computers are not available. To carry out hand calculation, sometimes it is more convenient to calculate from the back or centre to gain better accuracy.

42.2.6 Runge Phenomenon

sec:runge

When Carl David Tolmé Runge (1856–1927) was exploring the behaviour of errors when using polynomial interpolation to approximate certain functions, he found that oscillation happened at the edges of an interval that occurs when using polynomial interpolation with polynomials of high degree over a set of equi-spaced interpolation points [Schlömlich et al., 1901].

We will investigate the “Runge phenomenon” through the study of a particular function R in the following example.

eg:runge

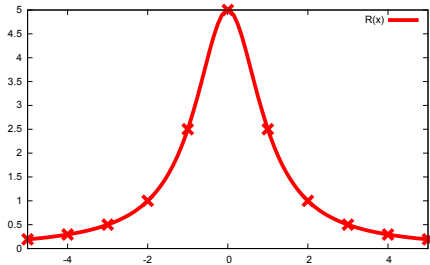
Example 42.25. Given $R(x) = \frac{5}{1+x^2}$, $-5 \leq x \leq 5$. Divide the interval $[-5, 5]$ to 10 equal subintervals and let $P(x)$ be the interpolating polynomial of the function R .

1. What is the order of $P(x)$? What is the possible degree of $P(x)$?

Solution: There are 11 points over the 10 equal subintervals, therefore the order is 11 and the most likely degree for $P(x)$ is 10.

2. Use your favourite mathematical software to plot $R(x)$ over the interval $[-5, 5]$.

Solution:



3. Write a program to plot $P(x)$ and $|R(x) - P(x)|$.

Solution: We will be using our interpolation module `textbook_interp` and Fortran plotting module `plplot`:

```

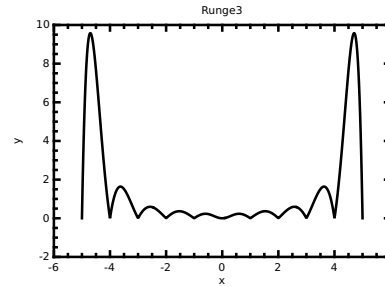
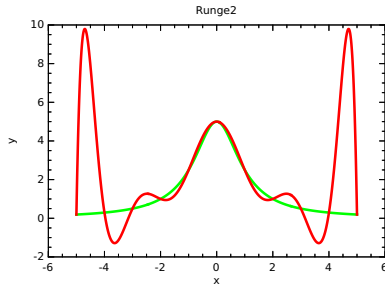
program runge2
  use iso_c_binding, only : c_double
  use plplot ! -lplplotf95d -lplplotf95cd
  use textbook_interp, only : lagrange
  real(c_double), parameter :: ileft=-5d0, irigh=5d0, dx=1d-2
  real(c_double), external :: R
  integer, parameter :: N = 10, NN = nint((irigh-ileft)/dx)
  real(c_double) :: xvals(0:N), yvals(0:N)
  real(c_double) :: xr(0:NN), yr(0:NN), ir(0:NN), er(0:NN)
  integer :: j, info
  xvals = (/(-5d0+j, j=0,N)/)
  yvals = (/(R(xvals(j)), j=0,N)/)
  do j=0,NN
    xr(j) = ileft + j*dx
    yr(j) = R(xr(j))
    ir(j) = lagrange(xvals, yvals, xr(j), info)
    er(j) = dabs(yr(j)-ir(j))
  end do
  call plscolbg(255,255,255)
  call plparseopts( PL_PARSE_FULL )
  call plsfnam("runge2.eps")
  call plinit
  call plscol0(15,0,0,0) ! Changing background and front colour
  call plcol0(15) !15=black
  ! Create a labelled box to hold the plot.
  call plenv(-6d0, 6d0, -2d0, 10d0, 0, 0 )
  call pllabb( "x", "y", "Runge2" )
  call plcol0(3) ! 3=green
  call plwidth(5d0)
  call plline(xr, yr) ! The function R
  call plcol0(1) ! 1=red
  call plline(xr, ir) ! The interpolation with 11 points
  call plend
  call plscolbg(255,255,255)
  call plsfnam("runge3.eps")
  call plinit
  call plscol0(15,0,0,0)
  call plcol0(15) !15=black
  call plwidth(5d0)
  call plenv(-6d0, 6d0, -2d0, 10d0, 0, 0 )
  call pllabb( "x", "y", "Runge3" )
  call plline(xr, er)

```

```

call plend
end program runge2
function R(x)
  use iso_c_binding, only : c_double
  real(c_double) :: x, R
  R = 5d0/(1d0+x**2)
end function

```



The Runge phenomenon is an important discovery because it shows that going for higher order polynomial does not improve accuracy!

42.3 Hermite Interpolation

sec:hermite

Osculating polynomials generalise the Lagrange polynomials. They not only agree in value with a given function at specified points, which is the idea of collocation, but their *derivatives* up to some order must also match the derivatives of a given function.

Definition 42.6. Let $x_0, \dots, x_n$ be $n+1$ distinct values in $[a, b]$ and m_i be a nonnegative integer associated with x_i , for $i = 0, 1, \dots, n$. Suppose that $f \in C^m[a, b]$, where $m = \max_{0 \leq i \leq n} m_i$. The *osculating polynomial* approximating f is the polynomial $P(x)$ of least degree such that

$$\frac{d^k P(x_i)}{dx^k} = \frac{d^k f(x_i)}{dx^k}, \quad \text{for each } k = 0, 1, \dots, m_i \text{ and } i = 0, 1, \dots, n.$$

Remark 42.26. The degree of this osculating polynomial is at most $M = n + \sum_{i=0}^n m_i$.

When $n = 0$, the osculating polynomial approximating f is the m_0 th Taylor polynomial for f at x_0 (by taking the limit of the divided differences as $h \rightarrow 0$). When $m_i = 0$, the osculating polynomial becomes the n th Lagrange polynomial interpolating f on $x_0, x_1, \dots, x_n$.

When $m_i = 1$ for $i = 0, \dots, n$ the osculating polynomials become the *Hermite polynomials*. An Hermite polynomial not only agrees with the function f but also the first derivative at specified points. Suppose that $n+1$ data points and values of function, $(x_0, f(x_0)), (x_1, f(x_1)), \dots, (x_n, f(x_n))$, are given, $x_i \in [a, b]$, $f \in C^1[a, b]$. The Lagrange polynomial agreeing with f will generally have degree n . Hermite polynomial agree with derivative of f at $x_0, \dots, x_n$. This additional $n+1$ conditions raise the expected degree of Hermite polynomial to $2n+1$.

thm:hermite

Theorem 42.27. Suppose $f \in C^1[a, b]$ and that $x_0, \dots, x_n$ in $[a, b]$ are distinct. The unique polynomial of least degree agreeing with f and f' at $x_0, \dots, x_n$ is the polynomial of degree at most $2n+1$ given by

$$H_{2n+1}(x) := H_{2n+1}(x; x_0, \dots, x_n, f) = \sum_{j=0}^n f(x_j) H_{n,j}(x) + \sum_{j=0}^n f'(x_j) \hat{H}_{n,j}(x) \quad (42.11)$$

e:hermite

where

$$H_{n,j}(x) = [1 - 2(x - x_j)L'_{n,j}(x_j)]L_{n,j}^2(x), \quad \hat{H}_{n,j}(x) = (x - x_j)L_{n,j}^2(x)$$

and $L_{n,j}(x)$ is the Lagrange basis [\(42.4\)](#).

Moreover, if $f \in C^{2n+2}[a, b]$, then

$$f(x) = H_{2n+1}(x) + \underbrace{\frac{f^{(2n+2)}(\xi(x))}{(2n+2)!}(x-x_0)^2 \cdots (x-x_n)^2}_{\text{Error term}}$$

for some $\xi(x)$ in the interval (a, b) .

[eg:hermite](#)

Example 42.28. [Burden and Faires, 2015, p138] Find the Hermite polynomial that meets the specification listed below and find an approximation to $f(1.5)$.

k	x_k	$f(x_k)$	$f'(x_k)$
0	1.3	0.6200860	-0.5220232
1	1.6	0.4554022	-0.5698959
2	1.9	0.2818186	-0.5811571

Solution:

$$H_5(x) = 0.6200860H_{2,0}(x) + 0.4554022H_{2,1}(x) + 0.2818186H_{2,2}(x) \\ - 0.5220232\hat{H}_{2,0}(x) - 0.5698859\hat{H}_{2,1}(x) - 0.5811571\hat{H}_{2,2}(x)$$

where

$$L_{2,0}(x) = \frac{(x-1.6)(x-1.9)}{(1.3-1.6)(1.3-1.9)} = \frac{50}{9}x^2 - \frac{175}{9}x + \frac{152}{9}$$

$$L'_{2,0}(x) = \frac{100}{9}x - \frac{175}{9}$$

$$L_{2,1}(x) = \frac{(x-1.3)(x-1.9)}{(1.6-1.3)(1.6-1.9)} = -\frac{100}{9}x^2 + \frac{320}{9}x - \frac{247}{9}$$

$$L'_{2,1}(x) = -\frac{200}{9}x + \frac{320}{9}$$

$$L_{2,2}(x) = \frac{(x-1.3)(x-1.6)}{(1.9-1.3)(1.9-1.6)} = \frac{50}{9}x^2 - \frac{145}{9}x - \frac{104}{9}$$

$$L'_{2,2}(x) = \frac{100}{9}x - \frac{145}{9}$$

$$H_{2,0}(x) = \left[1 - 2(x-1.3)\left(\frac{100}{9}(1.3) - \frac{175}{9}\right)\right] \left(\frac{50}{9}x^2 - \frac{175}{9}x + \frac{152}{9}\right)^2 = (10x-12)\left(\frac{50}{9}x^2 - \frac{175}{9}x + \frac{152}{9}\right)^2$$

$$H_{2,1}(x) = 1 \cdot \left(-\frac{100}{9}x^2 + \frac{320}{9}x - \frac{247}{9}\right)^2$$

$$H_{2,2}(x) = 10(2-x) \cdot \left(\frac{50}{9}x^2 - \frac{145}{9}x - \frac{104}{9}\right)^2$$

$$\hat{H}_{2,0}(x) = (x-1.3)\left(\frac{50}{9}x^2 - \frac{175}{9}x + \frac{152}{9}\right)^2$$

$$\hat{H}_{2,1}(x) = (x-1.6)\left(-\frac{100}{9}x^2 + \frac{320}{9}x - \frac{247}{9}\right)^2$$

$$\hat{H}_{2,2}(x) = (x-1.9)\left(\frac{50}{9}x^2 - \frac{145}{9}x - \frac{104}{9}\right)^2$$

So at $x = 1.5$,

$$H_5(1.5) = 0.6200860(4/27) + 0.4554022(64/81) + 0.2818186(5/81) \\ - 0.5220232(4/405) - 0.5698859(-32/405) - 0.5811571(-2/405) = 0.5118277.$$

Although the Hermite formula [\(42.11\)](#) provides a complete description of the Hermite polynomials, the need to determine and evaluate the Lagrange polynomials and their derivatives

makes the procedure tedious even for small values of n . An alternative method for generating Hermite approximations has as its basis the Newton divided-difference formula: If $f \in C^1[a, b]$, $x_0, \dots, x_n$ are distinct in $[a, b]$, then

$$H_{2n+1}(x) = f[z_0] + \sum_{k=1}^{2n+1} f[z_0, \dots, z_k](x - z_0)(x - z_1) \cdots (x - z_{k-1})$$

where $z_{2k} = z_{2k+1} = x_k$ and $f[z_{2k}, z_{2k+1}] = f'(z_{2k}) = f'(x_k)$, for $k = 0, 1, \dots, n$.

The method for calculating Hermite interpolating polynomial $H_{2n+1}(x)$ based on “Newton divided-difference formula” is shown in the following Fortran program.

```

1  subroutine hermite_dd(xv,yv,yp,Q)
2      real(c_double), intent(in) :: xv(:), yv(:), yp(:)
3      real(c_double), allocatable :: z(:), Q(:, :)
4      integer :: n, i, j
5      n = size(xv)
6      allocate(z(2*n))
7      allocate(Q(2*n, 2*n))
8      Q = 0d0
9      do i=1,n
10         z(2*i-1) = xv(i)
11         z(2*i)   = xv(i)
12         Q(2*i-1,1) = yv(i)
13         Q(2*i, 1) = yv(i)
14         Q(2*i, 2) = yp(i)
15     end do
16     do i=1,n-1
17         Q(2*i+1,2) = (Q(2*i+1,1) - Q(2*i,1)) / (z(2*i+1) - z(2*i))
18     end do
19     do i=3,2*n
20         do j=3,i
21             Q(i,j) = (Q(i,j-1) - Q(i-1,j-1)) / (z(i) - z(i-j+1))
22         end do
23     end do
24     call show_array(n,Q)
25 end subroutine hermite_dd

```

Let us analyse how the method works through an example.

Example 42.29. Find the Hermite interpolating polynomial in forward divided-difference form and recalculate $f(1.5)$ in Example 42.28, now using divided-difference formulation.

Solution: We can use a program to generate the following divided-difference table:

1.3	0.6200860					
1.3	0.6200860	-0.5220232				
1.6	0.4554022	-0.5489459	-0.0897424			
1.6	0.4554022	-0.5698959	-0.0698333	0.0663637		
1.9	0.2818186	-0.5786121	-0.0290539	0.0679657	0.0026699	
1.9	0.2818186	-0.5811571	-0.0084834	0.0685684	0.0010045	-0.0027758

From the table’s diagonal, we can construct the Hermite polynomial for the data points:

$$\begin{aligned}
 H(x) = & 0.6200860 - 0.5220232(x - 1.3) - 0.0897424(x - 1.3)^2 \\
 & + 0.0663637(x - 1.3)^2(x - 1.6) + 0.0026699(x - 1.3)^2(x - 1.6)^2 \\
 & - 0.0027758(x - 1.3)^2(x - 1.6)^2(x - 1.9).
 \end{aligned}$$

When $x = 1.5$, $H(1.5) = 0.51182772$.

Example 42.30. A switching path between parallel railroad tracks is to be a cubic polynomial joining positions $(0, 0)$ and $(4, 2)$ and tangent to the lines $y = 0$, $y = 2$. Find the Hermite’s polynomial for this curve.

Solution. It is easy to obtain the divided-difference table

0		0			
0		0	0		
4		2	$\frac{1}{2}$	$\frac{1}{8}$	
4		2	0	$-\frac{1}{8}$	$-\frac{1}{16}$

and the Hermite polynomial

$$H(x) = \frac{1}{8}x^2 - \frac{1}{16}x^2(x-4) = \frac{x^2}{16}(6-x)$$

□

Remark 42.31. In Example 42.30, the significance of this switching path is that it provides a smooth journey. Being tangent to both of the parallel tracks, there are no sudden changes of direction, no corners. However, there are discontinuities in the curvature.

Remark 42.32. One can choose piecewise cubic polynomials that satisfy Hermite interpolation conditions (PCHIP or Piecewise Cubic Hermite Interpolating Polynomials in Octave/Matlab). That is, the function values and derivatives are specified at each nodal point. If we don't actually have derivative values prescribed at the nodal points, then we can assign these values to satisfy additional constraints. We gain this flexibility at the cost of some differentiability; piecewise cubic Hermite interpolants are in general not twice continuously differentiable.

Example 42.33 (Tutorial 4, Q8). Construct an approximating Hermite polynomial for the following specification.

x_k	y_k	y'_k
8.3	17.56492	3.116256
8.6	18.50515	3.151762

The data were generated using $f(x) = x \ln x$. Calculate the absolute error for $f(8.4)$.

Solution. First, the divided-difference table is

8.3	17.5649204			
8.3	17.5649204	3.1162560		
8.6	18.5051498	3.1340961	0.0594668	
8.6	18.5051498	3.1517620	0.0588865	-0.0019346

From here, we can form the Hermite polynomial in Newton basis:

$$H(x) = 17.5649204 + 3.1162560(x-8.3) + 0.0594668(x-8.3)^2 - 0.0019346(x-8.3)^2(x-8.6).$$

and find $H(8.4) = 17.8771445372$.

The error is found to be $|H(8.4) - 8.4 \ln 8.4| = 1.791933851 \times 10^{-6}$.

□

42.4 Piecewise Polynomial Interpolation & Univariate Spline

Both interpolating polynomial and Hermite polynomial approximate arbitrary functions on closed intervals by the use of polynomials. However, using a single polynomial to approximate a function has problems: (a) the error in evaluating the polynomial $P(x) = a_0 + \dots + a_n x^n$ can be considerable, i.e. when both n and x are large, any error in x will be amplified greatly in raising x to the n th power; (b) high-degree polynomials have oscillatory nature, this is called the Runge phenomenon (Section 42.2.6).

So, instead of using a single polynomial to represent a given function over the entire interval, joining together several "polynomial segment of lower degree" may be a better idea. This is called *piecewise polynomial approximation*.

The derivatives of the piecewise polynomial approximation will be discontinuous at data points x_j . One way to tackle this issue is to derive the coefficients of the piecewise polynomials by using information from all data points, i.e., nonlocal information, to guarantee global smoothness

in the interpolated function to some order of derivatives. This leads to a special type of piecewise polynomial approximation called the *spline approximation* or *spline interpolation*.

42.4.1 Piecewise Linear Interpolation & Linear Spline

The simplest piecewise polynomial approximation is the *piecewise-linear approximation*, which is a piecewise linear function which passes through the given set of data points

$$\{(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)\}$$

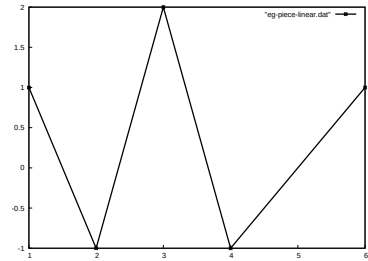
where $x_i < x_j$ when $i < j$. The following example illustrates two possible piecewise-linear approximations.

Example 42.34. Given 5 points $(1, 1)$, $(2, -1)$, $(3, 2)$, $(4, -1)$, $(6, 1)$. Write down two different piecewise-linear approximations for the points.

Solution: The simplest piecewise-linear approximation is to joint the points:

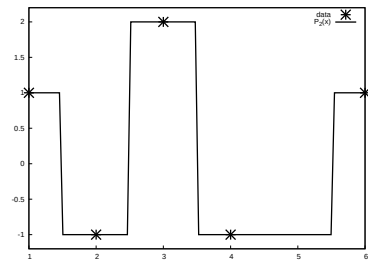
$$P_1(x) = \begin{cases} -2(x-1) + 1 & 1 \leq x < 2, \\ 3(x-2) - 1 & 2 \leq x < 3, \\ -3(x-3) + 2 & 3 \leq x < 4, \\ (x-4) - 1 & 4 \leq x \leq 6. \end{cases}$$

NumPy implemented this in `numpy/lib/function_base.py` and we can calculate $P_1(x)$ using `interp(linspace(1, 6, 100), [1, 2, 3, 4, 6], [1, -1, 2, -1, 1])`



Another possible piecewise-linear approximation is

$$P_2(x) = \begin{cases} 1 & 1 \leq x < 1.5, \\ -1 & 1.5 \leq x < 2.5, \\ 2 & 2.5 \leq x < 3.5, \\ -1 & 3.5 \leq x < 5.5, \\ 1 & 5.5 \leq x \leq 6. \end{cases}$$



The above example demonstrates that piecewise-linear approximation is not unique! However, the first piecewise-linear approximation is a *linear spline* (a function obtain by joining points continuously) which can be written in the following “B-spline representation” and widely used:

$$P_1(x) = 1\phi_1(x) - 1\phi_2(x) + 2\phi_3(x) - 1\phi_4(x) + 5\phi_5(x) \quad (42.12)$$

where

$$\phi_1(x) = \begin{cases} \frac{x-2}{1-2}, & 1 \leq x \leq 2, \\ 0, & \text{otherwise} \end{cases} \quad \phi_2(x) = \begin{cases} \frac{x-1}{2-1}, & 1 \leq x \leq 2, \\ \frac{x-3}{2-3}, & 2 \leq x \leq 3, \\ 0, & \text{otherwise} \end{cases} \quad \phi_3(x) = \begin{cases} \frac{x-2}{3-2}, & 2 \leq x \leq 3, \\ \frac{x-4}{3-4}, & 3 \leq x \leq 4, \\ 0, & \text{otherwise} \end{cases}$$

$$\phi_4(x) = \begin{cases} \frac{x-3}{4-3}, & 3 \leq x \leq 4, \\ \frac{x-6}{4-6}, & 4 \leq x \leq 6, \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad \phi_5(x) = \begin{cases} \frac{x-4}{6-4}, & 4 \leq x \leq 6, \\ 0, & \text{otherwise.} \end{cases}$$

Here $\phi_i(x)$ form the basis of the linear spline. They are applied in finite element methods, refer to https://en.wikiversity.org/wiki/Introduction_to_finite_elements and <http://hplgit.github.io/INF5620/doc/pub/sphinx-fem/> for more details.

42.4.2 Scientific Data Visualisation

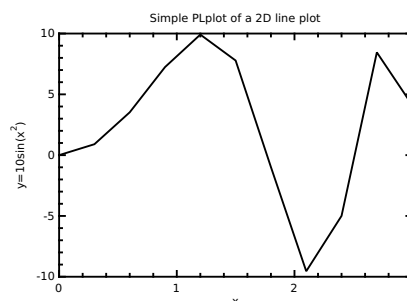
The use of computer-graphics methods in the analysis of data is called *data visualisation*. When the data comes from scientific and engineering observations or computations, we have *scientific data visualisation*. When the data comes from business, finance and social science studies, we have *business data visualisation*.

Scientific data sets are classified according to their type (scalar, vectors, tensors and multi-variate data) and spatial distribution [Hearn and Baker, 2004]. In this topic as well as topics 5 and 7, we discuss one-dimensional spatial data. In Topic 8, we will meet with two-dimensional spatial data, therefore, it is appropriate to discuss the software tools to visualise scientific data.

We will first mention the software library `plplot`, which supports many programming languages, such as Fortran.

```
1 program eg_plplot2d ! Simple demo of a 2D line plot
2   use plplot !-I/usr/lib/fortran/modules/plplot -lplplotf95d -lplplotf95cd
3
4   integer, parameter :: NSIZE = 11
5   integer             :: i
6   real(kind=plflt) :: x(NSIZE), y(NSIZE), xmin, xmax, ymin, ymax, delta
7
8   xmin = 0._plflt; xmax = 3._plflt; ymin = -10._plflt; ymax = 10._plflt
9   delta = 1/real(NSIZE-1,plflt)
10  x = (((xmax-xmin)*i*delta,i=0,NSIZE-1)/) ! x=[0:1/(NSIZE-1):1]
11  y = ymax * sin(x**2) ! y=10*sin(x.^2)
12
13  ! Initialize plplot by processing (command line) arguments
14  call plscolbg(255,255,255)
15  call plparseopts( PL_PARSE_FULL )
16  call plsdev("pdfqt") ! pdfcairo, epscairo, pscairo, pngcairo
17  call plsfnam("eg-plplot2d.pdf")
18  !call plsdiori(1.0d0) ! rotation: 0.1 = 9'
19  call plinit
20
21  ! Changing background and front colour & width
22  call plscol0(15,0,0,0)
23  call plcol0(15)
24  call plwidth(3.5_plflt)
25  ! Create a labelled box to hold the plot and plot the data
26  call plenv( xmin, xmax, ymin, ymax, 0, 0 )
27  call pllabb( "x", "y=10sin(x#u2#d)", "Simple PLplot of a 2D line plot" )
28  call plline( x, y )
29
30  call plend ! Close PLplot library
31 end program eg_plplot2d
```

The Fortran program generates a linear spline for the function $f(x) = 10 \sin(x^2)$ over the points $x = 0, 0.3, \dots, 2.7, 3$.



Linear spline (Subsection [42.4.1](#)) is commonly used in data visualisation software. Since we only pick 11 points to calculate $y = 10 \sin(x^2)$ in the interval $[0, 3]$, the linear spline shown is not smooth. The curve will look smooth if more points are used in the calculation.

Instead of writing programs to visualise data, scientists find it more convenient to store the data into a file and use a *data visualisation software* to visualise the data. Matlab is one of

the most widely use data visualisation software and LabView is another popular commercial for engineers. However, for less complicated data, gnuplot, a free to use software, is popular among scientists.

Gnuplot was published by Colin D. Kelly and Thomas Williams as an open source program to the sci.math USENET newsgroup at the fall of 1987 Janert [2010]. It is used in many open source scientific software such as Julia's Gaston graphics package, Octave, etc.

By using `gnuplot`, it is easy to produce the plots for the two piecewise-linear approximations to Example 42.34 as follows.

```

1 set border linewidth 0.2
2 set xtics nomirror
3 set ytics nomirror
4 set terminal postscript eps linewidth 5 enhanced
5 set output "eg-piece-linear1.eps"
6 plot "eg-piece-linear.dat" w lp pt 3
7 set output "eg-piece-linear2.eps"
8 P2(x) = x<1.5 ? 1 : (x<2.5 ? -1 : (x<3.5 ? 2 : (x<5.5 ? -1 : \
9   (x <= 6 ? 1 : 0))))
10 set xrange [-1.2:2.2]
11 plot [1:6] "eg-piece-linear.dat" w p ps 3 pt 3 t "data", P2(x) w l lt 1 t "P_2(x)"
12 #pause -1

```

When processing the solution of PDE (Topic 8), we often need to visualise a 2-D data. Gnuplot provides 3-D functions to handle these kind of visualisation.

Example 42.35. The PDE $u_t = u_{xx} + (x - t^2)e^{xt}$ with boundary conditions $u(0, t) = 1$ and $u(1, t) = e^t$ and initial condition 1 has a solution $u(x, t) = e^{xt}$. Plot the function u for the range $(x, t) \in [0, 1] \times [0, 3]$ with gnuplot.

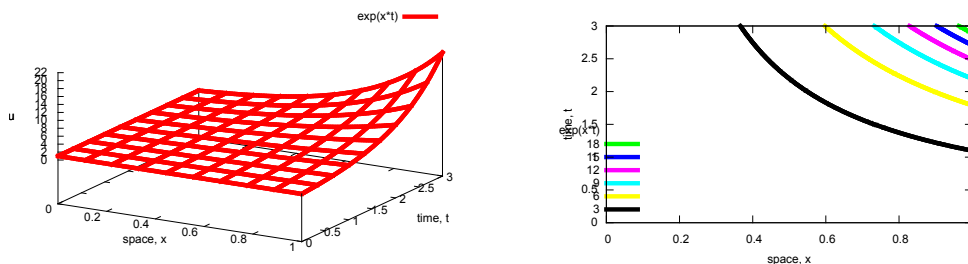
Solution: By using the `splot` function:

```

1 u(x,t) = exp(x*t)
2 set border linewidth 0.2
3 set xlabel "space, x"
4 set ylabel "time, t"
5 set zlabel "u"
6 set terminal pdfcairo font "Arial,12" linewidth 12
7 set output "eg-gnuplot3d.pdf"
8 splot [0:1][0:3] u(x,y) title "exp(x*t)"
9 set output "eg-gnuplot3d2.pdf"
10 set view map
11 unset surface
12 set contour
13 set cntrparam levels incremental 0, 3, 21
14 set isosamples 150
15 set key at 0,1.5 left top
16 splot [0:1][0:3] u(x,y) title "exp(x*t)"
17 #pause -1

```

we can produce a surface $u = u(x, t)$ in 3D plot and contour plot as follows:

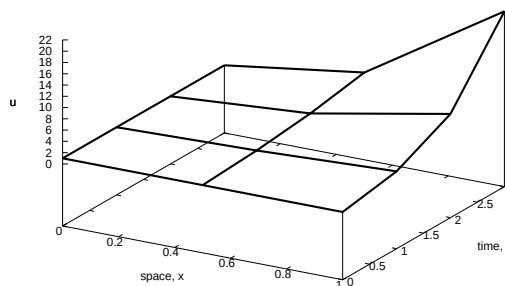


A sample 2-D “grid” data format is listed below.

```
# Julia 0.3: for x=0:0.5:1, t=0:3 u=exp(t*x); println("$x $t $u") end
# x,y,z = x,t,u
0.0 0 1.0
0.0 1 1.0
0.0 2 1.0
0.0 3 1.0

0.5 0 1.0
0.5 1 1.6487212707001282
0.5 2 2.718281828459045
0.5 3 4.4816890703380645

1.0 0 1.0
1.0 1 2.718281828459045
1.0 2 7.38905609893065
1.0 3 20.085536923187668
```



The following gnuplot commands are used to plot the 2-D data.

```
set dgrid3d 4,3
splot "eg-gnuplot3d3.dat" u 1:2:3 w l t ""
```

42.4.3 Piecewise Cubic Interpolation & Cubic Spline

cubic-spline-fn

Linear spline avoids the Runge phenomenon found in Lagrange and Hermite interpolations. However, it is not “smooth” where at the “corner” points, the interpolation function is not differentiable. To look for an interpolant which is differentiable, we need to consider higher degree piecewise polynomials. *Hermite cubic interpolant* is found to be most useful in practise! [Kahaner et al., 1989]

However, there is no unique Hermite cubic interpolant. Matlab’s `pchip` implements a piecewise Hermite cubic interpolant with the derivatives at the x_i are chosen in such a way that preserves the shape of the data and respects monotonicity.

For each interval $[x_i, x_{i+1}]$, the two endpoints x_i and x_{i+1} determine an Hermite polynomial

$$H_3(x; x_i, x_{i+1}) = f(x_i)H_{1,0}(x) + f(x_{i+1})H_{1,0}(x) + f'(x_i)\hat{H}_{1,0}(x) + f'(x_{i+1})\hat{H}_{1,1}(x).$$

Unless the “sufficiently smooth” function f is known, otherwise, we normally do not know $f'(x_i)$ and $f'(x_{i+1})$. In this case, we may use the following estimate according to Kahaner et al. [1989]:

$$f'(x_i) \approx \frac{h_{i-1}\delta_i + h_i\delta_{i-1}}{h_{i-1} + h_i}$$

where $y_i = f(x_i)$, $h_i = x_{i+1} - x_i$ and $\delta_i = \frac{y_{i+1} - y_i}{h_i}$.

Another popular piecewise interpolation is the *cubic spline (interpolation)*. It is a piecewise-polynomial approximation which uses cubic polynomials between each successive pair of nodes.

Spline interpolation is currently used in diverse domain of numerical methods including data smoothing (e.g. it is used in a financial software library QuantLib `ql/math/interpolations/cubicinterpolation.hpp`), computer aided geometric design, the numerical solution of differential and integral equations, etc. Advanced theory and discussions can be found in Phillips [2003], de Boor [2001], Dierckx [1993] and Prenter [1975].

Both linear splines and cubic splines are special cases of the general univariate spline defined below.

def:dierckx:p3

Definition 42.7. Let $\mathcal{P}_k$ be the set of polynomials of degree less than or equal to k . A function S , defined on a finite interval $[a, b]$, is called a (*univariate*) *spline (function)* of degree $k > 0$, having as *knots* the strictly increasing sequence x_j , $j = 0, 1, \dots, n$ ($x_0 = a$, $x_n = b$), if the following two conditions are satisfied:

1. On each knot interval $[x_j, x_{j+1}]$, $S(x)$ is given by a polynomial of degree k at most, i.e.

$$S|_{[x_j, x_{j+1}]} \in \mathcal{P}_k, \quad j = 0, 1, \dots, n-1.$$

2. The function S and its derivatives up to order $k-1$ are all continuous on $[a, b]$. This is denoted as $S \in C^{k-1}[a, b]$.

In the definition above, item 1. says that a spline is formed of piecewise polynomials; Item 2. says that the spline must be smooth, i.e. it can be differentiated $k-1$ times.

When $k=1$, the spline function S in Definition [42.7](#) is a linear spline. It is only continuous and may not be differentiable at every point.

When $k=3$, the spline function S becomes a *cubic spline* and the cubic spline function can be differentiated twice!

Definition 42.8. Given a function f defined on $[a, b]$ and a set of nodes $a = x_0 < x_1 < \dots < x_n = b$, a *cubic spline interpolant* S for f is a function that satisfies the following conditions:

1. For $j = 0, 1, \dots, n-1$,

- (a) $S(x)$ is a piecewise-polynomial such that

$$S_j(x) := S|_{[x_j, x_{j+1}]}(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3,$$

- (b) $S_j(x_j) = f(x_j)$ and $S_j(x_{j+1}) = f(x_{j+1})$.

2. For each $j = 0, 1, \dots, n-2$,

- (a) $S_{j+1}(x_{j+1}) = S_j(x_{j+1})$;

- (b) $S'_{j+1}(x_{j+1}) = S'_j(x_{j+1})$;

- (c) $S''_{j+1}(x_{j+1}) = S''_j(x_{j+1})$.

3. One of the following sets of boundary conditions is satisfied:

- (a) $S''(x_0) = S''(x_n) = 0$ (free of natural boundary);

- (b) $S'(x_0) = f'(x_0)$ and $S'(x_n) = f'(x_n)$ (clamped boundary).

The splines defined in Definition [42.7](#) is not unique. There are infinitely many of them as pointed in Prenter [1975, Section 4.2]. By adding item 3(a) or item 3(b), one obtains a unique cubic spline.

According to the definition, if we have $n+1$ control points, then we have n curve sections with a total of $4n$ polynomial coefficients to be determined. At each of the $n-1$ interior control points, we have 4 boundary conditions. When the free boundary conditions (item 3(a) in Definition [42.8](#)) are imposed, the spline is called a *natural spline*, and its graph approximates the shape that a long flexible rod would assume if forced to go through the data points $\{(x_i, y_i)\}_{i=0}^n$.

The following theorem shows that there is a unique interpolant for the free boundary conditions.

Theorem 42.36. If f is defined at $a = x_0 < x_1 < \dots < x_n = b$, then f has a unique natural spline interpolant S on the nodes $x_0, \dots, x_n$, that is, a spline interpolant that satisfies the boundary conditions $S''(a) = 0$ and $S''(b) = 0$.

Let $h_j = x_{j+1} - x_j$ and $a_j = f(x_j)$, $j = 0, \dots, n$, c_j can be obtained from the following

linear equation:

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & \cdots & 0 \\ h_0 & 2(h_0 + h_1) & h_1 & \ddots & & \vdots \\ 0 & h_1 & 2(h_1 + h_2) & h_2 & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & h_{n-2} & 2(h_{n-2} + h_{n-1}) & h_{n-1} \\ 0 & \cdots & \cdots & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ \vdots \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{3}{h_1}(a_2 - a_1) - \frac{3}{h_0}(a_1 - a_0) \\ \vdots \\ \vdots \\ \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_{n-2}}(a_{n-1} - a_{n-2}) \\ 0 \end{bmatrix} \quad (42.13)$$

e:cspline-n

The coefficients b_j and d_j ($j = 0, \dots, n-1$) can be obtained from

$$b_j = \frac{1}{h_j}(a_{j+1} - a_j) - \frac{h_j}{3}(2c_j + c_{j+1}), \quad (42.14)$$

e:ns:b

$$d_j = \frac{c_{j+1} - c_j}{3h_j}. \quad (42.15)$$

e:ns:d

We observe the first tridiagonal matrix mentioned in Topic 2 in the linear algebra problem (42.13). The algorithm to calculate the coefficients of a natural cubic spline is listed below.

```
1 subroutine csplinen(n,xv,a,b,c,d)
2   intent(in) :: n, xv, a ! a = yv
3   integer :: n, i
4   real(c_double), intent(out) :: b(n), c(n), d(n)
5   real(c_double) :: xv(n), a(n), h(n), alp(n), mu(n), z(n), t
6   do i=1,n-1
7     h(i) = xv(i+1)-xv(i)
8   end do
9   do i=2,n-1
10    alp(i) = 3d0/h(i)*(a(i+1)-a(i)) - 3d0/h(i-1)*(a(i)-a(i-1))
11  end do
12  ! Using chasing method (Crout factorisation) to solve
13  ! a tridiagonal linear system (equivalent to xGTSV?)
14  mu(1) = 0d0; z(1) = 0d0; c(n) = 0d0
15  do i=2,n-1 ! Forward elimination (Topic 2)
16    t = 2*(xv(i+1)-xv(i-1))-h(i-1)*mu(i-1)
17    mu(i) = h(i)/t
18    z(i) = (alp(i)-h(i-1)*z(i-1))/t
19  end do
20  do i=n-1,1,-1 ! Backward substitution (Topic 2)
21    c(i) = z(i)-mu(i)*c(i+1)
22    b(i) = (a(i+1)-a(i))/h(i) - h(i)*(c(i+1)+2*c(i))/3.
23    d(i) = (c(i+1)-c(i))/(3*h(i))
24  end do
25 end subroutine csplinen
```

The implementation of natural cubic spline in QuantLib (ql/math/interpolations/cubicinterpolation.hpp) is based on object-oriented programming technique which is complicated for beginners:

```
1 class CubicNaturalSpline : public CubicInterpolation {
2   public:
3     /*! \pre the \f{x} values must be sorted. */
4     template <class I1, class I2>
5     CubicNaturalSpline(const I1& xBegin,
6                       const I1& xEnd,
7                       const I2& yBegin)
8     : CubicInterpolation(xBegin, xEnd, yBegin,
9                         Spline, false,
10                        SecondDerivative, 0.0,
11                        SecondDerivative, 0.0) {}
12 };
```

Remark 42.37. Note that the program uses a kind of LU factorisation called *Crout factorisation* on a tri-diagonal matrix:

$$\underbrace{\begin{bmatrix} d_1 & u_1 & & & \\ \ell_2 & d_2 & u_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ell_{n-1} & d_{n-1} & u_{n-1} \\ & & & \ell_n & d_n & u_n \end{bmatrix}}_A = \underbrace{\begin{bmatrix} 1 & 0 & \cdots & & \\ \ell'_1 & 1 & 0 & \cdots & \\ & \ddots & \ddots & \ddots & \\ & & \ell'_{n-1} & 1 & 0 \\ & & & \ell'_n & 1 \end{bmatrix}}_L \underbrace{\begin{bmatrix} d'_1 & u_1 & & & \\ & d'_2 & u_2 & & \\ & & \ddots & \ddots & \\ & & & d'_{n-1} & u_{n-1} \\ & & & & d'_n \end{bmatrix}}_U$$

From this we can derive the relations: $\begin{cases} d'_1 = d_1, & \ell'_1 = \ell_1/d'_1 \\ d'_i = d_i - \ell'_i u_{i-1}, & \ell'_{i+1} = \ell_{i+1}/d'_i. \end{cases}$
and solve $A\mathbf{x} = \mathbf{b}$ where A is tridiagonal with the “chasing” algorithm below:

```

1 ! Naive ``Chasing`` algorithm for tridiagonal linear algebra problem
2 ! based on Crout factorisation (LU factorisation).
3 subroutine chasing(N,D,L,U,B,x)
4     use iso_c_binding, only : c_double
5     integer, intent(in) :: N
6     real(c_double) :: D(N), L(N-1), U(N-1), B(N), y(N), LP(N), DP(N)
7     real(c_double), intent(out) :: x(N)
8     ! Crout factorisation
9     DP = D
10    do i=2,N
11        LP(i) = L(i-1)/DP(i-1)
12        DP(i) = DP(i) - LP(i)*U(i-1)
13    end do
14    ! Forward substitution
15    y(1) = B(1)
16    do i=2,N
17        y(i) = B(i) - LP(i)*y(i-1)
18    end do
19    ! Backward substitution
20    x(N) = y(N)/DP(N)
21    do i=N-1,1,-1
22        x(i) = (y(i)-U(i)*x(i+1))/DP(i)
23    end do
24 end subroutine chasing

```

Further analysis allow us to simplify the above algorithm further into

$$s_0 = t_0 = 0, \quad s_i = \frac{b_i - \ell_i s_{i-1}}{d_i - \ell_i t_{i-1}}, \quad t_i = \frac{u_i}{d_i - \ell_i t_{i-1}}, \quad i = 1, 2, \dots, n$$

and

$$x_i = s_i - t_i x_{i+1}, \quad i = n-1, \dots, 2, 1$$

which was used in function csplines(xv,yv) (lines 14–24).

Compare to LU factorisation and Cholesky factorisation which take $\approx \frac{n^3}{3}$ and $\approx \frac{n^3}{6}$ flops respectively, tri-diagonal Crout factorisation takes $\approx 10n$ flops. A significant reduction in computing when n is large.

Example 42.38. [Burden and Faires, 2015, p150] Use a natural spline and the data points $(0, 1)$, $(1, e)$, $(2, e^2)$, and $(3, e^3)$ to (a) form a new approximating function $S(x)$. Then (b) compare the integrals of $f(x) = e^x$ and $S(x)$ on the interval $[0, 3]$.

Solution: (a) $n = 3$, $h_0 = h_1 = h_2 = 1$, $a_0 = 0$, $a_1 = e$, $a_2 = e^2$, and $a_3 = e^3$. We obtain

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 3(e^2 - 2e + 1) \\ 3(e^3 - 2e^2 + e) \\ 0 \end{bmatrix}.$$

by substituting the values of h_i and a_i into (42.13). Solving the linear equation gives

$$c_0 = c_3 = 0, \quad c_1 = 0.756852643, \quad c_2 = 5.83006675.$$

By applying formulae (42.14) and (42.14), we obtain

$$\begin{aligned} b_0 &= 1.465997614, \quad b_1 = 2.222850257, \quad b_2 = 8.809769651 \quad \text{and} \\ d_0 &= 0.0252284214, \quad d_1 = 1.691071368, \quad d_2 = -1.943355583. \end{aligned}$$

Therefore,

$$S(x) = \begin{cases} 1 + 1.46600x + 0.25228x^3, & 0 \leq x < 1, \\ 2.71828 + 2.22285(x-1) + 0.75685(x-1)^2 + 1.69107(x-1)^3, & 1 \leq x < 2, \\ 7.38906 + 8.80977(x-2) + 5.83007(x-2)^2 - 1.94336(x-2)^3, & 2 \leq x \leq 3. \end{cases}$$

(b) The cubic natural spline approximates the exact integral only to 2 effective figures.

$$\int_0^3 e^x dx = e^3 - 1 = 19.08553692$$

$$\int_0^3 S(x) dx = \int_0^1 S_0(x) dx + \int_1^2 S_1(x) dx + \int_2^3 S_2(x) dx = 19.55228750$$

Example 42.39 (Tutorial 4, Q9). Determine the natural cubic spline S that interpolates the data $f(0) = 0$, $f(1) = 1$, and $f(2) = 2$.

Solution: First, for $i = 0, 1$, $h_i = 1$ and $a_i = f(i) = i$ for $i = 0, 1, 2$,

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 2(1+1) & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{3}{1}(f(2) - f(1)) - \frac{3}{1}(f(1) - f(0)) \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow c_i = 0.$$

For $j = 0, 1$, $b_j = \frac{1}{h_j}(a_{j+1} - a_j) - \frac{h_j}{3}(2c_j + c_{j+1}) = 1$, $d_j = \frac{c_{j+1} - c_j}{3h_j} = 0$.

Hence,

$$S(x) = \begin{cases} S_0(x) = 0 + 1(x-0) = x, & x \in [0, 1] \\ S_1(x) = 1 + 1(x-1) = x, & x \in [1, 2] \end{cases} = x.$$

Example 42.40. Given 5 points $(1, 1)$, $(2, -1)$, $(3, 2)$, $(4, -1)$, $(6, 1)$.

1. Write down the cubic spline for the points.

Solution. First, we calculate $h = [1, 1, 1, 2]$ and formulate the linear algebra problem

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 \\ 0 & 0 & 1 & 6 & 2 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{3}{1}(2 - (-1)) - \frac{3}{1}(-1 - 1) \\ \frac{3}{1}(-1 - 2) - \frac{3}{1}(2 - (-1)) \\ \frac{3}{2}(1 - (-1)) - \frac{3}{1}(-1 - 2) \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 15 \\ -18 \\ 12 \\ 0 \end{bmatrix}.$$

Solving it gives us

$$c_0 = c_4 = 0, \quad c_1 = 5.40698, \quad c_2 = -6.62791, \quad c_3 = 3.10465.$$

By applying formulae (42.14) and (42.14) and substituting into S , we obtain

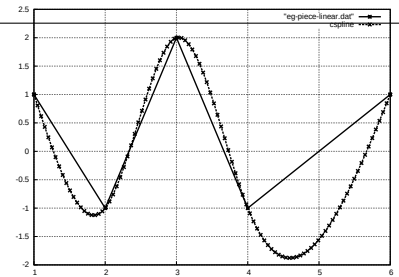
$$S(x) = \begin{cases} 1 - 3.8023256(x-1) + 1.8023256(x-1)^3 & 1 \leq x < 2 \\ -1 + 1.6046512(x-2) + 5.4069767(x-2)^2 - 4.0116279(x-2)^3 & 2 \leq x < 3 \\ 2 + 0.3837209(x-3) - 6.6279070(x-3)^2 + 3.2441860(x-3)^3 & 3 \leq x < 4 \\ -1 - 3.1395349(x-4) + 3.1046512(x-4)^5 - 0.5174419(x-4)^3 & 4 \leq x \leq 6. \end{cases}$$

□

2. In gnuplot, the plot option `csplines` applies a unique transformation and calculates the natural cubic splines sample points. Write a gnuplot script to plot the cubic spline for the 5 points.

Solution:

```
1 set border linewidth 0.2
2 set grid
3 set xtics nomirror
4 set ytics nomirror
5 set terminal postscript eps linewidth 5
6 set output "eg-piece-cubic.eps"
7 plot "eg-piece-linear.dat" w lp pt 3, \
8     "" smooth csplines t "cspline" w lp
9 #pause -1
```



3. There are many Fortran software library for data interpolation, such as `pppack` (`cubspl`), `slatec` (`polint`, `pchip`), `dierckx` (`curfit`, `splev`). There are also some C/C++ libraries, such as GNU Scientific Library. Write GNU Scientific Library to calculate the cubic spline for the given points.

Solution:

```
1 /* Interpolation with Gnu Scientific Library */
2 #include <gsl/gsl_errno.h>
3 #include <gsl/gsl_spline.h> // -lgsl -lgslcblas
4 #include <stdio.h>
5 #include <errno.h>
6 #include <stdlib.h>
7
8 int main() {
9     FILE *fp;
10     const int max_num_pts = 100, num_subintervals = 200;
11     int num_pts, i;
12     char *fname="eg-piece-linear.dat";
13     double xi, yi, xd, *x_array, *y_array;
14     gsl_spline *spline_ptr;
15     gsl_interp_accel *accel_ptr;
```

```

16
17     x_array = calloc(max_num_pts, sizeof(double));
18     y_array = calloc(max_num_pts, sizeof(double));
19     fp = fopen(fname, "r");
20     num_pts = 0;
21     if (fp == NULL) {
22         perror(fname);
23         exit(1);
24     }
25     while (fscanf(fp, "%lf%lf", &xi, &yi) != EOF) {
26         if (num_pts < max_num_pts) {
27             x_array[num_pts] = xi;
28             y_array[num_pts] = yi;
29             num_pts++;
30         } else {
31             printf("Only 100 points are process\n");
32             break;
33         }
34     }
35     fclose(fp);
36
37     accel_ptr = gsl_interp_accel_alloc();
38     spline_ptr = gsl_spline_alloc(gsl_interp_cspline, num_pts);
39     gsl_spline_init(spline_ptr, x_array, y_array, num_pts);
40
41     xi = x_array[0]; // Assuming data is sorted
42     xd = (x_array[num_pts-1]-xi)/num_subintervals;
43     printf("%lf %lf\n", xi, y_array[0]);
44     for(i=0; i<num_subintervals-1; i++){
45         xi += xd;
46         yi = gsl_spline_eval(spline_ptr, xi, accel_ptr);
47         printf("%lf %lf\n", xi, yi);
48     }
49     printf("%lf %lf\n", x_array[num_pts-1], y_array[num_pts-1]);
50
51     gsl_spline_free(spline_ptr);
52     gsl_interp_accel_free(accel_ptr);
53 }

```

fares:3.5:q11

Example 42.41 (Tutorial 4, Q10). A natural cubic spline S on $[0, 2]$ is defined by

$$S(x) = \begin{cases} S_0(x) = 1 + 2x - x^3, & \text{if } 0 \leq x < 1, \\ S_1(x) = a + b(x-1) + c(x-1)^2 + d(x-1)^3, & \text{if } 1 \leq x \leq 2. \end{cases}$$

Find a , b , c and d .

Solution. $S_0(1) = 2 = S_1(1) = a \Rightarrow a = 2$.

$S'_0(1) = 2 - 3 = S'_1(1) = b \Rightarrow b = -1$

$S''_0(1) = -3(2) = S''_1(1) = 2c \Rightarrow c = -3$

Natural spline condition at end point $S''_1(2) = 2c + 6d = 0 \Rightarrow d = 1$. □

Clamped boundary conditions (item 3(b) in Definition [42.8](#)) lead to more accurate approximations for a smooth function f since they include more information about f , i.e. the values of the derivative at the endpoints or an accurate approximation to those values. We have a similar result to the theorem for natural boundary conditions.

thm:cspline

Theorem 42.42. *If f is defined at $a = x_0 < x_1 < \dots < x_n = b$ and differentiable at a and b , then f has a unique clamped spline interpolant S on the nodes $x_0, \dots, x_n$, that is, a spline interpolant that satisfies the boundary conditions $S'(a) = f'(a)$ and $S''(b) = f'(b)$.*

Let $h_j = x_{j+1} - x_j$ and $a_j = f(x_j)$ for $j = 0, \dots, n-1$. The coefficients c_j can be obtained from the following linear equation:

$$\begin{bmatrix} 2h_0 & h_0 & 0 & \dots & \dots & 0 \\ h_0 & 2(h_0 + h_1) & h_1 & \ddots & & \vdots \\ 0 & h_1 & 2(h_1 + h_2) & h_2 & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & h_{n-2} & 2(h_{n-2} + h_{n-1}) & h_{n-1} \\ 0 & \dots & \dots & 0 & h_{n-1} & 2h_{n-1} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ \vdots \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} \frac{3}{h_0}(a_1 - a_0) - 3f'(a) \\ \frac{3}{h_1}(a_2 - a_1) - \frac{3}{h_0}(a_1 - a_0) \\ \vdots \\ \vdots \\ \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_{n-2}}(a_{n-1} - a_{n-2}) \\ 3f'(b) - \frac{3}{h_{n-1}}(a_n - a_{n-1}) \end{bmatrix} \quad (42.16)$$

e:cspline-c

Coefficients b_j and d_j ($j = 0, \dots, n-1$) can be obtained from

$$b_j = \frac{1}{h_j}(a_{j+1} - a_j) - \frac{h_j}{3}(2c_j + c_{j+1}), \quad \text{and} \quad (42.17)$$

e:cs:b

$$d_j = \frac{c_{j+1} - c_j}{3h_j}. \quad (42.18)$$

e:cs:d

A proof that this cubic spline is unique is given in Prenter [1975, Theorem 4.1].

The solution to the cubic spline problem with the boundary conditions $S'(x_0) = f'(x_0)$ and $S'(x_n) = f'(x_n)$ can be obtained by applying the a similar algorithm shown below.

```

1  ! Cubic Spline with Clamped Boundary Conditions
2  subroutine csplinec(n,xv,a,fp0,fpN,b,c,d)
3      intent(in) :: n, xv, a, fp0, fpN ! a=yv
4      integer :: n, i
5      real(c_double) :: xv(n), a(n), fp0, fpN, h(n), alp(n), mu(n), z(n), t
6      real(c_double), intent(out) :: b(n), c(n), d(n)
7      do i=1,n-1
8          h(i) = xv(i+1)-xv(i)
9      end do
10     alp(1) = 3*(a(2)-a(1))/h(1) - 3*fp0
11     alp(n) = 3*fpN - 3*(a(n)-a(n-1))/h(n-1)
12     do i=2,n-1
13         alp(i) = 3d0/h(i)*(a(i+1)-a(i)) - 3d0/h(i-1)*(a(i)-a(i-1))
14     end do
15     ! Using Crout factorisation to solve a tridiagonal problem
16     mu(1) = 0.5d0
17     z(1) = alp(1)/(2*h(1))
18     do i=2,n-1
19         t = 2*(xv(i+1)-xv(i-1))-h(i-1)*mu(i-1)
20         mu(i) = h(i)/t
21         z(i) = (alp(i)-h(i-1)*z(i-1))/t
22     end do
23     z(n) = (alp(n)-h(n-1)*z(n-1))/(h(n-1)*(2-mu(n-1)))
24     c(n) = z(n)
25     do i=n-1,1,-1
26         c(i) = z(i)-mu(i)*c(i+1)
27         b(i) = (a(i+1)-a(i))/h(i) - h(i)*(c(i+1)+2*c(i))/3.
28         d(i) = (c(i+1)-c(i))/(3*h(i))
29     end do
30 end subroutine csplinec

```

Example 42.43. [Burden and Faires, 2015, p156] Use a clamped spline and the data points $(0,1)$, $(1,e)$, $(2,e^2)$, and $(3,e^3)$ to (a) form a new approximating function $S(x)$. Then (b) compare the integrals of $f(x) = e^x$ and $S(x)$ on the interval $[0,3]$.

Solution: (a) $n = 3$, $h_0 = h_1 = h_2 = 1$, $a_0 = 0$, $a_1 = e$, $a_2 = e^2$, and $a_3 = e^3$. Since $f'(x) = e^x$, so $f'(0) = 1$ and $f'(3) = e^3$. So we have

$$\begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 3(e-2) \\ 3(e^2-2e+1) \\ 3(e^3-2e^2+e) \\ 3e^2 \end{bmatrix}.$$

Solving the above linear equation gives

$$c_0 = 0.4446822, \quad c_1 = 1.2654809, \quad c_2 = 3.3508725, \quad c_3 = 9.4081480.$$

By formulae (42.17) and (42.18), we obtain the coefficients

$$b_0 = 1.0000000, \quad b_1 = 2.7101631, \quad b_2 = 7.3265164$$

$$d_0 = 0.2735996, \quad d_1 = 0.6951305, \quad d_2 = 2.0190918.$$

and

$$\text{So } S(x) = \begin{cases} 1 + x + 0.4446822x^2 + 0.2735996x^3, & 0 \leq x < 1, \\ 2.7182817 + 2.7101631(x-1) + 1.2654809(x-1)^2 + 0.6951305(x-1)^3, & 1 \leq x < 2, \\ 7.3890562 + 7.3265164(x-2) + 3.3508725(x-2)^2 + 2.0190918(x-2)^3, & 2 \leq x \leq 3. \end{cases}$$

(b)

$$\int_0^3 S(x)dx = \int_0^1 S_0(x)dx + \int_1^2 S_1(x)dx + \int_2^3 S_2(x)dx = 19.059645.$$

- Absolute error for natural spline: $|19.08553692 - 19.55229| = 0.46675$
- Absolute error for clamped spline: $|19.08553692 - 19.06677| = 0.02589$

So for integration purposes the clamped spline is superior because the boundary conditions for clamped spline are exact.

fares:3.5:q7d

Example 42.44 (Tutorial 4, Q11). Construct the clamped cubic spline using the following data.

x	0.1	0.2	0.3	0.4
$f(x)$	-0.62049958	-0.28398668	0.00660095	0.24842440
$f'(x)$	3.58502082		2.16529366	

Solution. First, for $j = 0, 1, 2$, $h_j = 0.1$ and $a_i = f(x_i)$ for $i = 0, 1, 2, 3$,

$$\begin{bmatrix} 2(0.1) & 0.1 & 0 & 0 \\ 0.1 & 2(0.1+0.1) & 0.1 & 0 \\ 0 & 0.1 & 2(0.1+0.1) & 0.1 \\ 0 & 0 & 0.1 & 2(0.1) \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} \frac{3}{0.1}(a_1 - a_0) - 3f'(a) \\ \frac{3}{0.1}(a_2 - a_1) - \frac{3}{0.1}(a_1 - a_0) \\ \frac{3}{0.1}(a_3 - a_2) - \frac{3}{0.1}(a_2 - a_1) \\ 3f'(b) - \frac{3}{0.1}(a_3 - a_2) \end{bmatrix} = \begin{bmatrix} 0 \\ -1.3777581 \\ -1.4629254 \\ 0 \end{bmatrix}$$

Solving it gives

$$c_0 = -2.14984079, \quad c_1 = -2.29707303, \quad c_2 = -2.43944811, \quad c_3 = -2.57438855,$$

$$b_0 = 3.58502082, \quad b_1 = 3.14032944, \quad b_2 = 2.66667733,$$

$$d_0 = -0.49077413, \quad d_1 = -0.47458360, \quad d_2 = -0.44980147$$

$$S_0(x) = -0.62049958 + 3.58502082(x-0.1) - 2.14984079(x-0.1)^2 - 0.49077413(x-0.1)^3$$

$$S_1(x) = -0.28398668 + 3.14032944(x-0.2) - 2.29707303(x-0.2)^2 - 0.47458360(x-0.2)^3$$

$$S_2(x) = 0.00660095 + 2.66667733(x-0.3) - 2.43944811(x-0.3)^2 - 0.44980147(x-0.3)^3$$

□

The following example demonstrates that splines can be easily derived from the Definition 42.7. def:dierckx:p3

Example 42.45 (April 2012 Final Exam Q4(a)). Construct a spline function $S(x)$ of degree two and having three nodes at $(1, 5)$, $(2, 7)$, $(3, 9)$ and $S'(1) = 0$. (10 marks)

Solution. From definition,

$$S(x) = \begin{cases} S_1(x) = a_1 + b_1(x-1) + c_1(x-1)^2, & 1 \leq x < 2, \\ S_2(x) = a_2 + b_2(x-2) + c_2(x-2)^2, & 2 \leq x < 3. \end{cases}$$

$$S'(1) = 0 \Rightarrow b_1 = 0$$

$$S_1(1) = 5 \Rightarrow a_1 = 5$$

$$S_1(2) = S_2(2) = 7, a_2 = 7 \text{ and } a_1 + b_1 + c_1 = a_2 \dots\dots\dots (e1)$$

$$S_2(3) = 9 \Rightarrow a_2 + b_2 + c_2 = 9 \dots\dots\dots (e2)$$

$$\text{Solving (e1) \& (e2) leads to } c_1 = 2$$

$$S'_1(2) = S'_2(2) \Rightarrow b_1 + 2c_1 = b_2 \Rightarrow b_2 = 4 \Rightarrow c_2 = 9 - 7 - 4 = -2 \quad \square$$

Finally, let us compare the interpolating polynomial, Hermite polynomial, linear spline and cubic spline using Example 42.25. eg:runge

Example 42.46. Write a program to compare the interpolating polynomial, Hermite polynomial, linear spline and natural cubic spline for the Runge function $R(x)$ in Example 42.25 for the range $[-4, 4]$. Explain the result. eg:runge

Solution: We will make use of Fortran and gnuplot:

```

1  program runge4
2      use iso_c_binding, only : c_double
3      use textbook_interp
4      implicit none
5      real(c_double), external :: RR, Rprime
6      real(c_double), parameter :: l=-4d0, u=4d0, h=1d-2
7      integer, parameter :: N=8, NN=int((u-l)/h)
8      real(c_double), dimension(N+1) :: xvals, yvals, ypvals, b, c, d
9      real(c_double) :: x,rungeA,intrpA,osculA,spln1A,spln3A,grad,dx
10     real(c_double), allocatable :: Q(:, :)
11     integer :: i, j, stat
12     xvals = (/ (-4d0+(i-1)*1d0, i=1,N+1) /)
13     yvals = (/ (RR(xvals(i)), i=1,N+1) /)
14     ypvals = (/ (Rprime(xvals(i)), i=1,N+1) /)
15     call hermite_dd(xvals,yvals,ypvals,Q)
16     call csplinen(N+1,xvals,yvals,b,c,d)
17     print 200, "x", "y", "Lagrange", "Hermite", "Spline1", "Spline3"
18     do i=0,NN
19         x = l+i*h          ! Interpolation points
20         rungeA = RR(x)      ! Exact function value
21         ! Compute interpolating polynomial value
22         intrpA = lagrange(xvals, yvals, x, stat)
23         ! Compute Hermite polynomial value using Horner's method
24         osculA = Q(2*N+2,2*N+2)
25         do j=2*N+1,1,-1    ! Hermite Q is 2(N+1) x 2(N+1)
26             osculA = (x-xvals(int((j+1)/2)))*osculA + Q(j,j)
27         end do ! Can be numerically unstable with the wrong index
28         do j=1,N          ! Linear Search --- Slow
29             if (xvals(j) <= x .and. x < xvals(j+1)) then
30                 ! Compute linear spline
31                 grad = (yvals(j+1)-yvals(j))/(xvals(j+1)-xvals(j))
32                 spln1A = yvals(j) + grad*(x-xvals(j))
33                 ! Compute cubic spline
34                 dx = x-xvals(j)
35                 spln3A = yvals(j)+dx*(b(j)+dx*(c(j)+dx*d(j)))
36                 exit
37             end if

```

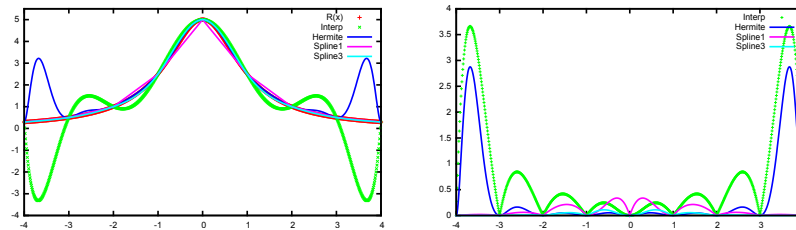


```

38     end do
39     if (xvals(N+1) <= x) then
40         spln1A=yvals(N+1)
41         spln3A=yvals(N+1)
42     end if
43     print 100, x, rungeA, intrpA, osculaA, spln1A, spln3A, &
44         abs(intrpA-rungeA), abs(osculaA-rungeA), abs(spln1A-rungeA), &
45         abs(spln3A-rungeA)
46 end do
47 100 format(f10.5,1x,f10.6,1x,f10.6,1x,f10.6,1x,f10.6,1x,f10.6, &
48         1x,f10.6,1x,f10.6,1x,f10.6,1x,f10.6)
49 200 format(a10, 1x, a10, 1x, a10, 1x, a15, 1x, a10, 1x, a10)
50 end program runge4
51 function RR(x)
52     use iso_c_binding, only : c_double
53     real(c_double) :: x, RR
54     RR = 5d0/(1d0+x**2)
55 end function RR
56 function Rprime(x)
57     use iso_c_binding, only : c_double
58     real(c_double) :: x, Rprime
59     Rprime = -10d0*x/(1d0+x**2)**2
60 end function Rprime

```

Plotting the data generated from the above Fortran program



we can see that both interpolating polynomial and Hermite polynomial differ largely from the original function R . Linear spline is better but natural cubic is the best.

A Java implementation of the interpolation techniques mentioned so far can be found in Artisynt's `maspack/interpolation` directory. Alglib provides spline interpolation in various languages, see <http://www.alglib.net/interpolation/spline3.php> for details. Interpolation techniques are applied in QuantLib and R to fit term structure, see `ql/termstructures/yield/nonlinearfittingmethods.hpp` and <https://CRAN.R-project.org/package=termstrc> for details.

42.5 Application: Static Term Structure Models

Interpolation methods can only be applied in relatively simple financial model such as the *static* term structure models. The static term structure models only fit the term structure (cross-section data). The simplest static models are based on spline-curve fitting and the more advanced static models are exponential-polynomial families. An authoritative introduction to these static models are given in [Filipović, 2009, Chapter 3]. Implementation of some of the static term structure models are given in QuantLib's `ql/termstructures/yield/nonlinearfittingmethods.hpp`. For example, the Nelson and Siegel (1987) yield curve model is

$$y(\tau) = \beta_1 + \beta_2 \left[\frac{1 - e^{-\tau/\lambda}}{\tau/\lambda} \right] + \beta_3 \left[\frac{1 - e^{-\tau/\lambda}}{\tau/\lambda} - e^{-\tau/\lambda} \right] \quad (42.19)$$

e:nelson-siege

belongs to the exponential-polynomial family and was used by some European countries.

By using static models with panel data (involving both term-structure and time), the parameters in the model change with time. It is possible to study the parameters using time-series models, however, it is unable to obtain parameter estimates for the risk-adjusted processes that are necessary for the pricing of financial products.

To interpolate (42.19), nonlinear regression and more complicated techniques from Topic 3 and are beyond this subject.

42.6 Parametric Curve & Bézier Curve

sec:bezier

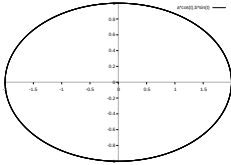
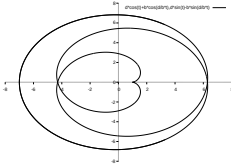
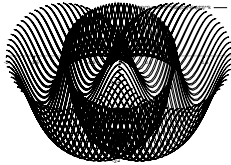
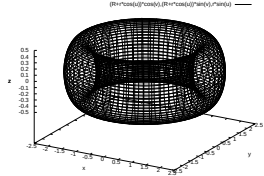
In previous sections, we have only considered interpolation techniques for a single-variable function f which are used in the approximation of functions for numerical integration (Topics 5 and 6). In this section, we will study the “generalised interpolation” of parametric curves $\mathbf{x} = \varphi(t)$, which is useful in geometric modelling, computer graphics [Bartels et al., 1987] and computer typesetting.

Definition 42.9. A *parametric curve* is a mapping of an interval $[a, b]$ or $\mathbb{R}$ into the n -dimensional space $\mathbb{R}^n$:

$$\varphi : [a, b] \rightarrow \mathbb{R}^n, \quad t \mapsto \mathbf{x}(t).$$

In computer graphics, we normally deal with parametric curves with $n = 2$ and $n = 3$.

Example 42.47 (Wikipedia). A list of some well-known 2-D and 3-D parametric curves.

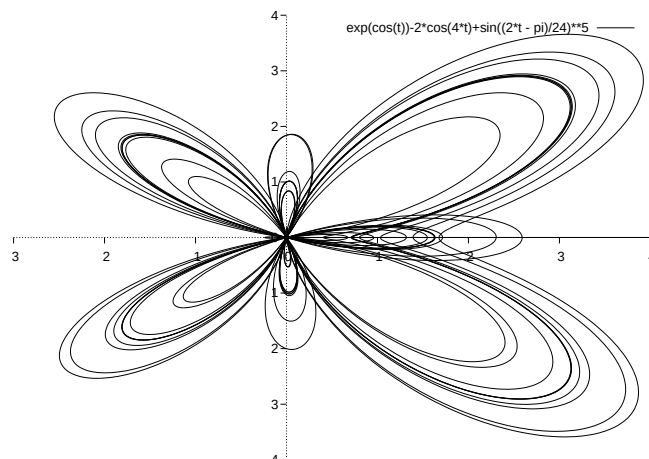
Elliptic curves. $x = a \cos t,$ $y = b \sin t.$  $a = 2, b = 1$	Let $d = a - b$, $k = \frac{a}{b}$ $x = d \cos t + b \cos\left(\frac{d}{b}t\right)$ $y = d \sin t - b \sin\left(\frac{d}{b}t\right)$  $a = 1, b = 4$	$x = \cos(at) - [\cos(bt)]^j$ $y = \sin(ct) - [\sin(dt)]^k$  $a = c = 1, b = 60,$ $d = 120, j = 3, k = 4$	Torus. $x = (R + r \cos u) \cos t,$ $y = (R + r \cos u) \sin t,$ $z = r \sin u.$  $R = 2, r = 0.5$

Example 42.48 (Wikipedia [2013]). The butterfly curve is a transcendental plane curve discovered and patterned by Temple H. Fay. The curve is given by the following parametric equations:

$$x = \sin(t) \left(e^{\cos(t)} - 2 \cos(4t) - \sin^5\left(\frac{t}{12}\right) \right)$$

$$y = \cos(t) \left(e^{\cos(t)} - 2 \cos(4t) - \sin^5\left(\frac{t}{12}\right) \right)$$

and a plot for $t \in [0, 2\pi]$ looks like



Interpolating the parametric curves φ in $\mathbb{R}^n$ can be achieved by using either the interpolating polynomials or splines discussed earlier. We need to introduce a parameter t on an interval $[t_0, t_n]$ with $t_0 < t_1 < \dots < t_n$, and construct interpolating polynomials or splines for each component $x^{(j)}$ of φ at the knots t_i , $i = 0, 1, \dots, n$. An example for $n = 2$ is given below.

eg:spline

Example 42.49. Given the points $(1, 1)$, $(2, 1)$, $(2, 3)$ and $(1, 2)$. Find the “cubic spline curve” passing through them.

Solution: These points can be regarded as points of a virtual function φ . By default, we normally pick $t_0 = 0$ and $t_n = 1$ and divide the interval uniformly leading to the following table:

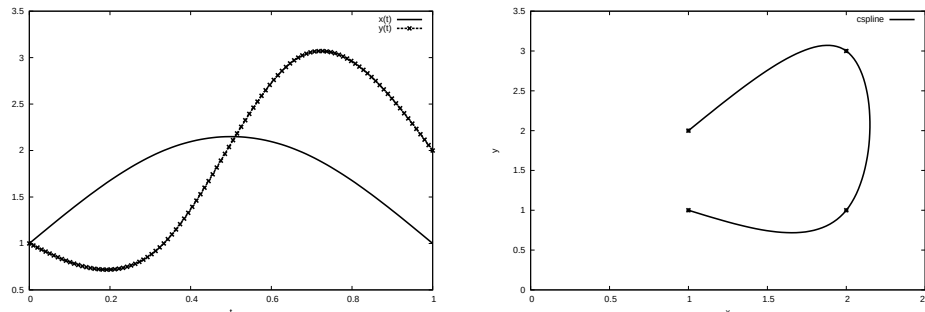
t_i	0	1/3	2/3	1
x_i	1	2	2	1
y_i	1	1	3	2

Using the natural spline algorithm (i.e. the `csplines` function), we obtain

$$x(t) = \begin{cases} 1 + 3.6t - 5.4t^3, & 0 \leq t < \frac{1}{3} \\ 2 + 1.8(t - \frac{1}{3}) - 5.4(t - \frac{1}{3})^2, & \frac{1}{3} \leq t < \frac{2}{3} \\ 2 - 1.8(t - \frac{2}{3}) - 5.4(t - \frac{2}{3})^2 + 5.4(t - \frac{2}{3})^3, & \frac{2}{3} \leq t \leq 1 \end{cases}$$

$$y(t) = \begin{cases} 1 - 2.2t + 19.8t^3, & 0 \leq t < \frac{1}{3} \\ 1 + 4.4(t - \frac{1}{3}) + 19.8(t - \frac{1}{3})^2 - 45(t - \frac{1}{3})^3, & \frac{1}{3} \leq t < \frac{2}{3} \\ 3 + 2.6(t - \frac{2}{3}) - 25.2(t - \frac{2}{3})^2 + 25.2(t - \frac{2}{3})^3, & \frac{2}{3} \leq t \leq 1. \end{cases}$$

If we sketch x and y against the parameter t , we will obtain the curves on the left of the following figures. If we plot y against x , we obtain the figure on the right.



Osculating polynomial and splines in their usual form are not suitable for the characterisation of parametric curves in computer modelling because an alteration in the location of one “point” (which is called *knot* in the computational geometry language, see Definition 42.7) requires the re-calculation of all coefficients of interpolating polynomial or the spline according to either (42.3), (42.5), (42.13) or (42.16). They are also numerically unstable (related to rounding error in Topic 1) in their original form.

Computer graphics, geometric modelling, etc. require rapid generation of smooth curves that can be drawn and modified quickly. Spline curves and some special osculating polynomials such as cubic Hermite polynomial (42.11) are used.

A *cubic Hermite polynomial* joins two points P_i and P_{i+1} with a cubic function

$$P(t) = t^3 \mathbf{a} + t^2 \mathbf{b} + t \mathbf{c} + \mathbf{d}$$

such that

$$P(0) = P_i, \quad P(1) = P_{i+1}, \quad P'(0) = Q_i, \quad P'(1) = Q_{i+1}.$$

It is fast and easy to solve for $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$:

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \\ \mathbf{c} \\ \mathbf{d} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 3 & 2 & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{P}_i \\ \mathbf{P}_{i+1} \\ \mathbf{Q}_i \\ \mathbf{Q}_{i+1} \end{bmatrix}.$$

Therefore a Hermite curve in $\mathbb{R}^n$ joining $\mathbf{P}_i$ and $\mathbf{P}_{i+1}$ with the “slope” $\mathbf{Q}_i$ and $\mathbf{Q}_{i+1}$ respectively is

$$\mathbf{P}(t) = \begin{bmatrix} t^3 & t^2 & t & 1 \end{bmatrix} \begin{bmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{P}_i \\ \mathbf{P}_{i+1} \\ \mathbf{Q}_i \\ \mathbf{Q}_{i+1} \end{bmatrix}.$$

For most problems in computer graphics, it is useful to generate curves without need to enter curve slopes $\mathbf{Q}_i$ and $\mathbf{Q}_{i+1}$. Spline curves are the generalisation of spline functions (Definition 42.7) which uses “control points” instead of collocation points to generate curves. We will limit the parameter domain to $[0, 1]$ without loss of generality.

Definition 42.10. Let $\mathbf{P}_i \in \mathbb{R}^n, i = 0, 1, \dots, n$ be a set of *control points* and let $F_i : [0, 1] \rightarrow \mathbb{R}, i = 0, \dots, n$ be a set of *blending functions*, a *spline curve* (associated with the control points and blending functions) or *generalised interpolation* [Thévenaz et al., 2000] is defined as

$$\mathbf{S}(t) = \sum_{i=0}^n F_i(t) \mathbf{P}_i. \quad (42.20)$$

e:spline-curve

This definition is so general that it includes too many possibilities. In computer graphics, useful spline curves need to be *affine invariant* ($\sum_{i=0}^n F_i(t) = 1$), being convex hull ($F_i(t) \geq 0$ for all i) and some regularity constraints.

One popular spline curve is the Bézier curve proposed by the French engineers Pierre Bézier (from Renault) and Paul de Casteljau (from Citroën). Together with the spline curves based on B-splines, they are now the most important mathematical representations of curves and surfaces used in computer graphics and CAD. Cubic Bézier curve segment is employed in the PostScript page-description language exclusively in the construction of curved paths, including very close approximations to circles.

def:bezier

Definition 42.11. A *Bézier curve* of degree n (or *order* $n + 1$) is specified by a sequence of $n + 1$ points $\mathbf{P}_i, i = 0, \dots, n$, which are called the *control points*:

$$\mathbf{B}_n(t) = \sum_{i=0}^n \binom{n}{i} t^i (1-t)^{n-i} \mathbf{P}_i, \quad 0 \leq t \leq 1. \quad (42.21)$$

e:bezier

The polygon obtained by joining the control points with line segments in the prescribed order is called the *control polygon*. Many software packages (such as gnuplot and PostScript) provide special functions to display cubic Bézier curve, i.e. (42.21) with $n = 3$:

$$\mathbf{B}_3(t) = \begin{bmatrix} t^3 & t^2 & t & 1 \end{bmatrix} \begin{bmatrix} -1 & 3 & -3 & 1 \\ 3 & -6 & 3 & 0 \\ -3 & 3 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{P}_0 \\ \mathbf{P}_1 \\ \mathbf{P}_2 \\ \mathbf{P}_3 \end{bmatrix}.$$

Control of the shape of a Bézier curve is facilitated by the fact that the control polygon reflects the basic shape of the curve. In many drawing and CAD packages the control points

of a Bézier curve may be specified by clicking with a mouse at the desired locations within a document window. The control points are visible on the computer screen, and modification of a control point is executed with a simple click and drag operation of the mouse. As an aid to design, many packages display the changing curve as a control point is modified [Marsh, 2005].

One excellent property of Bézier curve is that when a point P_k is changed to Q_k , $B_n(t)$ will need only minor changes as follows:

$$\tilde{B}_n(t) = B_n(t) + (Q_k - P_k)C_k^n(1-t)^{n-k}t^k$$

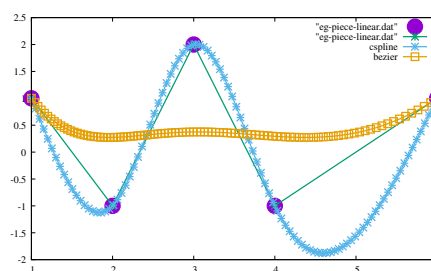
instead of solving a linear algebra problem as in spline function (Section [42.4.3](#)).

Example 42.50. Study the difference between splines and Bézier curve for the 5 given points $(1, 1)$, $(2, -1)$, $(3, 2)$, $(4, -1)$, $(6, 1)$ from Example [42.40](#) using gnuplot.

Solution: By looking the manual of the plot function in gnuplot, we discover that it supports the fitting of data using Bézier curve for $n = 5$ points:

```
1 set border linewidth 0.2
2 set terminal pdfcairo font "Times,12" linewidth 2
3 set output "eg-bezier.pdf"
4 plot "eg-piece-linear.dat" w p pt 7 ps 2, \
5      "eg-piece-linear.dat" w lp pt 3, \
6      "eg-piece-linear.dat" smooth csplines title "cspline" w lp, \
7      "eg-piece-linear.dat" smooth bezier title "bezier" w lp
8 #pause -1
```

From the figure on the right, it is easy to see that both linear spline and cubic pass through all the given data points. However, for Bézier curve, the internal knots are just control points and are not part of the Bézier curve.



Now, let us calculate the polynomial characterising the Bézier curve for a given data.

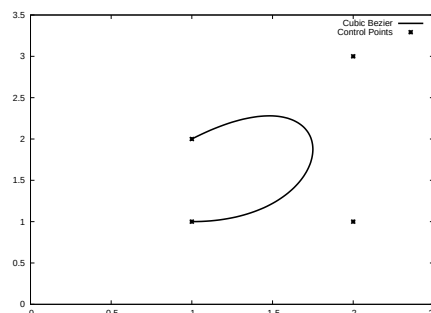
[eg:bezier](#)

Example 42.51. Given the points $(1, 1)$, $(2, 1)$, $(2, 3)$ and $(1, 2)$. Find the cubic Bézier curve with the given points as control points and sketch it.

Solution:

By Definition [42.11](#), the Bézier curve is

$$B_3(t) = (1-t)^3 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + 3t(1-t)^2 \begin{bmatrix} 2 \\ 1 \end{bmatrix} + 3t^2(1-t) \begin{bmatrix} 2 \\ 3 \end{bmatrix} + t^3 \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$



Example 42.52 (Tutorial 4, Q12). Construct a cubic Bézier curve with the *control polygon* $(1, 1)$, $(1.5, 1.25)$, $(7, 3)$ and $(6, 2)$.

Solution.

```

\begin{tikzpicture}[scale=.7]
  \filldraw [gray] (1,1) circle [radius=2pt]
    (1.5,1.25) circle [radius=2pt]
    (6,2) circle [radius=2pt]
    (7,3) circle [radius=2pt];
  \draw (1,1) .. controls (1.5,1.25) and (7,3) .. (6,2);
\end{tikzpicture}

```



$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = (1-t)^3 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + 3t(1-t)^2 \begin{bmatrix} 1.5 \\ 1.25 \end{bmatrix} + 3t^2(1-t) \begin{bmatrix} 7 \\ 3 \end{bmatrix} + t^3 \begin{bmatrix} 6 \\ 2 \end{bmatrix}.$$

The cubic Bézier curve (and all Bézier curves in general) don't actually go through the "control points". This does not affect computer graphic designers to use the control points to create beautiful curves and surfaces. The deciding factor for engineers to choose Bézier curves is a recursive property that makes them extremely efficient to compute. There is an algorithm attributed to de Casteljau that draws them very quickly. To construct the degree n Bézier curve for $n + 1$ points $P_0, \dots, P_n$, we can apply the following *de Casteljau's algorithm*:

1. All given control points are arranged into a column, which is the left-most one in a table. For each pair of adjacent control points, draw a south-east bound arrow and a north-east bound arrow, and write down a new point at the intersection of the two adjacent arrows. For example, if the two adjacent points are (i, j) and $(i, j + 1)$, the new point is $(i + 1, j)$. The south-east (resp., north-east) bound arrow means multiplying $1 - u$ (resp., u) to the point at its tail, (i, j) (resp., $(i, j + 1)$), and the new point is the sum.
2. From the initial column, column 0, we compute column 1; from column 1 we obtain column 2 and so on. Eventually, after n applications we shall arrive at a single point $(n, 0)$ and this is the point on the curve.

De Casteljau's algorithm (which has a close relation to the divided-difference algorithm in Section [42.2.4](#) ^{[sec:newton-f](#)} in Fortran is listed below:

```

1 function decasteljau(n,t,P)
2   intent(in) :: n, t, P ! n=number of points, degree=n-1
3   integer :: n,i,j
4   real(c_double) :: t, P(n), Q(n), decasteljau
5   Q(1:n) = P(1:n)
6   do i=2,n
7     do j=1,n-i+1
8       Q(j) = (1-t)*Q(j) + t*Q(j+1)
9     end do
10  end do
11  decasteljau = Q(1)
12 end function decasteljau

```

Remark 42.53. de Casteljau's algorithm takes $n(n + 1)/2$ affine combinations, each of which requires two multiplications and one addition. Why would we consider using such a slow algorithm? The main reason is that the code for de Casteljau's algorithm is simpler to write than is the code for forward differencing, especially if we want arbitrary degrees.

While speed is the primary advantage of forward differencing, de Casteljau's algorithm is easily fast enough to draw curves at interactive rates. Thus, the complicated fast forward differencing algorithm is unlikely to be chosen over de Casteljau's algorithm for drawing curves [Mann, 2006].

Various works by Donald Knuth explain clearly the role of Bézier curves in font design. Knuth [1986, pp123–131] explains extremely clearly the author's implementation of Bézier curves in his program MetaFont. For the admittedly rare programmer who wishes to build his own implementation (at the level of pixels), or for anyone who wants to see what attention to detail in first class work really amounts to, this is the best resource available.

A generalisation of Bézier curve is the “B-spline”. For de Boor [1986], a *B-spline* is just a “basis” (i.e. the blending functions in (42.20) for a spline curve. However, due to some kind of confusion, computer graphics researchers treated spline curves with B-spline basis as “B-spline”. And because computer graphics researchers get more funding and published papers, their terminology is now the standard and Henriksen and Erleben [2002] adopted the name “B-spline” for “spline” with a “B-spline basis function” defined below.

Definition 42.12. Let $m = n + 1 + k$. The *knot vector* $T = \{t_0, t_1, \dots, t_m\}$ is a non-decreasing sequence such that $t_i \in [0, 1]$. The “knots” $t_{k+1}, \dots, t_{m-k-1}$ are called *internal knots*.

A *normalised B-spline basis function* (this is the name used by computer graphics researchers but de Boor [1986] insisted to call it *B-spline*) is defined as

$$N_{i,k+1}(t) = \frac{t - t_i}{t_{i+k} - t_i} N_{i,k}(t) + \frac{t_{i+k+1} - t}{t_{i+k+1} - t_{i+1}} N_{i+1,k}(t) \quad (42.22)$$

for $k > 0$ and when $k = 0$:

$$N_{i,1}(t) = \begin{cases} 1, & \text{if } t_i \leq t < t_{i+1} \text{ and } t_i < t_{i+1}; \\ 0, & \text{otherwise.} \end{cases}$$

The index $k + 1$ is called the *order* of the B-spline. When a B-spline has an order $k + 1$, then the polynomial representation has a *degree* of $\leq k$.

Definition 42.13. When there are $n + 1$ control points $P_0, \dots, P_n$. Then the curve defined by

$$C(t) = \sum_{i=0}^n N_{i,k+1}(t) P_i, \quad 2 \leq k \leq n + 1 \quad (42.23)$$

is called (*uniform*) *B-spline (curve)*. When there is no internal knots, (42.23) becomes a *Bézier curve*.

A B-spline curve of degree m with n control points consist of $n - m$ Bézier curve segments. These segments all have C^2 continuity at the join points. Any Bézier curve of arbitrary degree can be converted in to a B-spline and any B-spline can be converted in to one or more Bézier curves.

In Example 42.34, it is possible to write the spline as a linear combination of B-splines as in Equation (42.12) and with a linear scaling, it is possible to write it in the form (42.20) with (42.22) and $k = 1$. According to Dierckx [1993, p10], all spline curves can be written as a linear combination of basis functions (42.22) just like the equivalence of the linear spline $P_1(x)$ in Example 42.34 and Equation (42.12).

B-spline curves are polynomial curves. While they are flexible and have many nice properties for curve design, they are not able to represent the simplest curve: the circle. Mathematically, circles can only be represented with rational functions, i.e., functions that are quotients of two polynomials. To cope with circles, ellipses and many other curves that cannot be represented by polynomials, we need an extension of B-spline curves to rational curves using homogeneous coordinates and this leads to the *Non-Uniform Rational B-Splines (NURBS)*.

GNU Scientific library provides the subroutines to handle B-splines and the manual is available at http://www.gnu.org/software/gsl/manual/html_node/Basis-Splines.html. Another B-spline library in C++ is available from <https://github.com/ncareol/bspline> (BSD) and Boost.Math. The following program from Boost documentation illustrates the complexity of merging various B-spline and the numerical methods from Topic 3.

```

2 #include <iostream>
3 #include <random>
4 #include <boost/random.hpp>
5 #include <boost/math/interpolators/cubic_b_spline.hpp>
6 #include <boost/math/tools/roots.hpp>
7
8 int main()
9 {
10     std::vector<double> v(500); // We begin with an array of samples
11     double step{0.01};         // And decide on a stepsize
12
13     // Initialize the vector with a function we'd like to interpolate:
14     for (size_t i = 0; i < v.size(); ++i) { v[i] = sin(i*step); }
15     // We could define an arbitrary start time, but for now we'll just use 0:
16     boost::math::cubic_b_spline<double>
17         spline(v.data(), v.size(), 0 /* start time */, step);
18
19     // Now we can evaluate the spline wherever we please.
20     std::mt19937 gen;
21     boost::random::uniform_real_distribution<double> absissa(0, v.size()*step);
22     for (size_t i = 0; i < 10; ++i) {
23         double x = absissa(gen);
24         std::cout << "sin(" << x << ") = " << sin(x)
25             << ", spline interpolation gives " << spline(x) << "\n";
26         std::cout << "cos(" << x << ") = " << cos(x)
27             << ", spline derivative interpolation gives "
28             << spline.prime(x) << "\n";
29     }
30
31     // The next example is less trivial:
32     // We will try to figure out when the population of the United States
33     // crossed 100 million. Since the census is taken every 10 years,
34     // the data is equally spaced, so we can use the cubic b spline.
35     // Data taken from https://en.wikipedia.org/wiki/United_States_Census
36     // We'll start at the year 1860:
37     double t0{1860}, time_step{10};
38     std::vector<double> population{
39         31443321, /* 1860 */      39818449, /* 1870 */
40         50189209, /* 1880 */      62947714, /* 1890 */
41         76212168, /* 1900 */      92228496, /* 1910 */
42         106021537, /* 1920 */     122775046, /* 1930 */
43         132164569, /* 1940 */     150697361, /* 1950 */
44         179323175}; /* 1960 */
45
46     // An eyeball estimate indicates that the population crossed 100
47     // million around 1915. Let's see what interpolation says:
48     boost::math::cubic_b_spline<double> p(population.data(),
49         population.size(), t0, 10);
50
51     // Now create a function which has a zero at p = 100,000,000:
52     auto f = [=](double t){ return p(t) - 100000000; };
53
54     // Boost includes a bisection algorithm, which is robust, though not
55     // as fast as some others we provide, but lets try that first. We
56     // need a termination condition for it, which takes the two endpoints
57     // of the range and returns either true (stop) or false (keep going),
58     // we could use a predefined one such as
59     // boost::math::tools::eps_tolerance<double>, but that
60     // won't stop until we have full double precision which is overkill,
61     // since we just need the endpoint to yield the same month. While
62     // we're at it, we'll keep track of the number of iterations required
63     // too, though this is strictly optional:
64     auto termination = [](double left, double right){
65         double left_month {std::round((left - std::floor(left)) * 12 + 1)};
66         double right_month{std::round((right - std::floor(right)) * 12 + 1)};
67         return (left_month==right_month) &&
68             (std::floor(left)==std::floor(right));

```



```

69     };
70     std::uintmax_t iterations = 1000;
71     auto result = boost::math::tools::bisect(
72         f, 1910.0, 1920.0, termination, iterations);
73     auto time = result.first; // termination condition ensures that both
74                             // endpoints yield the same result
75     auto month = std::round((time - std::floor(time))*12 + 1);
76     auto year = std::floor(time);
77     std::cout << "The population of the US surpassed 100 million on the ";
78     std::cout << month << "th month of " << year << std::endl;
79     std::cout << "Found in " << iterations << " iterations" << std::endl;
80
81     // Since the cubic B spline offers the first derivative, we could equally
82     // have used Newton iterations, this takes "number of bits correct" as
83     // a termination condition - 20 should be plenty for what we need,
84     // and once again, we track how many iterations are taken:
85     auto f_n = [=](double t) {
86         return std::make_pair(p(t) - 100000000, p.prime(t));
87     };
88     iterations = 1000;
89     time = boost::math::tools::newton_raphson_iterate(f_n, 1910.0,
90         1900.0, 2000.0, 20, iterations);
91     month = std::round((time - std::floor(time))*12 + 1);
92     year = std::floor(time);
93     std::cout << "The population of the US surpassed 100 million on the ";
94     std::cout << month << "th month of " << year << std::endl;
95     std::cout << "Found in " << iterations << " iterations" << std::endl;
96 }

```

42.7 Application: Computer Graphics and CAD

There are more than the generation of formula for B-splines in computer graphics. Mathematical problems such as derivatives, curvatures, intersections, etc. are not discussed here but found on in the Internet such as <https://pomax.github.io/bezierinfo/>.

So where are the applications of these mathematical ideas? They are used in font design (e.g. <https://www.youtube.com/watch?v=jAdspOtgciQ>, <https://cormullion.github.io/blog/2018/06/21/bezier.html>). Note that TTF uses quadratic Bézier curves, while OTF uses cubic Bézier curves, see https://www.reddit.com/r/typography/comments/8rqwx5/questions_regarding_otf_vs_ttf/), 2-D vector graphics [Casselmann, 2005], desktop publishing (e.g. <https://www.youtube.com/watch?v=S-NXz7I5dQc>, <https://www.youtube.com/watch?v=48tFB-sjHgY>, <https://www.youtube.com/watch?v=-cF0sAzigyQ>), web layout engine (e.g. Gecko, WebKit, Blink, refer to https://en.wikipedia.org/wiki/Comparison_of_browser_engines for more), computer game engines. These are the applications that are closer to everyday life.

Remark 42.54. Interesting Project: *Curves in Postscript Programming Language*

For specialised applications, they are used in geometric modelling, computer aided design (CAD) software, computer numerical control (CNC), robotics, etc. They are coming into business as fast prototyping and industrial revolution 4.0. Bartels et al. [1987] gave a detail discussion of the application of spline curves in CAD.

Much of the objects studied in this topic are 1-D and can only be used in simple curve fitting and the ODE-BVP problem. For applications in CAD and PDE, 2-D and 3-D geometric objects are required. There are many important CAD program such as AutoCAD in engineering, here we will introduce one with programming feature and freely available for download — OpenSCAD (based on CGAL). It is a software which provides very primitive CAD design using programming to handle geometric objects. For example, writing a OpenSCAD script

```

1 module example005()
2 {
3     N = 6;

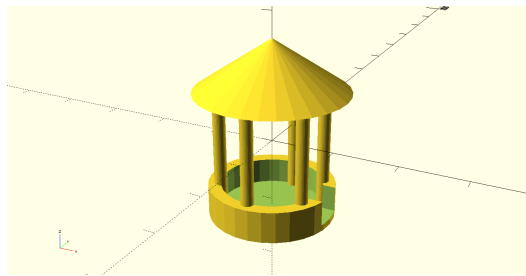
```

```

4     translate([0, 0, -120]) {
5         difference() {
6             cylinder(h = 50, r = 100);
7             translate([0, 0, 10]) cylinder(h = 50, r = 80);
8             translate([100, 0, 35]) cube(50, center = true);
9         }
10        for (i = [0:5]) {
11            echo(360*i/N, sin(360*i/N)*80, cos(360*i/N)*80);
12            translate([sin(360*i/N)*80, cos(360*i/N)*80, 0 ])
13                cylinder(h = 200, r=10);
14        }
15        translate([0, 0, 200])
16            cylinder(h = 80, r1 = 120, r2 = 0);
17    }
18 }
19
20 example005();

```

it is possible to produce the following geometric object rendering



Lloyd N. Trefethen, Spectral Methods in MATLAB (2000). John P. Boyd, Chebyshev and Fourier Spectral Methods (2nd ed., 2001). Chebfun Project: <https://www.chebfun.org/>

To work on:

Graphics:

2-D plots: Plotting a set of data, Plotting a function, Plotting parametric curves

Line styles, markers and colours, Labels, title, legends, Axis control Overlay plots and subplots, Specialized 2-D plots

Plotting 2D vector fields

3-D plots : Curve plot, contour, surface, mesh Colour bar and colour shading

Plotting 3D vector fields ?

References:

Giordano, F.R., Fox, W.P., Horton, S.B. and Weir, M.D., 2009. A first course in mathematical modeling. 4th ed. Singapore: Brooks/Cole Cengage Learning.

2. Higham, D. and Higham, A., 2008. MATLAB guide. 2nd ed. New York: SIAM.

3. Herniter, M.E., 2001. Programming in MATLAB. New York: Thomson/BrooksCole.

4. Chen, K., Giblin, P.J. and Irving, A., 1999. Mathematical explorations with MATLAB. Cambridge: Cambridge U. Press

Chapter 43

Nonlinear Systems

In Topic 2, we have discussed three basic “linear” problems, i.e. (a) the solution of linear system; (b) least square problem; and (c) eigenvalue problem. The generalisation of (a) is a *nonlinear system*:

$$\mathbf{f}(\mathbf{x}) = \mathbf{0}, \quad \mathbf{x} \in D \subset \mathbb{R}^n \quad (43.1) \quad \text{e:nle-n}$$

where $\mathbf{f} = (f_1, f_2, \dots, f_n)^T$ is a function from D to $\mathbb{R}^n$ and $\mathbf{0} = (0, 0, \dots, 0)^T$ is a zero vector. The vector $\mathbf{x}^* = (x_1^*, x_2^*, \dots, x_n^*) \in D$ such that $\mathbf{f}(\mathbf{x}^*) = \mathbf{0}$ is called the *(real) root* (or *solution*) of function $\mathbf{f}$. Note that Topic 2 Equation (2.1) is a special case of (43.1) with $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$.

The generalisation of (b) maybe the *learning machine* of a *training set* $\{(Z^1, D^1), \dots, (Z^P, D^P)\}$:

$$\min_{W \in \mathbb{R}^k} \sum_{i=1}^P C(D^p, M(Z^p, W)) \quad (43.2) \quad \text{e:learning-machine}$$

where Z^p and D^p , $p = 1, \dots, P$ are respectively the input patterns and the corresponding desired outputs, W is the parameters of the learning machine. When $C(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$, $D^p = \mathbf{y}_p$, $Z^p = \mathbf{x}_p$, $W = \{\mathbf{A}\}$ and $M(Z^p, W) = \mathbf{A}\mathbf{x}_p$, Equation (43.2) becomes Topic 2 Equation (2.2).

$$\min_{(\mathbf{A}, \mathbf{b}) \in \mathbb{R}^{n^2}} \sum_{i=1}^P \|\mathbf{y}^p - \mathbf{A}\mathbf{x}_p\|_2^2.$$

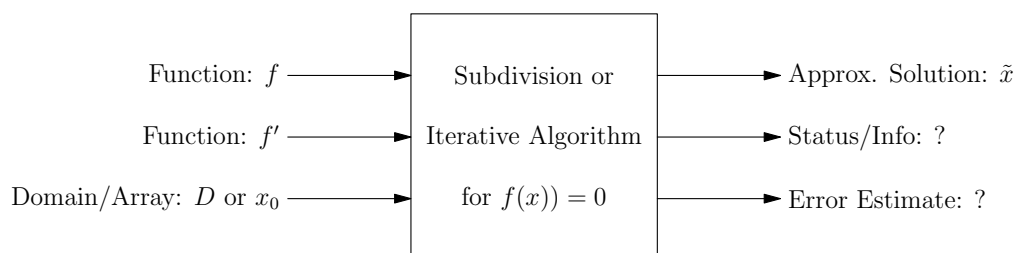
The generalisation of (c) eigenvalue problem is associated with PDE (Topic 8) and operator theory and is beyond an undergraduate level.

There is no standard method to find all the solutions of a nonlinear system (43.1). To under the complexity of finding a solution, all textbooks will start with a special case of (43.1) with $n = 1$. When $n = 1$, we have a *(nonlinear) equation*:

$$f(x) = 0, \quad x \in D \subset \mathbb{R} \quad (43.3) \quad \text{e:nle}$$

where f is a function from D to $\mathbb{R}$. The value x^* such that $f(x^*) = 0$ is called a *(real) root* (or *solution*) of the function f . In this topic, we will discuss two basic numerical techniques for finding approximate solutions to (43.3). The Excel function =rate(N, -PMT, PV, FV) probably uses Newton method to find the annuity $FV = PV + \sum_{j=1}^N \text{PMT}(1+i)^j$. We will touch on the generalisation of root-finding methods to a *system of equations* in Section 43.4.

There is no algebraic formula which gives us all the roots of a nonlinear equation (43.3). Mathematicians have studied several iterative techniques which allow us to “locate” and “approximate” some of the roots in Section 43.1, Section 43.2.2 and Section 43.2.3. These techniques can be regarded as the following black box.



In Section 43.6, we will try to explore techniques to solve “nonlinear” least square problem (43.2).

43.1 The Bisection Method

It is impossible to find a root from a general function f in (43.3). If limit ourselves to “continuous” functions (Calculus I), there is an effective method, called the *bisection method*, to find the real zeros of a one-dimensional continuous function. It is a root-finding algorithm which works by repeatedly diving the interval in (43.3) in half and then selecting the subinterval in which the root exists, which is theoretically guaranteed by the *Intermediate Value Theorem* in Subsection 43.1.1.

43.1.1 Properties of Continuous Functions

Theoretically, to find a root for the equation (43.3), we must “confirm” that a root exists. Otherwise, we will be looking for something that does not exists. The Intermediate Value Theorem confirms that for a continuous function satisfying a simple condition, a root exists.

Definition 43.1 (Continuous Function, Calculus I). Given a function $f : [a, b] \rightarrow \mathbb{R}$ and an element $c \in [a, b]$, the function f is said to be *continuous* at the point c if the following holds: For any number $\varepsilon > 0$, there exists some number $\delta > 0$ such that for all $x \in [a, b]$ with $c - \delta < x < c + \delta$, the value of $f(x)$ satisfies

$$f(c) - \varepsilon < f(x) < f(c) + \varepsilon.$$

A function f which is continuous at every point in $[a, b]$ is said to be *continuous over* $[a, b]$ and we denote the set of all continuous functions over $[a, b]$ as $C[a, b]$. So, instead of writing down “a function f is continuous”, mathematicians would sometimes write “ $f \in C[a, b]$ ”.

A theorem which guarantees the existence of a root for (43.3) when f is continuous.

Theorem 43.1 (Intermediate Value Theorem). Given a function $f \in C[a, b]$, when $f(a)$ and $f(b)$ have **opposite signs**, f will have at least a root in $[a, b]$, i.e.

$$\exists x^*(x^* \in [a, b] \wedge f(x^*) = 0).$$

43.1.2 Implementation

The bisection is implemented in Python Scipy module `scipy/optimize/Zeros/bisect.c`. It takes a continuous function f and the end points of an interval $[a, b]$ as input and then calculates $f(a)$ and $f(b)$ to see if they have opposite signs or any one of them is zero. If any of them is zero, we have found the zero and return, otherwise we locate the root in either $[a, \frac{a+b}{2}]$ or $[\frac{a+b}{2}, b]$. We will then continue with the sub-interval containing the root and continue to sub-divide the interval until we get a small enough sub-interval containing the root. The bisection method in Fortran 2003 is shown below.

```
1 subroutine bisect(f, a, b, root, info, ndiv)
2   procedure(func) :: f
3   double precision, intent(in) :: a, b
4   double precision, intent(out) :: root
5   integer, intent(out), optional :: info, ndiv
6   double precision :: left, h, fa, fb, fm, mid
7   integer :: iter
8   fa = f(a); fb = f(b)
9   if (sign(1d0, fa)*sign(1d0, fb)>0.0) then
10      root = 0d0 ! No solution
```

```

11         if(present(info)) info = -1
12         return
13     end if
14     if (fa .eq. 0d0 .or. fb .eq. 0d0) then
15         if (fa .eq. 0d0) root = a
16         if (fb .eq. 0d0) root = b
17         if(present(info)) info = 0
18         if(present(ndiv)) ndiv = 0
19         return
20     end if
21     left = a; h = b-a
22     do iter=1, ceiling((log(b-a)-log(accur))/log(2d0))
23         h = 0.5d0*h; mid = left+h; fm = f(mid)
24         if (sign(1d0,fa)*sign(1d0,fm)>0.0) left = mid
25         if (fm .eq. 0d0 .or. abs(fm)<accur .or. abs(h)<accur) then
26             root = mid
27             if(present(info)) info = 0
28             if(present(ndiv)) ndiv=iter
29             return
30         end if
31     end do
32     if (present(info)) info = iter ! Root Not Found After MaxIter
33 end subroutine bisect

```

Remark 43.2. The Fortran function $\text{sign}(A, B)$ returns the value of A with the sign of B . The reason to avoid using $f(a)f(b) > 0$ to see if $f(a)$ and $f(b)$ are of the same sign is to prevent an overflow (or underflow) when both $f(a)$ and $f(b)$ are large (or small).

The bisection method converges to a root x^* of f with the absolute error for the bisection method is at most $\frac{b-a}{2^n}$ after n steps. If ϵ is the desired accuracy, then to achieve an accuracy of at least ϵ in a minimum of N steps, we have

$$\frac{b-a}{2^N} < \epsilon \Rightarrow \ln(b-a) - N \ln 2 < \ln \epsilon \Rightarrow \frac{\ln(b-a) - \ln \epsilon}{\ln 2} < N.$$

The error is halved at every step, so the method converges linearly, which is quite slow. On the positive side, the method is guaranteed to converge if $f(a)$ and $f(b)$ have different sign.

eg:201401test1

Example 43.3 (Jan 2014 Test 1). Given $f(x) = 6x^3 - 7x^2 - 20x$ on the interval $[-2, 3]$.

- (a) Show that there is at least one root for $f(x)$ in the interval $[-2, 3]$. (2 marks)

Solution:

$f(-2) = -36 < 0$; $f(3) = 39 > 0$ [1.5 marks]

Since f is continuous, by Intermediate Value Theorem, f has a root in the interval $[-2, 3]$.

.....[0.5 mark]

- (b) Estimate at least how many bisections is required to obtain a root with an accuracy of 10^{-6} . (1 mark)

Solution. $n \geq \log(\frac{3-(-2)}{10^{-6}})/\log 2 \approx 22.2535$.

So 23 bisections are required to get an accuracy of 10^{-6} [1 mark]

□

- (c) Explain why bisection converges linearly. (1 mark)

Solution: By using the knowledge from Topic 2, the sequence $\{x_k\}$ for the function over the interval $[a, b]$, $a < b$ should satisfy

$$|x_k - x^*| \leq \frac{b-a}{2^k}, \quad k = 1, 2, 3, \dots$$

This says that x^* is either in the interval (a) $[x_k, x_k + \frac{b-a}{2^k}]$ or (b) $[x_k - \frac{b-a}{2^k}, x_k]$.

We will only consider case (a), case (b) can be studied in a similar way.

In case (a), assume $x^* \neq x_k$, otherwise the bisection method will stop. Therefore, $x_{k+1} = x_k + \frac{b-a}{2^{k+1}}$ and either (i) $x_k < x^* \leq x_{k+1}$ or (ii) $x_{k+1} < x^* \leq x_{k+1} + \frac{b-a}{2^{k+1}}$.

If $x_{k+1} = x^*$, then $|x_{k+1} - x^*| = 0 \leq c|x_k - x^*|$ where $0 < c < 1$.

Assume $x_{k+1} \neq x^*$, case (i) implies

$$(1) \ x_k < x_{k+2} < x^* < x_{k+1} \quad \text{or} \quad (2) \ x_k < x^* < x_{k+2} < x_{k+1}.$$

Case (1) implies $|x_{k+1} - x^*| \leq |x_k - x_{k+1}| = c|x_k - x^*|$ where $c = \frac{x^* - x_k}{x_{k+1} - x_k} < 1$.

Case (2) implies $|x_{k+2} - x^*| \leq c|x_{k+1} - x^*|$ where $c = \frac{x_{k+2} - x^*}{x_{k+1} - x^*} < 1$.

Case (ii) is similar to (i).

We can see that the analysis of convergence is very complicated. Trying to apply the idea of “approximation”, we assume $|x_k - x^*| \approx \frac{b-a}{2^k}$ (x_k should not be too close to x^* , otherwise, we get our answer x^*) and so

$$\left| \frac{x_{k+1} - x^*}{x_k - x^*} \right| \leq \left| \frac{\frac{b-a}{2^{k+1}}}{\frac{b-a}{2^k}} \right| = \frac{1}{2} \Rightarrow |x_{k+1} - x^*| \leq \frac{1}{2} |x_k - x^*|$$

By definition, $x_k \rightarrow x^*$ converges linearly. [1 mark]

eg:poly-bisect

Example 43.4. To find the root of the function $f(x) = x^5 - 3x + 7$ using bisection method for the range $[-3, 3]$ to the accuracy 10^{-7} using our Fortran module, one just need write a Fortran program to call setaccu and bisect as follows:

```
1 program bisect_eg
2   use textbook_root1
3   implicit none
4   double precision :: theroot
5   double precision, external :: f
6   integer :: info, cnt
7   call setaccu(1.0d-7)
8   call bisect(f, -3.0d0, 3.0d0, theroot, info, cnt)
9   write(*,*) "the root", theroot, "is found in", cnt, "bisections"
10 end program bisect_eg
11 function f(x)
12   double precision :: x, f
13   f = x**5 - 3.0*x + 7.0
14 end function f
```

Estimate theoretically how many steps we need and find the root numerically with the Fortran program.

Solution. Since the range is $[-3, 3]$, in theory, it will take us n steps to get

$$\frac{3 - (-3)}{2^n} < 10^{-7} \Rightarrow n > \frac{\log(6 \times 10^7)}{\log 2} \approx 25.8385.$$

So theoretically, we need 26 subdivisions. The Fortran calculation tells us that the root -1.6416942179203033 is found in 26 bisections. Using Python: `from scipy import optimize; optimize.bisect(lambda x: x**5-3.*x+7, -3, 3, xtol=1e-7)` $\square$

t2:q1

Example 43.5 (Tutorial 3, Q2). Show that $f(x) = \sqrt{x} - \cos x$ has a root in the interval $[0, 1]$ and use bisection method to find the third bisection x_3 .

Solution. $f(0) = -1$, $f(1) = 1 - \cos 1 > 0$. f is a continuous functions over $[0, 1]$ since $\sqrt{x}$ and $\cos x$ are continuous. By Intermediate Value Theorem, f has a root in $[0, 1]$.

$$x_1 = 0.5. \ f(0.5) = \sqrt{0.5} - \cos 0.5 = -0.17047578 < 0. \dots\dots\dots [0.5, 1]$$

$x_2 = 0.75$. $f(0.75) = 0.13433653 > 0$ [0.5, 0.75]
The third bisection $x_3 = 0.625$. $f(x_3) = -0.02039370 < 0$ [0.625, 0.75] $\square$

Let us apply the Fortran module to study an example related to financial mathematics.

eg:toi:tut3

Example 43.6 (Theory of Interest). Deposits of \$1,000 are made into an account at the end of each month for two years. The monthly rate of interest in the first year is j and the monthly rate of interest in the second year is $j + 0.002$. The balance in the account at the end of two years (just after the final deposit and monthly interest) is \$26,038. Find j using the following Fortran program:

```

1 program toi_bisect
2   use textbook_root1
3   double precision :: theroot
4   double precision, external :: f
5   integer :: cnt
6   call bisect(f,0d0,1d0,theroot,ndiv=cnt)
7   print *, "the root", theroot, "is found in ", cnt, "bisections"
8 end program toi_bisect
9 function f(j)
10  double precision :: j, f, fv    ! fv should never be initialised here
11  integer :: i
12  fv = 0d0
13  do i=11,0,-1
14    fv = fv + 1000.0*(1+j)**i*(1+j+0.002)**12 + 1000.0*(1+j+0.002)**i
15  end do
16  f = 26038d0 - fv
17 end function f

```

Solution. According to the time-value-principle,

$$\begin{aligned}
\$26,038 = \$1,000[(1+j)^{11} + \cdots + (1+j)^1 + 1](1+j+0.002)^{12} \\
+ \$1,000[(1+j+0.002)^{11} + \cdots + (1+j+0.002)^1 + 1].
\end{aligned}$$

The root is 5.5017471313476562E-003 found in 20 bisections. $\square$

ver:aims:ex2q1

Example 43.7 (Tutorial 3, Q3). Explain why the equation $e^{-x} = x$ has a solution on the interval $[0, 1]$. Use bisection to find the root to 4 decimal places (with the help of computer software). Can you prove that there are no other roots?

Solution. Rewrite the equation as $f(x) := e^{-x} - x = 0$. It is a continuous function because it is a linear combination of exponential and polynomial functions. Since $f(0) = 1 > 0$ and $f(1) \approx -0.6321 < 0$, by the Intermediate Value Theorem, f has a root in the interval $[0, 1]$.

We estimate $\left\lceil \frac{\ln \frac{1-0}{10^{-4}}}{\ln 2} \right\rceil = 14$ bisections are required and the exact calculation is as follows

$a^{(k)}$	$b^{(k)}$	$f(a^{(k)})$	$f(b^{(k)})$
0.0	1.0	1.0	-0.6321
0.5	1.0	0.1065	-0.6321
0.5	0.75	0.1065	-0.2776
0.5	0.625	0.1065	-0.0897
0.5625	0.625	0.0073	-0.0897
0.5625	0.59375	0.0073	-0.0415
0.5625	0.578125	0.0073	-0.0172
0.5625	0.5703125	0.0073	-0.0050
0.56640625	0.5703125	0.0012	-0.0050
0.56640625	0.568359375	0.0012	-0.0019
0.56640625	0.5673828125	0.0012	-0.0004
0.56689453125	0.5673828125	0.0004	-0.0004
0.567138671875	0.5673828125	7.24×10^{-6}	-0.0004
0.567138671875	0.5672607421875	7.24×10^{-6}	-0.0002
0.567138671875	0.56719970703125	7.24×10^{-6}	-8.84×10^{-5}

The root to 4 decimal places is $x_{15} = 0.56719970703125 \approx 0.5671$.
The root is unique because $f'(x) = -e^{-x} - 1 < 0$ over $(0, 1)$ indicating that f is decreasing monotonously. $\square$

43.2 Newton(-Raphson) Method & Secant Method

The bisection method mentioned in the previous section is based on the continuity of the function f in Equation (43.3). When a function f is not only continuous but also *differentiable*, the tangent to the graph of the function should play a role in the approximating the solution.

We will first study the properties of differentiable functions and then study the Newton-Raphson method (or Newton's Method), which is one of the most useful and best known algorithms that relies on the continuity of the derivatives f' and f'' . This method is attributed to Sir Isaac Newton (1643–1727) and Joseph Raphson (1648–1715), for details of its history, one can read the article from http://en.wikipedia.org/wiki/Newton's_method.

43.2.1 Properties of Differentiable Functions & Taylor's Series

Definition 43.2 (Differentiable Function, Calculus I). Given a function $f : [a, b] \rightarrow \mathbb{R}$ and an element $c \in (a, b)$, f is said to be *differentiable* at the point c if the limit

$$\lim_{x \rightarrow c} \frac{f(x) - f(c)}{x - c}$$

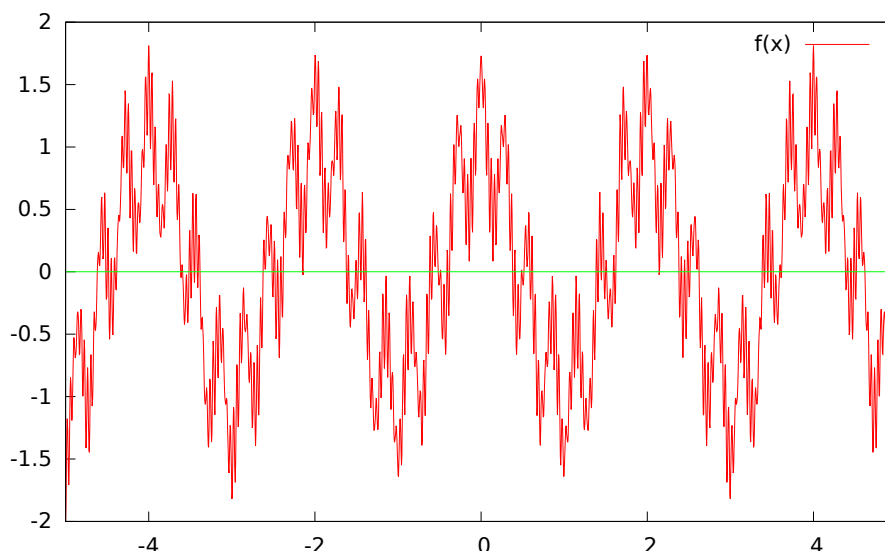
exists. A function f is said to be *differentiable* on $[a, b]$ if f is differentiable for all $x \in (a, b)$.

A continuous function may not be differentiable. In fact, mathematicians discovered that there are a lot more non-differentiable continuous function than differentiable continuous function.

Example 43.8. Around 1870, a German mathematician Weierstrass found that the “Weierstrass function” defined below:

$$f(x) = \sum_{n=0}^{\infty} a^n \cos(b^n \pi x), \quad 0 < a < 1, \quad b > \frac{1 + 3\pi/2}{a}$$

is continuous but nowhere differentiable. The following is a sketch of $f(x)$ when $a = \frac{1}{2}$ and $b = 7$:



Such “weird” nowhere-differentiable functions are good demonstrations for the functions found in financial mathematics. Such functions cannot be differentiated and calculus theory cannot be applied to them, therefore one needs a completely new theory to understand their behaviour and this made financial mathematics theory to be abstract and complex!

However, science and engineering models are normally “smooth” and all the theory from Calculus I, II and III can be applied.

Definition 43.3. A function f is said to be *continuously differentiable* if f is differentiable on $[a, b]$ and f' is a continuous on (a, b) . Continuously differentiable functions are said to be of class C^1 . A function f is of class C^2 if the first and second derivative of f both exist and are continuous. More generally, a function f is said to be *of class C^k* if the first k derivatives $f', f'', \dots, f^{(k)}$ all exist and are continuous. If derivatives $f^{(n)}$ are continuous for all positive integers n , the function is *smooth* or, equivalently, *of class C^∞* .

Note that it is possible to have a differentiable function in which the derivative is not a continuous function as shown below.

Example 43.9. The function

$$f(x) = \begin{cases} x^2 \sin(1/x) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

is differentiable at 0, since

$$f'(0) = \lim_{h \rightarrow 0} \left(\frac{h^2 \sin(1/h) - 0}{h} \right) = 0$$

exists. However, for $x \neq 0$,

$$f'(x) = 2x \sin(1/x) - \cos(1/x)$$

which has no limit as $x \rightarrow 0$. Therefore the derivative f' is not continuous at $x = 0$.

For a differentiable function, the mean value theorem below is a useful tool for deriving error estimates of many numerical methods.

Theorem 43.10 (Mean Value Theorem). *Let f be continuous over (a, b) and possess a derivative at each x for $a < \theta < b$. Then there is at least a number $\theta \in (a, b)$ such that*

$$f(b) - f(a) = (b - a)f'(\theta).$$

A sufficiently differentiable function f can be expressed as a Taylor polynomial as follows.

Theorem 43.11 (Taylor's Approximation). Suppose $f \in C^n[a, b]$ and $f^{(n+1)}$ exists on $[a, b]$. Let x_0 be a number in $[a, b]$. For all $x \in [a, b]$, there exists a number $\xi(x)$ between x_0 and x with

$$f(x) = \underbrace{\sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k}_{\text{Taylor polynomial}} + R_n(x) \quad (43.4)$$

e:taylor-approx

where

$$R_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!} (x - x_0)^{n+1}.$$

If f is of class C^∞ , then $n \rightarrow \infty$ and Taylor's approximation takes the form of *Taylor's series*:

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k.$$

When $x_0 = 0$, a Taylor's series is also called a *Maclaurin series* or a *power series*:

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \frac{f^{(4)}(0)}{4!}x^4 + \dots$$

Example 43.12. writing down the first five non-zero terms of the Maclaurin series of $f(x) = e^x \cos x$.

Solution: After calculating $\frac{f^{(k)}(0)}{k!}$ for $k = 0, 1, 2, 3, 4, 5$, we obtain the Maclaurin series:

$$f(x) = 1 + x - \frac{x^3}{3} - \frac{x^4}{6} - \frac{x^5}{30} + \dots$$

Similar to vectors, we can perform addition, subtraction, scalar multiplication on power series. We can also multiply and divide two power series. We can also substitute a function or a power series into a power series. A comprehensive explanation is found in Knuth [1998a, Section 1.2.9] and a brief introduction was given in UECM2033 ODE Topic 3.

In general, to make the operations on power series make sense from an analytic point of view, one needs to study the radius of convergence as well.

The manipulation of Taylor's series is carried out by computer algebra systems (CAS) such as Maple in modern day practice. However, even CAS can make mistakes, so one has to be careful when using CAS.

43.2.2 Newton's Method and Implementation

sec:newton

The derivation and geometrical meaning of Newton's method can be found in Burden and Faires [2015, Section 2.3] and we will just investigate how to perform calculation with them.

Definition 43.4. The *Newton's method* of a general equation (43.3) is an iteration process below:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, 2, \dots; \quad x_0 = \text{initial guess.} \quad (43.5)$$

e:newton

A Fortran implementation of the Newton's iteration for (43.5) is given below:

le:newton

```

1  subroutine newton(f, df, x0, root, info, nsteps)
2      double precision, intent(in) :: x0
3      double precision, intent(out) :: root
4      integer, intent(out), optional :: info, nsteps
5      double precision :: deltax, x, jac
6      procedure(func) :: f, df
7      x = x0
8      deltax = f(x)
9      do iter=1,maxIter
10         jac = df(x)
11         if (df(x)==0d0) then ! Ignore Fortran Warning
12             if (present(info)) info = 2
13             return
14         else if (abs(deltax)<accu) then
15             root = x
16             if (present(info)) info = 0
17             if (present(nsteps)) nsteps=iter-1
18             return
19         end if
20         deltax = f(x)/jac
21         x = x - deltax
22     end do
23     if (present(info)) info = 1 ! Root Not Found After MaxIter
24 end subroutine newton

```

By considering $T(x) = x - \frac{f(x)}{f'(x)}$, we find that Newton's method can be regarded as a fixed-point iteration method. If T satisfies the conditions in Fixed-Point Theorem 43.43, then a fixed point x^* exists. Mathematicians have proven that for suitable initial guess x_0 and when the $|f(x^*)| < 1$, then a unique solution exists according to Banach Fixed Point Theorem. brouwer-fp

Once we know that a solution exists, we want to know how fast does Newton's method converge? By using the Taylor's polynomial, we can analyse the convergence property of Newton's method by expanding the Taylor's polynomial at the fixed-point $x = x_k$:

$$f(x^*) = f(x_k) + f'(x_k)(x^* - x_k) + \frac{f''(\xi)}{2}(x^* - x_k)^2 = 0. \quad (43.6)$$

If $f'(x_k) \neq 0$ and f is C^2 , we rearrange the Taylor polynomial approximation (43.6) to e:newton-taylor

$$f(x_k) + f'(x_k)(x^* - x_k) = -\frac{f''(\xi)}{2}(x^* - x_k)^2 \Rightarrow \frac{f(x_k)}{f'(x_k)} + x^* - x_k = -\frac{f''(\xi)}{2f'(x_k)}(x^* - x_k)^2.$$

Therefore,

$$|x_{k+1} - x^*| = \left| x_k - \frac{f(x_k)}{f'(x_k)} - x^* \right| = \left| \frac{f''(\xi)}{2f'(x_k)}(x^* - x_k)^2 \right| \Rightarrow \boxed{|x^* - x_{k+1}| \leq \frac{M}{2|f'(x_k)|} |x^* - x_k|^2}.$$

where $M > 0$ is a minimum value such that $|f''(x)| \leq M$, for all $x \in [a, b]$. When $M < 2|f'(x)|$, we will have a convergence.

Example 43.13 (Tutorial 3, Q4). Let $f(x) = -x^3 - \cos x$ and $x_0 = -1$. Use Newton's method to find x_2 . Could $x_0 = 0$ be used?

Solution. $f'(x) = -3x^2 + \sin x$

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = -1 - \frac{-(-1)^3 - \cos(-1)}{-3(-1)^2 + \sin(-1)} = -0.88033290$$

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = -0.88033290 - \frac{-(-0.88033290)^3 - \cos(-0.88033290)}{-3(-0.88033290)^2 + \sin(-0.88033290)} = -0.86568416$$

Since $f'(0) = 0$, the Newton method will fail. □

Example 43.14. (a) Explain what does the following Fortran program do? (b) What will the value of x be when the variable $iter$ is 2 in the subroutine `newton` at line 22? (c) Finally, comment on the output of the program.

```

1 program newton_eg
2   use textbook_root1
3   implicit none
4   double precision :: theroot
5   double precision, external :: f, df
6   integer :: info, cnt
7   call setaccu(1.0d-7)
8   call newton(f,df,0d0,theroot,info,cnt)
9   write(*,*) "the root", theroot, "is found in", cnt, "iterations"
10 end program newton_eg
11 function f(x)
12   double precision :: x, f
13   f = x**5 - 3.0*x + 7.0
14 end function f
15 function df(x)
16   double precision :: x, df
17   df = 5.*x**4 - 3.
18 end function df

```

Solution. (a) The Fortran program find the root of the function $f(x) = x^5 - 3x + 7$ to an accuracy of 7 decimal places using Newton's method with an initial value $x_0 = 0$.

(b) To apply the Newton's method, we first need to find the derivative of f :

$$f'(x) = 5x^4 - 3.$$

We can then construct the Newton's iteration as follows

$$x_0 = 0; \quad x_{k+1} = x_k - \frac{x_k^5 - 3x_k + 7}{5x_k^4 - 3}, \quad k = 0, 1, 2, \dots$$

In which we can perform the calculation for the first few steps:

$$x_1 = x_0 - \frac{x_0^5 - 3x_0 + 7}{5x_0^4 - 3} = \frac{7}{3}$$

$$x_2 = \frac{7}{3} - \frac{\left(\frac{7}{3}\right)^5 - 3\left(\frac{7}{3}\right) + 7}{5\left(\frac{7}{3}\right)^4 - 3} = 1.85702545$$

(c) The root -1.6416942631448190 is found in 12 iterations. It is faster than the bisection method in Example 43.4, i.e. 26 iterations. □

eg:jhmathews:1

Example 43.15 (April 2018 Final Exam Question 2(b)). Formulate the Newton-Raphson method for $e^{-x} + \cos(x + 2) = 0$ and estimate the root in one-iteration using the initial guess $x_0 = 1$. Perform all your calculations to 6 decimal places. (4 marks)

Solution. Let $f(x) = e^{-x} + \cos(x + 2)$. Then

$$f'(x) = -e^{-x} - \sin(x + 2) \dots\dots\dots [1 \text{ mark}]$$

$$\text{The Newton-Raphson method is } x_{k+1} = x_k - \frac{e^{-x} + \cos(x + 2)}{-(e^{-x} + \sin(x + 2))} \dots\dots\dots [1 \text{ mark}]$$

When $x_0 = 1$, we have

$$x_1 = 1 - \frac{e^{-1} + \cos(2)}{-(e^{-1} + \sin(2))} = 1 - \frac{-0.622113}{-0.508999} = -0.222228 \quad [1+1=2 \text{ marks}]$$

□

ein-godunov:p53

Example 43.16 (Tutorial 3, Q5). Compare the effectiveness of the bisection method and Newton's method for the equation $x^3 - 2x - 2 = 0$ which has a single root $x^* \approx 1.76929235$ between $x = -4$ and $x = 2$ to an accuracy of 10^{-8} .

Solution. The step estimation formula: $\frac{2 - (-4)}{2^n} = 10^{-8} \Rightarrow n \approx 29.16$.
 implies that bisection method requires 30 bisections to reach the desire accuracy.
 For Newton's method,

- Trying with $x_0 = 2$, the root 1.7692923542386998 is found in 4 iterations;
- Trying with $x_0 = -4$, the root 1.7692923542386314 is found in 18 iterations.

□

rem:newton-cons

Remark 43.17. Newton's method is an efficient method. However, it has some limitations, which will be stated below:

1. Newton's method converges slowly at a multiple root and the sequence of iterates $\{x_k\}$ exhibits linear convergence.
2. It is prone to odd behaviour depending on the initial guess. E.g. $f(x) = xe^{-x}$, $x_0 = 1.5$ diverges very slowly; $f(x) = x^3 - x - 3$, $x_0 = 0$ diverges to a periodic sequence; $f(x) = \arctan(x)$, $x_0 = 2$ diverges to oscillating sequence; there are examples which diverge to chaotic sequences [Mathews and Fink, 2004].
3. When it works, Newton's method is fine but unless local structure of the function is known (e.g. the Jacobian), the unmodified Newton's method is to be avoided.

In Remark 43.17, one of the limitation of Newton's method is the linear convergence when "multiple root" happens. For multiple roots, $f'(x^*) = 0$ and in Equation (43.6), if $x_k \approx x^*$, $f'(x_k) \approx f'(x^*) = 0$. Therefore, the quadratic convergence will fail. The following example demonstrates this situation.

eg:jhmathews:1

Example 43.18. Use Newton's method to find the multiple root $x^* = 2$ of $f(x) = (x-2)^6(x+1)$ numerically using the starting value $x_0 = 0$. In how many steps can we obtain an approximate solution to an accuracy of 10^{-6} ? Explain why the iteration converges slowly by studying the convergence rate (Recall Topic 1 Section 1.2.3). [Hint: $\lim_{k \rightarrow \infty} \frac{x_{k+1} - x^*}{x_k - x^*}$]

Solution: First, let's find the derivative of f : $f'(x) = 6(x-2)^5(x+1) + (x-2)^6$.

Now, we will set up the Newton's iteration:

$$x_{k+1} = x_k - \frac{(x_k - 2)^6(x_k + 1)}{6(x_k - 2)^5(x_k + 1) + (x_k - 2)^6} = x_k - \frac{(x_k - 2)(x_k + 1)}{7x_k + 4}, \quad k \geq 0; \quad x_0 = 0.$$

Using the Fortran subroutine newton, the root 1.9999955454651872 is found in 70 iterations.

This is a slow convergence. Assume that $x_k \rightarrow 2$ as $k \rightarrow \infty$. This can be verify numerically. To study the convergence rate, we investigate

$$\left| \frac{x_{k+1} - x^*}{x_k - x^*} \right| = \left| \frac{x_{k+1} - 2}{x_k - 2} \right| = \left| \frac{x_k - \frac{(x_k - 2)(x_k + 1)}{7x_k + 4} - 2}{x_k - 2} \right| = \left| 1 - \frac{x_k + 1}{7x_k + 4} \right| = \left| \frac{6x_k + 3}{7x_k + 4} \right| \rightarrow \frac{5}{6}$$

since $x_k \rightarrow 2$ as $k \rightarrow \infty$. Therefore, the Newton's method has a linear convergence.

Example 43.19 (April 2014 Final Exam Question 1(b)). Use Newton's method to find the multiple root x^* of $f(x) = 3(2x - 3)^5(4x + 3)$ numerically using the starting value $x_0 = 0$.

1. Find the approximation x_2 of the second iteration. (5 marks)

Solution. First, the derivative of f is

$$f'(x) = 3[10(2x - 3)^4(4x + 3) + 4(2x - 3)^5]. \quad [1 \text{ mark}]$$

Now, we will set up the iteration:

$$x_0 = 0$$

$$x_{k+1} = x_k - \frac{3(2x_k - 3)^5(4x_k + 3)}{3[10(2x_k - 3)^4(4x_k + 3) + 4(2x_k - 3)^5]} = x_k - \frac{(2x_k - 3)(4x_k + 3)}{48x_k + 18}, \quad k \geq 0$$

[1 mark]

$$x_1 = 0 - \frac{(-3)(3)}{18} = \frac{1}{2}$$

[1 mark]

$$x_2 = \frac{1}{2} - \frac{(1-3)(2+3)}{24+18} = \frac{31}{42} = .73809524$$

[2 marks]

□

2. Explain why the iteration converges slowly to $\frac{3}{2}$ by studying the convergence rate. (6 marks)

Solution: To study the convergence rate, we investigate

$$\left| \frac{x_{k+1} - x^*}{x_k - x^*} \right| = \left| \frac{x_{k+1} - \frac{3}{2}}{x_k - \frac{3}{2}} \right| = \left| \frac{x_k - \frac{(2x_k-3)(4x_k+3)}{48x_k+18} - \frac{3}{2}}{x_k - \frac{3}{2}} \right|$$

[2 marks]

$$= \left| 1 - \frac{2(4x_k + 3)}{48x_k + 18} \right|$$

[2 marks]

since $x_k \rightarrow \frac{3}{2}$ and $\left| \frac{x_{k+1} - x^*}{x_k - x^*} \right| \rightarrow \frac{4}{5}$ as $k \rightarrow \infty$. Therefore, the Newton's method has a linear convergence. [2 marks]

In the above two examples, we have encountered multiple root problem, which causes Newton's method's convergence to be linear. Sometimes, the Steffensen's method may be applied to improve the convergence rate as shown below.

Example 43.20 (Tutorial 3, Q6). Use Newton's method to construct a linearly convergent sequence $\{x_k\}$ which converges slowly to the multiple root $x^* = \sqrt{2}$ of $f(x) = (x^2 - 2)\sin(x^2 - 2)$ with a starting value $x_0 = 1$. Then use the Aitken Δ^2 process to construct $\{q_k\}$ which converges faster to the root $x^* = \sqrt{2}$.

Solution. The Newton's method is

$$x_0 = 1; \quad x_{k+1} = x_k - \frac{(x_k^2 - 2)\sin(x_k^2 - 2)}{2x_k \sin(x_k^2 - 2) + 2x_k(x_k^2 - 2)\cos(x_k^2 - 2)}$$

and the Aitken's Acceleration is

$$q_k = x_k - \frac{(x_{k+1} - x_k)^2}{x_{k+2} - 2x_{k+1} + x_k}, \quad k = 0, 1, 2, 3, \dots$$

Then

#	x_k	x_k - x^*	q_k	q_k - x^*
0:	1.0000000000	-0.4142135624	1.3761908759	-0.0380226865
1:	1.3044895246	-0.1097240378	1.4110417375	-0.0031718248
2:	1.3625247184	-0.0516888440	1.4136143258	-0.0005992366
3:	1.3889502100	-0.0252633524	1.4140849840	-0.0001285784
4:	1.4017074304	-0.0125061319	1.4141839490	-0.0000296133
5:	1.4079896902	-0.0062238722	1.4142064692	-0.0000070932
6:	1.4111086647	-0.0031048977	1.4142118275	-0.0000017349
7:	1.4126628414	-0.0015507210	1.4142131334	-0.0000004289
8:	1.4134386299	-0.0007749324	1.4142134557	-0.0000001066
9:	1.4138262027	-0.0003873597	1.4142135358	-0.0000000266
10:	1.4140199091	-0.0001936533	1.4142135557	-0.0000000066
11:	1.4141167424	-0.0000968200	1.4142135607	-0.0000000017
12:	1.4141651540	-0.0000484083	1.4142135620	-0.0000000004

```
13: 1.4141893586 -0.0000242038
14: 1.4142014606 -0.0000121018
```

$\therefore x_{14} = 1.4142014606; q_{12} = 1.4142135620.$

□

eg:jhmathews:4

Example 43.21 (Tutorial 3, Q7). Use the combined Newton's method and Steffensen's acceleration method to find numerical approximations to the multiple root $x^* = \sqrt{2}$ of the function $f(x) = (x^2 - 2)\sin(x^2 - 2)$. Show details of the computations for the starting value $x_0 = 1$. Compare the number of iterations for the two methods.

Solution.

#	x_k	$x_k - x^*$	q_k	$q_k - x^*$
0:	1.0000000000	-0.4142135624	1.3761908759	-0.0380226865
0:	1.3044895246	-0.1097240378		
0:	1.3625247184	-0.0516888440		
1:	1.3761908759	-0.0380226865	1.4139058150	-0.0003077474
1:	1.3955011069	-0.0187124555		
1:	1.4049244064	-0.0092891560		
2:	1.4139058150	-0.0003077474	1.4142135456	-0.0000000168
2:	1.4140597054	-0.0001538569		
2:	1.4141366381	-0.0000769243		
3:	1.4142135456	-0.0000000168	1.4142135624	-0.0000000000
3:	1.4142135540	-0.0000000084		
3:	1.4142135582	-0.0000000042		

$q_3 = 1.4142135624$

□

Example 43.22. The following Fortran program

```
1 program newton_slow_fp2
2   use textbook_root1
3   implicit none
4   double precision :: theroot
5   double precision, external :: T
6   integer :: info, cnt
7   call fixedpt(T,0d0,theroot,info,cnt)
8   print *, "The root ", theroot, " is found in ", cnt, " steps"
9 end program newton_slow_fp2
10 function T(x)
11   double precision :: x, T
12   T = x - (2d0*x-3d0)*(4d0*x+3d0)/(48d0*x+18d0)
13 end function T
```

calls the subroutine fixedpt. Explain what does the program do.

Solution. The Fortran program applies the Steffensen's method to the Newton's iteration in Example 43.18.

The root $x^* = 2$ with an accuracy of 10^{-6} is found in [5] steps. Compare to Example 43.18 which requires 70 steps, this is a tremendous improvement.

□

Finally, we mention two interesting applications of Newton's method.

um2e:p334:25.12

Example 43.23. Use Newton's method to derive a fixed-point iterative formula for determining square roots $x = \sqrt{q}$ where q is a non-negative real number.

Solution: For the equation $f(x) = x^2 - q = 0$. The Newton's method leads to

$$x_{k+1} = x_k - \frac{x_k^2 - q}{2x_k} = x_k - \frac{x_k}{2} + \frac{q}{2x_k} = \frac{x_k}{2} + \frac{q}{2x_k}, \quad k = 0, 1, \dots; \quad x_0 = 1$$

Remark 43.24. Despite the numerical methods in this subject are using double precision most of the time, single precision arithmetic is important in computer graphics. Interestingly, Newton method and single precision arithmetic were used to evaluate *fast inverse square root* (<https://>

en.wikipedia.org/wiki/Fast_inverse_square_root) in Quake III Arena (a first-person shooter video game) as follows:

$$\begin{aligned}\ln \frac{1}{\sqrt{x}} &= -\frac{1}{2} \ln x = -0.5 \cdot (e + \ln(1+m)) \\ &\approx -0.5 \cdot (e + m + \sigma) = \underbrace{-0.5(m + \sigma)}_{(\dagger) \text{ need a good approx.}} - 0.5 \cdot e\end{aligned}\quad (43.7)$$

where $x = 2^e \cdot (1+m)$, $0 \leq m = 0.m_1 \cdots m_{23} < 1$ (Topic 1 Section 1.3.2), σ is a free parameter used to tune the approximation.

Let x be a given positive number and $t = \frac{1}{\sqrt{x}}$. Then $x = \frac{1}{t^2}$. If we define a function $f(t) = \frac{1}{t^2} - x$, the Newton method associated with f is

$$t_{k+1} = t_k - \frac{f(t_k)}{f'(t_k)} = t_k - \frac{\frac{1}{t_k^2} - x}{-2t_k^{-3}} = t_k + 0.5(t_k + t_k^3 x) = t_k(1.5 + 0.5xt_k^2) \quad (43.8)$$

Note that e is an integer in (43.7), division by 2 could be replaced by a shifting. $(\dagger)$ was chosen by some trial and error as pointed out in the Wikipedia. Apparently, the logarithmic value (43.7) is used as an initial value for the first order (i.e. only $k = 0$) Newton approximation (43.8) was used. The C++ implementation (<https://betterexplained.com/articles/understanding-quakes-fast-inverse-square-root/>) is

```
1 #include <iostream>
2 #include <cmath>
3
4 float InvSqrt(float x){
5     float xhalf = 0.5f * x;
6     int i = *(int*)&x;          // store floating-point bits in integer
7     i = 0x5F3759DF - (i >> 1); // initial guess for Newton's method
8     float t = *(float*)&i;      // convert new bits into float
9     t = t*(1.5f - xhalf*t*t);   // One round of Newton's method
10    return t;
11 }
12
13 int main() {
14     const float dx = 0.2f;
15     for(int i=1; i<=100; i++){
16         float x = i*dx;
17         float y = 1.0f/std::sqrt(i*dx);
18         float yapprox = InvSqrt(i*dx);
19         std::cout << x << " " << y << " " << yapprox << " ";
20         std::cout << yapprox-y << "\n";
21     }
22 }
```

Relevant informations: Fast Inverse Square Root — A Quake III Algorithm (https://www.youtube.com/watch?v=p8u_k2LIzYo) and https://www.reddit.com/r/programming/comments/kmcntc/quake_iiis_fast_inverse_square_root_explained_20/

43.2.3 Secant Method and Implementation

According to Remark 43.17, one limitation of the Newton's method is the need to calculate the derivative. When the expression of a function f is complicated, it is a tedious job to differentiate f . To differentiate f , there are two options, one option is to use a computer algebra software (such as Maple or Mathematica) to do the calculation, another option is to use an approximation of $f'(x_k)$ since $x_k - x_{k-1} \rightarrow 0$ as $k \rightarrow \infty$ if x_k converges:

$$f'(x_k) \approx \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}.$$

With this approximation, the Newton iteration (43.5) becomes the *Secant Method*:

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k), \quad k = 1, 2, \dots \quad (43.9) \quad \text{e:secant}$$

So, we have eliminated the need to differentiate f . But now x_{k+1} is dependent on two previous values and so we need **two** initial guess. But in practice, it is unwise to ask the user to simply key in two initial guesses x_0 and x_1 . An implementation in Python's SciPy library was adapted to our Fortran 2003 implementation.

```

1  subroutine secant(f, x0, root, info, nsteps)
2      double precision, intent(in) :: x0
3      double precision, intent(out) :: root
4      integer, intent(out), optional :: info, nsteps
5      double precision :: newx, deltax, x0t, x1t
6      procedure(func) :: f
7      x0t=x0
8      if (x0 > 0) then
9          x1t = x0*(1 + 1e-4) + 1e-4
10     else
11         x1t = x0*(1 + 1e-4) - 1e-4
12     end if
13     newx=x1t
14     deltax=f(x1t)
15     do iter=1,maxIter
16         if (abs(deltax) < accu) then
17             root = newx
18             if(present(info)) info = 0
19             if(present(nsteps)) nsteps=iter
20             return
21         end if
22         deltax = (x1t-x0t)/(f(x1t)-f(x0t))*f(x1t)
23         newx = newx - deltax
24         x0t = x1t
25         x1t = newx
26     end do
27     if (present(info)) info = 1 ! Root Not Found After MaxIter
28 end subroutine secant

```

Example 43.25 (Tutorial 3, Q8). Let $f(x) = x^2 - 6$. Use secant method with $x_0 = 3$, $x_1 = 2$ to find x_3 .

Solution. With the two given initial guess, we can compute x_2 and x_3 by using (43.9): e:secant

$$x_2 = x_1 - \frac{x_1 - x_0}{f(x_1) - f(x_0)} f(x_1) = 2 - \frac{2 - 3}{f(2) - f(3)} f(2) = 2.4$$

$$x_3 = x_2 - \frac{x_2 - x_1}{f(x_2) - f(x_1)} f(x_2) = 2.4 - \frac{2.4 - 2}{f(2.4) - f(2)} f(2.4) = 2.45454545$$

□

eg:poly-secant

Example 43.26. Find the root of the function $f(x) = x^5 - 3x + 7$ using secant method with an initial value $x_0 = 0$. In how many steps do we obtain an accuracy of 10^{-7} ? How does it compare to Newton's method?

Solution: The root -1.6416942631448148 is found in 19 iterations. Compare to Newton's method which takes 13 iterations, the convergence is not too bad.

43.3 Finite Difference Approximation

The most important principle about taking derivatives is to **avoid it if we can!** So what if we cannot avoid it? For example in solving partial differential equations. Then we will have to try

to approximate the k -order derivative $f^{(k)}(x_0)$ by using a linear combination of $f(x_0)$ and its neighbouring values:

$$f^{(k)}(x_0) \approx \sum_{i=-k}^k f(x_0 + ih)w_i(h), \quad (43.10)$$

e:diff-approx

where $w_i(h)$ are rational functions of h . This is the basic idea of *numerical differentiation*.

How can we find the right $w_i(h)$? Since Taylor's approximation (43.4) provides an approximation to the neighbouring values of a function, we can start from (43.4). In Section 43.3.1, we will study some properties of finite difference and the derivation of the approximation for numerical differentiation. Then we will discuss how the accuracy can be improved using extrapolation. Later, it will be applied to Newton's method (43.20) in Section 43.4.3.

43.3.1 Finite Difference

sec:fd

Definition 43.5. Given a sequence $y_0, y_1, \dots, y_k, \dots$, we can define the *first (forward) difference* to be

$$\Delta y_k = y_{k+1} - y_k, \quad k = 0, 1, \dots, n-1.$$

We can define the *second (forward) difference* to be

$$\Delta^2 y_k = \Delta y_{k+1} - \Delta y_k = y_{k+2} - 2y_{k+1} + y_k, \quad k = 0, 1, \dots, n-2.$$

In general, we can define the *mth (forward) difference* to be

$$\Delta^m y_k = \Delta^{m-1} y_{k+1} - \Delta^{m-1} y_k.$$

y:aitken-fdform

Example 43.27. Rewrite the Aitken's Acceleration formula (Topic 2)

$$q_k = s_k - \frac{(s_{k+1} - s_k)^2}{s_{k+2} - 2s_{k+1} + s_k}$$

in terms of the finite difference notation.

Solution. $q_k = s_k - \frac{(\Delta s_k)^2}{\Delta^2 s_k}.$

□

Example 43.28. Write a script to calculate the forward difference of the sequence $1^2, 2^2, \dots, 10^2$.

Solution: The forward difference is

$$\Delta([1, 4, 9, 16, 25, 36, 49, 64, 81, 100]) = [3, 5, 7, 9, 11, 13, 15, 17, 19]$$

If you calculate it with Python with NumPy, then

```
1 import numpy as np
2 ylist = np.array([x**2 for x in range(1,11)])
3 delt = np.diff(ylist)
4 print("delta({ylist})={delt}".format(ylist=ylist,delt=delt))
```

If you try it with R, then

```
1 ylist = c(1:10)^2
2 delt = diff(ylist)
3 cat("delta(",ylist,")=",delt,"\n")
```

If you like Octave (or MATLAB), then the script will be similar to

```
1 ylist = [1:10].^2;
2 delt = diff(ylist);
```

```
3 disp(strcat("delta(", num2str(ylist), ")\n= ", num2str(delt)))
```

If we apply the notion of finite difference to a function f , we would have the following definition.

Definition 43.6. Let $h > 0$. Given a function f at a finite and $x_{k+1} - x_k = h$. The *forward difference notation* for f is

$$\Delta_h f(x) = f(x + h) - f(x).$$

The *backward difference notation* for f is

$$\nabla_h f(x) = f(x) - f(x - h).$$

The h is sometimes ignored if it is clear from the context.

The *difference table* is the standard format for displaying finite differences. Its diagonal pattern makes each entry, except for the x_k, y_k , the difference of its two nearest neighbours to the left.

Example 43.29. Sketch the difference table for $y = f(x)$ with $x_0, \dots, x_4$.

Solution:

x_0	y_0				
		Δy_0			
x_1	y_1		$\Delta^2 y_0$		
		Δy_1		$\Delta^3 y_0$	
x_2	y_2		$\Delta^2 y_1$		$\Delta^4 y_0$
		Δy_2		$\Delta^3 y_1$	
x_3	y_3		$\Delta^2 y_2$		
		Δy_3			
x_4	y_4				

To approximate $f'(x_0)$ with a finite difference formula (e:diff-approx 43.10), we suppose first that $x_0 \in (a, b)$ and $f \in C^2[a, b]$. Let $h \neq 0$ be a sufficiently small number and $x = x_0 + h$. Then the Taylor's approximation (e:taylor-approx 43.4) becomes

$$f(x_0 + h) = f(x_0) + f'(x_0)h + \frac{f''(\xi(x_0, h))}{2!}h^2. \quad (43.11) \quad \text{e:fd}$$

Rearranging the above equation leads to

$$f'(x_0) = \frac{f(x_0 + h) - f(x_0)}{h} - \frac{h}{2}f''(\xi(x_0, h)) = \frac{\Delta_h f(x_0)}{h} + O(h).$$

This formula in the box above is known as the *forward-difference formula* when $h > 0$ and as the *backward-difference formula* when $h < 0$.

Example 43.30 (Tutorial 3, Q9). Use the forward-difference and backward-difference formulae to determine each missing entry in the following tables.

	x	$\sin x$	$\frac{\Delta \sin x}{\Delta x}$	$\frac{\nabla \sin x}{\Delta x}$	$ \frac{\Delta \sin x}{\Delta x} - \cos x $	$\frac{h}{2} \max f''(\xi) $
1.	0.5	0.47943	0.85210	-	0.02548	0.02823
	0.6	0.56464	0.79580	0.85210	0.02954	0.03221
	0.7	0.64422	-	0.79580	-	-

2. Let $f(x) = e^x - 2x^2 + 3x - 1$.

x	$f(x)$	$\frac{\Delta f(x)}{\Delta x}$	$\frac{\nabla f(x)}{\Delta x}$	$ \frac{\nabla f(x)}{\Delta x} - f'(x) $	$\frac{h}{2} \max f''(\xi) $
0.0	0.00000	3.70700	-	-	-
0.2	0.74140	3.15200	3.70700	0.28560	0.30000
0.4	1.3718	-	3.15200	0.26018	0.27786

Example 43.31. Differentiate $f(x) = e^x \sin x$ using the forward-difference formula at $x = 1.9$ starting with $h = 0.05$ and halving h each time. Fill in the blanks in following table.

h	$\Delta_h f(1.9)/h$	Error
0.05	4.05010229	-0.11528028
0.05/2	4.10956065	-0.05582193
0.05/4	4.13791992	-0.02746266
0.05/8	4.15176255	-0.01362003
0.05/16	4.15860029	-0.00678229
0.05/32	4.16199835	-0.00338422
0.05/64	4.16369220	-0.00169038
0.05/128	4.16453782	-0.00084476
0.05/256	4.16496031	-0.00042227
0.05/512	4.16517147	-0.00021111

Explain how the “Error”s in the table are obtained.

Solution: The “Error”s in the table are obtained by finding the difference

$$\frac{\Delta_h f(1.9)}{h} - f'(1.9)$$

where $f'(1.9) = e^{1.9} \sin(1.9) + e^{1.9} \cos(1.9) = 4.1653825786581$.

43.3.2 Richardson’s Extrapolation

In 1911, L. F. Richardson introduced the *Richardson’s extrapolation* as a *sequence acceleration method* to improve the rate of convergence of a sequence with an error term of a predictable form (e.g. it depends on the step size h). For example, if $N(h)$ is a formula involving step size $h > 0$ that approximates an unknown value M as follow:

$$M = N(h) + K_1 h^{k_1} + K_2 h^{k_2} + K_3 h^{k_3} + \cdots = N(h) + K_1 h^{k_1} + O(h^{k_2}). \quad (43.12) \quad \text{eq:4.1}$$

Here, k_i are known constants such that $h^{k_i} > h^{k_{i+1}}$. The *Big-O notation* (or *Landau notation*) $O(h^{k_2})$ indicates that there is a positive real number M and a real number h_0 such that

$$|K_2 h^{k_2} + K_3 h^{k_3} + \cdots| \leq M |h^{k_2}|, \quad \text{for all } h > h_0.$$

Substitute h with the step size $\frac{h}{2}$ in (eq:4.1), we have

$$M = N\left(\frac{h}{2}\right) + K_1 \left(\frac{h}{2}\right)^{k_1} + O(h^{k_2}).$$

Subtracting (eq:4.1) from (43.12) the above, we have

$$(2^{k_1} - 1)M = 2^{k_1} N\left(\frac{h}{2}\right) - N(h) + O(h^{k_2}) \Rightarrow M = \frac{2^{k_1} N\left(\frac{h}{2}\right) - N(h)}{2^{k_1} - 1} + O(h^{k_2}).$$

By this process, we have achieved a *better approximation* of M by subtracting the largest term in the error which was $O(h^{k_1})$. This process can be repeated to remove more error terms to get even better approximations.

Let $N_1(h) \equiv N(h)$ and $N_2(h) = \frac{2^{k_1} N\left(\frac{h}{2}\right) - N(h)}{2^{k_1} - 1}$. Then

$$M = N_2(h) + \tilde{K}_2 h^{k_2} + O(h^{k_3}). \quad (43.13) \quad \text{eq:4.2}$$

Again, we substitute h with the step size $\frac{h}{2}$ in (43.13) in order to eliminate the h^{k_2} term.

$$M = N_2\left(\frac{h}{2}\right) + \tilde{K}_2\left(\frac{h}{2}\right)^{k_2} + O(h^{k_3}). \quad (43.14) \quad \text{eq:4.3}$$

Similarly, subtracting (43.13) from 2^{k_2} times (43.14) gives

$$(2^{k_2} - 1)M = 2^{k_2}N_2\left(\frac{h}{2}\right) - N_2(h) + O(h^{k_3}) \Rightarrow M = \underbrace{\frac{2^{k_2}N_2\left(\frac{h}{2}\right) - N_2(h)}{2^{k_2} - 1}}_{\text{denoted as } N_3(h)} + O(h^{k_3}).$$

Continuing this process, we could obtain a recurrence relation

$$\begin{aligned} N_1(h) &= N(h) \\ N_i(h) &= \frac{2^{k_{i-1}}N_{i-1}\left(\frac{h}{2}\right) - N_{i-1}(h)}{2^{k_{i-1}} - 1} \quad i = 2, 3, \dots \end{aligned} \quad (43.15) \quad \text{e:richardson}$$

such that

$$M = N_i(h) + O(h^{k_i}), \quad i = 1, 2, \dots$$

eg:richardson1

Example 43.32. Assuming $f \in C^\infty$, the centred-difference formula for $f'(x_0)$ is $\frac{f(x_0+h)-f(x_0-h)}{2h}$.

1. Derive the error formula and state the Richardson extrapolation table for the centred-difference formula.

Solution: The Taylor series for f at x_0 is

$$\begin{aligned} f(x_0 + h) &= f(x_0) + f'(x_0)h + \frac{f''(x_0)}{2}h^2 + \frac{f'''(x_0)}{6}h^3 + \frac{f^{(4)}(x_0)}{24}h^4 + O(h^5) \\ f(x_0 - h) &= f(x_0) - f'(x_0)h + \frac{f''(x_0)}{2}h^2 - \frac{f'''(x_0)}{6}h^3 + \frac{f^{(4)}(x_0)}{24}h^4 + O(h^5) \end{aligned}$$

Subtracting them and after rearranging leads to the error formula

$$f'(x_0) = \underbrace{\frac{f(x_0 + h) - f(x_0 - h)}{2h}}_{N_1(h)} + \underbrace{\left(-\frac{f'''(x_0)}{6}\right)}_{K_1}h^2 + O(h^4). \quad (43.16) \quad \text{eq:4.15}$$

Note that $k_1 = 2$, $k_2 = 4$ (and in general, $k_i = 2i$, $i = 1, 2, \dots$). According to Richardson relation (43.15), we have

$$\begin{aligned} N_1(h) &= \frac{f(x_0 + h) - f(x_0 - h)}{2h} \\ N_i(h) &= \frac{2^{2(i-1)}N_{i-1}\left(\frac{h}{2}\right) - N_{i-1}(h)}{2^{2(i-1)} - 1} = N_{i-1}\left(\frac{h}{2}\right) + \frac{N_{i-1}\left(\frac{h}{2}\right) - N_{i-1}(h)}{2^{2(i-1)} - 1}, \quad i = 2, 3, \dots \end{aligned} \quad (43.17) \quad \text{e:richardson-cd}$$

2. Implement the Richardson extrapolation for centred-difference formula in Fortran 2003.

Solution:

```
1 function deriv_r(f, x, stepsize, tol)
2   procedure(func) :: f
3   double precision, intent(in) :: x, stepsize, tol
4   integer :: st, i
5   double precision :: N(maxStage,maxStage), h, aux, deriv_r
6   h = stepsize
7   do st=1,maxStage
8     N(st,1) = (f(x+h)-f(x-h))/(2*h)
9     do i=2,st
10      aux = (N(st,i-1)-N(st-1,i-1))/(4**(i-1)-1)
11      N(st,i)=N(st,i-1)+aux
12    end do
```

```

13      !print *, "h=", h
14      !print *, "N=", N(st,1:st)
15      if ((st>2) .and. abs(N(st,st)-N(st,st-1)) < tol) then
16          deriv_r = N(st,st)
17          return
18      end if
19      h = h/2
20  end do
21 end function deriv_r

```

Example 43.33. For the function $f(x) = x^2 \cos x$, find an approximation to $f'(1.0)$ using central difference formula with $h = 0.1$ and Richardson extrapolation to an accuracy of $O(h^8)$ (or order 4) to obtain an accuracy to 10^{-8} .

Solution: Using the Fortran subroutine in Example [eg:richardson1](#) 43.32, we can obtain

	$N_1(h)$	$N_2(h)$	$N_3(h)$	$N_4(h)$
Value of h	$\frac{f(1+h) - f(1-h)}{2h}$	1st extrapolation	2nd extrapolation	3rd extrapolation
0.1	0.22673616312855249			
0.05	0.23603092059474085	0.23912917308347029		
0.025	0.23835774147932920	0.23913334844085865	0.23913362679801786	
0.0125	0.23893964251058719	0.23913360952100651	0.23913362692634971	0.23913362692838672

The above table is generated using Formula ([e:richardson-cd](#) 43.17):

$$\begin{aligned}
 N_1(0.1) &= \frac{f(1.1) - f(0.9)}{0.2} \approx \frac{0.54885131 - 0.50350407}{0.2} = 0.2267362 \\
 N_1(0.05) &= \frac{f(1.05) - f(0.95)}{0.1} \approx \frac{0.54857208 - 0.52496899}{0.1} = 0.2360309 \\
 N_1(0.025) &= \frac{f(1.025) - f(0.975)}{0.05} \approx \\
 N_1(0.0125) &= \frac{f(1.0125) - f(0.9875)}{0.025} \approx \\
 N_2(0.1) &= \frac{4N_1(0.05) - N_1(0.1)}{4-1} \approx \frac{4 \times 0.2360309 - 0.2267362}{3} = 0.23912913 \\
 N_2(0.05) &= \frac{4N_1(0.025) - N_1(0.05)}{4-1} \approx \\
 N_2(0.025) &= \frac{4N_1(0.0125) - N_1(0.025)}{4-1} \approx \\
 N_3(0.1) &= \frac{16N_2(0.05) - N_2(0.1)}{16-1} \approx \frac{16 \times 0.23913335 - 0.23912913}{15} = 0.23913363 \\
 N_3(0.05) &= \frac{16N_2(0.025) - N_2(0.05)}{16-1} \approx \\
 N_4(0.1) &= \frac{64N_2(0.05) - N_2(0.1)}{64-1} \approx \frac{64 \times 0.23913363 - 0.23913363}{63} = 0.23913363
 \end{aligned}$$

Example 43.34 (Tutorial 3, Q12). Let $f(x) = xe^x$. By using the centre difference and Richardson's extrapolation, find the approximations of accuracy $O(h^2)$, $O(h^4)$, and $O(h^6)$ for $f'(2.0)$ with $h = 0.2$.

Solution. Note that $k_1 = 2$, $k_2 = 4$, $k_3 = 6$, etc.,

$$f'(x_0) = \underbrace{\frac{f(x_0+h) - f(x_0-h)}{2h}}_{N_1(h)} + \underbrace{\left(-\frac{f'''(x_0)}{3!}\right)h^2}_{K_1} + \underbrace{\left(-\frac{f^{(5)}(x_0)}{5!}\right)h^4}_{K_2} + \underbrace{\left(-\frac{f^{(7)}(x_0)}{7!}\right)h^6}_{K_3} + \dots$$

and

	$O(h^2)$	$O(h^4)$	$O(h^6)$
h	$N_1(h)$		
$h/2$	$N_1(h/2)$	$N_2(h) = \frac{2^2 N(h/2) - N(h)}{2^2 - 1}$	
$h/4$	$N_1(h/4)$	$N_2(h/2) = \frac{2^2 N(h/4) - N(h/2)}{2^2 - 1}$	$N_3(h) = \frac{2^4 N_2(h/2) - N_2(h)}{2^4 - 1}$

The approximations required are

$$\begin{aligned}
 N_1(h) &= \frac{19.85503 - 10.88937}{2(0.2)} = 22.414161 \\
 N_1(h/2) &= 22.228787 \\
 N_1(h/4) &= 22.182565 \\
 N_2(h) &= 22.166995 \\
 N_2(h/2) &= 22.167157 \\
 N_3(h) &= \boxed{22.167168}.
 \end{aligned}$$

□

43.4 System of Equations

sec:system

When we have more than one variables or one equations, we have a *system of equations* (43.1), which can be “classified” to three kinds of problems similar to what we encountered in Topic 2:

(a) n variables and n equations where $n \geq 1$:

$$\begin{aligned}
 f_1(x_1, x_2, \dots, x_n) &= 0, \\
 f_2(x_1, x_2, \dots, x_n) &= 0, \\
 &\dots \\
 f_n(x_1, x_2, \dots, x_n) &= 0.
 \end{aligned} \tag{43.18}$$

e:system

The abstract representation of the system of nonlinear equations (43.18).

(b) More equations than variables. There is no solution in general but it could be formulated as a generalisation of least-square problem or linear regression problem. This can be considered to be nonlinear regression problems in statistics or the machine learning problem (43.2).

(c) $n + k$ variables and n equations where $n, k \geq 1$ (more variables than equations).

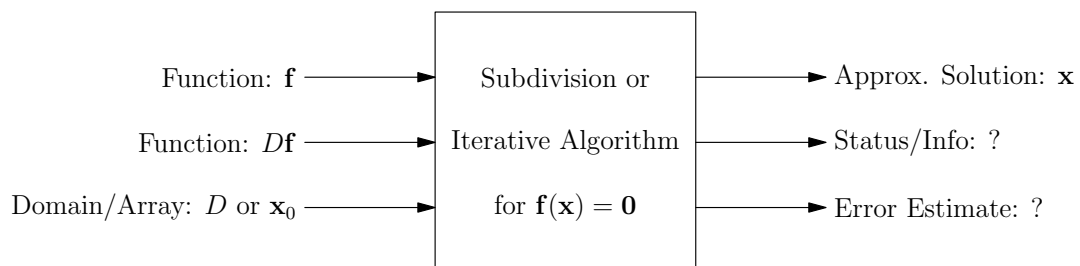
$$\mathbf{f}(\mathbf{x}; \lambda_1, \dots, \lambda_k) = \mathbf{0}. \tag{43.19}$$

e:nle-n+k

This is related to continuation problem and manifold problem which is applied in computer graphics and nonlinear control.

The task of finding roots for a system of equations (43.18) (or (43.1)) is very important in science and engineering because these roots correspond to certain physical states. For example, the roots may represent equilibrium states of a system of ordinary differential equations (Topic 6).

Bisection method, Newton’s method and secant method for equation (43.3) can be generalised to system of equations (43.1) and all these numerical methods can be regarded as a black box below.



43.4.1 Bisection Methods

The generalisation of bisection method to handle system of equations (43.1) involves rather complicated mathematical concepts (such as *topological index*, etc.). One generalisation, the *characteristic bisection method*, was implemented in ACM TOMS Algo 666 CHABIS [Vrahatis, 1988].

Example 43.35 (April 2014 Final Exam Question 4(c)). Consider the equation

$$\begin{cases} x^2 + y^2 = 1, \\ x^3 - y = y. \end{cases}$$

Use CHABIS to solve the system.

Solution: To use CHABIS (which is single precision), download the Fortran from <http://netlib.org/toms/> and then compile with the function to be solved

```
1 !http://iprc.soest.hawaii.edu/users/furue/improve-fortran.html
2 module fnc_mod
3     implicit none
4     integer, private :: nfcall=0
5     public :: incnt, getcnt, reset, fnc
6 contains
7     subroutine reset()
8         nfcall = 0
9     end subroutine reset
10    subroutine incnt()
11        if(nfcall.ge.100000) then
12            print *, "Maximum function iteration reached. Stop!"
13            stop
14        end if ! There is a bug in CHABIS: sometimes it will not stop
15    end subroutine incnt
16    function getcnt()
17        integer :: getcnt
18        getcnt = nfcall
19    end function getcnt
20    function fnc(x,iflag)
21        integer :: iflag
22        real :: x(2), fnc
23        nfcall = nfcall+1
24        call incnt()
25        if (iflag.eq.1) then
26            fnc = x(1)**2 + x(2)**2 - 1.0
27        else if (iflag.eq.2) then
28            fnc = x(1)**3 - x(2)
29        else
30            fnc = fnc/0.0
31        end if
32    end function fnc
33 end module fnc_mod
```

and the main driver

```
1 ! ALGORITHM 666, COLLECTED ALGORITHMS FROM ACM.
2 ! THIS WORK PUBLISHED IN TRANSACTIONS ON MATHEMATICAL SOFTWARE,
3 ! VOL. 14, NO. 4, PP. 330-336.
4 program bisect_main
5     use fnc_mod, only : getcnt, fnc
6     implicit none
7     integer, parameter :: n = 2, lwa = 2*n + (6*n+1)*2**n
8     real :: x0(n), h(n), as(n), vas(n), wa(lwa)
9     real, parameter :: delta = 0.625e-1, epsilo = 1.0e-6
10    integer :: icon, j, inf1, inf2, nfcall
11
12    ! Set the starting values
```



```

13      x0 = (/ -2e3, -2e3 /)
14      h = (/ 2e3, 2e3 /)
15      print 9999, (x0(j),h(j),j=1,n)
16      ! Call the Bisection Subroutine
17      icon = 1
18      call intsub(fnc,n,x0,h,delta,epsilo,icon,inf1,as,vas,inf2,wa,lwa)
19      if (inf1 .eq. 0) then
20          print 9998
21      elseif (inf1 .eq. 2 .and. icon .ne. 1) then
22          print 9997, n
23      else
24          print 9996, delta, epsilo, (as(j), vas(j), j=1,n)
25      endif
26      nfcall = int(real(getcnt())/real(n))
27      print 9995, inf1, inf2, nfcall
28
29 9999 format(//2x,'Initial Guess:',17x,'Stepsizes:'//2(f16.7,16x,f16.7/))
30 9998 format(//5x, '*** Improper Input Parameters ***' //)
31 9997 format(//5x, '*** The characteristic', i2,          &
32      '-Polyhedron has not been completed ***' //)
33 9996 format(//2x, 'Delta=',f20.18 //2x, 'Epsilo=', f20.18 &
34      ///2x, 'Final Approximate Solution:', 5x,          &
35      'Verification of the Solution:' //9(f24.18,9x,f24.18/))
36 9995 format(//2x, 'Exit Parameters: Inf1=',i1,1X,', ' ,Inf2=',&
37      i1//2x, 'Number of Function Calls: NFCALL=',i4)
38 end program bisect_main

```

Another generalisation is based on triangulation in n -dimensional space [Stynes, 1979]. Wood [1992] generalise it solve optimisation problems. These generalisations are not only complex but uses a lot of memory space and are only used in special computer graphics problem.

43.4.2 Newton's Method

Definition 43.7. The *Newton's method* for a system of equations (43.1) is

$$\mathbf{x}_{k+1} = \mathbf{x}_k - [D\mathbf{f}(\mathbf{x}_k)]^{-1}\mathbf{f}(\mathbf{x}_k), \quad k = 0, 1, 2, \dots; \quad \mathbf{x}_0 = \text{initial guess vector} \quad (43.20)$$

where $D\mathbf{f}(\mathbf{x})$ is the Jacobian matrix of the function $\mathbf{f}$ at $\mathbf{x}$. In particular, we obtain (43.5) when $n = 1$.

Newton's method is fixed-point iteration method with $T(\mathbf{x}) = \mathbf{x} - D\mathbf{f}(\mathbf{x})^{-1}\mathbf{f}(\mathbf{x})$ (when $n = 1$, $T(x) = x - \frac{f(x)}{f'(x)}$). If T satisfies the conditions in Fixed-Point Theorem 43.43, then a fixed point $\mathbf{x}^*$ exists. Mathematicians have proven that for suitable initial point $\mathbf{x}_0$ and when the spectral radius of the Jacobian $D\mathbf{f}(\mathbf{x}^*)$ is less than 1, then a unique solution exists according to Banach Fixed Point Theorem. For “analytical” function $\mathbf{f}$, the convergence rate for this method can be fast if a good starting vector $\mathbf{x}_0$ is chosen and the “derivative” of the fixed-point $\mathbf{x}^*$ is not zero.

With Fortran's support for array operations and the use of LAPACK subroutines from Topic 2, it is possible to implement the Newton's method for system of equations (43.1) as follows.

```

1  subroutine newton(f, n, x0, root, info, nsteps, df)
2      integer, intent(in) :: n
3      real(c_double), intent(in) :: x0(n)
4      real(c_double), intent(out) :: root(n)
5      integer, intent(out), optional :: info, nsteps
6      optional :: df
7      procedure(funcn) :: f
8      procedure(jacob) :: df
9      interface
10         function dnm2(n, dx, incx)

```

```

11         use iso_c_binding, only : c_double
12         integer :: n, incx
13         real(c_double) :: dx(n), dnrms2
14     end function dnrms2
15 end interface
16 real(c_double) :: deltax(n), x(n), jac(n,n)
17 integer :: p(n), tmp=0 ! for dgesv info.
18 x = x0
19 deltax = f(n, x)
20 do iter=1,maxIter
21     if (present(df)) then
22         jac = df(n, x)
23     else
24         call fdjac(f, n, x, jac)
25     end if
26     !!! Checking if Jacobian is singular. To implement in future.
27     !if (matrix_norm(df(x))==0.0) then
28     !    if (present(info)) info = 2
29     !    return
30     !end if
31     if (dnrms2(n,deltax,1)<accu) then ! Taking 2-norm (BLAS. Topic 2)
32         root = x
33         if (present(info)) info = 0
34         if (present(nsteps)) nsteps=iter-1
35         return
36     end if
37     !!! Solving Jac*dx = f(x) (LAPACK. Topic 2)
38     deltax = f(n,x)
39     call dgesv(n,1,jac,n,p,deltax,n,tmp)
40     x = x - deltax
41 end do
42 if (present(info)) info = 1 ! Root Not Found After MaxIter
43 end subroutine newton

```

Example 43.36 (April 2014 Final Exam Question 4(c)). Consider the equation

$$\begin{cases} \frac{1}{2}x^2 + \ln y = 1.3, \\ \frac{1}{2}y^2 + \ln x = 0.825. \end{cases}$$

1. Write down the Newton's iteration formula for solving the above equation. (5 marks)

Solution: First, the equation need to be rearranged into $\begin{pmatrix} f \\ g \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ form below:

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} \frac{1}{2}x^2 + \ln y - 1.3 \\ \frac{1}{2}y^2 + \ln x - 0.825 \end{bmatrix} = \mathbf{0}. \quad [1 \text{ mark}]$$

Then we need to find the Jacobian

$$D\mathbf{f}(\mathbf{x}_k) = \begin{bmatrix} x & \frac{1}{y} \\ \frac{1}{x} & y \end{bmatrix} \quad [2 \text{ marks}]$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - [D\mathbf{f}(\mathbf{x}_k)]^{-1}\mathbf{f}(\mathbf{x}_k) \Rightarrow \begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - \begin{bmatrix} x_k & \frac{1}{y_k} \\ \frac{1}{x_k} & y_k \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{2}x_k^2 + \ln y_k - 1.3 \\ \frac{1}{2}y_k^2 + \ln x_k - 0.825 \end{bmatrix}$$

..... [2 marks]

2. Explain why Aitken's acceleration is not implemented for Newton's iteration. (2 marks)

Solution: Aitken's acceleration is used to accelerate iterations with linear convergence but Newton's iteration is theoretically having quadratic convergence and Aitken's accelerate will not improve the convergence. [2 marks]

eg:newton2b

Example 43.37. Estimate the points of intersection between the circle $x^2 + y^2 = 3$ and the hyperbola $xy = 1$ using Newton method with initial guess $(x_0, y_0) = (0.5, 1.5)$ by programming in Python.

Solution: The steps to formula a numerical solution for this problem are

1. Formulate $\mathbf{f}(\mathbf{x}) = \mathbf{0}$:
$$\begin{cases} x^2 + y^2 - 3 = 0 \\ xy - 1 = 0 \end{cases}$$

2. Formulate (43.20) (see page 1087):

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - \begin{bmatrix} 2x_k & 2y_k \\ y_k & x_k \end{bmatrix}^{-1} \begin{bmatrix} x_k^2 + y_k^2 - 3 \\ x_k y_k - 1 \end{bmatrix}, \quad \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix}$$

3. Python's SciPy module only implements the Newton's method for (43.3) in `scipy/optimize/zeros.py` but does not implement n -dimensional Newton method ($n > 1$), so we have to use n -dimensional fixed-point method instead:

```
import numpy
import scipy.optimize
def newton_f(vx):
    fx = numpy.zeros(2)
    fx[0] = vx[0]**2 + vx[1]**2 - 3.0
    fx[1] = vx[0]*vx[1] - 1.0
    jacobian = numpy.zeros((2,2))
    jacobian[0,:] = [2*vx[0], 2*vx[1]]
    jacobian[1,:] = [vx[1], vx[0]]
    vx_update = vx - numpy.linalg.solve(jacobian, fx)
    return vx_update
guess = numpy.array([0.5,1.5])
est = scipy.optimize.fixed_point(newton_f,guess,method='iteration')
print("Estimation of the intersection point = ", est)
```

Running the script produces

```
Estimation of the intersection point = [ 0.61803399  1.61803399]
```

eg:newton-n

Example 43.38. You are given the following Fortran 2008 program:

```
1 program newton_n
2   use textbook_rootn
3   implicit none
4   integer, parameter :: n = 2
5   real(c_double) :: x(n), root(n)
6   integer :: info, cnt
7   x = (/ .8, .6 /)
8   call newton(f, n, x, root, info, cnt, df)
9   !call newton(f, n, x, root, info, cnt)
10  write(*,*) "the root", root, "is found in", cnt, "iterations"
11  contains ! Using Fortran 2008: allow call to internal functions
12  function f(n, x)
13    use iso_c_binding, only : c_double
14    integer, intent(in) :: n
15    real(c_double), intent(in) :: x(n)
16    real(c_double) :: f(n)
17    f = (/ x(1)**2 + x(2)**2 - 1, x(1)**3 - x(2) /)
18  end function f
```

```

19 function df(n, x)
20     use iso_c_binding, only : c_double
21     integer, intent(in) :: n
22     real(c_double), intent(in) :: x(n)
23     real(c_double) :: df(n,n)
24     df = reshape((/2*x(1), 3*x(1)**2, 2*x(2), -1d0/), (/2, 2/))
25 end function df
26 end program newton_n

```

1. Read the program and explain its function.

Solution. The program finds a root to an accuracy of 10^{-7} for the equation

$$f_1(x, y) = x^2 + y^2 - 1 = 0$$

$$f_2(x, y) = x^3 - y = 0$$

with the initial point $(x_0, y_0) = (0.8, 0.6)$

□

2. Determine the root and the number of iterations.

Solution: By running the Fortran program, the root $(0.82603135765418689, 0.56362416216125855)$ is found in 4 iterations whether a Jacobian is provided or not.

Example 43.39 (April 2018 Final Exam Question 2(c)). Given

$$\begin{cases} x^3 + y = 3, \\ (x+1)(y+1) = 2. \end{cases}$$

Formulate the Newton-Raphson method for this problem and find the first iteration x_1 (you do not need to find y_1) using the initial guess $x_0 = y_0 = 0$. (3 marks)

Solution. First, let $\mathbf{x} = (x, y)$ and $\mathbf{f}(\mathbf{x}) = \begin{bmatrix} x^3 + y - 3 \\ (x+1)(y+1) - 2 \end{bmatrix}$.

The Newton-Raphson method to solve this problem is

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - \begin{bmatrix} 3x_k^2 & 1 \\ y_k + 1 & x_k + 1 \end{bmatrix}^{-1} \begin{bmatrix} x_k^3 + y_k - 3 \\ (x_k + 1)(y_k + 1) - 2 \end{bmatrix}, \quad k = 0, 1, 2, \dots$$

When $x_0 = y_0 = 0$, $\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} -3 \\ 1 - 2 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} -3 \\ -1 \end{bmatrix} \Rightarrow \boxed{x_1 = -2}$

[1+0.5=1.5 marks]

□

Example 43.40 (Tutorial 3, Q10). Write down the Newton iteration formula for

$$\begin{cases} x^2 + y^2 + 1 = 0, \\ x + y = 1. \end{cases}$$

Will the Newton iteration converge for any given initial point?

Solution. If the iteration converges, then a limit exists! Suppose the limit is $(x^*, y^*) \in \mathbb{R}^2$, then

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = - \begin{bmatrix} 2x^* & 2y^* \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} x^{*2} + y^{*2} + 1 \\ x^* + y^* - 1 \end{bmatrix} \Rightarrow \begin{bmatrix} x^{*2} + y^{*2} + 1 \\ x^* + y^* - 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

However, $x^2 + y^2 + 1 > 0$ for all $(x, y) \in \mathbb{R}^2$ and has no solution. A contradiction to the existence of a limit. Therefore, the Newton's method will not converge because there is no root to converge to.

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - \begin{bmatrix} 2x_k & 2y_k \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} x_k^2 + y_k^2 + 1 \\ x_k + y_k - 1 \end{bmatrix} \quad \square$$

t2:DrGoh:nm3

Example 43.41 (Tutorial 3, Q11). Let $\mathbf{x} = (x, y, z)^T$. Write down the Newton's iteration for the following system of nonlinear equations

$$\begin{aligned} x + y + z &= 3 \\ x^2 + y^2 + z^2 &= 5 \\ e^x + xy - xz &= 1 \end{aligned}$$

and use the initial conditions $\mathbf{x}^{(0)} = (0.1, 1.2, 2.5)^T$ to find $\mathbf{x}^{(1)}$. If possible, with the help of computer software, find the solution.

Solution. Note that we need to change $F(\mathbf{x}) = \mathbf{c}$ to $F(\mathbf{x}) - \mathbf{c} = 0$. The Newton's iteration is

$$\begin{bmatrix} x^{(k+1)} \\ y^{(k+1)} \\ z^{(k+1)} \end{bmatrix} = \begin{bmatrix} x^{(k)} \\ y^{(k)} \\ z^{(k)} \end{bmatrix} - \begin{bmatrix} 1 & 1 & 1 \\ 2x^{(k)} & 2y^{(k)} & 2z^{(k)} \\ e^{x^{(k)}} + y^{(k)} - z^{(k)} & x^{(k)} & -x^{(k)} \end{bmatrix}^{-1} \begin{bmatrix} x^{(k)} + y^{(k)} + z^{(k)} - 3 \\ (x^{(k)})^2 + (y^{(k)})^2 + (z^{(k)})^2 - 5 \\ e^{x^{(k)}} + x^{(k)}(y^{(k)} - z^{(k)}) - 1 \end{bmatrix}$$

When $k = 0$, we can find

$$\begin{aligned} \mathbf{x}^{(1)} &= \begin{bmatrix} 0.1 \\ 1.2 \\ 2.5 \end{bmatrix} - \begin{bmatrix} 1 & 1 & 1 \\ 2(0.1) & 2(1.2) & 2(2.5) \\ e^{0.1} + 1.2 - 2.5 & 0.1 & -0.1 \end{bmatrix}^{-1} \begin{bmatrix} 0.1 + 1.2 + 2.5 - 3 \\ 0.1^2 + 1.2^2 + 2.5^2 - 5 \\ e^{0.1} + 0.1(1.2 - 2.5) - 1 \end{bmatrix} \\ &= \begin{bmatrix} 0.0033981 \\ 0.8783420 \\ 2.1182599 \end{bmatrix} \end{aligned}$$

Using a Fortran 2008 program (calling the `textbook_rootn` module), after 16 iterations, we obtain the root $(2.9289541769953056 \times 10^{-7}, 0.99999941420916416, 2.0000002928954181)$. $\square$

43.4.3 Approximation of Jacobian and Newton's Method of n -Dimension

sec:newton-n

The forward-difference formula (43.11) can be extended to n -dimensional space as the approximation of Jacobian:

$$D\mathbf{f}(\mathbf{x}_0) \approx \begin{bmatrix} \frac{f_1(\mathbf{x}_0 + h\mathbf{e}_1) - f_1(\mathbf{x}_0)}{h} & \cdots & \frac{f_1(\mathbf{x}_0 + h\mathbf{e}_n) - f_1(\mathbf{x}_0)}{h} \\ \vdots & \ddots & \vdots \\ \frac{f_n(\mathbf{x}_0 + h\mathbf{e}_1) - f_n(\mathbf{x}_0)}{h} & \cdots & \frac{f_n(\mathbf{x}_0 + h\mathbf{e}_n) - f_n(\mathbf{x}_0)}{h} \end{bmatrix} \quad (43.21) \quad \text{e:fdjac}$$

where $\mathbf{e}_i = (0, \dots, 0, 1, 0, \dots, 0)$ where $(\mathbf{e}_i)_i = 1$ and $(\mathbf{e}_i)_j = 0$ when $j \neq i$.

The h in (43.21) cannot be chosen to be too small. Dennis and Schnabel [1996, 49] suggests that the use of $h = \sqrt{\epsilon_{\min}}|x|$, where $\epsilon_{\min} = 2^{-t} = 2^{-53} \approx 2.22 \times 10^{-16}$ is the IEEE double precision floating point's computer machine epsilon (Topic 1).

The Fortran implementation of the forward-difference formula (43.21) is listed below and was used in the subroutine `newton` (line 24) when Jacobian is not available.

```
1 subroutine fdjac(f, n, x, jacmat)
2   procedure(funcn) :: f
3   integer, intent(in) :: n
4   real(c_double) :: x(n)
5   real(c_double), intent(out) :: jacmat(n,n) ! jacobian matrix.
```

```

6      !integer, intent(out), optional :: iflag
7      real(c_double) :: fvec(n), fvec(n), eps, temp, h
8      integer :: j
9      eps = sqrt(epsilon(1d0))    ! 53-bit mantissa, b^{-(p-1)}=2^{-52}
10     do j = 1, n
11         fvec = f(n,x)
12         temp = x(j)
13         h = eps*abs(temp)    ! Dennis, Schnabel, SIAM 1996
14         if (h == 0.0D+00) h = eps
15         x(j) = temp + h
16         fvec = f(n,x)
17         x(j) = temp
18         jacmat(1:n,j) = (fvec(1:n)-fvec(1:n))/h
19     end do
20 end subroutine fdjac

```

An alternative technique for approximating Jacobian with analytic expressions is the automatic differentiation technique which is introduced in Bücker et al. [2006].

43.4.4 Variations of Newton's Method and SUNDIALS' KINSOL

The finite difference approximation (43.21) can yield quite accurate approximations to the Jacobian but at the cost of additional function evaluations per Newton iteration [Beers, 2006]. In practice, variations of Newton iteration (43.20) such as (a) the *modified Newton method*:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - D\mathbf{f}(\mathbf{x}_0)^{-1}f(\mathbf{x}_k)$$

or (b) *Broyden's Method* [Broyden, 1965], a kind of *quasi-Newton methods*, which is a generalisation of Secant method [Beers, 2006, Chapter 2]:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - B_k^{-1}f(\mathbf{x}_k)$$

are used. The matrix B_k is some “good” approximation to Jacobian matrix.

Let B_k be an *estimate* of $D\mathbf{f}$ and we update the solution estimate by the rule

$$B_k \Delta \mathbf{x}_k = -\mathbf{f}(\mathbf{x}_k), \quad (43.22)$$

$$\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k + \Delta \mathbf{x}_k. \quad (43.23)$$

At $\mathbf{x}_{k+1}$, we need an estimate B_{k+1} for the Jacobian. Broyden [1965] proposed to obtain B_{k+1} from B_k , $\mathbf{f}(\mathbf{x}_k)$ and $\mathbf{f}(\mathbf{x}_{k+1})$ using the *integral mean value theorem* for $\mathbf{f}$:

$$\mathbf{f}(\mathbf{x}_{k+1}) = \mathbf{f}(\mathbf{x}_k) + \underbrace{\left[\int_0^1 D\mathbf{f}(\mathbf{x}_k + s\Delta \mathbf{x}_k) ds \right]}_{\text{matrix}} \Delta \mathbf{x}_k = \mathbf{f}(\mathbf{x}_k) + B_{k+1} \Delta \mathbf{x}_k.$$

which leads to

$$B_{k+1} \Delta \mathbf{x}_k = \mathbf{f}(\mathbf{x}_{k+1}) - \mathbf{f}(\mathbf{x}_k) \quad (43.24)$$

Subtracting (43.22) from (43.23), we have

$$(B_{k+1} - B_k) \Delta \mathbf{x}_k = \mathbf{f}(\mathbf{x}_{k+1}). \quad (43.25)$$

Broyden [1965] suggests a few updates for B_{k+1} which satisfy (43.25) with proofs that B_{k+1} is an acceptable estimate to the Jacobian at $\mathbf{x}_{k+1}$:

1. $B_{k+1} = B_k + \frac{\mathbf{f}(\mathbf{x}_{k+1})(\Delta \mathbf{x}_k)^T}{(\Delta \mathbf{x}_k)^T \Delta \mathbf{x}_k}$ or $H_{k+1} = H_k - \frac{(\Delta \mathbf{x}_k + H_k(\mathbf{f}(\mathbf{x}_{k+1}) - \mathbf{f}(\mathbf{x}_k)))(\Delta \mathbf{x}_k)^T H_k}{(\Delta \mathbf{x}_k)^T H_k(\mathbf{f}(\mathbf{x}_{k+1}) - \mathbf{f}(\mathbf{x}_k))}$;
2. $H_{k+1} = H_k - \frac{(\Delta \mathbf{x}_k + H_k(\mathbf{f}(\mathbf{x}_{k+1}) - \mathbf{f}(\mathbf{x}_k)))(\mathbf{f}(\mathbf{x}_{k+1}) - \mathbf{f}(\mathbf{x}_k))^T}{(\mathbf{f}(\mathbf{x}_{k+1}) - \mathbf{f}(\mathbf{x}_k))^T (\mathbf{f}(\mathbf{x}_{k+1}) - \mathbf{f}(\mathbf{x}_k))}$

where $H_k = B_k^{-1}$, $k = 0, 1, 2, \dots$, $B_0 = I$. A detail theoretical study can be found in Kelly [1995, Chapter 7]

A well-known C software library (with Fortran interface) for solving system of equations (43.1) as well as ODE-IVP (topic 6) is SUNDIALS, which stands for a “SUite of Nonlinear and Differential/ALgebraic Equation Solvers”. Part of this software library, KINSOL, is a solver for nonlinear system (43.1) which includes a *Newton-Krylov solver* as well as *Picard and fixed point solvers*, both of which can be accelerated with *Anderson acceleration*. KINSOL’s Newton solver employs the Inexact Newton method. As this solver is intended mainly for large systems, four *iterative methods* (Topic 2) are provided to solve the resulting linear systems — GMRES, Bi-CGStab, TFQMR, and FGMRES. These are Krylov methods, implemented with scaling and preconditioning, and can be used with all versions of the NVECTOR module. For the sake of convenience to users with smaller systems, KINSOL (used with the serial NVECTOR module) also includes direct (dense, banded, and sparse) linear solvers for the linear systems. In this case the nonlinear iteration is a *modified Newton method*.

Example 43.42. You are working in a company with only C compiler and uses SUNDIALS 2.5.0 library. By looking up the relevant information from the Internet, explain how you can solve Example 43.37.

Solution: Looking up for “User Documentation for KINSOL 2.5.0” (the latest version as of February 2019 is 4.1.0. There are updates in this version and 2.5.0 will need modification to work in 4.1.0) from the Internet, one can implement a solution for Example 43.37:

```

1 // Constraints are imposed to make all components of the solution
2 // positive.
3 #include <stdio.h>
4 #include <stdlib.h>
5 #include <math.h>
6
7 #include <kinsol/kinsol.h>
8 #include <kinsol/kinsol_dense.h>
9 #include <nvector/nvector_serial.h>
10 #include <sundials/sundials_types.h>
11 #include <sundials/sundials_math.h>
12
13 // Problem Constants
14 #define NVAR    2
15 #define NEQ     3*NVAR
16
17 #define FTOL    RCONST(1.e-5) // function tolerance
18 #define STOL    RCONST(1.e-5) // step tolerance
19
20 typedef struct {
21     realtype lb[NVAR];
22     realtype ub[NVAR];
23 } *UserData;
24
25 /* Functions Called by the KINSOL Solver */
26 static int func(N_Vector u, N_Vector f, void *user_data);
27
28 /* Private Helper Functions */
29 static int SolveIt(void *kmem, N_Vector u, N_Vector s, int glstr, int mset);
30 static void PrintHeader(int globalstrategy, realtype fnormtol,
31                        realtype scsteptol);
32 static void PrintOutput(N_Vector u);
33 static void PrintFinalStats(void *kmem);
34 static int check_flag(void *flagvalue, char *funcname, int opt);
35
36 int main() {
37     UserData data=NULL;
38     realtype fnormtol, scsteptol;

```

```

39 N_Vector u1=NULL, u=NULL, s=NULL, c=NULL;
40 int glstr, mset, flag;
41 void *kmem=NULL;
42
43 // User data
44 data = (UserData)malloc(sizeof *data);
45 data->lb[0] = RCONST(1e-6); // x lower bound > 0
46 data->ub[0] = RCONST(10); // x upper bound
47 data->lb[1] = RCONST(1e-6); // y lower bound > 0
48 data->ub[1] = RCONST(20); // y upper bound
49
50 // Initial Guess
51 const realtype x=0.5, y=1.5;
52 u1 = N_VNew_Serial(NEQ); // Create serial vectors of length NEQ
53 NV_Ith_S(u1,0) = x;
54 NV_Ith_S(u1,1) = y;
55 NV_Ith_S(u1,2) = x - data->lb[0];
56 NV_Ith_S(u1,3) = x - data->ub[0];
57 NV_Ith_S(u1,4) = y - data->lb[1];
58 NV_Ith_S(u1,5) = y - data->ub[1];
59
60 s = N_VNew_Serial(NEQ);
61 N_VConst_Serial(RCONST(1.0),s); /* no scaling */
62
63 kmem = KINCreate();
64 if (check_flag((void *)kmem, "KINCreate", 0)) return(1);
65
66 flag = KINSetUserData(kmem, data);
67 if (check_flag(&flag, "KINSetUserData", 1)) return(1);
68
69 // Setup Constraint
70 c = N_VNew_Serial(NEQ);
71 NV_Ith_S(c,0) = RCONST(0); /* no constraint on x1 */
72 NV_Ith_S(c,1) = RCONST(0); /* no constraint on x2 */
73 NV_Ith_S(c,2) = RCONST(1); /* l1 = x1 - x1_min >= 0 */
74 NV_Ith_S(c,3) = -RCONST(1); /* L1 = x1 - x1_max <= 0 */
75 NV_Ith_S(c,4) = RCONST(1); /* l2 = x2 - x2_min >= 0 */
76 NV_Ith_S(c,5) = -RCONST(1); /* L2 = x2 - x2_max <= 0 */
77 flag = KINSetConstraints(kmem, c);
78 if (check_flag(&flag, "KINSetConstraints", 1)) return(1);
79
80 // Setup Tolerance for Solver
81 fnormtol=FTOL; scsteptol=STOL;
82 flag = KINSetFuncNormTol(kmem, fnormtol);
83 if (check_flag(&flag, "KINSetFuncNormTol", 1)) return(1);
84 flag = KINSetScaledStepTol(kmem, scsteptol);
85 if (check_flag(&flag, "KINSetScaledStepTol", 1)) return(1);
86
87 // Setup system and solution vector $u$
88 u = N_VNew_Serial(NEQ);
89 flag = KINInit(kmem, func, u);
90 if (check_flag(&flag, "KINInit", 1)) return(1);
91
92 // Call KINDense to specify the linear solver based on LAPACK/BLAS
93 flag = KINDense(kmem, NEQ);
94 if (check_flag(&flag, "KINDense", 1)) return(1);
95
96 // Print out the problem size, solution parameters, initial guess.
97 glstr = KIN_NONE;
98 PrintHeader(glstr, fnormtol, scsteptol);
99
100 printf("\n-----\n");
101 printf("\nInitial guess, [x,y] = ");
102 PrintOutput(u1);
103
104 N_VScale_Serial(RCONST(1),u1,u);

```



```

105     glstr = KIN_NONE; // or KIN_LINESEARCH;
106     mset = 1; // 1=Exact Newton; 0 = Inexact Newton
107     SolveIt(kmem, u, s, glstr, mset);
108
109     /* Free memory */
110     N_VDestroy_Serial(u);
111     N_VDestroy_Serial(s);
112     N_VDestroy_Serial(c);
113     KINFree(&kmem);
114     free(data);
115     return(0);
116 }
117
118
119 static int SolveIt(void *kmem, N_Vector u, N_Vector s, int glstr, int mset)
120 {
121     int flag;
122
123     printf("\n");
124     printf(mset==1 ? "Exact Newton" : "Modified Newton");
125     printf(glstr == KIN_NONE ? "\n" : " with line search\n");
126
127     flag = KINSetMaxSetupCalls(kmem, mset);
128     if (check_flag(&flag, "KINSetMaxSetupCalls", 1)) return(1);
129
130     flag = KINSol(kmem, u, glstr, s, s);
131     if (check_flag(&flag, "KINSol", 1)) return(1);
132
133     printf("Solution:\n  [x,y] = ");
134     PrintOutput(u);
135
136     PrintFinalStats(kmem);
137
138     return(0);
139 }
140
141
142 // Example 3.4.9
143 static int func(N_Vector u, N_Vector f, void *user_data) {
144     realtype *udata, *fdata;
145     realtype x, l1, L1, y, l2, L2;
146     UserData data;
147
148     data = (UserData)user_data;
149
150     udata = NV_DATA_S(u);
151     x = udata[0];
152     y = udata[1];
153     l1 = udata[2];
154     L1 = udata[3];
155     l2 = udata[4];
156     L2 = udata[5];
157
158     fdata = NV_DATA_S(f);
159     fdata[0] = x*x + y*y - RCONST(3.0);
160     fdata[1] = x*y - RCONST(1.0);
161     fdata[2] = l1 - x + data->lb[0];
162     fdata[3] = L1 - x + data->ub[0];
163     fdata[4] = l2 - y + data->lb[1];
164     fdata[5] = L2 - y + data->ub[1];
165
166     return(0);
167 }
168
169 // PRIVATE FUNCTIONS
170

```

```

171 // Print problem description
172 static void PrintHeader(int globalstrategy, realtype fnormtol,
173                        realtype scsteptol)
174 {
175     printf("\nExample 3.4.4\n");
176     printf("Tolerance parameters:\n");
177 #if defined(SUNDIALS_EXTENDED_PRECISION)
178     printf("  fnormtol = %12.8Lg\n  scsteptol = %12.8Lg\n",
179 #elif defined(SUNDIALS_DOUBLE_PRECISION)
180     printf("  fnormtol = %12.8lg\n  scsteptol = %12.8lg\n",
181 #else
182     printf("  fnormtol = %12.8g\n  scsteptol = %12.8g\n",
183 #endif
184         fnormtol, scsteptol);
185 }
186
187 static void PrintOutput(N_Vector u) {
188 #if defined(SUNDIALS_EXTENDED_PRECISION)
189     printf(" %10.8Lg %10.8Lg\n", NV_Ith_S(u,0), NV_Ith_S(u,1));
190 #elif defined(SUNDIALS_DOUBLE_PRECISION)
191     printf(" %10.8lg %10.8lg\n", NV_Ith_S(u,0), NV_Ith_S(u,1));
192 #else
193     printf(" %10.8g %10.8g\n", NV_Ith_S(u,0), NV_Ith_S(u,1));
194 #endif
195 }
196
197 static void PrintFinalStats(void *kmem) {
198     long int nni, nfe, nje, nfeD;
199     int flag;
200
201     flag = KINGetNumNonlinSolvIters(kmem, &nni);
202     check_flag(&flag, "KINGetNumNonlinSolvIters", 1);
203     flag = KINGetNumFuncEvals(kmem, &nfe);
204     check_flag(&flag, "KINGetNumFuncEvals", 1);
205
206     flag = KINDlsGetNumJacEvals(kmem, &nje);
207     check_flag(&flag, "KINDlsGetNumJacEvals", 1);
208     flag = KINDlsGetNumFuncEvals(kmem, &nfeD);
209     check_flag(&flag, "KINDlsGetNumFuncEvals", 1);
210
211     printf("Final Statistics:\n");
212     printf("  nni = %5ld    nfe = %5ld \n", nni, nfe);
213     printf("  nje = %5ld    nfeD = %5ld \n", nje, nfeD);
214 }
215
216 /*
217  * Check function return value...
218  *   opt == 0 means SUNDIALS function allocates memory so check if
219  *   returned NULL pointer
220  *   opt == 1 means SUNDIALS function returns a flag so check if
221  *   flag >= 0
222  *   opt == 2 means function allocates memory so check if returned
223  *   NULL pointer
224  */
225 static int check_flag(void *flagvalue, char *funcname, int opt) {
226     int *errflag;
227
228     // Check if SUNDIALS function returned NULL pointer - no memory allocated
229     if (opt == 0 && flagvalue == NULL) {
230         fprintf(stderr,
231             "\nSUNDIALS_ERROR: %s() failed - returned NULL pointer\n\n",
232             funcname);
233         return(1);
234     }
235     // Check if flag < 0
236     else if (opt == 1) {

```

```

237     errflag = (int *) flagvalue;
238     if (*errflag < 0) {
239         fprintf(stderr,
240             "\nSUNDIALS_ERROR: %s() failed with flag = %d\n\n",
241             funcname, *errflag);
242         return(1);
243     }
244 }
245 // Check if function returned NULL pointer - no memory allocated
246 else if (opt == 2 && flagvalue == NULL) {
247     fprintf(stderr,
248         "\nMEMORY_ERROR: %s() failed - returned NULL pointer\n\n",
249         funcname);
250     return(1);
251 }
252 return(0);
253 }

```

A variation called the Powell's hybrid method is implemented in the MINPACK library, which provides Fortran subroutines for solving $\min\{\sum_{i=1}^m f_i(\mathbf{x})^2 : \mathbf{x} \in \mathbb{R}^n\}$ and (43.3). The subroutines `hybrd`¹, `lmdif` use forward-difference method (43.21) to find Jacobian. While subroutines `lmder`, `lmstr`, `hybrj` require the user to supply a subroutine to calculate the Jacobian matrix. Since many open source numerical computing system such as Python (SciPy), Octave and R have employed MINPACK, it is good to know that the convergence is superlinear (Topic 2).

43.5 Fixed-Point Iteration

As mentioned in Topic 2, majority of the solution of numerical problems need to be translated into some kind of *iteration processes*. One important form of iteration process is the *fixed point iteration method*

$$\mathbf{x}_{k+1} = T(\mathbf{x}_k), \quad k = 0, 1, 2, 3, \dots; \quad \mathbf{x}_0 = \text{initial guess vector} \quad (43.26)$$

where $T : D \rightarrow \mathbb{R}^n$ is a mapping of a domain in n -dimensional space.

A generalisation of the fixed point iteration process is the *multi-point iteration method*:

$$\mathbf{x}_{k+m} = T(\mathbf{x}_{k+m-1}, \dots, \mathbf{x}_{k+1}, \mathbf{x}_k), \quad k = 0, 1, 2, 3, \dots; \quad \mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{m-1} = \text{initial guess vectors} \quad (43.27)$$

The Newton method in Section 43.2.2 is a fixed point iteration while the secant method in Section 43.2.3 is a 2-point iteration method.

An iteration process generates a sequence $\mathbf{x}_0, \mathbf{x}_1, \dots$. If the next point $\mathbf{x}_{k+1}$ and the current point $\mathbf{x}_k$ is “close” enough for some integer k , then we consider ourselves getting a “fixed point” defined below.

Definition 43.8. The point $\mathbf{x}^* \in D$ is called a *fixed point* of T if $T(\mathbf{x}^*) = \mathbf{x}^*$.

43.5.1 Existence Theorem

The need to rewrite a numerical problem into an iterative form (43.26) is because mathematicians found a general theorem characterising the existence of fixed point for continuous functions.

Theorem 43.43 (Brouwer Fixed-Point Theorem). Assume that $T : D \rightarrow D$ is continuous on $D \subset \mathbb{R}^n$ where D is a compact and convex set. Then T has a fixed point $\mathbf{p}$, i.e. $\exists \mathbf{p} (\mathbf{p} \in D \wedge T(\mathbf{p}) = \mathbf{p})$.

¹Powell's hybrid algorithm

This theorem says that for a continuous mapping T of D to itself, T has at least one fixed point but it does not tell us how many T have. On the other hand, for a special kind of mapping, the contraction mapping, it is possible to have a unique fixed point.

Definition 43.9. Let d be a “distance function” on $D \subset \mathbb{R}^n$. A map $T : D \rightarrow D$ is called a *contraction mapping* on D if there exists $\theta \in [0, 1)$ such that

$$d(T(\mathbf{x}), T(\mathbf{y})) \leq \theta d(\mathbf{x}, \mathbf{y})$$

for all $\mathbf{x}, \mathbf{y}$ in D .

Theorem 43.44 (Banach Fixed Point Theorem; Contraction Mapping Principle). *Let $T : D \rightarrow D$ be a contraction mapping and (D, d) is a complete metric space. Then T admits a unique fixed-point $\mathbf{x}^*$ in D . Furthermore, $\mathbf{x}^*$ can be found as follows: start with an arbitrary element $\mathbf{x}_0$ in $\mathbb{R}^n$ and define a sequence $\{\mathbf{x}_k\}$ by (43.26), then $\mathbf{x}_k \rightarrow \mathbf{x}^*$.*

A special case of this theorem for one-dimensional space is a function T defined over an interval $D = [a, b]$ such that $|T'(x)| \leq \theta < 1$ for all $x \in (a, b)$. Leonardo of Pisa (well-known as Fibonacci) happened to solve a polynomial of degree 3 in 1225. We do not know how he managed to find a solution. However, with the help of modern computer, a fixed-point iteration takes 23 steps to reach a fixed point of 8 decimal places as shown below.

Example 43.45. In the year 1225 Fibonacci studied the equation $x^3 + 2x^2 + 10x - 20 = 0$ and produced $x = 1.368808107$. This is amazing since he used hand to perform the calculation. Rewrite the above equation into the following iterative form

$$x_{k+1} = \frac{20}{x_k^2 + 2x_k + 10}, \quad k = 0, 1, 2, \dots \quad (43.28)$$

By using $x_0 = 1$ and the fixed point iteration program below to find out how many iterations will you obtain Fibonacci's result? Explain about the convergence of the iteration. [Hint: You may use the Fortran program below to help you in your calculation.]

```

1 program fixedpt
2   double precision, external :: T
3   double precision :: v0, v1
4   integer :: iter
5   v0 = 1d0
6   do iter=1,50
7     v1 = T(v0)
8     if (abs(v1-v0) < 1d-8) then
9       print *, "Fixed point with prec 1e-8 = ", v1, " after ", &
10        & iter, " steps"
11       exit ! do loop
12     end if
13     v0 = v1
14   end do
15 end program fixedpt
16 function T(x)
17   double precision :: x, T
18   T = 20d0/(x**2+2d0*x+10d0)
19 end function
```

Solution: Fixed point 1.3688081107179968 with a precision of 10^{-8} is obtained after 23 iterations.

The rate of convergence may be estimated from

$$e_n = x^* - x_n = T(x^*) - T(x_{n-1}) = T'(\xi)(x^* - x_{n-1}) = T'(\xi)e_{n-1}$$

which compares the n^{th} error e_n with the preceding error. As n increases we may take $T'(x^*)$ as an approximation to $T'(\xi)$, assuming the existence of this derivative. Then

$$\frac{e_n}{e_{n-1}} = T'(\xi) \approx T'(x^*) = -\frac{40(x^* + 1)}{(x^{*2} + 2x^* + 10)^2} \approx -0.44$$

making each error about -0.44 times the previous. This suggests that the two or three iterations will be required for each new correct decimal place.

To illustrate that fixed point iteration process (43.26) is used in practice, we illustrate it with Python. eg: fixed-it

Example 43.46. Use SciPy to solve Example 43.45. eg: schaum2e:p330:25.2

Solution: SciPy provides a function `fixed_point` which can be used to find a fixed point if the iteration process converges:

```
1 def T(x): return 20.0/(x**2+2*x+10)
2 from scipy.optimize import fixed_point # in minpack.py
3 v0 = 1.0
4 v1 = fixed_point(T, v0, xtol=1e-8, maxiter=50, method="iteration")
5 print("Fixed point with prec 1e-8 = {v1}".format(v1=v1))
```

Example 43.47. Is it possible to use SciPy to solve Example 43.45 with Steffensen's method (Topic 2)? eg: schaum2e:p330:25.2

Solution: Yes. The function `fixed_point` has a method "del2" which refers to Steffensen's method:

```
1 def T(x): return 20.0/(x**2+2*x+10)
2 from scipy.optimize import fixed_point
3 v1 = fixed_point(T, 1.0, xtol=1e-8, maxiter=4, method="del2")
4 print("Fixed point with prec 1e-8 = {v1}".format(v1=v1))
```

43.5.2 Convergence Acceleration

Mathematically, it is interesting to find ways to generalise the Aitken's acceleration (Topic 2) from $n = 1$ (Example 43.27): eg: aitken-fdform

$$q_k = s_k - (s_{k+1} - s_k)(s_{k+2} - 2s_{k+1} + s_k)^{-1}(s_{k+1} - s_k) = s_k - (\Delta s_k)(\Delta^2 s_k)^{-1}(s_{k+1} - s_k)$$

to $T: \mathbb{R}^n \rightarrow \mathbb{R}^n$ with $n > 1$. This is given in Henrici [1982, Chapter 5] and is stated in a clearer manner in Noda [1987]:

$$\mathbf{q}_k = \mathbf{x}_k - [\mathbf{x}_{k+1} - \mathbf{x}_k, \dots, \mathbf{x}_{k+n} - \mathbf{x}_{k+n-1}][\mathbf{x}_{k+2} - 2\mathbf{x}_{k+1} + \mathbf{x}_k, \dots, \mathbf{x}_{k+n+1} - 2\mathbf{x}_{k+n} + \mathbf{x}_{k+n-1}]^{-1}(\mathbf{x}_{k+1} - \mathbf{x}_k).$$

So far, it is not clear what we can do with this generalised formula.

43.5.3 Fortran Module

In Topic 2, we learned that there are standard libraries LAPACK and UMFPACK for solving linear algebra problems but there is no library for simple iterative methods. A simple implementation in Fortran module was carried out. For nonlinear problems, there are standard libraries such as MINPACK, SUNDIALS for solving system of equations. However, there is no simple implementation of the textbook algorithms for solving nonlinear equation (43.3), eg: nle therefore, we construct a Fortran module with the declarations below.

```

module textbook_root1
  implicit none
  abstract interface                                ! Type definition
    function func(x)
      double precision :: x, func
    end function func
  end interface
  integer :: iter=0, maxIter=100, maxStage=10 ! Type declarations
  double precision :: accu=1d-6
  private :: iter, maxIter, maxStage, accu
  public :: setaccu, setmiter, fixedpt, bisect, secant, newton
contains

  subroutine setaccu(p) ! set accuracry requirement
    double precision, intent(in) :: p
    accu = p
  end subroutine setaccu

  subroutine setmiter(m) ! set maximum iteration allowed
    integer, intent(in) :: m
    maxIter = m
  end subroutine setmiter

  ...

```

Some of the common variables such `accu`, `maxIter` control the desired accuracy and iteration steps are similar to Topic 2's Fortran module in iteration methods section.

Complex Roots

Finding roots for a general equation is very important in science and engineering because these roots usually correspond to a certain physical state. For example, the roots may represent equilibrium states of a differential equation. The complex roots of a transfer function determines the transient behaviour of a system which is important in filter theory and feedback control design.

There are no algebraic technique which allows one to find the roots a any given function exactly in general. For example, $x^5 - 3x + 7 = 0$, $x - 10^x + 2 = 0$ cannot be solved algebraically and this has been proved using the mathematical techniques from complex analysis and algebraic geometry. So mathematicians have to come up with techniques which allow us to “search” or “approximate” the roots. Some of these techniques will be investigate in Sections 43.1 to 43.5.3.

A.1. Zeroes of Polynomials

We learned that $x^2 + 1 = 0$ has no solution over $\mathbb{R}$. In complex analysis, it is known that any polynomial of positive integer degree with have a solution.

Theorem 43.48 (Fundamental Theorem of Algebra). *For any polynomial $p(z) = a_0 + a_1z + \cdots + a_nz^n$ of positive degree, i.e. $n \geq 1$, $p(z)$ always has a complex root.*

One a root z_0 is found, the original degree n polynomial can be divided by $z - z_0$ and an $(n - 1)$ -degree polynomial can be obtained:

$$p(z) = a_0 + a_1z + \cdots + a_nz^n = (z - z_0)(b_0 + b_1z + \cdots + b_{n-1}z^{n-1})$$

the roots of the $(n - 1)$ -degree polynomial $b_0 + b_1z + \cdots + b_{n-1}z^{n-1}$ is also the roots of the original polynomial $p(z)$. By the Fundamental Theorem of Algebra, the $(n - 1)$ -degree polynomial has a root. By recursion, all roots (may be repeated) of $p(z)$ can be obtained.

A.2. Complex Dynamical Systems

Newton’s method can also be generalised to work with complex numbers. Because it is prone to odd behaviour, the study of what happens to the initial guess of Newton’s method on a complex plain leads to a beautiful topic of complex dynamical system as follows.

Example 43.49 (Complex Dynamical System). Investigate the region $\{x + yi : -1 \leq x, y \leq 1\}$ where the function $f(z) = z^4 - 1$ converges.

Solution: First note that $f(z) = 0$ has four solutions $z = \pm 1, \pm i$. Mathematicians discovered that by numerical calculation, we can find that these four solutions has a “symmetric” attracting basin. In order to calculate the attracting, we use the following Python program to perform calculation.

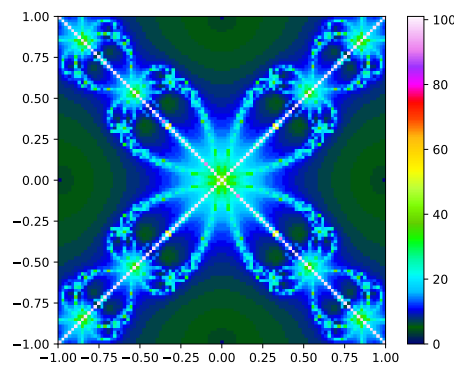
```
1 import numpy as np
2
3 def f(z): return z**4-1.
4 def df(z): return 4*z**3
5
6 N = 100; maxIter = 100; prec = 1e-7
7 x0, x1, y0, y1 = -1.,1.,-1.,1.
8 dx, dy = (x1-x0)/N, (y1-y0)/N
9 img = np.zeros((N+1,N+1))
10 for i in range(N+1):
11     for j in range(N+1):
12         z = x0+i*dx + 1j*(y0+j*dy)
13         error = abs(f(z))
14         for k in range(maxIter):
15             if(error<prec): break
16             dfz = df(z)
```

```

17         if abs(dfz) < prec:
18             k = maxIter+1
19             break
20         deltaz = f(z)/dfz
21         z = z - deltaz
22         error = abs(deltaz)
23     img[i,j] = k
24
25 from matplotlib.pyplot import imshow, show, cm, colorbar, savefig
26 imshow(img, extent=[x0,x1,y0,y1], interpolation='nearest',\
27         origin="lower", cmap=cm.gist_ncar)
28 colorbar()
29 savefig("newtonc.pdf") #show()

```

The calculation indicates that when $z = a(1 \pm i)$, the Newton's iteration diverges. Otherwise, it converges at a different speed, and the speed determines the colour of the figure.



Müller's Method

sec:muller

According to Press et al. [2007], Müller's method is sometimes also used for finding complex zeros of analytic functions (not just polynomials) in the complex plane, for example in the IMSL routine ZANLY.

In this section we will introduce another “derivative-free” method — Müller's method, a method first presented by David E. Muller in 1956 Muller [1956], which is a generalisation of the secant method of root finding by using quadratic three-point interpolation, see Faires and Burden [2003].

Muller's method is a generalisation of the secant method, in the sense that it does not require the derivative of the function. It is an iterative method that requires three starting points $(x_0, f(x_0))$, $(x_1, f(x_1))$ and $(x_2, f(x_2))$. A parabola

$$f(x) = a(x - x_2)^2 + b(x - x_2) + c$$

is constructed that passes through the three points, i.e. the constants a , b , and c can be determined from

$$f(x_i) = a(x_i - x_2)^2 + b(x_i - x_2) + c, \quad i = 0, 1, 2$$

i.e.

$$\begin{aligned}
 c &= f(x_2), \\
 b &= \frac{(x_0 - x_2)^2[f(x_1) - f(x_2)] - (x_1 - x_2)^2[f(x_0) - f(x_2)]}{(x_0 - x_2)(x_1 - x_2)(x_0 - x_1)} \\
 a &= \frac{(x_1 - x_2)[f(x_0) - f(x_2)] - (x_0 - x_2)[f(x_1) - f(x_2)]}{(x_0 - x_2)(x_1 - x_2)(x_0 - x_1)}
 \end{aligned}$$

and the quadratic formula is used to find a root of the quadratic for the next approximation:

$$x_3 = x_2 - \frac{2c}{b + \operatorname{sgn}(b)\sqrt{b^2 - 4ac}}.$$

It has been proved that near a simple root Müller's method converges faster than the secant method and almost as fast as Newton's method. The method can be used to find real or complex zeros of a function and can be programmed to use complex arithmetic.

```

1 module textbook_rootc
2   use iso_c_binding, only : c_double
3   implicit none
4   public :: muller
5   integer, private :: iter=0, maxIter=100
6   real(c_double), private :: accu=1d-6
7 contains
8 subroutine muller(f, x0, x1, x2, root, info, nsteps)
9   complex(c_double), intent(in) :: x0, x1, x2
10  complex(c_double), intent(out) :: root
11  integer, intent(out), optional :: info, nsteps
12  interface
13    function f(x)
14      use iso_c_binding, only : c_double
15      complex(c_double), intent(in) :: x
16      complex(c_double) :: f
17    end function f
18  end interface
19  complex(c_double) :: h1,h2,d1,d2,a,b,d,e,h,x,x0t,x1t,x2t,y0,y1,y2
20  x0t = x0; x1t = x1; x2t = x2
21  do iter=1,maxIter
22    h1 = x1t-x0t
23    h2 = x2t-x1t
24    y0 = f(x0t)
25    y1 = f(x1t)
26    y2 = f(x2t)
27    d1 = (y1-y0)/h1
28    d2 = (y2-y1)/h2
29    a = (d2-d1)/(h2+h1)
30    b = d2+h2*a
31    d = sqrt(b**2-4.*a*y2) ! may need complex calculation
32    if (abs(b-d)<abs(b+d)) then
33      e = b+d
34    else
35      e = b-d
36    end if
37    h = -2.0*y2/e
38    x = x2t+h
39    if (abs(h)<accu) then
40      root = x
41      if(present(info)) info = 0
42      if(present(nsteps)) nsteps=iter
43      return
44    end if
45    x0t = x1t
46    x1t = x2t
47    x2t = x
48  end do
49  if (present(info)) info = 1 ! Root Not Found After MaxIter
50 end subroutine muller
51 end module textbook_rootc

```

This method can be used in finding the roots of polynomials, but can also be used to find *complex roots of analytic functions*.

eg:poly-muller

Example 43.50. Find the root of the function $f(x) = x^5 - 3x + 7$ using Müller's method with initial values $x_0 = 0$, $x_1 = -0.6$, $x_2 = -1.2$. The Fortran implementation is given below.

```

1 program muller_eg
2   use iso_c_binding, only : c_double_complex
3   use textbook_rootc
4   implicit none
5   complex(c_double_complex) :: theroot, x0, x1, x2

```

```

6      complex(c_double_complex), external :: f
7      integer :: info, cnt
8      x0 = ( 0.0d0,0.0d0)
9      x1 = (-0.6d0,0.0d0)
10     x2 = (-1.2d0,0.0d0)
11     call muller(f,x0,x1,x2,theroot,info,cnt)
12     write(*,*) "the root", theroot, "is found in", cnt, "iterations"
13 end program muller_eg
14 function f(x)
15     use iso_c_binding, only : c_double_complex
16     complex(c_double_complex) :: x, f
17     f = x**5 - 3.*x + 7.
18 end function f

```

In how many steps do we obtain an accuracy of 10^{-6} ? How does it perform compare to Newton's method? Give a possible reason for the performance. What is the major problem with Müller's method?

Solution:

The root -1.6416942631448153 is found in 7 iterations. It is faster than Newton's method because the initial values are chosen nicely.

If different initial values $x_0 = 0, x_1 = -0.5, x_2 = -1.0$, a different root $1.1457738431332329 + 0.66544401299692324i$ is found in 10 iterations.

A major problem with Müller's method the sensitivity of final solution with respect to the *three initial values*. It is difficult to decide what initial values would be appropriate.

Remark 43.51. The R function `uniroot` (which is `zbrent` in the IMSL libraries) is based on an algorithm of Richard Brent that uses a combination of linear interpolation, inverse quadratic interpolation, and bisection to find a root of a univariate function in an interval whose endpoints evaluate to values with different signs. The R function `polyroot` is based on the Traub-Jenkins algorithm to find the roots of a univariate polynomial.

To work on examples from MathematicsAndTechnology

Let (x, y, z) be the unknown position of the receiver on the ground and let (a_i, b_i, c_i) , be the known position of three satellite P_i , $i = 1, 2, 3$. Then (x, y, z) must satisfy

$$(x - a_i)^2 + (y - b_i)^2 + (z - c_i)^2 = c^2 t_i^2.$$

We can solve them for the position (x, y, z) ...

43.6 Optimisation and Machine Learning

sec:optim

Optimisation is very important in real-world problems, especially in economics in which we try to maximise the investment. Optimisation problems can be roughly categorised to (a) continuous problems; (b) discrete problems; and (c) mixed problems. These problems can be further categorised into (a) unconstrained nonlinear programming problems (NLP) [Diwekar, 2008, Chapter 3], convex optimisation [Ben-Tal and Nemirovski, 2001, Boyd and Vandenberghe, 2004], constrained optimisation [Bertsekas, 1996], linear programming, quadratic programming, (b) combinatorial optimisation problems [Hochstättler and Schliep, 2010, Papadimitriou and Steiglitz, 1982, Hromkovič, 2001], discrete optimisation problems [Diwekar, 2008, Chapter 4], etc. A set of problems in local and global optimisation problems is available in Floudas and Pardalos [1999] and **practical guides** can be found from Betts [2010] and Moré and Wright [1993].

The theory and methods of optimisation are applied in economics [Luptáčík, 2010], financial portfolio management [Cornuejols and Tütüncü, 2007, Würtz et al., 2009], estimating american option pricing, statistical inference Gentle [2009, 241], engineering control [Bertsekas, 1995a, 9], etc. The older standard library for nonlinear optimisation MINPACK is used in Python Scipy

module and GNU Octave (a MATLAB clone). For newer optimisation libraries, we have COIRO Python library, etc.

Since this subject is about “continuous models”, we will start with simple *nonlinear programming minimisation problem*:

$$\operatorname{argmin}_{\mathbf{x} \in D} F(\mathbf{x}). \quad (43.29)$$

The scalar-valued function F is called the *objective function*. The variable $\mathbf{x}$ is called the *decision variable*. The domain $D \subset \mathbb{R}^n$ of the decision variables is called the *feasible set*. If $D = \mathbb{R}^n$, the optimisation problem (43.29) is said to be *unconstrained*. Otherwise, (43.29) is said to be *constrained*. Closely related to optimisation problem is the solution of general nonlinear problem (43.3) of n -dimension.

Mathematically, we can regard (43.29) as a generalisation of nonlinear system (43.1) with $F(\mathbf{x}) = \sum_{i=1}^{\infty} (f_i(\mathbf{x}))^2$ [Dennis and Schnabel, 1996, Chapter 1].

When $D = \mathbb{R}^n$ and F is C^2 , the first three terms of the Taylor series expansion at $\mathbf{x}$ is

$$F(\bar{\mathbf{x}}) = F(\mathbf{x}) + \underbrace{(\mathbf{g}(\mathbf{x}))^T (\bar{\mathbf{x}} - \mathbf{x})}_{\text{directional derivative along } \mathbf{p}} + \frac{1}{2} \underbrace{(\bar{\mathbf{x}} - \mathbf{x})^T H(\mathbf{x}) (\bar{\mathbf{x}} - \mathbf{x})}_{\text{curvature in the direction } \mathbf{p}},$$

where $\mathbf{g}(\mathbf{x}) = (\frac{\partial F}{\partial x_1}, \dots, \frac{\partial F}{\partial x_n})^T$ is the gradient of F and $H(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 F}{\partial x_1^2} & \frac{\partial^2 F}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 F}{\partial x_1 \partial x_n} \\ \frac{\partial^2 F}{\partial x_2 \partial x_1} & \frac{\partial^2 F}{\partial x_2^2} & \cdots & \frac{\partial^2 F}{\partial x_2 \partial x_n} \\ \vdots & \cdots & \ddots & \vdots \\ \frac{\partial^2 F}{\partial x_n \partial x_1} & \frac{\partial^2 F}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 F}{\partial x_n^2} \end{bmatrix}$

is the Hessian of F . Note that $\mathbf{p} := \bar{\mathbf{x}} - \mathbf{x}$.

At minimum point $\mathbf{x}^*$, the Newton form of the gradient $\mathbf{g}$ is

$$\mathbf{g}(\mathbf{x}^*) = \mathbf{0}, \quad \mathbf{p}^T H^* \mathbf{p} > 0 \Rightarrow \mathbf{x}_{k+1} = \mathbf{x}_k - H(\mathbf{x}_k)^{-1} \mathbf{g}(\mathbf{x}_k). \quad (43.30)$$

The Hessian $H(\mathbf{x}_k)$ is required to be positive definite for a minimisation problem in which Cholesky decomposition may be employed to solve (43.30) or Broyden methods (Section 43.4.4) may allow us to avoid solving (43.30).

43.6.1 Machine Learning

Machine learning (ML) has developed into a huge subject including *statistical machine learning* [Hastie et al., 2008, Duda et al., 2001] and *graphical models*. *Bayesian networks* and *Markov random fields* are two major graphical models whereas statistical machine learning consists of various learning models such as K-means clustering, naïve Bayes classifier, decision trees, boosting, random trees, expectation maximisation (EM), K-nearest neighbours, neural networks, support vector machine (SVM), etc. Despite the abundance of learning models, the goal of ML is to turn data into information structure by extracting rules or patterns from the data [Bradski and Kaehler, 2008, Chapter 13].

If the data has labels (desired output), the data (and the learning process) is called *supervised* and if there is no label associated with the data, the data and the learning process are said to be *unsupervised*. In the learning process has both supervised and unsupervised data, it is said to be *semi-supervised learning*.

Supervised data can be *categorical* or *numeric*. In the former case, the learning process is called *classification* while in the later case, it is called *regression*.

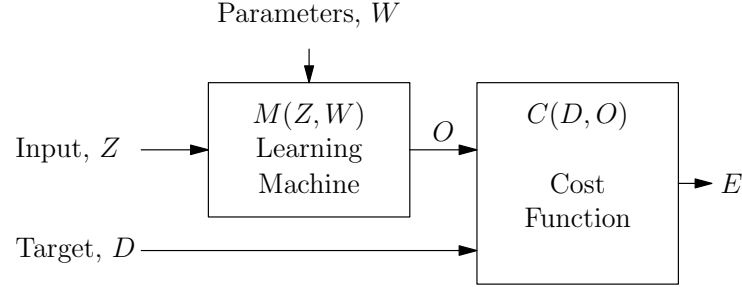
Statistical machine learning will be explored in *UECM3993 Predictive Modelling* (using R language) and more traditional knowledge representation system (e.g. fuzzy system) and neural network will be explored in *UECS2053/UECS2153 Artificial Intelligence*.

For many practitioners, ML is also regarded as *data mining* or *predictive analytics* because ML is just a kind of “mathematical model” with appropriate parameters which imitates human in performing classification.

To demonstrate that ML is just a “mathematical model” (and no way near human intelligence), we will study the *neural network*, which is a special case of the *gradient-based learning machine*:

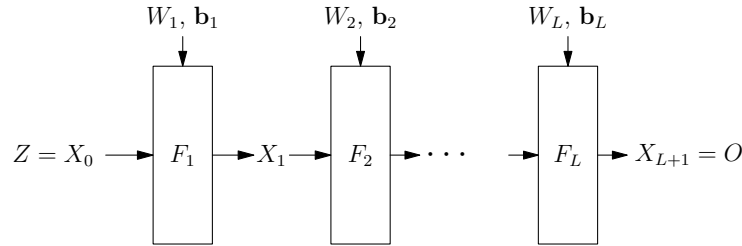
$$E = C(D, M(Z, W)) \quad (43.31)$$

where E is the discrepancy between the desired output D (for the input Z) and the output O of the learning machine with parameters W . It can be shown as the following block diagram:



This may be regarded as the generalisation of fitting data with a “nonlinear model” (the learning machine) as mentioned in page 1065.

The most popular learning machine $M(Z, W)$ is a multi-layer (deep) neural network (the one used in image recognition is a special case with “convolution” network connection, see <http://cs231n.github.io/optimization-2/> for information about CNN) which can be visualised as follows:



Here $W_1, W_2, \dots, W_L$ are matrices which need to be determined and we have a “chain” of functions:

$$\begin{aligned} X_1 &= F_1(W_1 X_0 + \mathbf{b}_1) = F_1(W_1 Z + \mathbf{b}_1) \\ X_2 &= F_2(W_2 X_1 + \mathbf{b}_2) = F_2(W_2 F_1(W_1 X_0 + \mathbf{b}_1) + \mathbf{b}_2) \\ &\dots \end{aligned}$$

$$O = X_{L+1} = F_L(W_L X_{L-1} + \mathbf{b}_L) = F_L(W_L F_{L-1}(W_{L-1} \dots F_2(W_2 F_1(W_1 Z + \mathbf{b}_1) + \dots) + \mathbf{b}_{L-1}) + \mathbf{b}_L). \quad (43.32)$$

The parameters $W_i, \mathbf{b}_i$ are normally obtained from a training process and all F_i are normally the same activating function $\sigma(x)$.

43.6.2 ML Software Libraries and A Little Bit About Training

There are a few popular machine learning software libraries such as Google’s TensorFlow [Abadi et al., 2015], University of Montréal’s Theano [Theano Development Team, 2016], Facebook’s PyTorch [Paszke et al., 2017], University of California, Berkeley’s Caffe [Jia et al., 2014], Microsoft Cognitive Toolkit (CNTK, <https://github.com/Microsoft/cntk>), Apache MXNet [Chen et al., 2015], etc. There are also other higher level machine learning libraries build on top of them such as Keras, a Python deep learning library (<https://keras.io/>) and Lasagne (<https://github.com/Lasagne/Lasagne>, based on Theano).

These libraries provide the software interface to build models and algorithms for *handwritten character recognition*, *image recognition* (e.g. Google’s Inception Image Classifier), *machine translation* (which is applied in google translate), *natural language processing* (for summarising texts), *sentiment analysis*, *speech recognition* (e.g. Baidu’s DeepSpeech), etc.

To explain how to work with the multi-layer neural network (43.32), I will use the Google image classification Inception-v3 model on TensorFlow as an example. The archive of the model parameters W_i for Inception-v3 model can be downloaded from <http://download.tensorflow.org/models/image/imagenet/inception-2015-12-05.tgz>. The following files (under Apache 2.0 License) can be extracted from the archive:

- `classify_image_graph_def.pb`: Binary representation of the GraphDef protocol buffer which contains the model and the parameters W_i s of the neural network;
- `imagenet_2012_challenge_label_map_proto.pbtxt`: Text representation of a protocol buffer mapping a label to synset UID;
- `imagenet_synset_to_human_label_map.txt`: Map from synset UID to human-readable string D_i ;
- `cropped_panda.jpg`: test image of size 100×100 for the Python program.

Putting them in appropriate directory and it is possible to use TensorFlow to load the parameters and construct the model for image recognition as follows.

```

1 # https://www.tensorflow.org/tutorials/images/image_recognition
2 #
3 # Copyright 2015 The TensorFlow Authors. All Rights Reserved.
4 # Licensed under the Apache License, Version 2.0
5 #
6 # Modified by Liew How Hui (February 2019)
7
8 """Simple image classification with Google Inception-v3 trained on
9 ImageNet 2012 Challenge data set.
10
11 This program creates a graph from a saved GraphDef protocol buffer,
12 and runs inference on an input JPEG image. It outputs human readable
13 strings of the top 5 predictions along with their probabilities.
14 """
15
16 import argparse, os.path, re, sys
17 import numpy as np
18 import tensorflow as tf
19
20 FLAGS = None # Define a global variable
21
22 class NodeLookup(object):
23     """Converts integer node ID's to human readable labels."""
24
25     def __init__(self, label_lookup_path=None, uid_lookup_path=None):
26         if not label_lookup_path:
27             label_lookup_path = os.path.join(
28                 FLAGS.model_dir, 'imagenet_2012_challenge_label_map_proto.pbtxt')
29         if not uid_lookup_path:
30             uid_lookup_path = os.path.join(
31                 FLAGS.model_dir, 'imagenet_synset_to_human_label_map.txt')
32         self.node_lookup = self.load(label_lookup_path, uid_lookup_path)
33
34     def load(self, label_lookup_path, uid_lookup_path):
35         """Loads a human readable English name for each softmax node.
36
37         Args:
38             label_lookup_path: string UID to integer node ID.
39             uid_lookup_path: string UID to human-readable string.
40
41         Returns:
42             dict from integer node ID to human-readable string.
43         """
44         if not tf.gfile.Exists(uid_lookup_path):
45             tf.logging.fatal('File does not exist %s', uid_lookup_path)

```

```

46     if not tf.gfile.Exists(label_lookup_path):
47         tf.logging.fatal('File does not exist %s', label_lookup_path)
48
49     # Loads mapping from string UID to human-readable string
50     proto_as_ascii_lines = tf.gfile.GFile(uid_lookup_path).readlines()
51     uid_to_human = {}
52     p = re.compile(r'[n\d]*[ \S,]*')
53     for line in proto_as_ascii_lines:
54         parsed_items = p.findall(line)
55         uid = parsed_items[0]
56         human_string = parsed_items[2]
57         uid_to_human[uid] = human_string
58
59     # Loads mapping from string UID to integer node ID.
60     node_id_to_uid = {}
61     proto_as_ascii = tf.gfile.GFile(label_lookup_path).readlines()
62     for line in proto_as_ascii:
63         if line.startswith('  target_class:'):
64             target_class = int(line.split(': ')[1])
65         if line.startswith('  target_class_string:'):
66             target_class_string = line.split(': ')[1]
67             node_id_to_uid[target_class] = target_class_string[1:-2]
68
69     # Loads the final mapping of integer node ID to human-readable string
70     node_id_to_name = {}
71     for key, val in node_id_to_uid.items():
72         if val not in uid_to_human:
73             tf.logging.fatal('Failed to locate: %s', val)
74             name = uid_to_human[val]
75             node_id_to_name[key] = name
76
77     return node_id_to_name
78
79 def id_to_string(self, node_id):
80     if node_id not in self.node_lookup: return ''
81     return self.node_lookup[node_id]
82
83 def main(_):
84     image = (FLAGS.image_file if FLAGS.image_file else
85             os.path.join(FLAGS.model_dir, 'cropped_panda.jpg'))
86
87     if not tf.gfile.Exists(image): tf.logging.fatal('File does not exist %s', image)
88     image_data = tf.gfile.GFile(image, 'rb').read()
89
90     # Creates graph from saved graph_def.pb.
91     with tf.gfile.GFile(os.path.join(
92         FLAGS.model_dir, 'classify_image_graph_def.pb'), 'rb') as f:
93         graph_def = tf.GraphDef()
94         graph_def.ParseFromString(f.read())
95         _ = tf.import_graph_def(graph_def, name='')
96
97     with tf.Session() as sess:
98         # Some useful tensors:
99         # 'softmax:0': A tensor containing the normalized prediction across
100         # 1000 labels.
101         # 'pool_3:0': A tensor containing the next-to-last layer containing 2048
102         # float description of the image.
103         # 'DecodeJpeg/contents:0': A tensor containing a string providing JPEG
104         # encoding of the image.
105         # Runs the softmax tensor by feeding the image_data as input to the graph.
106         softmax_tensor = sess.graph.get_tensor_by_name('softmax:0')
107         predictions = sess.run(softmax_tensor, {'DecodeJpeg/contents:0': image_data})
108         predictions = np.squeeze(predictions)
109
110         node_lookup = NodeLookup() # Creates node ID --> English string lookup.
111
112         top_k = predictions.argsort()[::-1]

```

```

113     for node_id in top_k:
114         human_string = node_lookup.id_to_string(node_id)
115         score = predictions[node_id]
116         print('%s (score = %.5f)' % (human_string, score))
117
118 if __name__ == '__main__':
119     parser = argparse.ArgumentParser()
120     parser.add_argument(
121         '--model_dir',
122         type=str,
123         default='/tmp/imagenet',
124         help="""\
125         Path to classify_image_graph_def.pb,
126         imagenet_synset_to_human_label_map.txt, and
127         imagenet_2012_challenge_label_map_proto.pbtxt.\
128         """)
129
130     parser.add_argument(
131         '--image_file',
132         type=str,
133         default='',
134         help='Absolute path to image file.')
135
136     parser.add_argument(
137         '--num_top_predictions',
138         type=int,
139         default=5,
140         help='Display this many predictions.')
141
142 FLAGS, unparsed = parser.parse_known_args() # update global variable
143 tf.app.run(main=main, argv=[sys.argv[0]] + unparsed)

```

How does Google obtained the parameters W_i and stored them in `classify_image_graph_def.pb`? They used a set of input/output pairs called the *training set* $\{(Z^1, D^1), \dots, (Z^P, D^P)\}$ and an appropriate cost function $C(D, O) = \frac{1}{2} \|D - O\|_2^2$, where $\|\cdot\|_2$ is the 2-norm. The *average cost function* is defined as

$$E_{train}(W) = \frac{1}{P} \sum_{p=1}^P \frac{1}{2} \|D^p - M(Z^p, W)\|_2^2. \quad (43.33) \quad \text{e:err-train}$$

And then they “trained” the neural network, i.e. optimising (43.33), i.e. $\min_{W \in \mathbb{R}^k} E_{train}(W)$, using “chain rule” and *gradient descent algorithm*

1. Initialise $W^{(0)}$ by a “random” guess;
2. While $\|\nabla C\| > \epsilon$, calculate [Bishop, 2006, Chapter 5]

$$W^{(k+1)} \leftarrow W^{(k)} - \eta \nabla E_{train}, \quad \nabla E_{train} = \left(\frac{\partial E_{train}}{\partial W_1}, \dots, \frac{\partial E_{train}}{\partial W_L} \right)^T \quad (43.34) \quad \text{e:grad-descent}$$

to find a “local minimum” solution for (43.29). According to <https://stats.stackexchange.com/questions/253632/why-is-newtons-method-not-widely-used-in-machine-learning>, neither the Newton form (43.30) or Broyden methods mentioned earlier are used in machine learning because the Hessian is expensive to compute.

Textbooks on machine learning such as Goodfellow et al. [2016] are normally too advanced and therefore the abundance of resources on the Internet about the training of a neural network such as <https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example/>, <http://colah.github.io/posts/2015-08-Backprop/>, https://www.reddit.com/r/math/comments/4ieenr/calculus_and_backpropagation/, <https://eli.thegreenplace.net/2018/backpropagation-through-a-fully-connected-layer/>, <http://neuralnetworksanddeeplearning.com/chap2.html>, etc. can be very helpful.

The algorithm to train a neural network is roughly as follows:

1. Feed-forward calculation based on (43.32);
2. Back-propagation of the error at the output layer $L + 1$:

$$(\delta_L)_j = \frac{\partial C}{\partial (X_{L+1})_j} \times \sigma((X_L)_j) \quad \text{or in vector form: } \delta_L = \nabla_{X_{L+1}} C \odot \sigma(X_L) = (D - X_{L+1}) \odot \sigma(X_L)$$

where $\frac{\partial C}{\partial (X_{L+1})_j} = D_j - (X_{L+1})_j$ measures how fast the cost is changing as a function of the j th output activation; $\sigma((X_L)_j)$ measures how fast the activation function σ is changing at $(X_L)_j$; $\odot$ is the Hadamard product (or Schur product), which is the same as Octave/MATLAB's $.*$.

3. Back-propagation of the hidden layer for (i) the error term:

$$\delta_{\ell-1} = ((W_\ell)^T \delta_\ell) \odot \sigma(X_{\ell-1}),$$

- (ii) the rate of change of the cost with respect to any bias $\mathbf{b}_\ell$ in the network:

$$\frac{\partial C}{\partial (\mathbf{b}_\ell)_j} = (\delta_\ell)_j \quad \text{or in vector form: } \frac{\partial C}{\partial \mathbf{b}_\ell} = \delta_\ell$$

- (iii) the rate of change of the cost with respect to any weight in the network:

$$\frac{\partial C}{\partial (W_\ell)_{jk}} = (X_{\ell-1})_k (\delta_\ell)_j$$

where $\ell = L, L - 1, \dots, 2$.

4. With a step size η , update the weights based on

$$W_\ell \leftarrow W_\ell - \frac{\eta}{P} \sum_{p=1}^P \delta_\ell^{(p)} (X_{\ell-1}^{(p)})^T,$$

$$\mathbf{b}_\ell \leftarrow \mathbf{b}_\ell - \frac{\eta}{P} \sum_{p=1}^P \delta_\ell^{(p)}.$$

On YouTube, one can find a lot of a graphical presentation of how neural networks are trained, e.g. <https://www.youtube.com/watch?v=Ilq3gGewQ5U>. A few computer programs which can be helpful in helping one to learn simple neural network with back-propagation are listed below.

- <http://karpathy.github.io/neuralnets/> and <https://www.mladdict.com/neural-network-simulator> (JavaScript)
- <https://github.com/codeplea/genann> (C)
- <https://github.com/pim-book/programmers-introduction-to-mathematics> (Python)
- <https://github.com/kaszperro/deepy> (Python)
- <https://github.com/modern-fortran/neural-fortran> (Fortran 2008)

Chapter 44

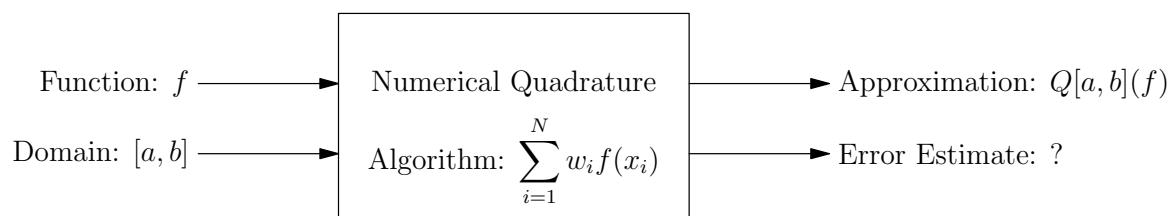
Numerical Integration

Much efforts in undergraduate STEM courses have been devoted to techniques for the analytic evaluation of integrals. However, more routine integrals in practical scientific work are incapable of being evaluated in *closed form*. Even if an expression can be derived for the evaluation of an integral, often this reveals itself only after inordinate amounts of error prone algebraic manipulation. Furthermore, successful analytic integration of rather simple functions can result in lengthy formulae. Thus many expressions for integrals which appear in books or tables as “exact” require approximations before they can be evaluated numerically. Thus seeking methods to approximate integrals directly based on integrands rather than by performing problem specific algebraic manipulations seems a better idea [Piessens et al., 1983].

In this topic, we investigate the numerical techniques for solving one-dimensional definite integral of f (left-hand side) by the approximation on the right-hand side of

$$\int_a^b f(x)dx \approx \sum_{i=1}^n w_i f(x_i) =: Q[a, b](f). \quad (44.1) \quad \text{e:quadrature}$$

In a block diagram:



Most numerical integration methods in this topic will be implemented in Fortran 2003. The C++ implementations are available in Boost 1.66 Math Toolkit and Flowers [2000]. However, one needs to beware that the Fortran implementation in this topic is very primitive and does not handle singularity such as $\int_0^1 \frac{1}{3x-1} dx$. *Quadrature* is used in statistics and finance such as (1) Bayesian analysis; (2) finding maximum likelihood estimates in models with qualitative dependent variables; (3) the expected value of a “continuous” random variable is an integral and quadrature method could be used Andricopoulos et al. [2003] (This paper used trapezoidal and Simpson rules to approximate option price, however, I find it very confusing. Fortunately, a Fortran pseudo-code was given in the paper to clarify the methodology). The methods discussed in this topic are not sufficient for modern finance, physics and engineering problems which require more sophisticated probability theory to perform “data filtering” (refer to Section ?? for examples). sec:svjd

In numerical methods, *numerical integration* is used to denote the solution process of an ODE-IVP (Topic 6). On the other hand, the process integrating a function numerically is also called “numerical integration”. To avoid confusion, numerical analyst used somewhat classical terms **quadrature** for univariate integral and **cubature** for numerical computation of a multiple integral e.g. $\iint_D f(x, y) dS$ [Stroud, 1971].

44.1 Evenly-Spaced Quadrature

sec:quad-basic

How is it possible for us to approximate a definite integral $\int_a^b f(x)dx$ by the right-hand term of (44.1)? We can first divide the interval $[a, b] \subset \mathbb{R}$ into *equal* subintervals $[x_i, x_{i+1}]$, $i = 0, \dots, n-1$ where $x_1 = a$, $x_{i+1} = x_j + h$, $x_n = b$ and $h = \frac{b-a}{n}$. Then, we will use a suitable “interpolation” polynomial $P(x)$ (Topic 4) to approximate the integrand f .

44.1.1 Basic Quadrature Rules

sec:basic-quad

If we are trying to approximate a function f with a polynomial, we will need to answer this question: *Is it possible for us to approximate any function f with a polynomial to any desirable accuracy?*

In science and engineering, most functions we encountered are continuous. A German mathematician Weierstrass shown that it is possible to approximate any given continuous function on a closed interval using polynomial.

Theorem 44.1 (Weierstrass Approximation Theorem). *Suppose that $f \in C^0[a, b]$. For each $\varepsilon > 0$, there exists a polynomial $P(x)$ defined on $[a, b]$, with the property that*

$$|f(x) - P(x)| < \varepsilon, \quad \forall x \in [a, b].$$

The Weierstrass approximation theorem just says that it is possible to approximate a continuous function up to any accuracy but does not tell us how can we find such a polynomial.

Suppose $f \in C^{m+1}[a, b]$ and recall Newton form from Topic 4:

$$P_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + \dots + f[x_0, \dots, x_n](x - x_0) \cdots (x - x_{n-1}) \quad (44.2)$$

e:newton-form

with an error estimate given by

$$f(x) - P_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!} (x - x_0) \cdots (x - x_{n-1}) \quad (44.3)$$

e:err-poly

where $\xi(x)$ is a number depending on x that lies in the smallest interval that contains all of $x, x_0, \dots, x_n$.

If the error estimate in (44.3) is very small, then we have the following approximation:

$$\int_a^b f(x)dx - \int_a^b P_n(x)dx = \int_a^b \frac{f^{(n+1)}(\xi(x))}{(n+1)!} (x - x_0) \cdots (x - x_{n-1})dx. \quad (44.4)$$

e:basic

We would like to extract the function $f^{(n+1)}(\xi(x))$ from (44.4). This could be achieved by applying the following theorem.

thm:mvt1

Theorem 44.2 (First Mean Value Theorem). *If $g \in C^0[a, b]$ and φ is an integrable function that does not change sign on the interval (a, b) , then there exists a number $\hat{x} \in (a, b)$ such that*

$$\int_a^b g(x)\varphi(x)dx = g(\hat{x}) \int_a^b \varphi(x)dx.$$

Let $g(x) = f^{(n+1)}(\xi(x))$ (which is C^0 because $f \in C^{n+1}[a, b]$) and $\varphi(x) = (x - x_0) \cdots (x - x_{n-1})$, then (44.4) becomes

$$\int_a^b f(x)dx = \int_a^b P_n(x)dx + \underbrace{\frac{f^{(n+1)}(\xi(\hat{x}))}{(n+1)!} \int_a^b (x - x_0) \cdots (x - x_{n-1})dx}_{\text{error estimate}}. \quad (44.5)$$

e:basic:n

When $\boxed{n=0}$, $P_n(x) = f[x_0]$. There are many choices for x_0 . If $f \in C^1[a, b]$ and we choose the midpoint $x_0 = \frac{a+b}{2}$, then $f[x_0] = f(\frac{a+b}{2})$ and we obtain a *basic midpoint (or rectangular) rule*:

$$\int_a^b f(x)dx = \int_a^b P_0(x)dx + \frac{f'(\xi(\hat{x}))}{1!} \int_a^b (x-a)dx = f(\frac{a+b}{2})(b-a) + \frac{f'(\xi(\hat{x}))}{2}(b-a)^2. \quad (44.6) \quad \text{e:basic:0}$$

When $\boxed{n=1}$, $P_n(x) = f[x_0] + f[x_0, x_1](x - x_0)$. If $f \in C^2[a, b]$ and we choose $x_0 = a$, $x_1 = b$, then $P_1(x) = f(a) + \frac{f(b) - f(a)}{b-a}(x-a)$ and we obtain the *basic trapezoidal rule*:

$$\begin{aligned} \int_a^b f(x)dx &= \int_a^b P_1(x)dx + \frac{f''(\xi(\hat{x}))}{2!} \int_a^b (x-a)(x-b)dx \\ &= f(a)(b-a) + \frac{f(b) - f(a)}{b-a} \left[\frac{(x-a)^2}{2} \right]_a^b + \frac{f''(\xi(\hat{x}))}{2} \left[\frac{a^3 - 3a^2b + 3ab^2 - b^3}{6} \right] \\ &= \frac{f(a) + f(b)}{2}(b-a) - \frac{f''(\xi)}{12}(b-a)^3. \end{aligned} \quad (44.7) \quad \text{e:basic:1}$$

When $\boxed{n=2}$, $P_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)^2$. If we choose $x_0 = a$, $x_1 = \frac{a+b}{2}$, $x_2 = b$, and if $f \in C^4[a, b]$, we will obtain the *basic Simpson rule*:

Theorem 44.3. [Burden and Faires, 2015, p196] If $f \in C^4[a, b]$, then there is a number $\xi \in (a, b)$ such that

$$\int_a^b f(x)dx = \frac{b-a}{6} \left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right] - \frac{f^{(4)}(\xi)}{2880}(b-a)^5. \quad (44.8) \quad \text{e:basic:2}$$

In general, by using the Lagrangian form of the interpolating polynomial on the points x_i , $i = 0, 1, \dots, n$, we will obtain an $(n+1)$ -point *closed Newton-Cotes formula*:

Theorem 44.4. [Burden and Faires, 2015, p198]

1. If n is even and $f \in C^{n+2}[a, b]$, then there is a $\xi \in (a, b)$ such that

$$\int_a^b f(x)dx = \sum_{k=0}^n c_k f(x_k) + \underbrace{\frac{h^{n+3} f^{(n+2)}(\xi)}{(n+2)!} \int_0^n t^2(t-1) \cdots (t-n)dt}_{\text{Error formula}}.$$

2. If n is odd and $f \in C^{n+1}[a, b]$, then there is a $\xi \in (a, b)$ such that

$$\int_a^b f(x)dx = \sum_{k=0}^n c_k f(x_k) + \underbrace{\frac{h^{n+2} f^{(n+1)}(\xi)}{(n+1)!} \int_0^n t(t-1) \cdots (t-n)dt}_{\text{Error formula}}.$$

Here $c_k = \int_a^b L_{n,k}(x)dx = \int_{x_0}^{x_n} \prod_{\substack{j=0 \\ j \neq k}}^n \frac{(x - x_j)}{(x_i - x_j)} dx$ and $L_{n,k}(x)$ ($k = 0, 1, \dots, n$) is the Lagrange polynomial.

Remark 44.5. The n -point formulae for odd integer n are of the same order of accuracy as the $(n+1)$ -point formula. Both have degree of precision of n . For example, Simpson's rule possesses degree of precision 3.

Let us study how to calculate with the basic quadrature rules with a simple example.

Example 44.6. Estimate the value of $\int_0^{10} x^3 dx$ using (a) basic midpoint rule; (b) basic trapezoidal rule; and (c) basic Simpson rule.

Solution: We know that $\int_0^1 x^3 dx = 2500$.

	(a)	(b)	(c)
n	0	1	2
$\int_0^1 P_n(x) dx$	$(\frac{0+10}{2})^3(10-0)$ $= 1250$	$\frac{0^3+10^3}{2}(10-0)$ $= 5000$	$\frac{10-0}{6} [0^3 + 4(\frac{0+10}{2})^3 + 10^3]$ $= 2500$
Abs. Error	1250	2500	0

The basic Simpson rule obtain exact integral according to Theorem 44.8, where the fourth derivative of x^3 is zero.

When the abscissas x_i in (44.1) does not include the boundary points a and b , a quadrature formula derived using interpolating polynomial is called an *open Newton-Cotes formula*. The midpoint rule is an example of open Newton-Cotes formula.

44.1.2 Composite Rules

Basic quadrature rules are unsatisfactory because $[a, b]$ can be large and when this happens, the size of the interval $h = b - a$ indicates that the error can be large due to the Runge phenomenon (Topic 4). So it is better to cut $[a, b]$ into n equal sub-intervals so that each sub-interval has

small length given by $h = \frac{b-a}{n}$ and then to apply the basic quadrature rules (Section 44.1.1) to each sub-interval. Summing up the basic quadrature rules for each sub-interval leads to the *composite rules* or the *extended rules* [Press et al., 2007]. In particular, we will obtain the composite midpoint rule, the composite trapezoidal rule, and the composite Simpson's rule as follows [Burden and Faires, 2015].

Applying the basic midpoint rule (44.6) to each sub-interval and sum them together, we obtain:

Theorem 44.7 (Composite Midpoint Rule). When $f \in C^2[a, b]$, let $x_i = a + ih$, there is a number $\mu \in (a, b)$ such that

$$\int_a^b f(x) dx = \underbrace{h \sum_{j=0}^{n-1} f\left(\frac{x_j + x_{j+1}}{2}\right)}_{Q[a,b](f) = M_{[a,b]}^n(f)} + \underbrace{\frac{(b-a)h^2}{6} f''(\mu)}_{\text{Error term}}.$$

The Fortran implementation of the composite midpoint rule $M_{[a,b]}^n(f)$ are given below respectively.

Midpoint in Fortran

```

1 function midpoint(f, a, b, n)
2   procedure(func1) :: f
3   intent(in) :: a, b, n
4   double precision :: a, b, mysum, h, midpoint
5   integer :: n, i
6   mysum = 0.0
7   h = (b-a)/n
8   do i=1,n
9     mysum = mysum + f(a + (i-0.5)*h)
10  end do

```

```

11     midpoint = mysum*h
12 end function

```

Note that the func1 is an interface to Fortran function:

```

1     abstract interface                                ! Type definition
2         function func1(x)                             ! One-Variable Function
3             double precision :: x, func1
4         end function func1
5     end interface

```

Applying the basic midpoint rule (44.7) to each sub-interval and sum them together, we obtain:

thm:trapz

Theorem 44.8 (Composite Trapezoidal Rule). *When $f \in C^2[a, b]$, there is a number $\mu \in (a, b)$ such that*

$$\int_a^b f(x)dx = \underbrace{\frac{h}{2} \left[f(a) + 2 \sum_{j=1}^{n-1} f(x_j) + f(b) \right]}_{Q[a,b](f)=T_{[a,b]}^n(f)} - \underbrace{\frac{(b-a)h^2}{12} f''(\mu)}_{\text{Error term}}. \quad (44.9)$$

e:trapz

The Fortran implementation of the composite trapezoidal rule is given below.

```

1 function trapInt(f, a, b, n)
2     procedure(func1) :: f
3     intent(in) :: a, b, n
4     double precision :: a, b, mysum, h, trapInt
5     integer :: n, i
6     mysum = f(a) + f(b)
7     h = (b-a)/n
8     do i=1,n-1
9         mysum = mysum + 2.0*f(a + i*h)
10    end do
11    trapInt = mysum*h/2.0
12 end function

```

If we apply basic Simpson's rule (44.8) to each sub-interval and sum them together, we obtain the (*composite*) *Simpson's rule*.

Theorem 44.9 (Simpson's Rule). *Let n be even. When $f \in C^4[a, b]$, there is a number $\mu \in (a, b)$ such that*

$$\int_a^b f(x)dx = \underbrace{\frac{h}{3} \left[f(a) + 2 \sum_{j=1}^{n/2-1} f(x_{2j}) + 4 \sum_{j=1}^{n/2} f(x_{2j-1}) + f(b) \right]}_{Q[a,b](f)=S_{[a,b]}^n(f)} - \underbrace{\frac{(b-a)h^4}{180} f^{(4)}(\mu)}_{\text{Error term}}.$$

The Fortran implementation of the composite Simpson's rule is given below.

```

1 function simpInt(f, a, b, n)
2     procedure(func1) :: f
3     intent(in) :: a, b, n
4     double precision :: a, b, mysum, h, simpInt
5     integer :: n, i
6     mysum = f(a)+f(b)
7     h = (b-a)/n
8     do i=1,n-1,2
9         mysum = mysum + 4.0*f(a + i*h)
10    end do

```

```

11      do i=2,n-2,2
12          mysum = mysum + 2.0*f(a + i*h)
13      end do
14      simpInt = mysum*h/3.0
15 end function

```

Before analysing any examples, let us define a function which is used in Matlab as a test function for quadrature.

Definition 44.1. The *hump function* is defined as

$$h(x) := \frac{1}{(x-0.3)^2 + 0.01} + \frac{1}{(x-0.9)^2 + 0.04} - 6.$$

We will be using the following functions as test cases for the composite rules. With the knowledge from Calculus, it is not difficult to obtain the exact values as shown below.

	(a)	(b)	(c)	(d)	(e)	(f)
Integral	$\int_0^{10} x^3 dx$	$\int_0^1 x^9 dx$	$\int_0^1 \sqrt{x} dx$	$\int_0^{2\pi} \sin^2 90x dx$	$\int_{-2}^2 x^3 e^x dx$	$\int_{-3}^3 h(x) dx$
Exact Solution	2500	$\frac{1}{10}$	$\frac{2}{3}$	π	$2e^2 + 38e^{-2}$	C

where the constant $C = 10 \tan^{-1} 33 + 10 \tan^{-1} 27 + 5 \tan^{-1} \frac{39}{2} + 5 \tan^{-1} \frac{21}{2} - 36$. These functions are implemented as a Fortran module `exmp_funcs` below.

```

1 module exmp_funcs
2     use iso_c_binding, only : c_double
3     use textbook_quad, only : func1
4     implicit none
5     public :: fa, fb, fc, fd, fe, h
6     public :: intfa, intfb, intfc, intfd, intfe, inth
7
8     real(c_double), parameter :: pi = 4.0*atan(1d0)
9     real(c_double), parameter :: intfa = 2.5d3
10    real(c_double), parameter :: intfb = 1d-1
11    real(c_double), parameter :: intfc = 2d0/3d0
12    real(c_double), parameter :: intfd = pi
13    real(c_double), parameter :: intfe = 2d0*exp(2d0) + 38d0*exp(-2d0)
14    real(c_double), parameter :: inth = 10d0*atan(33d0) &
15        + 10d0*atan(27d0) + 5d0*atan(39d0/2d0) &
16        + 5d0*atan(21d0/2d0) - 36d0
17 contains
18 function fa(x)
19     real(c_double) :: x, fa
20     fa = x**3
21 end function
22 function fb(x)
23     real(c_double) :: x, fb
24     fb = x**9
25 end function
26 function fc(x)
27     real(c_double) :: x, fc
28     fc = sqrt(x)
29 end function
30 function fd(x)
31     real(c_double) :: x, fd
32     fd = (sin(9d1*x))**2
33 end function
34 function fe(x)
35     real(c_double) :: x, fe
36     fe = x**3*exp(x)

```

```

37 end function
38 function h(x)
39     real(c_double) :: x, h
40     h = 1.0d0/((x-0.3d0)**2+1d-2) - 6d0
41     h = h + 1.0d0/((x-0.9d0)**2+4d-2)
42 end function
43 end module exmp_funcs

```

eg:cosx2b

Example 44.10. Use composite trapezoidal rule with $n = 4, 16, 256, 1024$ to approximated the following integrals:

	(a)	(b)	(c)	(d)	(e)	(f)
Integral	$\int_0^{10} x^3 dx$	$\int_0^1 x^9 dx$	$\int_0^1 \sqrt{x} dx$	$\int_0^{2\pi} \sin^2 90x dx$	$\int_{-2}^2 x^3 e^x dx$	$\int_{-3}^3 h(x) dx$

Solution:

```

N= 4: Q[ 0.000,10.000]      x^3=2656.250000000000 + 156.250000
N= 16: Q[ 0.000,10.000]     x^3=2509.765625000000 + 9.765625
N= 256: Q[ 0.000,10.000]    x^3=2500.038146972656 + 0.038147
N=1024: Q[ 0.000,10.000]    x^3=2500.002384185791 + 0.002384
N= 4: Q[ 0.000, 1.000]      x^9= 0.144260406494 + 0.044260
N= 16: Q[ 0.000, 1.000]      x^9= 0.102919036115 + 0.002919
N= 256: Q[ 0.000, 1.000]      x^9= 0.100011443929 + 0.000011
N=1024: Q[ 0.000, 1.000]      x^9= 0.100000715255 + 0.000001
N= 4: Q[ 0.000, 1.000]      sqrt(x)= 0.643283046243 + -0.023384
N= 16: Q[ 0.000, 1.000]      sqrt(x)= 0.663581196877 + -0.003085
N= 256: Q[ 0.000, 1.000]      sqrt(x)= 0.666616548977 + -0.000050
N=1024: Q[ 0.000, 1.000]      sqrt(x)= 0.666660362219 + -0.000006
N= 4: Q[ 0.000, 6.283] sin^2(90x)= 0.000000000000 + -3.141593
N= 16: Q[ 0.000, 6.283] sin^2(90x)= 3.141592653590 + 0.000000
N= 256: Q[ 0.000, 6.283] sin^2(90x)= 3.141592653590 + 0.000000
N=1024: Q[ 0.000, 6.283] sin^2(90x)= 3.141592653590 + 0.000000
N= 4: Q[-2.000, 2.000]      x^3e^x= 31.365285650064 + 11.444433
N= 16: Q[-2.000, 2.000]      x^3e^x= 20.684292517940 + 0.763440
N= 256: Q[-2.000, 2.000]      x^3e^x= 19.923848509235 + 0.002996
N=1024: Q[-2.000, 2.000]      x^3e^x= 19.921040185710 + 0.000187
N= 4: Q[-3.000, 3.000]      h(x)= -13.341385488644 + -23.061192
N= 16: Q[-3.000, 3.000]      h(x)= 10.395111306503 + 0.675305
N= 256: Q[-3.000, 3.000]      h(x)= 9.719788251808 + -0.000018
N=1024: Q[-3.000, 3.000]      h(x)= 9.719805525580 + -0.000001

```

$$n = 4: h = \frac{1-0}{4} = 0.25$$

$$\int_0^1 \sqrt{x} dx \approx \frac{0.25}{2} \left[\sqrt{0} + 2\sqrt{0.25} + 2\sqrt{0.5} + 2\sqrt{0.75} + \sqrt{1} \right]$$

$$= 0.64328305$$

In the composite trapezoidal rule example above, we observe that the more sub-intervals we have, the smaller the step size h and the more accurate approximation is obtained. In practice, composite trapezoidal rule is sufficient for basic estimation of a well-behaved integral.

eg:simpson

Example 44.11. Use composite Simpson's rule with $n = 4, 16, 256, 1024$ to approximated the following integrals:

	(a)	(b)	(c)	(d)	(e)	(f)
Integral	$\int_0^{10} x^3 dx$	$\int_0^1 x^9 dx$	$\int_0^1 \sqrt{x} dx$	$\int_0^{2\pi} \sin^2 90x dx$	$\int_{-2}^2 x^3 e^x dx$	$\int_{-3}^3 h(x) dx$

Solution:

```

N=   4: Q[ 0.000,10.000]      x^3=2500.000000000000 + 0.000000
N=  16: Q[ 0.000,10.000]      x^3=2500.000000000000 + 0.000000
N= 256: Q[ 0.000,10.000]      x^3=2500.000000000000 + 0.000000
N=1024: Q[ 0.000,10.000]      x^3=2500.000000000000 + 0.000000
N=   4: Q[ 0.000, 1.000]      x^9= 0.108688354492 + 0.008688
N=  16: Q[ 0.000, 1.000]      x^9= 0.100042131497 + 0.000042
N= 256: Q[ 0.000, 1.000]      x^9= 0.100000000652 + 0.000000
N=1024: Q[ 0.000, 1.000]      x^9= 0.100000000003 + 0.000000
N=   4: Q[ 0.000, 1.000]      sqrt(x)= 0.656526264793 + -0.010140
N=  16: Q[ 0.000, 1.000]      sqrt(x)= 0.665398188628 + -0.001268
N= 256: Q[ 0.000, 1.000]      sqrt(x)= 0.666646846203 + -0.000020
N=1024: Q[ 0.000, 1.000]      sqrt(x)= 0.666664189109 + -0.000002
N=   4: Q[ 0.000, 6.283]      sin^2(90x)= 0.000000000000 + -3.141593
N=  16: Q[ 0.000, 6.283]      sin^2(90x)= 3.141592653590 + 0.000000
N= 256: Q[ 0.000, 6.283]      sin^2(90x)= 3.141592653590 + 0.000000
N=1024: Q[ 0.000, 6.283]      sin^2(90x)= 3.141592653590 + 0.000000
N=   4: Q[-2.000, 2.000]      x^3e^x= 22.477125358234 + 2.556272
N=  16: Q[-2.000, 2.000]      x^3e^x= 19.934357390866 + 0.013504
N= 256: Q[-2.000, 2.000]      x^3e^x= 19.920853171348 + 0.000000
N=1024: Q[-2.000, 2.000]      x^3e^x= 19.920852961675 + 0.000000
N=   4: Q[-3.000, 3.000]      h(x)= -17.224495719220 + -26.944302
N=  16: Q[-3.000, 3.000]      h(x)= 16.328741452635 + 6.608935
N= 256: Q[-3.000, 3.000]      h(x)= 9.719832242706 + 0.000026
N=1024: Q[-3.000, 3.000]      h(x)= 9.719806677184 + -0.000000

```

$$n = 4: h = \frac{2\pi - 0}{4} = \frac{\pi}{2}; \quad n = 16: h = \frac{2\pi - 0}{16} = \frac{\pi}{8}.$$

0	$\frac{\pi}{8}$	$\frac{\pi}{4}$	$\frac{3\pi}{8}$	$\frac{\pi}{2}$	$\frac{5\pi}{8}$	$\frac{3\pi}{4}$	$\frac{7\pi}{8}$	π	$\frac{9\pi}{8}$	$\frac{5\pi}{4}$	$\frac{11\pi}{8}$	$\frac{3\pi}{2}$	$\frac{13\pi}{8}$	$\frac{7\pi}{4}$	$\frac{15\pi}{8}$	2π
0	$\frac{1}{2}$	1	$\frac{1}{2}$	0	$\frac{1}{2}$	1	$\frac{1}{2}$	0	$\frac{1}{2}$	1	$\frac{1}{2}$	0	$\frac{1}{2}$	1	$\frac{1}{2}$	0

$$\begin{aligned} \int_0^{2\pi} \sin^2(90x) dx &\approx \frac{\pi/2}{3} \left[\sin^2(0) + 4\sin^2\left(\frac{90\pi}{2}\right) + 2\sin^2(90\pi) + 4\sin^2\left(\frac{270\pi}{2}\right) + \sin^2(180\pi) \right] \\ &= \frac{\pi}{4} [0 + 0 + 0 + 0 + 0] = 0 \end{aligned}$$

$$\begin{aligned} \int_0^{2\pi} \sin^2(90x) dx &\approx \frac{\pi/8}{3} \left[0 + 4\left(\frac{1}{2}\right) + 2(1) + 4\left(\frac{1}{2}\right) + 2(0) + 4\left(\frac{1}{2}\right) + 2(1) + 4\left(\frac{1}{2}\right) + 2(0) \right. \\ &\quad \left. + 4\left(\frac{1}{2}\right) + 2(1) + 4\left(\frac{1}{2}\right) + 2(0) + 4\left(\frac{1}{2}\right) + 2(1) + 4\left(\frac{1}{2}\right) + 0 \right] \\ &= \frac{\pi}{24} \times 24 = \pi. \end{aligned}$$

Example 44.12 (Tutorial 5, Q2). Given $I = \int_0^3 \left[|x^2 - 2x| + \frac{1}{1+x} \right] dx$.

1. Calculate the exact value of I .

Solution.

$$\int_0^2 \left[-(x^2 - 2x) + \frac{1}{1+x} \right] dx + \int_2^3 \left[x^2 - 2x + \frac{1}{1+x} \right] dx = \frac{4}{3} + \ln 3 + \frac{4}{3} + \ln \frac{4}{3} = \frac{8}{3} + \ln 4$$

□

2. Use composite midpoint rule with $n = 6$ to estimate I .

Solution.

$$M_{[0,3]}^6(f) = \frac{3-0}{6}(1.2375+1.5089+1.3819+0.8011+0.8702+2.3292) = 4.0644 \text{ (Err: 0.011473)}$$

x	0.25	0.75	1.25	1.75	2.25	2.75
$f(x)$	1.2375	1.5089	1.3819	0.8011	0.8702	2.3292

□

3. Use composite trapezoidal rule with $n = 6$ to estimate I .

Solution.

$$\begin{aligned} T_{[0,3]}^6(f) &= \frac{3-0}{2} \left[1.0 + 2(1.416667 + 1.5 + 1.15 + 0.333333 + 1.535714) + 3.25 \right] \\ &= 4.030357 \text{ (Err: -0.022604)} \end{aligned}$$

x	0.0	0.5	1.0	1.5	2.0	2.5	3.0
$f(x)$	1.0	1.416667	1.5	1.15	0.333333	1.535714	3.25

□

4. Use composite Simpson's rule with $n = 6$ to estimate I .

Solution.

$$\begin{aligned} S_{[0,3]}^6(f) &= \frac{3-0}{3} \left[1.0 + 4 \times 1.416667 + 2 \times 1.5 + 4 \times 1.15 + 2 \times 0.333333 \right. \\ &\quad \left. + 4 \times 1.535714 + 3.25 \right] \\ &= 4.054365 \text{ (Err: 0.001404)} \end{aligned}$$

□

Example 44.13 (Tutorial 5, Q1). Determine the values of n and h required to approximate $\int_0^2 e^{2x} \sin 3x dx$ with the composite Trapezoidal rule to within 10^{-4} by using the theoretical estimate of the Error formula. Use numerical computing environment to estimate a better estimate for n .

Solution. From Theorem [44.8](#), ^{thm:trapz}

$$\begin{aligned} \text{Error} &= \left| \frac{(b-a)h^2}{12} f''(\mu) \right| = \frac{2h^2}{12} \left| 12e^{2\mu} \cos(3\mu) - 5e^{2\mu} \sin(3\mu) \right| \leq 10^{-4} \\ &\Rightarrow \frac{h^2}{6} e^4 \left| 12 \cos(3\mu) - 5 \sin(3\mu) \right| = \frac{h^2}{6} e^4 \left| \sqrt{12^2 + 5^2} \cos(3\mu + \theta) \right| \leq 10^{-4} \\ &\Rightarrow h^2 \leq \frac{6}{13e^4} \times 10^{-4} \approx 8.4534 \times 10^{-7} \\ &\Rightarrow h \leq 9.1942 \times 10^{-4} \Rightarrow n \geq \frac{2-0}{h} = 2176.3 \end{aligned}$$

where $\theta = \tan^{-1} \frac{5}{12}$.

The estimate of 700 is obtained from

```

1 program tut_composite
2   use textbook_quad
3   character(len=80) :: fstr = "e^{2x}sin(3x)"
4   real(c_double), external :: f
5   real(c_double) :: a=0d0, b=2d0, exact, qf, err
6   exact = (3d0 + exp(4d0)*(2d0*sin(6d0) - 3d0*cos(6d0)))/13d0
7   qf = trapInt(f,a,b,700)

```

```

8      err = qf - exact
9      write(*,100) a,b,fstr,qf,err
10     100 format("Q[",f6.3,"",f6.3,"] ",a13,"=",f16.12," (Err: ",f8.6,")")
11     end program tut_composite
12     function f(x)
13         use iso_c_binding, only : c_double
14         real(c_double) :: f, x
15         f = exp(2d0*x)*sin(3d0*x)
16     end function f

```

□

Composite trapezoidal rule and composite Simpson rule are sufficient for most well-behaved integral with finite interval. What happens when the interval is infinite? We will normally take large (or small) enough value to represent ∞ (or $-\infty$). Let us consider an example from mathematical finance, the Black-Scholes formula for the pricing of European call options.

Example 44.14. In mathematical finance, we learned that the “analytic” formula for the pricing of European call option is given by

$$\mathbb{E}[e^{-r\tau}(S - K)^+] = S_0 N(d_1) - e^{-r\tau} N(d_2)$$

where

- S_0 = spot price; K = strike price; τ = time to maturity;
- r = constant rate of risk-free interest; σ = constant volatility of stock price S ;
- $N(d) = P[Z \leq d] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^d e^{-t^2/2} dt$.

Given $S_0 = 100$, $K = 110$, $\sigma = 0.2$, $\tau = 0.5$ and $r = 0.03$.

1. Calculate $\mathbb{E}[e^{-r\tau}(S - K)^+]$ with the analytic formula.

Solution: We will use R to perform the calculation because R supports numerical probability function of many random variables:

```

spot = 100.0; strike = 110.0; vol = 0.2; tau = 0.5; r = 0.03
d1 = (log(spot/strike) + (r + vol^2/2.0)*tau)/(sqrt(tau)*vol)
d2 = d1 - sqrt(tau)*vol
C = spot*pnorm(d1) - strike*exp(-r*tau)*pnorm(d2)
cat("exp(-r*tau)*E(S-K)^+ = ", C, "\n")

```

The price is found to be 2.611902.

2. Use the trapezoidal rule `trapInt` and the Simpson rule `simpInt` with 1000 slices to estimate $\mathbb{E}[e^{-r\tau}(S - K)^+]$!

Solution: By definition,

$$\mathbb{E}[e^{-r\tau}(S - K)^+] = \int_{-\infty}^K 0 \cdot \rho(x) dx + \int_K^{\infty} (x - K) \rho(x) dx = \int_K^{\infty} (x - K) \rho(x) dx$$

where $\rho(x)$ is the log-normal density with mean $\log(S_0) + (r - \frac{1}{2}\sigma^2)\tau$ because S follows log-normal distribution.

It is impossible to integrate to ∞ , $\rho(x)$ would be extremely small when $x > 10K$. Therefore, we pick $b = 10K$ instead of ∞ :

```

deriv = function(spot, K, r, vol, tau, x) {
  avg = log(spot) + (r - 0.5*vol^2)*tau
  stdDev = vol*sqrt(tau)
  return((x-K)*dlnorm(x, meanlog=avg, sdlog=stdDev)*exp(-r*tau))
}
trapInt = function(f, a, b, n){
  h = (b-a)/n; xi = a+seq(1,n-1)*h

```

```

mysum = f(a)+f(b)+2.0*sum(f(xi))
return(h*mysum/2.0)
}
simpInt = function(f, a, b, n){
  h = (b-a)/n
  xi_odd = a+seq(1,n-1,by=2)*h
  xi_evn = a+seq(2,n-2,by=2)*h
  mysum = f(a)+f(b)+4.0*sum(f(xi_odd))+2.0*sum(f(xi_evn))
  return(h*mysum/3.0)
}

spot = 100.0; strike = 110.0; vol = 0.2; tau = 0.5; r = 0.03
f = function(x){return(deriv(spot, strike, r, vol, tau, x))}
a = strike; b = 10*strike
cat("composite Trapezoidal = ", trapInt(f, a, b, 1000), "\n")
cat("composite Simpson      = ", simpInt(f, a, b, 1000), "\n")

```

Note that trapezoidal method and Simpson method are not available in R (because their accuracy is low) and we have to implement them in R. The computation shows that Simpson rule is more accurate:

- composite Trapezoidal = 2.610219
- composite Simpson = 2.611903

We will encounter more application of integration in Loss Model (an important subject in general insurance) where the numerical estimation of moments of *claims of severity with deductible* K has the form $\mathbb{E}[(S - K)^+]$.

44.1.3 Romberg Method

Richardson extrapolation (Topic 3) can be used to improve the accuracy of composite trapezoidal rule. This leads to *Romberg method*, which uses a sequence of composite trapezoidal formulas of accuracy of $N(h_k)$ and then applies the Richardson extrapolation to improve the accuracy.

The first step in Romberg process is to calculate the composite trapezoidal rules (44.9) with $n = 2^{k-1}$ and $h = \frac{b-a}{n} = \frac{b-a}{2^{k-1}} =: h_k$ for $k = 1, 2, \dots$.

$$N(h_k) = \frac{h_k}{2} \left[f(a) + f(b) + 2 \left(\sum_{i=1}^{2^{k-1}-1} f(a + ih_k) \right) \right], \quad k = 1, 2, \dots \quad (44.10)$$

To speed up computation, we let $R_{k,1} = N(h_k)$, we note that $h_{k+1} = \frac{h_k}{2}$ and then we use the relation between $N(h_k)$ and $N(h_{k+1})$ to rewrite (44.10) as

$$R_{k+1,1} = N(h_{k+1}) = \frac{R_{k,1}}{2} + h_{k+1} \left[\sum_{i=1}^{2^{k-1}} f(a + (2i-1)h_{k+1}) \right] = T_{[a,b]}^{2^k}(f), \quad k = 2, 3, \dots, n. \quad (44.11)$$

To apply Richardson extrapolation, we must first have approximation method with its error term in the form $\sum_{i=1}^{\infty} K_i h^i$, i.e.

$$\underbrace{\int_a^b f(x) dx}_M = \underbrace{T_{[a,b]}^{2^{k-1}}(f)}_{N_1(h)} + \sum_{i=1}^{\infty} K_i h_k^i, \quad k = 2, 3, \dots, n.$$

Therefore,

$$\int_a^b f(x)dx - R_{k-1,1} = \sum_{i=1}^{\infty} K_i h_{k-1}^{2i} = K_1 h_{k-1}^2 + \sum_{i=2}^{\infty} K_i h_{k-1}^{2i} = O(h_{k-1}^2), \quad (44.12) \quad \text{eq:4.4}$$

$$\int_a^b f(x)dx - R_{k,1} = \sum_{i=1}^{\infty} K_i h_k^{2i} = \frac{K_1 h_{k-1}^2}{4} + \sum_{i=2}^{\infty} \frac{K_i h_{k-1}^{2i}}{4^i}. \quad (44.13) \quad \text{eq:4.5}$$

Subtracting (44.12) from four times (44.13), we have

$$\begin{aligned} 3 \int_a^b f(x)dx - 4R_{k,1} + R_{k-1,1} &= 4 \sum_{i=2}^{\infty} \frac{K_i h_{k-1}^{2i}}{4^i} - \sum_{i=2}^{\infty} K_i h_{k-1}^{2i}, \\ \Rightarrow \int_a^b f(x)dx - \underbrace{\frac{4R_{k,1} - R_{k-1,1}}{3}}_{R_{k,2}} &= \sum_{i=2}^{\infty} \frac{K_i}{3} \left(\frac{1 - 4^{i-1}}{4^{i-1}} \right) h_{k-1}^{2i} = O(h_{k-1}^4). \end{aligned}$$

Mathematicians discovered that $N_2(h_k) = R_{k,2}$ is just the Simpson's rule! A similar analysis on $R_{k,2}$ will generate a formula for $R_{k,3} = \frac{16R_{k,2} - R_{k-1,2}}{15}$. A general formula is

$$R_{k,j} = \underbrace{\frac{4^{j-1}R_{k,j-1} - R_{k-1,j-1}}{4^{j-1} - 1}}_{\text{hand calc.}} = \underbrace{R_{k,j-1} + \frac{R_{k,j-1} - R_{k-1,j-1}}{4^{j-1} - 1}}_{\text{computer calc.}}, \quad 2 \leq j \leq k \leq n. \quad (44.14) \quad \text{e:romberg}$$

We illustrate how Romberg method works with examples below.

Example 44.15 (April 2014 Final Exam Question 2(b)). Use Romberg integration (44.14) to find the following integral up to $n = 3$ stages:

$$I = \int_0^3 |(x+1)(x-2)(x-3)| dx.$$

and find the absolute error $|I - R_{3,3}|$. (10 marks)

Solution: Let $f(x) = |(x+1)(x-2)(x-3)|$. Let us calculate $R_{1,1}$ until $R_{3,1}$:

- $R_{1,1} = \frac{3-0}{2}[6+0] = 9$ (Basic Trapezoidal Rule) [1 mark]
- $R_{2,1} = \frac{R_{1,1}}{2} + \frac{3-0}{2}[f(0 + \frac{3-0}{2})] = \frac{9}{2} + \frac{3}{2}[1.875] = 7.3125$ (Using (44.11)) [1 mark]
- $R_{3,1} = \frac{R_{2,1}}{2} + \frac{3}{2^2}[f(0 + \frac{3-0}{2^2}) + f(0 + 3\frac{3-0}{2^2})] = \frac{7.3125}{2} + \frac{3}{4}[(4.921875 + 0.609375)] = 7.8046875$ (Using (44.11)) [2 marks]

Thus

- $R_{2,2} = \frac{4R_{2,1} - R_{1,1}}{4-1} = 6.75$ [1 mark]
- $R_{3,2} = \frac{4R_{3,1} - R_{2,1}}{4-1} = 7.96875$ [1 mark]
- $R_{3,3} = \frac{4^2 R_{3,2} - R_{2,2}}{4^2 - 1} = 8.05$ [2 marks]

Integrating I gives

$$\begin{aligned} I &= \int_0^2 (x+1)(x-2)(x-3)dx - \int_2^3 (x+1)(x-2)(x-3)dx \\ &= \int_0^2 (x^3 - 4x^2 + x + 6)dx - \int_2^3 (x^3 - 4x^2 + x + 6)dx \\ &= \left[\frac{x^4}{4} - \frac{4x^3}{3} + \frac{x^2}{2} + 6x \right]_0^2 - \left[\frac{x^4}{4} - \frac{4x^3}{3} + \frac{x^2}{2} + 6x \right]_2^3 \\ &= \frac{22}{3} - \left(-\frac{7}{12}\right) \end{aligned} \quad [1 \text{ mark}]$$

The absolute error $|I - R_{33}| = |7.91667 - 8.05| \approx 0.13333$ [1 mark]

Example 44.16 (Tutorial 5, Q3). Use Romberg integration to compute $R_{4,4}$, i.e. of order $O(h^8)$, for the integral $\int_0^3 \left[|x^2 - 2x| + \frac{1}{1+x} \right] dx$.

Solution. The computation of the first column is as follows:

$$\begin{aligned} R_{1,1} &= \frac{3-0}{2} [1 + 3.25] = 6.375 \\ R_{1,2} &= \frac{6.375}{2} + \frac{3-0}{2} (1.15) = 4.9125 \\ R_{1,3} &= \frac{4.9125}{2} + \frac{3-0}{4} (f(0.75) + f(2.25)) \\ &= 2.45625 + 0.75(1.50892857 + 0.87019231) = 4.24059066 \text{ (Err} = 0.009667) \end{aligned}$$

$$\begin{array}{cccc} 6.375000 & & & \\ 4.912500 & 4.425000 & & \\ 4.240591 & 4.016621 & 3.989396 & \\ 4.104157 & 4.058680 & 4.061484 & 4.062628 \end{array}$$

□

Example 44.17 (April 2018 Final Exam Question 4). Let n be a non-negative integer. The n -moment of an exponentially distributed random variable X with parameter λ is

$$\mathbb{E}[X^n] = \int_0^\infty x^n \lambda e^{-\lambda x} dx.$$

Suppose $\lambda = 2$ and you are using the interval 0 to 10 to approximate the 2-moment:

$$Q_f = \int_0^{10} 2x^2 e^{-2x} dx.$$

- (a) Construct a Romberg table of order $O(h^{10})$ for Q_f with an accuracy of 7 decimal places for all calculations.

$$\begin{array}{cccccc} R_{1,1} & & & & & \\ R_{2,1} & R_{2,2} & & & & \\ R_{3,1} & R_{3,2} & R_{3,3} & & & \\ R_{4,1} & R_{4,2} & R_{4,3} & R_{4,4} & & \\ R_{5,1} & R_{5,2} & R_{5,3} & R_{5,4} & R_{5,5} & \end{array}$$

where

$$R_{i,j} = \frac{4^{j-1} R_{i,j-1} - R_{i-1,j-1}}{4^{j-1} - 1}, \quad 1 \leq i \leq j \leq 5.$$

[**Warning:** Marks will be deducted for low precision calculation but ZERO marks will be awarded for calculations with less than 4 decimal places.] (15 marks)

Solution: Quite a few students did not write down the steps and calculate using the wrong answer, they receive zero in the end. To get partial credit, always write down the steps. Let $f(x) = 2x^2 e^{-2x}$. For $h = \frac{10-0}{16} = 0.625$,

x_i	0	0.625	1.25	1.875	2.5	3.125
$f(x_i)$	0.0	0.2238319	0.2565156	0.1653592	0.0842243	0.0377042
x_i	3.75	4.375	5.0	5.625	6.25	6.875
$f(x_i)$	0.0155555	0.0060661	0.0022700	0.0008231	0.0002911	0.0001009
x_i	7.5	8.125	8.75	9.375	10	
$f(x_i)$	0.0000344	0.0000116	0.0000038	0.0000013	0.0000004	

..... [1 mark]

$$\begin{aligned}
R_{1,1} &= \frac{10-0}{2}(0.0 + 0.0000004) = 0.0000020 \\
R_{2,1} &= \frac{10-0}{4}(0.0 + 2(0.0022700) + 0.0000004) = 0.0113510 \quad [1 \text{ mark}] \\
R_{3,1} &= \frac{10-0}{8}(0.0 + 2(0.0842243 + 0.0022700 + 0.0000344) + 0.0000004) \\
&= 0.2163223 \quad [1 \text{ mark}] \\
R_{4,1} &= \frac{10-0}{16}(0.0 + 2(0.2565156 + 0.0842243 + 0.0155555 + 0.0022700 \\
&\quad + 0.0002911 + 0.0000344 + 0.0000038) + 0.0000004) \\
&= 0.4486186 \quad [1 \text{ mark}] \\
R_{5,1} &= \frac{10-0}{32}(0.0 + 2(0.2238319 + \cdots + 0.0000013) + 0.0000004) \\
&= 0.4954957 \quad [1 \text{ mark}]
\end{aligned}$$

Therefore,

$$\begin{aligned}
R_{2,2} &= \frac{4(0.0113510) - 0.0000020}{3} = 0.0151340 \quad [1 \text{ mark}] \\
R_{3,2} &= \frac{4(0.2163223) - 0.0113510}{3} = 0.2846461 \quad [1 \text{ mark}] \\
R_{4,2} &= \frac{4(0.4486186) - 0.2163223}{3} = 0.5260507 \quad [1 \text{ mark}] \\
R_{5,2} &= \frac{4(0.4954957) - 0.4486186}{3} = 0.5111214 \quad [1 \text{ mark}]
\end{aligned}$$

$$\begin{aligned}
R_{3,3} &= \frac{4^2(0.2846461) - 0.0151340}{15} = 0.3026136 \quad [1 \text{ mark}] \\
R_{4,3} &= \frac{4^2(0.5260507) - 0.2846461}{15} = 0.5421443 \quad [1 \text{ mark}] \\
R_{5,3} &= \frac{4^2(0.5111214) - 0.5260507}{15} = 0.5101261 \quad [1 \text{ mark}]
\end{aligned}$$

$$\begin{aligned}
R_{4,4} &= \frac{4^3(0.5421443) - 0.3026136}{63} = 0.5459464 \quad [1 \text{ mark}] \\
R_{5,4} &= \frac{4^3(0.5101261) - 0.5421443}{63} = 0.5096179 \quad [1 \text{ mark}]
\end{aligned}$$

$$\text{Final, } R_{5,5} = \frac{4^4(0.5096179) - 0.5459464}{255} = 0.5094754 \dots\dots\dots [1 \text{ mark}]$$

- (b) Use the composite Simpson rule with $n = 16$ to approximate Q_f . Write your answer to 7 decimal places. [Hint: Use the Romberg table from part (a).] (2 marks)

Solution. Composite Simpson rule with $n = 16$ is $R_{5,2} = 0.5111214$. .. [2 marks] $\square$

- (c) The n -moment of an exponentially distributed random variable X with parameter λ has an analytic expression

$$\mathbb{E}[X^n] = \frac{n!}{\lambda^2}.$$

Use this formula to find the *relative error* between the estimate of Q_f found in part (a) and the exact value $\mathbb{E}[X^2]$. (3 marks)

Solution. Exact value $\mathbb{E}(X^2) = \frac{2!}{2^2} = 0.5$ [1 mark]

From part (a), the relative error

$$\left| \frac{R_{5,5} - \mathbb{E}(X^2)}{\mathbb{E}(X^2)} \right| = \left| \frac{0.5094754 - 0.5}{0.5} \right| = 0.0189508 \quad [2 \text{ marks}]$$

□

After practising with the above examples, it is not difficult to write a computer program to calculate the Romberg method with (44.11) and (44.14). A naïve implementation of the Romberg method in Fortran 2003 is shown below:

```

1 function romberg(f, a, b, order, R)
2   procedure(func1) :: f
3   intent(in) :: a, b, order
4   intent(out) :: R
5   integer :: order, i, j
6   double precision :: romberg, a, b, R(order,order), h
7   R = 0d0
8   h = b-a
9   R(1,1) = h/2.0*(f(a)+f(b))
10  do i=2,order
11    R(i,1) = 0.5*R(i-1,1)
12    do j=1,2**(i-2) ! Trapezoidal Approximation
13      R(i,j) = R(i,j-1) + 0.5*h*f(a+(j-0.5)*h)
14    end do
15    do j=2,i ! Extrapolating
16      R(i,j) = R(i,j-1) + (R(i,j-1) - R(i-1,j-1)) / (4**(j-1) - 1)
17    end do
18    h = 0.5*h
19  end do
20  romberg = R(order,order)
21 end function

```

We can use this function to perform calculations for various integrals below.

eg:romberg

Example 44.18. Use Romberg integration with order $O(h^{12})$ to approximate the following integrals:

	(a)	(b)	(c)	(d)	(e)	(f)
Integral	$\int_0^{10} x^3 dx$	$\int_0^1 x^9 dx$	$\int_0^1 \sqrt{x} dx$	$\int_0^{2\pi} \sin^2 90x dx$	$\int_{-2}^2 x^3 e^x dx$	$\int_{-3}^3 h(x) dx$

Solution:

By compiling and running a Fortran program, we obtain the results below.

```

Q[ 0.000,10.000]      x^3=2500.000000000000 + 0.000000
5000.000
3125.000 2500.000
2656.250 2500.000 2500.000
2539.062 2500.000 2500.000 2500.000
2509.766 2500.000 2500.000 2500.000 2500.000
2502.441 2500.000 2500.000 2500.000 2500.000 2500.000
-----
Q[ 0.000, 1.000]      x^9= 0.100000000000 + 0.000000
0.500000
0.250977 0.167969
0.144260 0.108688 0.104736
0.111550 0.100646 0.100110 0.100037
0.102919 0.100042 0.100002 0.100000 0.100000
0.100732 0.100003 0.100000 0.100000 0.100000 0.100000
-----
Q[ 0.000, 1.000]      sqrt(x)= 0.666287699034 + -0.000379
0.500000
0.603553 0.638071
0.643283 0.656526 0.657757

```

```

0.658130    0.663079    0.663516    0.663608
0.663581    0.665398    0.665553    0.665585    0.665593
0.665559    0.666218    0.666273    0.666284    0.666287    0.666288
-----
Q[ 0.000, 6.283] sin^2(90x)=  3.143162892849 +  0.001570
0.000000
0.000000    0.000000
0.000000    0.000000    0.000000
3.141593    4.188790    4.468043    4.538964
3.141593    3.141593    3.071779    3.049617    3.043776
3.141593    3.141593    3.141593    3.142701    3.143066    3.143163
-----
Q[-2.000, 2.000]      x^3e^x=  19.920853014353 +  0.000000
116.059533
58.029767    38.686511
31.365286    22.477125    21.396500
22.934098    20.123702    19.966807    19.944114
20.684293    19.934357    19.921734    19.921019    19.920928
20.112356    19.921711    19.920867    19.920854    19.920853    19.920853
-----
Q[-3.000, 3.000]      h(x)=  6.319664957780 +  -3.400142
-34.442933
-1.692055    9.224905
-13.341385    -17.224496    -18.987789
-7.405779    -5.427244    -4.640760    -4.413030
10.395111    16.328741    17.779140    18.135012    18.223436
7.973559    7.166375    6.555550    6.377398    6.331290    6.319665
-----

```

Consider $f(x) = \sqrt{x}$: $h = 1 - 0 = 1$,

$$R_{1,1} = \frac{1}{2}(f(0) + f(1)) = 0.5$$

$$R_{2,1} = \frac{0.5}{2} + \frac{1}{2}\sqrt{0.5} = 0.60355339$$

$$R_{3,1} = \frac{0.60355339}{2} + \frac{1}{4}(\sqrt{0.25} + \sqrt{0.75}) = 0.64328305$$

$$R_{4,1} = \frac{0.64328305}{2} + \frac{1}{8}(\sqrt{0.125} + \sqrt{0.375} + \sqrt{0.625} + \sqrt{0.875}) = 0.65813022$$

Now, let's calculate the second "order":

$$R_{2,2} = \frac{4(0.60355339) - 0.5}{3} = 0.63807119$$

$$R_{3,2} = \frac{4(0.64328305) - 0.60355339}{3} = 0.65652627$$

$$R_{4,2} = \frac{4(0.65813022) - 0.64328305}{3} = 0.66307928$$

Now, let's calculate the third "order":

$$R_{3,3} = \frac{16(0.65652627) - 0.63807119}{15} = 0.65775661$$

$$R_{4,3} = \frac{16(0.66307928) - 0.65652627}{15} = 0.66351615$$

The forth "order":

$$R_{4,4} = \frac{64(0.66351615) - 0.65775661}{63} = 0.66360757$$

44.2 Gaussian Quadrature

In Sections [44.1](#) to [44.1.3](#), all of the numerical quadrature formulae are based on evenly spaced x -values; this means the x -values were predetermined. Gaussian quadrature is based on suitably chosen non-evenly spaced x -values (called the abscissas). Let us consider the case when there are two points, i.e. Equation [\(44.1\)](#) with $n = 2$:

$$\int_a^b f(x)dx \approx w_1 f(x_1) + w_2 f(x_2). \quad (44.15) \quad \text{e:quad-2term}$$

In the previous sections, $x_1 = a$ and $x_2 = b$. Gauss observed that if we remove the requirement that the function be evaluated at predetermined x -values, a two-term formula [\(44.15\)](#) will contain four parameters, i.e. x_1, x_2, w_1, w_2 . If f is replaced by appropriate polynomials, the approximation sign “ $\approx$ ” becomes an equal sign “ $=$ ”, we would be able to determine x_1, x_2, w_1 and w_2 . Formulas based on this principle are called *Gaussian quadrature formulae*. They can be applied only when f is known explicitly, so that can be evaluated at any desired value of x . Let us investigate Gauss’ idea by limiting ourselves to the following special case of [\(44.15\)](#), where $a = -1, b = 1$ and we use the variable t instead of x :

$$\int_{-1}^1 f(t)dt \approx w_1 f(t_1) + w_2 f(t_2).$$

Since w_1, w_2, t_1 and t_2 are unknown, we have to determine their values. Let $f(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3$ so that the $\approx$ becomes an equal sign as follows:

$$\int_{-1}^1 (a_0 + a_1 t + a_2 t^2 + a_3 t^3)dx = w_1(a_0 + a_1 t_1 + a_2 t_1^2 + a_3 t_1^3) + w_2(a_0 + a_1 t_2 + a_2 t_2^2 + a_3 t_2^3).$$

Since a_i are arbitrary, a system of nonlinear equations can be obtained below by choosing appropriate values:

$$\begin{aligned} a_0 \neq 0, a_1 = 0, a_2 = 0, a_3 = 0 &\Rightarrow \int_{-1}^1 1dt = 2 = w_1 + w_2 \\ a_0 = 0, a_1 \neq 0, a_2 = 0, a_3 = 0 &\Rightarrow \int_{-1}^1 tdt = 0 = w_1 t_1 + w_2 t_2, \\ a_0 = 0, a_1 = 0, a_2 \neq 0, a_3 = 0 &\Rightarrow \int_{-1}^1 t^2 dt = \frac{2}{3} = w_1 t_1^2 + w_2 t_2^2, \\ a_0 = 0, a_1 = 0, a_2 = 0, a_3 \neq 0 &\Rightarrow \int_{-1}^1 t^3 dt = 0 = w_1 t_1^3 + w_2 t_2^3. \end{aligned} \quad (44.16) \quad \text{e:nonlinear:ga}$$

Solving the system of nonlinear equations [\(44.16\)](#) gives us $w_1 = w_2 = 1, t_2 = -t_1 = \sqrt{\frac{1}{3}}$. From this, we obtain a *two-point Gaussian quadrature* or a *Gauss(-Lagrange) quadrature with $n = 2$* :

$$\int_{-1}^1 f(t)dt \approx f(-\sqrt{\frac{1}{3}}) + f(\sqrt{\frac{1}{3}}). \quad (44.17) \quad \text{e:gauss-lagran}$$

Note that [\(44.17\)](#) gives the exact value of the integral of any cubic polynomial over $[-1, 1]$.

We can generalise Gauss-Lagrange quadrature with $n = 2$ to general n :

$$\int_{-1}^1 f(t)dt \approx w_1 f(t_1) + w_2 f(t_2) + \cdots + w_n f(t_n). \quad (44.18) \quad \text{e:quad:gauss}$$

This formula is *exact* for functions $f(t)$ that are polynomials of degree $2n - 1$ or less! By extending the method we used previously for the 2-point formula, we obtain a system of $2n$ nonlinear equations:

$$w_1 t_1^k + \cdots + w_n t_n^k = \begin{cases} 0, & \text{for } k = 1, 3, 5, \dots, 2n - 1; \\ \frac{2}{k+1}, & \text{for } k = 0, 2, 4, \dots, 2n - 2. \end{cases} \quad (44.19) \quad \text{e:nonlinear:ga}$$

To determine the values of t_i and w_i , we need to solve the system of nonlinear equations (44.19) (Topic 3). However, this nonlinear system is difficult solved numerically!

In the analysis of linear partial differential equation (UECM3003), mathematicians found the special functions which are orthogonal. By applying the theory of special functions, in particular, n th(-degree) Legendre polynomial $P_n(x)$, the points $t_1, t_2, \dots, t_n$ needed to produce an integral approximation formula that gives exact results for any polynomial of degree less than $2n$ are the roots of $P_n(x)$, mathematicians established the following result.

Theorem 44.19 (Burden and Faires [2015]). Suppose that $t_1, \dots, t_n$ are the roots of the n th Legendre polynomial

$$P_n(x) = \begin{cases} 1, & n = 0, \\ \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n, & n > 0 \end{cases}$$

and that for each $i = 1, \dots, n$, the numbers w_i are defined by

$$w_i = \int_{-1}^1 L_{n,i}(t) dt = \int_{-1}^1 \prod_{\substack{j=1 \\ j \neq i}}^n \frac{t - t_j}{t_i - t_j} dt.$$

If $P(t)$ is any polynomial of degree less than $2n$, then $\int_{-1}^1 P(t) dt = \sum_{i=1}^n w_i P(t_i)$. The right-hand side is called the Gauss-Legendre Quadrature.

Remark 44.20. The Legendre polynomials can alternatively be defined by recursion:

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_{n+1}(x) = \frac{(2n+1)}{n+1} x P_n(x) - \frac{n}{n+1} P_{n-1}(x), \quad n \geq 1$$

and w_i is related to Legendre polynomial as follows [Abramowitz and Stegun, 1964, page 887, item 25.4.29]:

$$w_i = \frac{2}{(1 - t_i^2)[P'_n(t_i)]^2}.$$

Using the properties of Legendre polynomial, the results are shown below:

No. of terms, n	Values of t_i	Weighting factor, w_i	Degree
1	0	2	1
2	$\pm \sqrt{1/3}$	1	3
3	$\pm \sqrt{3/5}, 0$	$\frac{5}{9}, \frac{8}{9}$	5
4	$\pm \sqrt{(3 - 2\sqrt{6/5})/7}, \pm \sqrt{(3 + 2\sqrt{6/5})/7}$	$\frac{18+\sqrt{30}}{36}, \frac{18-\sqrt{30}}{36}$	7
5	$0, \pm \frac{1}{3}\sqrt{5 - 2\sqrt{10/7}}, \pm \frac{1}{3}\sqrt{5 + 2\sqrt{10/7}}$	$\frac{128}{255}, \frac{322+13\sqrt{70}}{900}, \frac{322-13\sqrt{70}}{900}$	9

These values can be obtained using `roots_legendre(n)` from the `scipy.special` module.

The “degree” in the table means the “degree of precision”. We usually say a quadrature formula (44.1) has *degree of precision* m if it is exact whenever $f(x)$ is a polynomial of degree $\leq m$ and it is not exact for $f(x) = x^{m+1}$. The x_i are called the *points* or *nodes* of the formula and the w_i are called *weights* or *coefficients* [Stroud and Secrest, 1966, p1].

So far we limit ourselves to the interval $[-1, 1]$, what about arbitrary finite interval $[a, b]$? If the limits of integration are from a to b instead of -1 to 1 , we could use the *change of variable*

by replacing the variable x in (44.1) by $x = \frac{(b-a)t + b + a}{2}$ so that $dx = \left(\frac{b-a}{2}\right) dt$, then

$$\int_a^b f(x) dx = \frac{b-a}{2} \int_{-1}^1 f\left(\frac{(b-a)t + b + a}{2}\right) dt \approx \frac{b-a}{2} \sum_{i=1}^n w_i f\left(\frac{(b-a)t_i + b + a}{2}\right). \quad (44.20)$$

A Python implementation of the n -point Gaussian-Legendre quadrature is as simple as

```
1 def gausslegendre(n,f,a,b):
2     from scipy.special import roots_legendre
3     s = (b-a)/2.0
4     c = (b+a)/2.0
5     t, w = roots_legendre(n)
6     return s*dot(w, [f(s*t[i]+c) for i in range(n)])
```

The Fortran implementation is much more complicated because there is a lack of standard library for finding the roots of Legendre functions as in Python. So we have to use a simple method introduced by Trefethen:

```
1  !!! https://people.maths.ox.ac.uk/trefethen/spectral.html
2  !!! Check: OK compare to https://pomax.github.io/bezierinfo/legendre-gauss.html
3  subroutine gauss(N, x, w, info)
4      integer, intent(in) :: N
5      double precision, intent(out) :: x(2*N-1), w(2*N-1)
6      integer, intent(out) :: info
7      double precision :: beta, T(N,N), WI(N)
8      double precision, allocatable :: WORK(:), VL(:,,:), VR(:,,:)
9      integer :: i, LDA, LDVL, LDVR, LWORK
10
11      LDA=N; LDVL=N; LDVR=N; LWORK=4*N
12      allocate(WORK(LWORK)); allocate(VL(LDVL,N)); allocate(VR(N,LDVR))
13      if (N .lt. 1) return
14      T = 0d0
15      do i=1,N-1
16          beta = 0.5d0/sqrt(1d0-(2d0*i)**(-2))
17          T(i,i+1) = beta; T(i+1,i) = beta
18      end do
19      ! 'N'x2 = Don't compute both left and right eigenvectors
20      ! x = real part of the eigenvalues
21      call dgeev('N','V',N,T,LDA,x,WI,VL,LDVL,VR,LDVR,WORK,LWORK,info)
22      do i=1,N ! WI(i) can be imaginary due to numerical error
23          !if (WI(i) .ne. 0d0) print *, "Imag part /= 0, something wrong"
24          w(i) = 2*VR(1,i)**2
25      end do
26  end subroutine
```

An alternative using Newton method is available from the Internet:

```
1  !!! https://rosettacode.org/wiki/Numerical_integration/Gauss-Legendre_Quadrature#Fortran
2  subroutine gauss2(n, tt, ww, info)
3      intent(in) :: n
4      intent(out) :: tt, ww, info
5      integer :: n, i, k, iter, info
6      double precision :: tt(n), ww(n), x, f, df, dx
7      double precision, allocatable :: p0(:), p1(:), p2(:)
8
9      if (n .lt. 1) then
10          info = 1
11          return
12      end if
13      ! Calculate the coefficients of Legendre Polynomials
14      p0 = [1d0]
15      p1 = [1d0, 0d0]
16      do k = 2, n
17          ! http://mathworld.wolfram.com/LegendrePolynomial.html
18          ! Using the Recurrence Relation
19          p2 = ((2*k-1)*[p1,0d0]-(k-1)*[0d0, 0d0,p0])/k
20          p0 = p1; p1 = p2
21      end do
22      do i = 1, n
23          x = cos(pi*(i-0.25d0)/(n+0.5d0))
24          do iter = 1, 1000
25              f = p1(1); df = 0d0
26              do k = 2, size(p1)
```

```

27         df = f + x*df
28         f = p1(k) + x*f
29     end do
30     dx = f / df
31     x = x - dx      ! Newton's Method
32     if (abs(dx)<10*epsilon(dx)) exit
33 end do
34 tt(i) = x
35 ww(i) = 2d0/((1-x**2)*df**2)
36 end do
37 info = 0
38 end subroutine

```

A third alternative is to use ACM TOMS 726. Unfortunately, the tests in TOMS 726 failed which indicates some problem in the Fortran program.

With either of the subroutines gauss or gauss2, it is possible to implement the Gauss-Legendre quadrature:

```

1 function gausslegendre(f,a,b,NN)
2   procedure(func1) :: f
3   double precision, intent(in) :: a, b
4   integer, intent(in) :: NN
5   double precision :: x(NN), w(NN), gausslegendre, s, c
6   integer :: i, info
7   s = (b-a)/2.0
8   c = (b+a)/2.0
9   call gauss(NN,x,w,info) ! Alternative: gauss2
10  if (info .ne. 0) then
11    s = -1.0
12    gausslegendre = sqrt(s) ! return Not a Number
13    return
14  end if
15  gausslegendre = 0.0d0
16  do i=1,NN
17    gausslegendre = gausslegendre + s*w(i)*f(s*x(i)+c)
18  end do
19 end function

```

Now, let us investigate how to calculate numerical integration using Gaussian quadrature.

gauss-lagrange-5

Example 44.21. Use Gauss(-Legendre) quadrature with $n = 2, 3, 4, 5$ to approximate the following integrals:

	(a)	(b)	(c)	(d)	(e)	(f)
Integral	$\int_0^{10} x^3 dx$	$\int_0^1 x^9 dx$	$\int_0^1 \sqrt{x} dx$	$\int_0^{2\pi} \sin^2 90x dx$	$\int_{-2}^2 x^3 e^x dx$	$\int_{-3}^3 h(x) dx$

Solution:

N= 2: Q[0.000,10.000]	x^3=2500.0000000000001 +	0.0000000
N= 3: Q[0.000,10.000]	x^3=2499.9999999999998 +	-0.0000000
N= 4: Q[0.000,10.000]	x^3=2500.0000000000001 +	0.0000000
N= 5: Q[0.000,10.000]	x^3=2499.9999999999999 +	-0.0000000
N= 2: Q[0.000, 1.000]	x^9= 0.0590277777778 +	-0.040972
N= 3: Q[0.000, 1.000]	x^9= 0.0955625000000 +	-0.004438
N= 4: Q[0.000, 1.000]	x^9= 0.099897959184 +	-0.000102
N= 5: Q[0.000, 1.000]	x^9= 0.1000000000000 +	-0.0000000
N= 2: Q[0.000, 1.000]	sqrt(x)= 0.673887338679 +	0.007221
N= 3: Q[0.000, 1.000]	sqrt(x)= 0.669179633899 +	0.002513
N= 4: Q[0.000, 1.000]	sqrt(x)= 0.667827645375 +	0.001161
N= 5: Q[0.000, 1.000]	sqrt(x)= 0.667296789695 +	0.000630
N= 2: Q[0.000, 6.283]	sin^2(90x)= 0.091356227350 +	-3.050236
N= 3: Q[0.000, 6.283]	sin^2(90x)= 2.139958463642 +	-1.001634
N= 4: Q[0.000, 6.283]	sin^2(90x)= 5.904606734878 +	2.763014
N= 5: Q[0.000, 6.283]	sin^2(90x)= 4.407939396298 +	1.266347

N=	2:	Q[-2.000, 2.000]	$x^3e^x=$	8.800114948051 +	-11.120738
N=	3:	Q[-2.000, 2.000]	$x^3e^x=$	18.570704788002 +	-1.350148
N=	4:	Q[-2.000, 2.000]	$x^3e^x=$	19.859938553634 +	-0.060914
N=	5:	Q[-2.000, 2.000]	$x^3e^x=$	19.919469359779 +	-0.001384
N=	2:	Q[-3.000, 3.000]	$h(x)=$	-29.292268174176 +	-39.012075
N=	3:	Q[-3.000, 3.000]	$h(x)=$	-4.582390409550 +	-14.302197
N=	4:	Q[-3.000, 3.000]	$h(x)=$	6.091602913727 +	-3.628204
N=	5:	Q[-3.000, 3.000]	$h(x)=$	-12.416732207298 +	-22.136539

(b) The computation process is illustrated for two cases:

When $n = 2$:

$$\int_0^1 x^9 dx = \frac{1}{2} \int_{-1}^1 \left(\frac{t+1}{2}\right)^9 dt \approx \frac{1}{2} \left[\left(\frac{-\sqrt{\frac{1}{3}}+1}{2}\right)^9 + \left(\frac{\sqrt{\frac{1}{3}}+1}{2}\right)^9 \right] = 0.05902778$$

When $n = 5$:

$$\begin{aligned} \int_0^1 x^9 dx = \frac{1}{2} \int_{-1}^1 \left(\frac{t+1}{2}\right)^9 dt \approx \frac{1}{2} \left\{ \frac{128}{255} \left(\frac{1}{2}\right)^9 + \frac{322+13\sqrt{70}}{900} \left[\left(\frac{-\frac{1}{3}\sqrt{5-2\sqrt{10/7}}+1}{2}\right)^9 \right. \right. \\ \left. \left. + \left(\frac{\frac{1}{3}\sqrt{5-2\sqrt{10/7}}+1}{2}\right)^9 \right] + \frac{322-13\sqrt{70}}{900} \left[\left(\frac{-\frac{1}{3}\sqrt{5+2\sqrt{10/7}}+1}{2}\right)^9 \right. \right. \\ \left. \left. + \left(\frac{\frac{1}{3}\sqrt{5+2\sqrt{10/7}}+1}{2}\right)^9 \right] \right\} = \frac{1}{2} \left[0.000980 + 0.478629(1.8558 \times 10^{-6} + 0.094304) \right. \\ \left. + 0.236927(1.1 \times 10^{-12} + 0.648941) \right] = 0.099935 \approx 0.1 \end{aligned}$$

If we study the above example carefully, we will discover that for $\int_0^{10} x^3 dx$, when $N \geq 2$, the Gaussian quadrature is exact. This is because Gaussian quadrature is exact for any polynomial of ≤ 3 . We observe that $\int_0^1 x^9 dx$ requires 5-point Gaussian quadrature to obtain exact value. However, our hand calculation with just 6 decimal places is insufficient to obtain an exact integral. This implies that double precision (i.e. much more decimal places) is necessary to get an exact integral! However, the error of the Hump function $h(x)$ is very large!

The earlier numerical quadrature methods in Section 44.1 to 44.1.3 seem to be quite good, why do we need to learn Gaussian quadrature? The main reason is because Gaussian quadrature is very useful in the finite element methods (FEM, to be mentioned in Topic 7 and Topic 8). According to Zienkiewicz et al. [2005], in order for FEM to converge, the integral of the volume of the finite element need to be exact. Since the approximation function in FEM is normally a polynomial over a finite element, a properly choose Gaussian quadrature will give exact integral. Besides that, Gaussian quadrature is also employed in the numerical solution of ODEs and PDEs by spectral methods [Trefethen, 2000].

If we are using C++ instead of Fortran, Boost.Math provides Gauss-Legendre quadrature routine. However, C++ is far more complicated compare to Fortran with template programming as illustrated below.

```

1 #include <boost/math/quadrature/gauss.hpp>
2 #include <cmath>
3
4 template<class F>
5 void compute_gauss(F f, double a, double b, const char *dstr, double exact){
6     double qf, err;
7     qf = boost::math::quadrature::gauss<double, 2>::integrate(f,a,b);
8     err = qf - exact;
9     printf("N=%2d: Q[%6.3f,%6.3f] %10s=%17.12f + %12.6f\n",
10         2, a, b, dstr, qf, err);
11     qf = boost::math::quadrature::gauss<double, 3>::integrate(f,a,b);
12     err = qf - exact;

```

```

13     printf("N=%2d: Q[%6.3f,%6.3f] %10s=%17.12f + %12.6f\n",
14           3, a, b, dstr, qf, err);
15     qf = boost::math::quadrature::gauss<double, 4>::integrate(f,a,b);
16     err = qf - exact;
17     printf("N=%2d: Q[%6.3f,%6.3f] %10s=%17.12f + %12.6f\n",
18           4, a, b, dstr, qf, err);
19     qf = boost::math::quadrature::gauss<double, 5>::integrate(f,a,b);
20     err = qf - exact;
21     printf("N=%2d: Q[%6.3f,%6.3f] %10s=%17.12f + %12.6f\n",
22           5, a, b, dstr, qf, err);
23 };
24
25 int main(){
26     auto fa = [](const double x) {return x*x*x;};
27     auto fb = [](const double x) {return x*x*x*x*x*x*x*x*x;};
28     auto fc = [](const double x) {return sqrt(x);};
29     auto fd = [](const double x) {double v=sin(90.*x); return v*v;};
30     auto fe = [](const double x) {return x*x*x*exp(x);};
31     auto h = [](const double x) {
32         return 1/((x-0.3)*(x-0.3)+0.01) + 1/((x-0.9)*(x-0.9)+0.04) - 6;
33     };
34     double exact_solution[] {
35         2500., 1/10., 2/3., M_PI, 2*exp(2) + 38*exp(-2),
36         10*atan(33) + 10*atan(27) + 5*atan(39/2.) + 5*atan(21/2.) - 36
37     };
38     compute_gauss(fa, 0.0, 10., "x^3", exact_solution[0]);
39     compute_gauss(fb, 0.0, 1.0, "x^9", exact_solution[1]);
40     compute_gauss(fc, 0.0, 1.0, "sqrt(x)", exact_solution[2]);
41     compute_gauss(fd, 0.0, 2*M_PI, "sin^2(90x)", exact_solution[3]);
42     compute_gauss(fe, -2.0, 2.0, "x^3e^x", exact_solution[4]);
43     compute_gauss(h, -3.0, 3.0, "h(x)", exact_solution[5]);
44 }

```

Let us practise with Gauss-Legendre quadrature method with another example below.

Example 44.22 (April 2014 Final Exam Question 4(a)). Use an appropriate linear transform and 3-point Gauss-Legendre quadrature

$$\int_{-1}^1 f(x)dx \approx \frac{5}{9}f(-\sqrt{\frac{3}{5}}) + \frac{8}{9}f(0) + \frac{5}{9}f(\sqrt{\frac{3}{5}})$$

to approximate the integral $\int_0^1 e^{-x^2/2}dx$ numerically. (6 marks)

Solution: The linear transform transforms the interval $[0, 1]$ to $[-1, 1]$:

$$0 \leq x \leq 1 \Rightarrow 0 \leq 2x \leq 2 \Rightarrow -1 \leq 2x - 1 \leq 1. \quad [1 \text{ mark}]$$

So let $t = 2x - 1 \Rightarrow x = \frac{t+1}{2}$, then

$$\int_0^1 e^{-x^2/2}dx = \int_{-1}^1 \exp(-(\frac{t+1}{2})^2/2) \frac{1}{2}dt = \frac{1}{2} \int_{-1}^1 \exp(-\frac{(t+1)^2}{8})dt \quad [2 \text{ marks}]$$

Therefore

$$\int_0^1 e^{-x^2/2}dx \approx \frac{1}{2} \left[\frac{5}{9} \exp(-\frac{(-\sqrt{\frac{3}{5}}+1)^2}{8}) + \frac{8}{9} \exp(-\frac{1^2}{8}) + \frac{5}{9} \exp(-\frac{(\sqrt{\frac{3}{5}}+1)^2}{8}) \right] \quad [1 \text{ mark}]$$

$$= \frac{1}{2} [0.55203850 + 0.78444169 + 0.37477260] = 0.85562640 \quad [2 \text{ marks}]$$

Example 44.23 (Tutorial 5, Q4). Use 2-point, 3-point and 4-point Gaussian quadrature to approximate the following integral:

$$I = \int_{-3}^7 8(x+3)^6 + 19(x-2)^3 + 9 dx.$$

You may want to use an interactive numerical computing environment to perform calculation for 4-point quadrature.

Solution. Using calculus, the exact value $I = 80000630/7$.

$$\begin{aligned} I &= \frac{7 - (-3)}{2} \int_{-1}^1 \left[8\left(\frac{10t+4}{2} + 3\right)^6 + 19\left(\frac{10t+4}{2} - 2\right)^3 + 9 \right] dt \\ &= \int_{-1}^1 \underbrace{5 \left[8(5t+5)^6 + 19(5t)^3 + 9 \right]}_f dt \approx Q(f) \end{aligned}$$

When $n = 2$,

$$Q(f) = f\left(-\sqrt{\frac{1}{3}}\right) + f\left(\sqrt{\frac{1}{3}}\right) = 1322.226752 + 9628397.40288 = 9629719.63;$$

$$I - Q(f) = -1798941.80$$

When $n = 3$,

$$\begin{aligned} Q(f) &= \frac{5}{9}f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9}f(0) + \frac{5}{9}f\left(\sqrt{\frac{3}{5}}\right) = \frac{5}{9}(-5392.03) + \frac{8}{9}(625045) + \frac{5}{9}(19525482.03) \\ &= 11400090.00; \end{aligned}$$

$$I - Q(f) = -28571.43$$

When $n = 4$,

$$\begin{aligned} Q(f) &= \frac{18 + \sqrt{30}}{36} (f(-0.33998104) + f(0.33998104)) \\ &\quad + \frac{18 - \sqrt{30}}{36} (f(-0.86113631) + f(0.86113631)) \\ &= 0.65214515(51245.97 + 3618539.77) + 0.34785485(-7533.66 + 25982242.60) \\ &= 11428661.43; \end{aligned}$$

$I - Q(f) \approx 0.025$ should be 0 but weren't 0 because of rounding error

□

Example 44.24 (Tutorial 5, Q5). Use 2-point Gaussian quadrature to find $I = \int_1^\infty \frac{1}{x^2(1+\ln x)} dx$ numerically by applying the transform $\int_a^\infty f(x) dx = \int_0^1 f\left(a + \frac{1-t}{t}\right) \frac{dt}{t^2}$, where a is finite.

Solution.

$$\begin{aligned} I &= \int_0^1 \frac{1}{\left(1 + \frac{1-t}{t}\right)^2 \left(1 + \ln\left(1 + \frac{1-t}{t}\right)\right)} \frac{dt}{t^2} = \int_0^1 \frac{1}{1 + \ln \frac{1}{t}} dt = \int_0^1 \frac{1}{1 - \ln t} dt \\ &= \frac{1}{2} \int_{-1}^1 \frac{1}{1 - \ln\left(\frac{t+1}{2}\right)} dt \approx \frac{1}{2} \left[\frac{1}{1 - \ln\left(\frac{-\sqrt{1/3+1}}{2}\right)} + \frac{1}{1 - \ln\left(\frac{\sqrt{1/3+1}}{2}\right)} \right] \\ &= \frac{0.39148770 + 0.80814560}{2} = 0.59981665 \end{aligned}$$

□

Gauss' work on Gaussian quadrature (44.18) was later generalised to integrals with integrable weight function $w(x)$ by many mathematicians:

$$I_w[a, b]f := \int_a^b w(x)f(x)dx \quad (44.21)$$

e:quadpack:p19

over the finite or infinite interval $[a, b] \subset \overline{\mathbb{R}}$. A *quadrature sum* yields an approximation [Piessens et al., 1983]

$$Q_n[a, b]f := \sum_{k=1}^n w_k f(x_k) \approx I_w[a, b]f. \quad (44.22)$$

e:quadpack:p19

The *integration error* is defined as $E[a, b]f := Q_n[a, b]f - I_w[a, b]f$.

For a given weight function $w(x) \geq 0$, integrable on $[a, b]$, and for which all moments given by $\mu_i = \int_a^b w(x)x^i dx$, $i \geq 0$ exists, and $\mu_0 > 0$, it is possible to define a sequence of polynomials $\{\omega_n(x)\}$, $n = 0, 1, 2, \dots$ where $\omega_n(x)$ is of exact degree n and which satisfies

$$\int_a^b w(x)\omega_m(x)\omega_n(x)dx = 0, \quad m \neq n. \quad (44.23)$$

This sequence of polynomials is called *orthogonal*. It can be show that the zeroes of $\omega_n(x)$, $n \geq 1$ are real simple and located in (a, b) .

The connection between Gaussian integration rules and orthogonal polynomials is described by the following theorem [Piessens et al., 1983, Theorem 2.1].

Theorem 44.25. Let $w(x) \geq 0$ be a weight function over $[a, b]$ for which a sequence of orthogonal polynomials $\{\omega_n(x)\}$ exists. if the nodes $x_1, \dots, x_n$ in (44.22) are the zeros of $\omega_n(x)$, the weights $w_1, \dots, w_n$ can be determined such that formula (44.22) yields **exact integration** for polynomials of degree $2n - 1$ or less! These weights are real and positive.

There are many slightly different definitions of the classical orthogonal polynomials in the literature. According to Abramowitz and Stegun [1964], the orthogonal polynomials and Gaussian quadrature formulae corresponding to the following special selection $[a, b]$ and $w(x)$ which are the *classical polynomials* and *classical Gaussian quadrature formulae* as listed below (from https://en.wikipedia.org/wiki/Gaussian_quadrature).

a	b	$w(x)$	classical orthogonal polynomials	quadrature
-1	1	1	Legendre polynomials $P_n(x)$	Gauss-Legendre
-1	1	$\frac{1}{\sqrt{1-x^2}}$	Chebyshev polynomials of the first kind $T_n(x)$	Gauss-Chebyshev quadrature of the first kind.
-1	1	$\sqrt{1-x^2}$	Chebyshev polynomials of the second kind $U_n(x)$	Gauss-Chebyshev quadrature of the second kind.
-1	1	$(1-x)^\alpha(1+x)^\beta$, $\alpha, \beta > -1$	Jacobi polynomials $P_n^{(\alpha, \beta)}(x)$	Gauss-Jacobi
0	∞	e^{-x}	Laguerre polynomials $L_n(x)$	Gauss-Laguerre
0	∞	$x^\alpha e^{-x}$, $\alpha > -1$	Generalised Laguerre polynomials $L_n^{(\alpha)}(x)$	Generalised Gauss-Laguerre
$-\infty$	∞	e^{-x^2}	Hermite polynomials $H_n(x)$	Gauss-Hermite

How do we know that Gaussian quadrature works? In other words, when n is large, is the approximation in (44.1) accurate? The following theorem provides a confirmation.

Theorem 44.26. The sequence of n -point Gauss rules satisfies

$$\lim_{n \rightarrow \infty} Q_n[a, b]f = I_w[a, b]f. \quad (44.24)$$

whenever f is bounded and Riemann integrable on $[a, b]$.

If $f \in C^{2n}$, the error estimate can be expressed as

$$I_w[a, b]f - Q_n[a, b]f = \frac{f^{(2n)}(\xi)}{(2n)!} \langle \omega_n, \omega_n \rangle$$

where ω_n is the degree n orthogonal polynomial with respect to the weight function $w(x)$.

Let us investigate some of the Gaussian quadrature formulae mentioned in the table above. First, let us begin with Gauss-Laguerre quadrature.

thm:gauss-laguerre

Theorem 44.27. When $w(x) = e^{-x}$, we obtain the Gauss-Laguerre Quadrature:

$$\int_0^\infty e^{-x} f(x) dx \approx \sum_{i=1}^n w_i f(x_i), \quad w_i = \frac{1}{x_i (\ell'_n(x_i))^2}$$

where x_i are the roots of n -th Laguerre polynomial $\ell_n(x) = \frac{1}{n!} e^x \frac{d^n}{dx^n} (e^{-x} x^n) = \sum_{k=0}^n \binom{n}{k} \frac{1}{k!} (-x)^k$.

The following example explains the calculation of 2-point formula.

Example 44.28. Derive the 2-point Gauss-Laguerre quadrature rule.

Solution: From Theorem [thm:gauss-laguerre](#) 44.27, $x_i, i = 1, 2$ are the solutions of

$$\ell_2(x) = \frac{1}{2!} e^x (e^{-x} x^2)'' = 0.5(x^2 - 4x + 2) = 0 \Rightarrow \begin{cases} x_1 = \frac{4 + \sqrt{4^2 - 4(2)}}{2} = 2 + \sqrt{2} \\ x_2 = \frac{4 - \sqrt{4^2 - 4(2)}}{2} = 2 - \sqrt{2} \end{cases}$$

and the corresponding w_i are

$$w_1 = \frac{1}{x_1 (\ell'_2(x_1))^2} = \frac{1}{x_1 (x_1 - 2)^2} = \frac{2 - \sqrt{2}}{4}, \quad w_2 = \frac{1}{x_2 (x_2 - 2)^2} = \frac{2 + \sqrt{2}}{4}.$$

In <http://mathworld.wolfram.com/Laguerre-GaussQuadrature.html>, we will find the values of x_i and w_i for $n = 2, 3, 4$ and 5-point formula. The following example demonstrates how to apply Gauss-Laguerre formula when the interval is not $[0, \infty)$.

am2e:p151:15.38

Example 44.29. Apply the 6-point Gauss-Laguerre method to approximate $\int_1^\infty (e^{-t}/t) dt$ where the x_i and w_i could be obtained from an advanced mathematical handbook:

x_k	0.22284660	1.18893210	2.99273633	5.77514357	9.83746742	15.98287398
w_k	0.45896467	0.41700083	0.11337338	0.01039920	0.00026102	0.00000090

Solution:
$$\int_1^\infty (e^{-t}/t) dt \stackrel{t=x+1}{=} \int_0^\infty \frac{e^{-(x+1)}}{x+1} dx = \int_0^\infty e^{-x} \left[\frac{1}{e(x+1)} \right] dx$$

$$\approx w_1 f(x_1) + \cdots + w_6 f(x_6) \approx 0.21918.$$

Example 44.30 (Tutorial 5, Q6). Use 3-point Gauss-Laguerre quadrature

$$\int_0^\infty e^{-x} f(x) dx \approx 0.711093 f(0.415775) + 0.278518 f(2.29428) + 0.0103893 f(6.28995)$$

to estimate (a) $\Gamma(1.8)$, (b) $\int_1^\infty e^{-x} x^2 dx$, (c) $\Gamma(5.0)$ where $\Gamma(\alpha) = \int_0^\infty e^{-z} z^{\alpha-1} dz$. Are the estimates exact?

Solution. (a) $\Gamma(1.8) = \int_0^\infty e^{-z} z^{0.8} dz \approx 0.35238150 + 0.54121572 + 0.04523807 = 0.93883529$
 (b) By a suitable transform $x \rightarrow t + 1$, we have

$$\begin{aligned} \int_1^\infty e^{-x} x^2 dx &= \int_0^\infty e^{-(t+1)} (t+1)^2 dt = e^{-1} \int_0^\infty e^{-t} (t+1)^2 dt \\ &= e^{-1} (1.42532821 + 3.02255552 + 0.55212242) = 5.00000615 e^{-1} \end{aligned}$$

$$(c) \Gamma(5) = \int_0^\infty e^{-z} z^4 dz \approx 0.02125005 + 7.7168302 + 16.262041 = 24.000121 \quad \square$$

Now, let us study another example of Gaussian quadrature, the Gauss-Hermite quadrature.

Theorem 44.31. When $w(x) = e^{-x^2}$, we obtain the Gauss-Hermite Quadrature:

$$\int_{-\infty}^{\infty} e^{-x^2} f(x) dx \approx \sum_{i=1}^n w_i f(x_i), \quad w_i = \frac{2^{n+1} n! \sqrt{\pi}}{[H'_n(x_i)]^2}$$

where x_i are the zeroes of the n th Hermite polynomial $H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} (e^{-x^2})$ [Wikipedia, 2015.02.28].

Example 44.32. Use the most appropriate Gaussian quadrature to find the **exact** integral for $\int_{-\infty}^{\infty} e^{-x^2} x^2 dx$.

Solution:

1. There is a weight function $w(x) = e^{-x^2}$ and the range is $(-\infty, \infty)$. Therefore, Gauss-Hermite quadrature should be used.
2. x^2 is a degree 2 polynomial. We know that a 2-point Gauss-Hermite can find the exact integral for polynomials of degree 3 or less.
3. According to Theorem 44.31, the second Hermite polynomial is

$$H_2(x) = e^{x^2} \frac{d^2}{dx^2} (e^{-x^2}) = 4x^2 - 2.$$

The zeros of H_2 are $x_k = \pm\sqrt{2}/2$.

4. The weights are

$$w_i = \frac{2^{2+1} 2! \sqrt{\pi}}{(8x_i)^2} = \frac{\sqrt{\pi}}{2}.$$

5. Therefore

$$\int_{-\infty}^{\infty} e^{-x^2} x^2 dx = \frac{\sqrt{\pi}}{2} \left(\frac{\sqrt{2}}{2}\right)^2 + \frac{\sqrt{\pi}}{2} \left(-\frac{\sqrt{2}}{2}\right)^2 = \frac{\sqrt{\pi}}{2}.$$

Example 44.33 (Tutorial 5, Q7). If X is a random variable which follows the normal distribution $N(\mu = 1, \sigma^2 = 2)$. Find the expected value

$$\mathbb{E}[(X + 3)^2] = \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) (x + 3)^2 dx$$

exactly using the Gauss-Hermite quadrature. [**Hint:** Use an appropriate transform $x = \sqrt{2}\sigma t + \mu$ and apply the results from Example 5.2.11]

Solution.

$$\begin{aligned} \mathbb{E}[(X + 3)^2] &= \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-t^2} (2t + 1 + 3)^2 (\sqrt{2}\sigma dt) = \int_{-\infty}^{+\infty} e^{-t^2} \left[\frac{(2t + 4)^2}{\sqrt{\pi}} \right] dt \\ &= \frac{\sqrt{\pi}}{2} \left[\frac{(2(-\frac{\sqrt{2}}{2}) + 4)^2}{\sqrt{\pi}} \right] + \frac{\sqrt{\pi}}{2} \left[\frac{(2(\frac{\sqrt{2}}{2}) + 4)^2}{\sqrt{\pi}} \right] \\ &= \frac{(-\sqrt{2} + 4)^2}{2} + \frac{(\sqrt{2} + 4)^2}{2} = \frac{2 - 8\sqrt{2} + 16 + 2 + 8\sqrt{2} + 16}{2} = \frac{18 + 18}{2} = 18. \end{aligned}$$

Example 44.34. Consider the following integrals from [Press et al., 2007, Chapter 4]:

$$\int_0^\infty x^{-3/2} \sin \frac{x}{2} e^{-x} dx = \sqrt{\pi(\sqrt{5}-2)}, \quad \int_0^\infty x^{-2/7} e^{-x^2} dx = \frac{1}{2} \Gamma\left(\frac{5}{14}\right)$$

Use suitable software tool and appropriate Gauss quadrature of order 2 to approximate them.

Solution: Let $f(x) = x^{-3/2} \sin \frac{x}{2}$ and $g(x) = |x|^{-2/7}$. Note that the left integral has the function e^{-x} , so Gauss-Laguerre quadrature would be appropriate. As for the right integral, we need to rewrite it so that Gauss-Hermite quadrature could be used:

$$\int_0^\infty x^{-2/7} e^{-x^2} dx = \frac{1}{2} \int_{-\infty}^\infty |x|^{-2/7} e^{-x^2} dx = \frac{1}{4} \Gamma\left(\frac{5}{14}\right)$$

Python will be a good choice (Fortran is also OK):

```

1 from math import gamma, sin, pi, sqrt
2 def f(x): return x**(-3.0/2)*sin(0.5*x)
3 x1 = 2+sqrt(2); x2 = 2-sqrt(2)
4 w1 = (2-sqrt(2))/4; w2 = (2+sqrt(2))/4
5 Q = w1*f(x1) + w2*f(x2)
6 print("Q[f]=", Q, "; Error=", Q-sqrt(pi*(sqrt(5)-2)), sep="")
7 def g(x): return abs(x)**(-2.0/7)
8 w = sqrt(pi)/2; x=sqrt(2)/2
9 Q = w*g(x) + w*g(-x)
10 print("Q[g]=", Q, "; Error=", Q-0.25*gamma(5/14.0), sep="")

```

and the output is

```

1 Q[f]=0.5726704954536302; Error=-0.2885085938542443
2 Q[g]=1.9569477102555468; Error=1.333632042778516

```

So far, we have only explored Gauss-Legendre, Gauss-Laguerre, Gauss-Hermite, what about other situations? According to Press et al. [2007, Chapter 4] (and Wikipedia:Gaussian_quadrature), the $\omega_n(x)$ in Theorem 44.25 can be constructed by the recurrence relation

$$\omega_{k+1}(x) = (x - a_k)\omega_k(x) - b_k\omega_{k-1}(x), \quad k = 0, 1, 2, \dots$$

where $p_{-1}(x) = 0$, $p_0(x) = 1$ and

$$a_j = \frac{\int_a^b w(x)x(\omega_j(x))^2 dx}{\int_a^b w(x)(\omega_j(x))^2 dx}, \quad j = 0, 1, 2, \dots; \quad b_k = \frac{\int_a^b w(x)(\omega_k(x))^2 dx}{\int_a^b w(x)(\omega_{k-1}(x))^2 dx}, \quad k = 1, 2, 3, \dots \quad (44.25)$$

The computation of Gaussian quadrature rules involves two steps: (a) the generation of orthogonal polynomials $\omega_0, \dots, \omega_n$ for finding a_j and b_k using (44.25); and (b) the determination of the zeros of $\omega_n(x)$ and the computation of the associated weights (by Golub-Welsch algorithm). For classical orthogonal polynomials, a_j and b_k are known, we only need to deal with step (b). For non-classical weight function $w(x)$, the construction of the associated set of orthogonal polynomials is not trivial and is discussed in Press et al. [2007, Section 4.6.3].

There are other variations and generalisations of Gaussian quadrature rules such as Gauss-Kronrod quadrature formula, Gauss-Lobatto quadrature formula, Gauss-Radau quadrature formula, etc. In these methods, some nodes are fixed and one needs to choose weights and extra abscissas to maximise the degree of exactness of the quadrature rule [Press et al., 2007, Section 4.6.4].

44.3 Adaptive Quadrature

Another class of unevenly-spaced quadrature method is based on the adaptive principle. A function may vary largely on some region and may vary little in other region. If the approximation

error is to be evenly distributed, a smaller step size is needed for the large-variation regions than for those with less variation. An efficient technique for this type of problem should predict the amount of functional variation and adapt the step size to the varying requirements. These methods are called *adaptive quadrature methods*.

Adaptive quadrature is a process in which the integrand f is approximated using static *quadrature rules* (either by basic quadrature or by Gaussian quadrature) on adaptively refined subintervals of the integration domain. Generally, adaptive algorithms are as efficient as traditional algorithms (Sections [44.1.2](#) and [44.1.3](#)) for “well-behaved” integrands, but are also effective for “badly behaved” integrands for which traditional algorithms fail.

Most adaptive quadrature algorithms are very complicated. In Burden and Faires [2015, Section 4.6], only the *adaptive Simpson’s method/rule* is introduced. It is a method of numerical integration proposed by G. F. Kuncir in 1962. It is probably the first recursive adaptive algorithm for numerical integration to appear in print, although more modern adaptive methods based on *Gauss-Kronrod quadrature* (which is a generation of Gaussian quadrature, which is used in QUADPACK) and *Clenshaw-Curtis quadrature* (which is based on fast Fourier transform and is used in MATLAB) are now generally preferred.

Adaptive Simpson’s method uses an estimate of the error we get from calculating a definite integral using Simpson’s rule. If the error exceeds a user-specified tolerance, the algorithm calls for subdividing the interval of integration in two and applying adaptive Simpson’s method to each subinterval in a recursive manner. The technique is usually much more efficient than composite Simpson’s rule since it uses fewer function evaluations in places where the function is well-approximated by a cubic function.

In the adaptive Simpson’s rule, when trapezoidal approximation differ by more than 15ε , we apply the Simpson’s rule technique individually to the subintervals $[a, \frac{a+b}{2}]$ and $[\frac{a+b}{2}, b]$. Then we use the error estimation procedure to determine if the approximation to the integral on each subinterval is within a tolerance of $\varepsilon/2$. If so, we sum the approximations to produce and approximation to $\int_a^b f(x)dx$ within the tolerance ε . If the approximation on one of the subintervals fails to be within the tolerance $\varepsilon/2$, then that subinterval is itself subdivided, and the procedure is reapplied to the two subintervals to determine if the approximation on each subinterval is accurate to within $\varepsilon/4$. This halving procedure is continued until each portion is within the required tolerance. The implementation in Fortran is shown below.

```

1  !!! http://en.wikipedia.org/wiki/Adaptive_quadrature
2  !!! http://en.wikipedia.org/wiki/Adaptive_Simpson%27s_method
3  function adaptiveSimpsons(f, a, b, tol, maxRecurDepth)
4      procedure(func1) :: f
5      intent(in) :: a, b
6      optional :: tol, maxRecurDepth
7      double precision :: a,b,tol,loc_tol,c,h,fa,fb,fc,S,adaptiveSimpsons
8      integer :: maxrec, maxRecurDepth
9      c = (a + b)/2d0
10     h = b - a
11     fa = f(a); fb = f(b); fc = f(c)
12     S = (h/6d0)*(fa + 4d0*fc + fb)    ! Basic Simpson rule
13     loc_tol = 1d-6
14     if (present(tol)) loc_tol = tol
15     maxrec = 10
16     if (present(maxRecurDepth)) maxrec = maxRecurDepth
17     adaptiveSimpsons = adaptSimpAux(f,a,b,loc_tol,S,fa,fb,fc,maxrec)
18 end function

```

The auxiliary function is implemented using recursion technique.

```

1  recursive function adaptSimpAux(f,a,b,tol,S,fa,fb,fc,bot) result(ret)
2      procedure(func1) :: f
3      double precision, intent(in) :: a, b, tol, S, fa, fb, fc
4      integer, intent(in) :: bot
5      double precision :: ret, c, h, d, e, fd, fe, S1, Sr, S2, left, righ
6      c = (a + b)/2d0
7      h = b - a
8      d = (a + c)/2d0

```

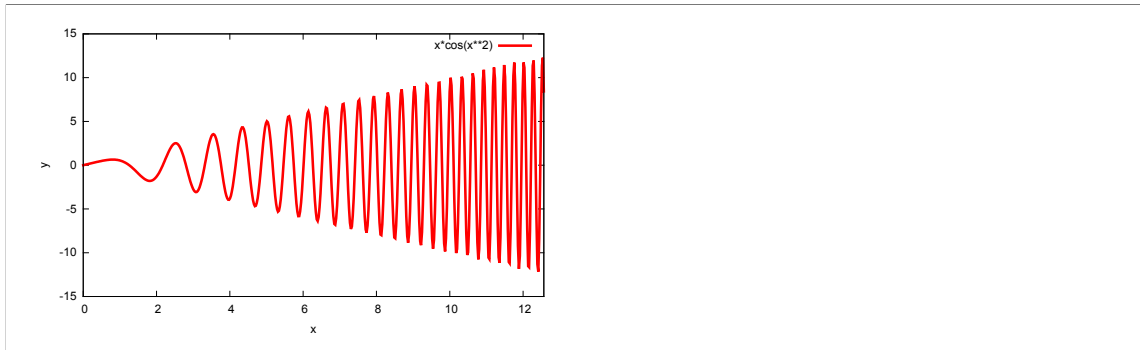
```

9      e = (c + b)/2d0
10     fd = f(d); fe = f(e)
11     S1 = (h/12d0)*(fa + 4d0*fd + fc)
12     Sr = (h/12d0)*(fc + 4d0*fe + fb)
13     S2 = S1 + Sr
14     if (bot <= 0 .or. dabs(S2 - S) <= 15d0*tol) then
15         ret = S2 + (S2 - S)/15d0
16         return
17     end if
18     left = adaptSimpAux(f, a, c, tol/2d0, S1, fa, fc, fd, bot-1)
19     righ = adaptSimpAux(f, c, b, tol/2d0, Sr, fc, fb, fe, bot-1)
20     ret = left + righ
21 end function

```

Example 44.35. Given a function $f(x) = x \cos(x^2)$.

1. Sketch the function f for the interval $[0, 4\pi]$.



2. Find the exact value of $I = \int_0^{4\pi} f(x)dx$.

Solution:
$$\int_0^{4\pi} f(x)dx = \frac{1}{2} \left[\sin(x^2) \right]_0^{4\pi} = \frac{1}{2} \sin(16\pi^2) \approx 0.3703 \ 2515 \ 9903 \ 29495$$

3. Estimate the integral I in 2. using the following quadrature and estimate each error:
 - (a) Simpson's rule with $n = 2^{15}$;
 - (b) Romberg method with 15 stages;
 - (c) Adaptive Simpson's method `adaptiveSimpsons` with an accuracy of 10^{-9} and at most 15 iteration

Solution: We can obtain the numerical solutions by using a Fortran program below.

```

program eg_adaptive2
  use iso_c_binding, only : c_double
  use textbook_quad, only : simpInt, romberg, adaptiveSimpsons, pi
  implicit none
  integer, parameter :: order=15
  real(c_double) :: a, b, exact, est1, est2, est3, R(order,order)
  real(c_double), external :: f
  a = 0d0; b = 4*pi
  exact = 0.5d0*sin(16d0*pi**2)
  est1 = simpInt(f, a, b, 2**15)
  est2 = romberg(f, a, b, order, R)
  est3 = adaptiveSimpsons(f, a, b, 1d-9, 15)
  print *, "Simpson Est. =", est1, " + ", (exact-est1)
  print *, "Romberg Est. =", est2, " + ", (exact-est2)
  print *, "Adaptive Est. =", est3, " + ", (exact-est3)
end program eg_adaptive2

function f(x)
  use iso_c_binding, only : c_double
  real(c_double) :: x, f
  f = x*cos(x**2)

```

```
end function
```

- (a) Simpson Est. = 0.3703251599207422, error = $-1.744721 \times 10^{-11}$
- (b) Romberg Est. = 0.3703251599032947, error = 2.220446×10^{-16}
- (c) Adaptive Simpson Est. = 0.3703251599032902, error = 4.773959×10^{-15}

With C++, we can use `boost/math/quadrature/trapezoidal.hpp`.

From the above example, it seems that Romberg quadrature is comparable to adaptive Simpson quadrature. Why would need adaptive quadrature since it is so complicated? Because they are usually faster, more accurate and more robust. Is accuracy important? It depends on the field of study.

In economics and finance, expected utility, discounted utility and profits over a long horizon, Bayesian posterior are integrals that require relatively rough estimate because the economics and finance data contain errors.

In engineering, engineers probably just need rough estimates based on the engineering tolerance. In science, some integrals are singular and accurate quadrature formula can help to uncover the properties of the integrals.

In physics, there are many “oscillatory” and “singular” integrals such as

$$\int_a^b f(x)e^{-ikg(x)}dx \int_0^\infty e^{-x} \frac{\sin(xy)}{xy} dx$$

where $y \rightarrow 0$, $f \in L^1[a, b]$, $k > 0$. Near singularity, many of the quadrature methods mentioned earlier will work poorly or even fail, adaptive methods or Filon quadrature may perform better.

In mathematics, Borwein and Borwein [2001, p86] found some interesting integrals

$$\begin{aligned} \int_0^\infty \frac{\sin x}{x} dx &= \frac{\pi}{2} \\ \int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin(x/3)}{x/3} dx &= \frac{\pi}{2} \\ &\dots \\ \int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin(x/3)}{x/3} \dots \frac{\sin(x/13)}{x/13} dx &= \frac{\pi}{2} \end{aligned}$$

but

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin(x/3)}{x/3} \dots \frac{\sin(x/13)}{x/13} \cdot \frac{\sin(x/15)}{x/15} dx = \frac{467807924713440738696537864469}{935615849440640907310521750000} \pi \neq \frac{\pi}{2}$$

which says that to experiment with mathematical theory, very accurate numerical integration is needed!

44.4 Quadrature Software

The computer function for quadrature in most open source numerical software (for example, SciPy, Octave and R) relies on the QUADPACK package (available from <http://www.netlib.org/quadpack/>). QUADPACK is a collection of Fortran subroutines of Gaussian quadrature with naming conventions below.

1st letter	2nd letter	3rd letter	4th letter ¹
Q: Quadra- ture	N: Non- adaptive integrator	G: General purpose user defined integrand	S: Singularities can be more readily in- tegrated P: Points of special difficult (like singu- larities) can be input. I: Infinite interval of integration
	A: Adaptive integrator	W: Weight function, e.g. certain explicit integrand forms	O: Oscillatory trigonometric weight functions $\cos(\omega x)$ and $\sin(\omega x)$ F: Fourier intergrals
			C: Cauchy principle value intergrals

The two general-purpose subroutines most suitable for use without further analysis of the integrand are QAGS for integration over a finite interval and QAGI for integration over an infinite interval. These two routines are used in

- Python: `scipy.integrate`
- Octave:
 - `quad`: Numerical integration based on Gaussian quadrature.
 - `quadv`: Numerical integration using an adaptive vectorized Simpson's rule.
 - `quadl`: Numerical integration using an adaptive Lobatto rule.
 - `quadgk`: Numerical integration using an adaptive Gauss-Konrod rule.
 - `quadcc`: Numerical integration using adaptive Clenshaw-Curtis rules.
- R:

```
integrate(f, lower, upper, ..., subdivisions = 100L,
          rel.tol = .Machine$double.eps^0.25, abs.tol = rel.tol,
          stop.on.error = TRUE, keep.xy = FALSE, aux = NULL)
```

QAGS uses global adaptive quadrature based on 21-point Gauss-Kronrod quadrature within each subinterval, with acceleration by Peter Wynn's ϵ -algorithm.

QAGI is the only general-purpose routine for infinite intervals, and maps the infinite interval onto (0,1) using a transformation then uses the same approach as QAGS, except with 15-point rather than 21-point Gauss-Kronrod quadrature. For an integral over the whole real line, the transformation used is $x = (1 - t)/t$:

$$\int_{-\infty}^{+\infty} f(x)dx = \int_0^1 \frac{dt}{t^2} \left(f\left(\frac{1-t}{t}\right) + f\left(-\frac{1-t}{t}\right) \right).$$

This is not be the best approach for all integrands: another transformation may be appropriate, or one might prefer to break up the original interval and use QAGI only on the infinite part.

Example 44.36. Integrate $x \cos(x^2)$ for the interval $[0, 4\pi]$ using standard subroutine from QUADPACK.

Solution: Since Python, R, Octave uses QUADPACK, we just use these numerical computing environment.

In Python:

```
1 from scipy.integrate import quad
2 from math import cos, pi
3 def f(x): return x*cos(x*x) # LiewHH 2014
4 print("I =", quad(f,0,4*pi,relabs=1e-12))
```

¹indicates the type of integrals or integrands the routine is especially designed for.

Output: I = (0.3703251599034056, 7.040423881161537e-10)

In R:

```
1 options(digits=16)    # LiewHH 2014
2 f = function(x) { x*cos(x*x) }
3 res = integrate(f,0,4*pi,rel.tol=1e-12)
4 cat("I =", res$value)
5 cat("; abs.err = ", res$abs.error, "\n")
```

Output: I = 0.3703251599033918; abs.err = 5.866907081416425e-13

In Octave:

```
1 function res = f(x)      % How different will MATLAB be?
2     res = x .* cos(x.*x); % x is a vector --- LiewHH 2014
3 endfunction
4 [q, ier, nfun, err] = quad("f",0,4*pi,relabs=1e-12);
5 format long
6 disp(["I = ", num2str(q, "%15.12f")])
7 disp(["err = ", num2str(err)])
```

Output: I = 0.370325159903 err = 7.0404e-10

We have not discussed anything on cubature because it is not in the syllabus, however Gaussian cubature is useful in the finite element analysis of 2- and 3-dimensional system. Another useful technique for estimating numerical integration is the Monte Carlo method, which will be discussed in *UECM2423 Statistical Simulation for Actuary and Finance*.

44.5 Hard Integration Problems

The adaptive Romberg quadrature and Gaussian quadrature cannot handle all possible problems in numerical quadrature. In this section we shall list some of the difficulties they encounter. Deuffhard and Hohmann [2003]

- **Discontinuous Integrands:** A common problem in numerical quadrature are discontinuities of the integrand f or its derivatives. Such integrands occur when a physical or engineering system is described by different models in different regions, which do not quite fit at the interface positions. If the jumps are known, then one should subdivide the integration interval at these positions and solve the arising subproblems separately. Otherwise the quadrature program reacts quite differently. Without any further preparation, a non-adaptive quadrature program yields incorrect results or does not converge. The jumps cannot be localised. An adaptive quadrature program, such as the adaptive Romberg method, freezes at the jumps. Thus the jumps can be localized and treated separately.
- **Impulses:** In principle, every quadrature program will fail if a function has small enough peaks. On the other hand, such integrands are pretty common: just think of the spectrum of a star whose entire radiation is to be computed. If the positions of the peaks are known, then one should subdivide the interval in a suitable way, and again compute the sub-integrals separately. Otherwise, there only remains the hope that the adaptive quadrature program does not “overlook” them.
- **Highly Oscillatory Integrands:** They are ill-conditioned from the relative error viewpoint. For example

$$f(t) = \cos(te^{4t^2}), \quad t \in [-1, 1].$$

The numerical quadrature is powerless against such integrands. They have to be prepared by analytical averaging over subintervals (pre-clarification of the structure of the inflection points of the integrand). A concrete example encountered in computational finance are given in Section [sec:svld](#) ??.

- **Weakly Singular Integrands:** A function f , which is integrable over the interval $[a, b]$, is called *weakly singular*, if one of its derivatives $f^{(k)}$ in $[a, b]$ does not exist. For example, the

functions $f(t) = t^\alpha g(t)$, where $g \in C^\infty[0, T]$ is an arbitrarily smooth function and $\alpha > -1$. When $\alpha = 1/2$ and $g(t) = \cos t$, $f'(t)$ has a pole at 0.

In the case of weakly singular integrands, adaptive quadrature programs usually tend to contract step size and order, and they therefore tend to be extremely slow. Nonadaptive quadrature algorithms, however, do not get slow, but usually false. The singularities can often be removed via a substitution.

- **Cubature:** It is complicated because the domain Ω can be very complicated as opposed to quadrature where the domain is just an interval. For finite regular domain, it may be possible to decompose the integral to “iterated integral” form, e.g.

$$\iint_{[0,1] \times [1,2]} \frac{1}{(2x+3y)^2} dA = \int_0^1 \int_{y=y_1(x)}^{y=y_2(x)} \frac{1}{(2x+3y)^2} dy dx, \quad y_1(x) = 1, \quad y_2(x) = 2$$

can be reduced to quadrature (using Python) as follows:

```

1 from scipy import integrate, log
2 def f(x,y): return 1.0/(2*x+3*y)**2 # (x,y) \in [0,1] x [1,2]
3 def g(x):
4     def y2(x): return 2.0
5     def y1(x): return 1.0
6     return integrate.quad(lambda y: f(x,y), y1(x), y2(x))[0]
7 print("Exact=", log(5/4.0)/6, " vs Cubature=", integrate.quad(g,0,1)[0])

```

The output is as follows:

```
Exact= 0.037190591885701625 vs Cubature= 0.03719059188570162
```

For a two-dimensional problem, we can see the computer code to be complex. For high-dimensional problem or a domain Ω which is large, Monte-Carlo method may be a good alternative.

- **Parameter- Dependent Integrands:** Often the integrand $f(x, \lambda)$ depends on a parameter $\lambda \in \mathbb{R}$. We thus have to solve an entire family of problems

$$I(\lambda) := \int_a^b f(x, \lambda) dx.$$

The most important class of examples for such parameter-dependent integrals is the multi-dimensional quadrature. Usually the integrand is differentiable with respect to λ , and so is therefore the integral $I(\lambda)$. Of course, one hopes that the approximation $\hat{I}(\lambda)$ inherits this property. Unfortunately, however, it turns out that just our best methods, the adaptive quadrature methods, do not have this property—in contrast to the simple non-adaptive quadrature formulas. There are essentially three possibilities to rescue the adaptive approach for parameter-dependent problems.

The first possibility consists of carrying out the quadrature for one parameter value, storing away the employed orders and step sizes, and use them again for all other parameter values. This is also called freezing of orders and step sizes. This can only be successful if the integrand qualitatively does not change too much in dependence of the parameter.

If, however, a peak varies with the parameter, and if this dependence is known, then one can employ parameter-dependent grids. One transforms the integral in dependence of λ in such a way that the integrand stays the same qualitatively (the movement of the peak is, e.g., counter-balanced) or, in dependence of A , one shifts the adaptive partitioning of the integration interval.

The last possibility requires a lot of insight into the respective problem. We choose a fixed grid adapted to the respective problem and integrate over this grid with a fixed quadrature formula (Newton-Cotes or Gauss). In order to do this, the qualitative properties of the integrand need to be largely known, of course.

Chapter 45

Non-stiff ODE-IVPs

ch:ode-ivp

In this topic, the numerical solution of the *initial value problem (IVP)* of *systems of ordinary differential equations (ODEs)* which are mostly “non-stiff” and “dissipative” will be investigated:

$$\begin{aligned}\dot{x}_i(t) &= f_i(t, x_1, \dots, x_n, \mu_1, \dots, \mu_\ell), \\ x_i(t_0) &= x_{i0}.\end{aligned}\tag{45.1}$$

e:ode-ivp0

where $i = 1, \dots, n$ and $(t, x_1, \dots, x_n) \in \Omega \subset \mathbb{R}^{n+1}$ and $\mu_1, \dots, \mu_\ell$ are *parameters (constants)* of the system. In vector notation (introduced by Giuseppe Peano in 1890), such IVP has the form

$$\dot{\mathbf{x}} = \mathbf{f}_\mu(t, \mathbf{x}), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \in \mathbb{R}^n.\tag{45.2}$$

e:ode-ivp

where $(t, \mathbf{x}) = (t, x_1, \dots, x_n)$ and $\mathbf{f}_\mu : \Omega \rightarrow \mathbb{R}^n$.

In engineering (e.g. process simulation [Ramirez, 1998]), the variable t in Equation (45.2) is a physical time where t_0 (normally it is set as zero) is the starting time so that $t_0 \leq t < +\infty$ and the vector $\mathbf{x}_0$ is the initial state of the system. The vector $\mathbf{x}$ normally characterises the *state* of the system and is usually called a *state vector*. In mathematical physics, for the ODE-BVP in Topic 7, t represents a *space variable* and is normally written as x instead of t and $\mathbf{y}$ is used in place of $\mathbf{x}$. When the system (45.2) is in *equilibrium*, $\dot{\mathbf{x}} = \mathbf{0}$, the determination of equilibrium state is reduced to a nonlinear equation problem in Topic 3.

The IVP (45.2) describes a function $\mathbf{x}(t)$ by its “time derivative” $\dot{\mathbf{x}}(t)$, so the solution process may be interpreted as an *integration*. Therefore, the numerical solution of IVP is also called *numerical integration of IVP*. The numerical integration methods can be classified into

$$(explicit) \text{ one-step method: } \quad \mathbf{x}_{i+1} = \Phi_h(t_i, \mathbf{x}_i),\tag{45.3}$$

e:one-step

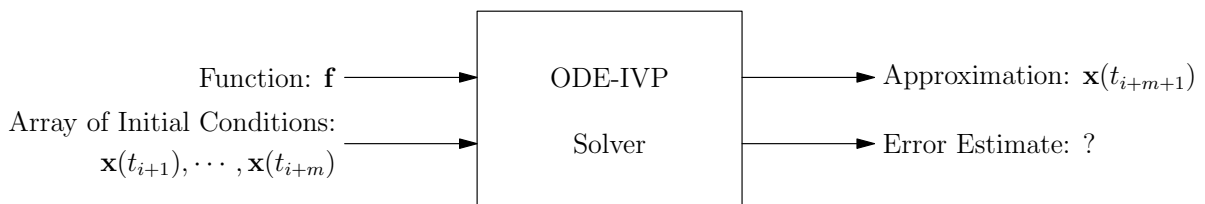
$$implicit \text{ one-step method: } \quad \mathbf{x}_{i+1} = \Phi_h(t_i, \mathbf{x}_{i+1}, \mathbf{x}_i),\tag{45.4}$$

e:one-step:i

$$multi\text{-}step \text{ method (multi-point iteration): } \quad \mathbf{x}_{i+1} = \Phi_h(t_i, \mathbf{x}_{i+1}, \mathbf{x}_i, \mathbf{x}_{i-1}, \dots, \mathbf{x}_{i+1-m}),\tag{45.5}$$

e:multi-step

Explicit Runge-Kutta methods, to be studied in Section 45.3, are popular explicit one-step methods for handling non-stiff problems. The Adams multi-step methods, to be studied in Section 45.5, are suitable for non-stiff systems. Implicit Runge-Kutta methods, which are not in our syllabus, are implicit one-step methods. They are used in the handling of *stiff systems* (Section 45.7). However, for large stiff problems, the most popular methods used are based on the so-called *backward differentiation formula (BDF)*, which are multi-step methods first implemented by C. W. Gear. All these solvers can be visualised as a block diagram below:



Some of these solvers are built into numerical computing software such as Python/Scipy [Blanco-Silva, 2013, Chapter 4], Octave/MATLAB [Quarteroni and Saleri, 2006, Chapter 7], Scilab [Gomez et al., 1999, Chapter 8], etc.

Much of the content in this topic mostly based on Burden and Faires [2015, Chapter 5]. For classic text, Henrici [1962] is a good candidate. For more advanced topics, Hairer et al. [1993], Hairer and Wanner [1996], and Hairer et al. [2002] are highly recommended. For the relation of IVP to nonlinear science, refer to Meiss [2003].

45.1 IVP Theory and Examples

sec:ivp-eg

Since higher-dimensional ODE-IVP is often encountered in science and engineering, we will investigate the solution of n -dimensional ODE system (45.2) instead of limiting ourselves to one-dimensional case. When we attempt to find a *solution* for an ODE-IVP (45.2), we are actually finding a mapping $\varphi : J \subset \mathbb{R} \rightarrow \mathbb{R}^n$ that satisfies (45.2). To make sure that a solution exists and is unique, mathematicians found that if the function $\mathbf{f}(t, \mathbf{x})$ in (45.2) meets the Lipschitz condition requirement below, then we can have a beautiful existence and uniqueness theory for IVP.

Definition 45.1. A function $\mathbf{f}(t, \mathbf{x})$ is said to satisfy a *Lipschitz condition* in the variable $\mathbf{x}$ on a set $D \subset \mathbb{R}^{n+1}$ if a constant $L > 0$ exists with

$$\|\mathbf{f}(t, \mathbf{x}_1) - \mathbf{f}(t, \mathbf{x}_2)\| \leq L\|\mathbf{x}_1 - \mathbf{x}_2\|, \quad \forall (t, \mathbf{x}_1), (t, \mathbf{x}_2) \in D.$$

where $\|\cdot\|$ denotes the vector norm (Topic 2). The value L is called a *Lipschitz constant* for $\mathbf{f}$.

The following theorem is the fundamental existence and uniqueness theorem for first-order IVP (45.2) or (45.1).

Theorem 45.1 (Picard(-Lindelöf) Theorem). *Suppose that $D = [a, b] \times \mathbb{R}^n$ and that $\mathbf{f}(t, \mathbf{x})$ is **continuous** on D . If $\mathbf{f}$ satisfies a Lipschitz condition on D in the variable $\mathbf{x}$, then the IVP (45.2) has a unique solution $\mathbf{x}(t)$ for some domain containing $(t_0, \mathbf{x}_0) \in D$.*

The Picard Theorem only guarantees that a local solution $\mathbf{x}(t)$ exists for some interval $(t_0 - \epsilon, t_0 + \epsilon) \subset [a, b]$. For example, let us consider the ODE $\dot{x} = 1 + x^2 = f(t, x)$, $x(0) = 0$, $(t, x) \in [-100, 100] \times [-1, 1] = D$. Since $f(t, x)$ is continuous and Lipschitz on D , the Picard Theorem guarantees a unique solution for the ODE around $t = 0$. The integration of the ODE leads to $x(t) = \tan t$ which has a maximum interval $t \in (-\frac{\pi}{2}, \frac{\pi}{2})$ instead of the interval $(-100, 100)$. There is no solution which is defined for the larger interval $(-100, 100)$.

In physics and engineering models, knowing that a solution exists for (45.2) is not enough, we also need to make sure that (45.2) the solution exists and has little changes under the perturbations or variation of (45.2). This leads to the introduction of the notion of “well-posed problem”.

Definition 45.2. The IVP (45.2) is said to be a *well-posed problem* if:

- A unique solution, $\mathbf{x}(t)$, to the problem exists, and
- There exist constants $\epsilon_0 > 0$ and $k > 0$ such that for any ϵ , with $\epsilon_0 > \epsilon > 0$, whenever $\delta(t)$ is continuous with $\|\delta(t)\| < \epsilon$ for all t in $[a, b]$, and when $|\delta_0| < \epsilon$, the IVP

$$\frac{dz}{dt} = \mathbf{f}(t, \mathbf{z}) + \delta(t), \quad a < t \leq b; \quad \mathbf{z}(a) = \boldsymbol{\alpha} + \delta_0, \quad (45.6)$$

eq:5.3

has a unique solution $\mathbf{z}(t)$ that satisfies $\|\mathbf{z}(t) - \mathbf{x}(t)\| < k\epsilon \quad \forall t \in [a, b]$.

The problem specified by Equation (45.6) is called a *perturbed problem* associated with the original problem (45.2). It assumes the possibility of an error $\delta(t)$ being introduced in the statement of the differential equation, as well as an error δ_0 being present in the initial condition.

Numerical methods will always be concerned with solving a perturbed problem since any round-off error introduced in the representation perturbs the original problem. Unless the original problem is well-posed, there is little reason to expect that the numerical solution to a perturbed problem will accurately approximate the solution to the original problem.

When the function $\mathbf{f}$ in (45.2) is Lipschitz, the system is well-posed as stated below.

Theorem 45.2. Suppose $D = [a, b] \times \mathbb{R}^n$. If $\mathbf{f}$ is continuous and satisfies a Lipschitz condition in the variable $\mathbf{x}$ on the set D , then the IVP (45.2) is well-posed.

Most engineering ODE-IVP models (that aroused in mechanical systems [Rao, 2018], system vibrations [Géradin and Rixen, 2015, Chapter 7], electric networks, chemical kinetics, control theory, etc.) satisfy the Lipschitz condition and this implies that for most IVPs, it is possible for us to obtain a good approximation to the original problem.

The purpose of numerical simulation [Cellier and Kofman, 2006] is to improve our understanding of the behaviour of the physical, chemical and other natural phenomena [Tiller, 2001]. In order to test the various numerical methods discussed in this topic, we will pick two examples, one from Burden and Faires [2015, p.263] which has an exact solution and one from population dynamics.

Example 45.3. [Burden and Faires, 2015] Solve the following ODE-IVP exactly:

$$x' = x - t^2 + 1, \quad 0 \leq t \leq 2, \quad x(0) = C. \quad (45.7)$$

where $C = 0.5$. Implement the right hand function in (45.7) using appropriate numerical software.

Solution: This is a first order linear non-homogeneous equation. We could apply the mathematical techniques we learn in *UECM1033 Calculus II* to solve this problem:

$$x' - x = 1 - t^2 \Rightarrow e^{-t}x' - e^{-t}x = e^{-t}(1 - t^2) \Rightarrow \frac{d}{dt}(e^{-t}x) = e^{-t}(1 - t^2)$$

Integrating from 0, we have

$$\left[e^{-\tau}x(\tau) \right]_0^t = \int_0^t e^{-\tau}(1 - \tau^2)d\tau \Rightarrow e^{-t}x(t) - x(0) = \left[e^{-\tau}(\tau^2 + 2\tau + 1) \right]_0^t = e^{-t}(t + 1)^2 - 1$$

Simplifying gives $x(t) = Ce^t + (t + 1)^2 - e^t = (t + 1)^2 - 0.5e^t$ since $C = 0.5$. The Fortran implementation is

```
1 function burden_faires1(neq,t,x)
2     integer, intent(in) :: neq
3     double precision :: burden_faires1(neq), t, x(neq)
4     burden_faires1 = x-t**2+1d0
5 end function
```

Example 45.4 (Lotka-Volterra (Predator-Prey) Model). According to http://en.wikipedia.org/wiki/Lotka-Volterra_equation, this model was initially proposed by Alfred J. Lotka “in the theory of autocatalytic chemical reactions” in 1910. This was effectively the logistic equation, which was originally derived by Pierre François Verhulst. In 1920 Lotka extended, via Kolmogorov, the model to “organic systems” using a plant species and a herbivorous animal species as an example and in 1925 he utilised the equations to analyse predator-prey interactions in his book on biomathematics arriving at the equations that we know today. Vito Volterra, who made a statistical analysis of fish catches in the Adriatic independently investigated the equations in 1926.

This model describes the population dynamics of biological systems in which two species interact, one as a *predator* and the other as *prey*. The populations change through time according to the pair of equations:

$$\begin{aligned}\frac{dx}{dt} &= x(\alpha - \beta y) \\ \frac{dy}{dt} &= -y(\gamma - \delta x)\end{aligned}\tag{45.8}$$

where x is the number of prey (for example, rabbits), y is the number of some predator (for example, foxes); $\frac{dy}{dt}$ and $\frac{dx}{dt}$ represent the growth rates of the two populations over time; t represents time; and $\alpha = \beta = \gamma = \delta = 1$ are parameters describing the interaction of the two species. The initial values are

$$x(0) = 2, \quad y(0) = 1.$$

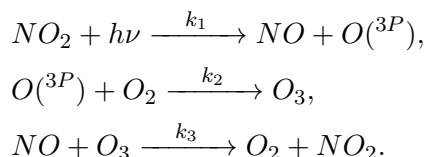
Implement the right hand function in (45.8) as a software function.

Solution: The following is an implementation of the right hand function of (45.8) in Fortran:

```
1      double precision, parameter :: alph=1d0, bet=1d0, gam=1d0, del=1d0
2      integer, intent(in) :: neq
3      double precision :: t, x(neq), f(neq)
4      f(1) = x(1)*(alph - bet*x(2))
5      f(2) = -x(2)*(gam - del*x(1))
6  end function
```

We look forward to have a *sustainable, green, automated and intelligent* future city. But how is it possible for us to have such a city if we are constantly creating waste and various air pollutants from our consumption? It is only possible if the speed of degradation of the waste by nature and the speed of decomposition of air pollutants are faster than the speed we generate them.

Consider a simple example from Roux [2015]. The following are basic reactions in any tropospheric air pollution model:



In this system $O(^3P)$ is the oxygen atom in its fundamental state. The first equation is the principal source for $O(^3P)$ by photolysis of nitrogen dioxide (NO_2) in the troposphere. The second equation above describes the formation of ozone (O_3). The third equation describes ozone destruction by reaction with nitrogen monoxide (NO) and formation of NO_2 . Normally, NO , NO_2 and $O(^3P)$ are in equilibrium and the ozone concentration is moderated. The formation of high ozone concentrations requires a mechanism that quickly consumes NO and regenerates NO_2 . In a polluted environment, this is realised by coupling the NO/NO_2 cycle with weathering hydrocarbons (HC). The HC attacked by a hydroxyl radical (OH) becomes an alkyl radical (R) which, by a sequence of quick reactions with air-oxygen, produces peroxy and hydroperoxyl radicals ((RO_2) and (HO_2) respectively) which, in turn, quickly oxidize NO to NO_2 with regeneration of the OH radical. These oxidation reactions from NO to NO_2 short-circuit that for NO by O_2 and the latter can accumulate.

The engine behind this complex chemistry is, therefore, the photolysis of participating chemical components (O_3 , NO_2 , etc.). In this way high levels of ozone can be produced during episodes of intensive solar irradiation. This condition is not sufficient, however. The precursors sources must be sufficiently important, especially in HC , and dynamical processes must favour the accumulation of precursors and products with weak dispersion and dilution. A good reference for the chemistry of the troposphere can be found in Seinfeld and Pandis [1998].

Let $[x]$ be the concentration of the chemical species x . Denote $c_1 = [O(^3P)]$, $c_2 = [NO]$, $c_3 = [NO_2]$, $c_4 = [O_3]$, then the ODE system associated with the air pollution model above can be written as:

$$\begin{aligned}\dot{c}_1 &= k_1 c_3 - k_2 c_1, \\ \dot{c}_2 &= k_1 c_3 - k_3 c_2 c_4, \\ \dot{c}_3 &= k_3 c_2 c_4 - k_1 c_3, \\ \dot{c}_4 &= k_2 c_1 - k_3 c_2 c_4.\end{aligned}\tag{45.9}$$

e:air-model

Note that oxygen in these equations is taken as constant, which is a reasonable assumption. Typically, the first reaction is photochemical and represents rapid changes in concentrations values at sunset and sunrise.

Example 45.5. Belousov-Zhbotinski chemical reaction (oregonator) can be described by the following set of equations proposed by the researchers from University of Oregon:

$$\begin{aligned}\dot{x}_1 &= s(x_2 - x_1 x_2 + x_1 - q x_1^2), \\ \dot{x}_2 &= \frac{1}{s}(-x_2 - x_1 x_2 + x_3), \\ \dot{x}_3 &= w(x_1 - x_3)\end{aligned}$$

where $s = 77.27$, $w = 0.161$, $q = 8.375 \cdot 10^{-6}$ and the initial condition is $\mathbf{x}(0) = (1, 2, 3)^T$. Implement the right hand function in (45.5) as a Fortran 90 function.

Solution:

```
1  subroutine f(neqn,t,x)
2  ! computed using double precision RADAU on an Alphaserer DS20E
3  ! 667 MHz EV67 processor & uround = 1.01d-19, rtol = h0 = 1.1d-18,
4  ! atol = 1.1d-40. At t = 360,
5  !      y(1) = 0.1000814870318523D+001
6  !      y(2) = 0.1228178521549917D+004
7  !      y(3) = 0.1320554942846706D+003
8      implicit none
9      integer :: neqn
10     double precision :: t,x(neqn),f(neqn)
11     f(1)=77.27D0*(x(2)+x(1)*(1.D0-8.375D-6*x(1)-x(2)))
12     f(2)=(x(3)-(1.D0+x(1))*x(2))/77.27D0
13     f(3)=0.161D0*(x(1)-x(3))
14 end
```

Example 45.6 (Tutorial 6, Q1). Implement the right hand function of the system of ODEs (45.9) with a computer program (e.g. Fortran) assuming that k_1 , k_2 , k_3 have default values 10^{-3} , 2.1×10^{-2} and 3.4×10^{-5} .

Solution. The implementation in Fortran is

```
1  function air_pollution(neq, t, x) result(f)
2      use iso_c_binding, only : c_double
3      real(c_double), parameter :: k1=1d-3,k2=2.1d-2,k3=3.4d-5
4      integer, intent(in) :: neq
5      real(c_double) :: t, x(neq), f(neq)
6      f(1) = k1*x(3) - k2*x(1)
7      f(2) = k1*x(3) - k3*x(2)*x(4)
8      f(3) = k3*x(2)*x(4) - k1*x(3)
9      f(4) = k2*x(1) - k3*x(2)*x(4)
10 end function
```

□

Another important application of ODE is celestial mechanics where we are interested to know the motions of astronomical bodies above the skies.

Example 45.7 (Kepler's Two Body Problem). The displacements of two celestial bodies of mass $m_1 = 2$ and $m_1 = 2$ on a plane that interact by gravitational force $G = 2$ is described by the following IVP [Gutzwiller, 1998]:

$$\begin{aligned} m_1 \ddot{\mathbf{q}}_1 &= \frac{Gm_1m_2}{\|\mathbf{q}_1 - \mathbf{q}_2\|^3}(\mathbf{q}_2 - \mathbf{q}_1), \\ m_2 \ddot{\mathbf{q}}_2 &= \frac{Gm_1m_2}{\|\mathbf{q}_2 - \mathbf{q}_1\|^3}(\mathbf{q}_1 - \mathbf{q}_2), \\ \mathbf{q}_1(0) &= (-1, 0)^T, \\ \mathbf{q}_2(0) &= (1, 0)^T, \\ \dot{\mathbf{q}}_1(0) &= (0, -1)^T, \\ \dot{\mathbf{q}}_2(0) &= (0, 1)^T. \end{aligned} \tag{45.10} \quad \text{e:kepler}$$

where $\mathbf{q}_i = (x_i, y_i)^T$, $i = 1, 2$. Implement the right hand function in (45.10) as a Fortran function. le:kepler

Solution:

```
1 function twobody(neq, t, x) result(f)
2   use iso_c_binding, only : c_double
3   real(c_double), parameter :: G=2d0, m1=2d0, m2=2d0
4   integer, intent(in) :: neq
5   real(c_double) :: t, x(neq), f(neq), distance2
6   distance2 = ((x(1)-x(5))**2+(x(3)-x(7))**2)**1.5
7   f(1) = x(2)
8   f(3) = x(4)
9   f(2) = G*m2/distance2 * (x(5)-x(1)) ! q1 = [x(1), x(3)]
10  f(4) = G*m2/distance2 * (x(7)-x(3)) ! q2 = [x(5), x(7)]
11  f(5) = x(6)
12  f(7) = x(8)
13  f(6) = G*m1/distance2 * (x(1)-x(5))
14  f(8) = G*m1/distance2 * (x(3)-x(7))
15 end function
```

ODE-IVP also plays an important part in the 3D collision detection and rigid body dynamics modelling [Millington, 2010] of 3-D role playing games, racing games, movie animations, etc.

45.2 Taylor Methods

If the solution $\mathbf{x}(t)$ to the IVP le:ode-ivp (45.2) has $n+1$ continuous derivatives, by Taylor's Approximation (Topic 3), we can expand $\mathbf{x}(t)$ in terms of its n th Taylor polynomial about t_i , $i = 0, 1, 2, \dots, N-1$:

$$\mathbf{x}(t_{i+1}) = \mathbf{x}(t_i) + h\mathbf{x}'(t_i) + \frac{h^2}{2}\mathbf{x}''(t_i) + \dots + \frac{h^n}{n!}\mathbf{x}^{(n)}(t_i) + \underbrace{\frac{h^{n+1}}{(n+1)!}\mathbf{x}^{(n+1)}(\xi_i)}_{\text{Local error term}} \tag{45.11} \quad \text{e:taylor-approx}$$

for some $\xi_i \in (t_i, t_{i+1})$. By substituting le:ode-ivp (45.2) into le:taylor-approx (45.11) and removing the error term, it is possible for us to obtain the “(Higher-Order) Taylor Method” as a numerical integration scheme for le:ode-ivp (45.2).

45.2.1 Euler Method

If $\mathbf{x}(t)$ has two continuous derivatives on $[a, b]$, then le:taylor-approx (45.11) could be written as

$$\mathbf{x}(t_{i+1}) = \mathbf{x}(t_i) + (t_{i+1} - t_i)\mathbf{x}'(t_i) + \frac{(t_{i+1} - t_i)^2}{2}\mathbf{x}''(\xi_i),$$

for some number $\xi_i \in (t_i, t_{i+1})$. Let $t_{i+1} - t_i = h = (b - a)/N$, we have

$$\mathbf{x}(t_{i+1}) = \mathbf{x}(t_i) + h\mathbf{x}'(t_i) + \frac{h^2}{2}\mathbf{x}''(\xi_i) = \mathbf{x}(t_i) + h\mathbf{f}(t_i, \mathbf{x}(t_i)) + O(h^2). \tag{45.12} \quad \text{eq:localerr}$$

If we remove the $O(h^2)$, then we obtain a *difference equation*

$$\mathbf{x}_{i+1} = \mathbf{x}_i + h\mathbf{f}(t_i, \mathbf{x}_i), \quad i = 0, 1, \dots, N-1. \quad (45.13)$$

e:euler-m

The above numerical method of finding approximations to $\mathbf{x}(t_i)$ is called the (*explicit*) *Euler Method*. The following is a simple implementation Euler method in Fortran:

```
1  procedure(funcn) :: f
2  integer, intent(in) :: neq
3  double precision, intent(in) :: x(neq), t, tout
4  double precision :: euler(neq), h
5  h = tout-t
6  euler = x + h*f(neq, t, x) ! Array Addition (Topic 2)
7  end function
```

The implementations in other programming languages are available from http://rosettacode.org/wiki/Euler_method.

Euler method is rarely used in practice. It is popular in numerical methods textbooks because of its simplicity and it is easier to perform mathematical analysis when compare to other numerical methods.

Example 45.8. Use the Euler method with $h = 0.2$ to solve the ODE-IVP (45.7) numerically.

1. Find x_2 .

Solution. From (45.7), we know that

$$x_0 = x(0) = 0.5, \quad t_0 = 0$$

Let $t_i = 0.2i, i = 0, 1, \dots, 10$. The Euler method difference equation is

$$x_{i+1} = x_i + h(x_i - t_i^2 + 1), i = 0, 1, \dots, 9.$$

When $i = 0, x_1 = x_0 + 0.2(x_0 - t_0^2 + 1) = 0.5 + 0.2(0.5 + 1) = 0.8$

When $i = 1, x_2 = x_1 + 0.2(x_1 - t_1^2 + 1) = 0.8 + 0.2(0.8 - 0.2^2 + 1) = 1.152 \quad \square$

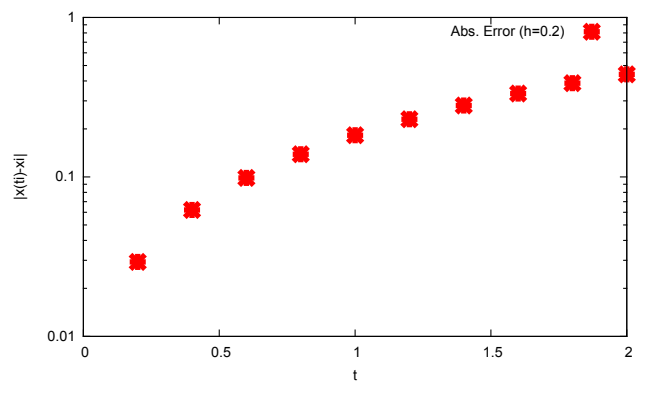
2. Calculate the values in the range $0 \leq t \leq 2$ with Euler method ($h = 0.2$) by using computer program and plot the absolute error against t_i .

Solution:

```
1  program eg_euler1
2  use iso_c_binding, only : c_double
3  use textbook_odeivp, only : euler, burden_faires1
4  implicit none
5  integer, parameter :: neq=1
6  integer :: i
7  real(c_double) :: h, tend, t, tnext, x(neq)
8  t = 0d0; tend=2d0; h=2d-1; x = [0.5d0]
9  print "(' #', a4, a15)", "t", "x"
10 do i=1,int((tend-t)/h)
11     print "(f5.2, f15.12)", t, x
12     tnext = t+h
13     x = euler(neq,burden_faires1,x,t,tnext)
14     t = tnext
15 end do
16 print "(f5.2, f15.12)", t, x
17 end program eg_euler1
```

The output of the program (left) and the absolute error (right) are shown below.

t_i	x_i
0.0	0.500000000000
0.2	0.800000000000
0.4	1.152000000000
0.6	1.550400000000
0.8	1.988480000000
1.0	2.458176000000
1.2	2.949811200000
1.4	3.451773440000
1.6	3.950128128000
1.8	4.428153753600
2.0	4.865784504320



From the above example, we find the *global error* $|x(t_i) - x_i|$ increases as t_i increases. In Burden and Faires [2015, p.269], the theorem characterising the global error of Euler method.

Theorem 45.9. Suppose that $n = 1$, f is continuous and satisfies a Lipschitz condition with constants L on $D = \{(t, x) | a \leq t \leq b, -\infty < x < \infty\}$ and that M exists with

$$|x''(t)| \leq M, \quad \forall t \in [a, b].$$

Let $x(t)$ denote the unique solution to the IVP (45.2) and $x_0, x_1, \dots, x_N$ be the approximations generated by Euler Method for some positive integer N . Then for each $i = 0, 1, \dots, N$,

$$|x(t_i) - x_i| \leq \frac{hM}{2L} [e^{L(t_i-a)} - 1]. \quad (45.14)$$

The local error of Euler method is $O(h^2)$, i.e. if the current step is exact, then the next step differs from the exact value by the order $O(h^2)$. On the other hand (45.14) indicates that the global error is $O(h)$.

Definition 45.3. We say a numerical method for solving sufficiently smooth systems (45.2) has an *order* p if

$$\|\mathbf{x}(t_i + h) - \mathbf{x}_{i+1}\| \leq Kh^{p+1}, \quad i = 0, \dots, N-1.$$

This means that the Taylor series for the exact solution $\mathbf{x}(t_i + h)$ and for $\mathbf{x}_i$ coincide up to and including the term h^p .

Note that the error grows slightly as the value of t_i increases. This controlled error growth is the consequence of the stability (Section 45.7) of Euler method, which implies that the error is expected to grow in no worse than a linear manner.

Example 45.10 (Tutorial 6, Q2). Use Euler's method with $h = 0.5$ to approximate the solution at $t = 2.5$ and $t = 3$ for the following IVP:

$$x' = 1 + (t - x)^2, \quad 2 \leq t \leq 3, \quad x(2) = 1.$$

Solution. Let $t_0 = 2$, $x_0 = 1$ and $t_1 = 2.5$. So

$$x(2.5) \approx x_1 = x_0 + 0.5(1 + (t_0 - x_0)^2) = 2.0000000$$

$$x(3.0) \approx x_2 = x_1 + 0.5(1 + (t_1 - x_1)^2) = 2.6250000$$

□

Let us learn from the following example how to perform numerical integration for higher dimension ODE-IVP.

Example 45.11. Use the Euler method with $h = 0.1$ to solve (45.8) for $0 \leq t \leq 10$.

1. Find the numerical solution $\mathbf{x}_1$ by hand.

Solution. The equation is $\dot{x} = x(1 - y), \quad x(0) = 2,$
 $\dot{y} = y(x - 1), \quad y(0) = 1.$

Hence, from $\mathbf{x}_{i+1} = \mathbf{x}_i + h\mathbf{f}(t_i, \mathbf{x}_i)$, we have

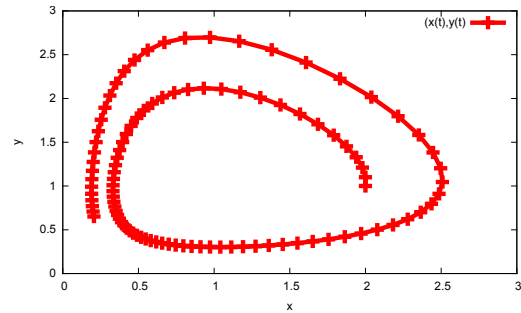
$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + 0.1 \begin{bmatrix} x_0(1 - y_0) \\ y_0(x_0 - 1) \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} + 0.1 \begin{bmatrix} 2(1 - 1) \\ 1(2 - 1) \end{bmatrix} = \begin{bmatrix} 2 \\ 1.1 \end{bmatrix}$$

□

2. Find the $\mathbf{x}_i$ for $i = 0, 1, \dots, 10$ by a numerical software and plot (x_i, y_i) .

Solution:

t_i	x_i	y_i
0.00	2.000000000000	1.000000000000
0.10	2.000000000000	1.100000000000
0.20	1.980000000000	1.210000000000
0.30	1.938420000000	1.328580000000
$\vdots$	$\vdots$	$\vdots$
9.70	0.192505507025	0.837597861612
9.80	0.195631837625	0.769962295554
9.90	0.200132107509	0.708028979877
10.00	0.205975385068	0.651396015081



A simplified real world simulation problem is far more complicated than what we have shown above. The following is an example of an n -body simulation in Fortran.

Example 45.12. Investigate what does the following Fortran program do.

```

1  ! The Computer Language Benchmarks Game
2  ! http://benchmarksgame.alioth.debian.org/u64q/nbody.html
3  !
4  !   contributed by Simon Geard, translated from Mark C. Williams nbody.java
5  !   modified by Brian Taylor, yuankun shi, Padraig O Conbhui
6  program nbody    ! Execute with an integer argument
7      implicit none
8      integer, parameter :: dp = kind(1.d0)
9      real(kind=dp), parameter :: tstep = 0.01d0
10     real(kind=dp), parameter :: PI = 3.141592653589793d0
11     real(kind=dp), parameter :: SOLAR_MASS = 4 * PI * PI
12     real(kind=dp), parameter :: DAYS_PER_YEAR = 365.24d0
13
14     type body
15         real(kind=dp) :: x, y, z, u, vx, vy, vz, vu, mass
16     end type body
17
18     type(body), parameter :: jupiter = body( &
19         4.84143144246472090d0, -1.16032004402742839d0, &
20         -1.03622044471123109d-01, 0d0, 1.66007664274403694d-03*DAYS_PER_YEAR, &
21         7.69901118419740425d-03*DAYS_PER_YEAR, &
22         -6.90460016972063023d-05*DAYS_PER_YEAR, 0.d0, &
23         9.54791938424326609d-04*SOLAR_MASS)
24
25     type(body), parameter :: saturn = body( &
26         8.34336671824457987d+00, &
27         4.12479856412430479d+00, &
28         -4.03523417114321381d-01, 0.d0, &
29         -2.76742510726862411d-03*DAYS_PER_YEAR, &
30         4.99852801234917238d-03*DAYS_PER_YEAR, &
31         2.30417297573763929d-05*DAYS_PER_YEAR, 0.d0, &
32         2.85885980666130812d-04*SOLAR_MASS)
33
34     type(body), parameter :: uranus = body( &

```

```

35      1.28943695621391310d+01, &
36      -1.51111514016986312d+01, &
37      -2.23307578892655734d-01, 0.d0,&
38      2.96460137564761618d-03*DAYS_PER_YEAR, &
39      2.37847173959480950d-03*DAYS_PER_YEAR, &
40      -2.96589568540237556d-05*DAYS_PER_YEAR, 0.d0,&
41      4.36624404335156298d-05*SOLAR_MASS )
42
43      type(body), parameter :: neptune = body( &
44          1.53796971148509165d+01, &
45          -2.59193146099879641d+01, &
46          1.79258772950371181d-01, 0.d0,&
47          2.68067772490389322d-03*DAYS_PER_YEAR, &
48          1.62824170038242295d-03*DAYS_PER_YEAR, &
49          -9.51592254519715870d-05*DAYS_PER_YEAR, 0.d0,&
50          5.15138902046611451d-05*SOLAR_MASS)
51
52      type(body), parameter :: sun = body(0.0d0, 0.0d0, 0.0d0, 0.0d0, 0.0d0, &
53          0.0d0, 0.d0, 0.d0, SOLAR_MASS)
54
55      integer, parameter :: nb = 5, N = (nb-1)*nb/2
56      real(kind=dp), parameter :: mass(nb) = (/ sun%mass, jupiter%mass, &
57          saturn%mass, uranus%mass, neptune%mass /)
58      integer :: num, i
59      character(len=8) :: argv
60      real(kind=dp) :: e, x(3,nb), v(3,nb)
61
62      x(:,1) = (/ sun%x, sun%y, sun%z /)
63      x(:,2) = (/ jupiter%x, jupiter%y, jupiter%z /)
64      x(:,3) = (/ saturn%x, saturn%y, saturn%z /)
65      x(:,4) = (/ uranus%x, uranus%y, uranus%z /)
66      x(:,5) = (/ neptune%x, neptune%y, neptune%z /)
67
68      v(:,1) = (/ sun%vx, sun%vy, sun%vz /)
69      v(:,2) = (/ jupiter%vx, jupiter%vy, jupiter%vz /)
70      v(:,3) = (/ saturn%vx, saturn%vy, saturn%vz /)
71      v(:,4) = (/ uranus%vx, uranus%vy, uranus%vz /)
72      v(:,5) = (/ neptune%vx, neptune%vy, neptune%vz /)
73
74      call getarg(1, argv)      ! GNU Fortran extension
75      read (argv,*) num        ! Command line arg: 50000000
76      call offsetMomentum(1, v, mass)
77      e = energy(x, v, mass)
78      write (*,'(f12.9)') e
79      do i = 1, num
80          call advance(tstep, x, v, mass)
81      end do
82      e = energy(x, v, mass)
83      write (*,'(f12.9)') e
84
85      contains
86
87      pure subroutine offsetMomentum(k, v, mass)
88          integer, intent(in) :: k
89          real(kind=dp), dimension(3,nb), intent(inout) :: v
90          real(kind=dp), dimension(nb), intent(in) :: mass
91          real(kind=dp), dimension(3) :: p
92
93          p = (/ (sum(v(i,:) * mass(:)), i=1,3) /)
94          v(:,k) = -p / SOLAR_MASS
95      end subroutine offsetMomentum
96
97      pure subroutine advance(tstep, x, v, mass)
98          real(kind=dp), intent(in) :: tstep
99          real(kind=dp), dimension(3,nb), intent(inout) :: x, v
100         real(kind=dp), dimension(nb), intent(in) :: mass
101         real(kind=dp) :: r(3,N),mag(N)

```

```

102     real(kind=dp) :: distance, d2
103     integer :: i, j, m
104
105     m = 1
106     do i = 1, nb
107         do j = i + 1, nb
108             r(:,m) = x(:,i) - x(:,j)
109             m = m + 1
110         end do
111     end do
112
113     do m = 1, N
114         d2 = sum(r(:,m)**2)
115         distance = 1/sqrt(real(d2))
116         distance = distance * (1.5d0 - 0.5d0 * d2 * distance * distance)
117         mag(m) = timestep * distance**3
118     end do
119
120     m = 1
121     do i = 1, nb
122         do j = i + 1, nb
123             v(:,i) = v(:,i) - r(:,m) * mass(j) * mag(m)
124             v(:,j) = v(:,j) + r(:,m) * mass(i) * mag(m)
125             m = m + 1
126         end do
127     end do
128
129     x = x + timestep * v
130 end subroutine advance
131
132 pure function energy(x, v, mass)
133     real(kind=dp), dimension(3,nb), intent(in) :: x, v
134     real(kind=dp), dimension(nb), intent(in) :: mass
135     real(kind=dp) :: energy, distance
136     real(kind=dp), dimension(3) :: d
137     integer :: i, j
138
139     energy = 0.0d0
140     do i = 1, nb
141         energy = energy + 0.5d0 * mass(i) * sum(v(:,i)**2)
142         do j = i + 1, nb
143             d = x(:,i) - x(:,j)
144             distance = sqrt(sum(d**2))
145             energy = energy - (mass(i) * mass(j)) / sqrt(sum(d**2))
146         end do
147     end do
148 end function energy
149
150 end program nbody

```

Note that the Fortran derived data type used in the program (first introduced in Fortran 90) is similar to C structures and also has some similarities with C++ classes. Pure is used to give the Fortran processor flexibility in the ordering of procedure invocations while still having a reasonably deterministic outcome from a particular stretch of code.

Solution: It is a program based on <http://janus.astro.umd.edu/HNBody/> which integrates the motion of particles in self-gravitating systems where the total mass is dominated by a single object; it is based on *symplectic integration techniques* in which two-body Keplerian motion is integrated exactly.

From http://janus.astro.umd.edu/HNBody/doc/html/integ_algo.html: Particles can be divided into three basic groups:

- HWP (Heavy Weight Particles), which always includes the dominant body, whose self- and external-interactions are always accounted for.

- LWPs (Light Weight Particles), whose interactions with HWPs (and, optionally, ZWPs) is computed, but whose self-potential is always ignored.
- ZWPs (Zero Weight Particles), whose self- and external-interactions are always neglected—i.e., the particles are treated as massless. They are always perturbed by HWPs, but may or may not be perturbed by LWPs (user-selectable).

This logical division of particles is useful whenever $M_H \gg M_L \gg M_Z$, where M_k is the total mass of particles in group k . If used to follow close encounters between bodies, the package is able to shift particles between the LWP and ZWP classes to maintain a balance between accuracy and efficiency.

For background information on the symplectic algorithms employed, see for example, Wisdom, Holman, & Touma (1996), *Symplectic Correctors*, Fields Institute Comm. Volume 10, pp. 217–244; Rauch, K. P., Hamilton, D. P. (2002), *The HNBody Package for Symplectic Integration of Nearly-Keplerian Systems*, Bulletin of American Astronomical Society Vol. 34, p.938

Runge-Kutta and Bulirsh-Stoer ODE integrators are also available; they have been included mainly for the purposes of testing and generating comparison runs.

The HWP/LWP system integration coordinates can be either Jacobi (barycentric) or canonical bodycentric (a.k.a. heliocentric, in the Solar System case). ZWPs are integrated as perturbed two-body orbits, and perturb nothing themselves; their integration coordinates can be either barycentric (specifically, the barycenter of the HWP/LWP subsystem) or canonical bodycentric, and their coordinates (as well as the timestep used) need not agree with those used for the other two groups. HWPs and LWPs always share the same coordinate choice and timestep.

Body-centric, bary-centric, and (time) regularised bary-centric coordinates are available for ZWPs; body-centric and Jacobi coordinates are available for the HWP/LWPs. An optional Modified body-centric coordinate choice (see IntegCoord) implements Touma & Wisdom’s form of the Hamiltonian splitting, which is exact for the two-body problem (unlike canonical bodycentric coordinates).

45.2.2 Higher-Order Taylor Method

Consider the general Taylor Method, which is based on (45.11). By combining with the successive differentiation of $\mathbf{x}(t)$ in (45.2):

$$\mathbf{x}'(t) = \mathbf{f}(t, \mathbf{x}(t)), \quad \mathbf{x}''(t) = \mathbf{f}'(t, \mathbf{x}(t)), \quad \dots, \quad \mathbf{x}^{(k)}(t) = \mathbf{f}^{(k-1)}(t, \mathbf{x}(t))$$

we obtain the general Taylor method stated below.

Method 45.1 (Taylor Method of Order n). Let $\mathbf{x}_0 = \mathbf{x}(t_0)$ (which is assumed by default for single-step methods). For $i = 0, 1, \dots, N - 1$,

$$\mathbf{x}_{i+1} = \mathbf{x}_i + h\mathbf{f}(t_i, \mathbf{x}_i) + \frac{h^2}{2}\mathbf{f}'(t_i, \mathbf{x}_i) + \dots + \frac{h^n}{n!}\mathbf{f}^{(n-1)}(t_i, \mathbf{x}_i) \quad (45.15)$$

where the local error is $\frac{1}{(n+1)!}\mathbf{x}^{(n+1)}(\xi_i)h^{n+1}$ for some ξ_i in (t_i, t_{i+1}) .

Calculating the k -th order derivative $\mathbf{f}^{(k)}(t, \mathbf{x}(t))$ is a major obstacle to Taylor method. Using computer algebra (such as Maxima, Maple, Mathematica, etc.) may solve part of the problem, but it remains cumbersome because of the complicated and long mathematical expression.

Error estimates for the Taylor Method is similar to Euler Method. If sufficient differentiability conditions are met ($\mathbf{x} \in C^{n+1}[a, b]$), an n th-order Taylor method (with step size h) will have *local (discretisation) error (LDE)*

$$\|\mathbf{x}(t_i) - \mathbf{x}_i\| = O(h^{n+1}),$$

and the *global error (GE)* $O(h^n)$, as well as the *local truncation error (LTE)* $O(h^n)$.

The LTE at a specified step measures the amount by which the exact solution to the differential equation fails to satisfy the difference equation being used for the approximation. Here is a formal definition.

Definition 45.4. Let $\mathbf{x}_0 = \mathbf{x}(t_0)$. The difference method

$$\mathbf{x}_{i+1} = \mathbf{x}_i + h\phi(t_i, \mathbf{x}_i), \quad i = 0, 1, \dots, N-1,$$

has a *local truncation error*

$$\tau_{i+1}(h) = \frac{\mathbf{x}(t_{i+1}) - (\mathbf{x}(t_i) + h\phi(t_i, \mathbf{x}(t_i)))}{h} = \frac{\mathbf{x}(t_{i+1}) - \mathbf{x}(t_i)}{h} - \phi(t_i, \mathbf{x}(t_i)), \quad i = 0, 1, \dots, N-1.$$

Example 45.13. Write down the difference equations by applying Taylor method of order two and four to the following IVP:

$$x' = x - t^2 + 1, \quad 0 \leq t \leq 2, \quad y(0) = 0.5.$$

Solution: We must find the first three derivatives of $f(t, x(t)) = x(t) - t^2 + 1$ w.r.t the variable t :

$$f'(t, x(t)) = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial x} \frac{dx}{dt} = -2t + 1 \cdot (x - t^2 + 1) = x - t^2 - 2t + 1$$

$$f''(t, x(t)) = \frac{\partial}{\partial t}(f') + \frac{\partial}{\partial x}(f') \frac{dx}{dt} = -2t - 2 + 1 \cdot (x - t^2 + 1) = x - t^2 - 2t - 1$$

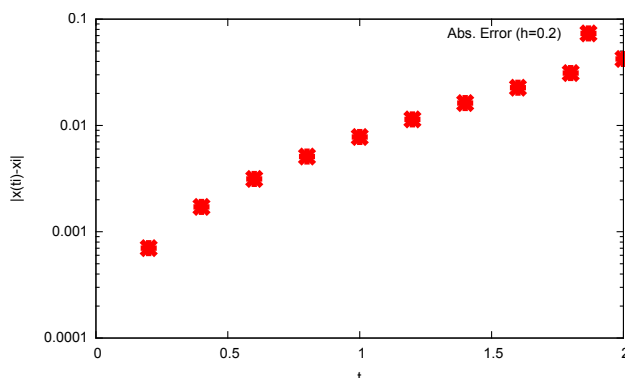
$$f'''(t, x(t)) = \frac{\partial}{\partial t}(f'') + \frac{\partial}{\partial x}(f'') \frac{dx}{dt} = -2t - 2 + 1 \cdot (x - t^2 + 1) = x - t^2 - 2t - 1.$$

So Taylor method of order two is

$$\begin{aligned} x_0 &= 0.5 \\ x_{i+1} &= x_i + hf(t_i, x_i) + \frac{h^2}{2}f''(t_i, x_i) \\ &= x_i + h(x_i - t_i^2 + 1) + \frac{h^2}{2}(x_i - t_i^2 + 1 - 2t_i) \quad i = 0, \dots, N-1 \end{aligned}$$

and it gives the computation table and absolute error plot below.

t_i	x_i
0.0	0.500000000000
0.2	0.830000000000
0.4	1.215800000000
0.6	1.652076000000
0.8	2.132332720000
1.0	2.648645918400
1.2	3.191348020448
1.4	3.748644584947
1.6	4.306146393635
1.8	4.846298600234
2.0	5.347684292286

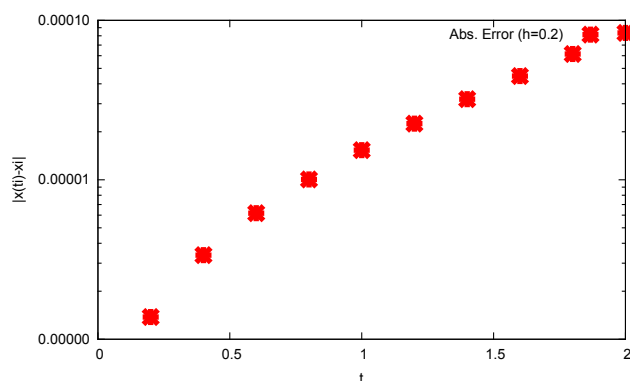


The Taylor method of order four is

$$\begin{aligned} x_0 &= 0.5 \\ x_{i+1} &= x_i + h \left[f(t_i, x_i) + \frac{h}{2}f'(t_i, x_i) + \frac{h^2}{6}f''(t_i, x_i) + \frac{h^3}{24}f'''(t_i, x_i) \right], \quad i = 0, 1, \dots, N-1 \\ &= x_i + h \left[\left(1 + \frac{h}{2} + \frac{h^2}{6} + \frac{h^3}{24}\right)(x_i - t_i^2) - \left(1 + \frac{h}{3} + \frac{h^2}{12}\right)(ht_i) + 1 + \frac{h}{2} + \frac{h^2}{6} + \frac{h^3}{24} \right]. \end{aligned}$$

and it gives the computation table and absolute error plot below.

t_i	x_i
0.0	0.500000000000
0.2	0.829300000000
0.4	1.214091020000
0.6	1.648946771828
0.8	2.127239587111
1.0	2.640874431697
1.2	3.179964030875
1.4	3.732432067310
1.6	4.283528527013
1.8	4.815237742894
2.0	5.30555379170



Example 45.14 (Jan 2014 Semester Test 2). Find the approximation $x(0.1)$ of the IVP below

$$x' = 3t + 2x, \quad 0 \leq t \leq 2, \quad x(0) = 2.$$

by using the Taylor method of order 3 with $h = 0.1$.

(3 marks)

Solution. Let $f(t, x) = 3t + 2x$. Then

$$f'(t, x) = 3 + 2x' = 3 + 2(3t + 2x) = 3 + 6t + 4x \quad [0.5 \text{ mark}]$$

$$f''(t, x) = 6 + 4x' = 6 + 4(3t + 2x) = 6 + 12t + 8x \quad [0.5 \text{ mark}]$$

$$\begin{aligned} x_{i+1} &= x_i + hf(t_i, x_i) + hf'(t_i, x_i) + \frac{h^2}{2} f''(t_i, x_i) + \frac{h^3}{6} f'''(t_i, x_i) \\ &= x_i + 0.1(3t_i + 2x_i) + \frac{0.1^2}{2}(3 + 6t_i + 4x_i) + \frac{0.1^3}{6}(6 + 12t_i + 8x_i) \end{aligned} \quad [1 \text{ mark}]$$

$$x_1 = 2 + 0.1(2 \cdot 2) + \frac{0.01}{2}(3 + 4 \cdot 2) + \frac{0.001}{6}(6 + 8(2)) = 2.45866667 \quad [1 \text{ mark}]$$

□

baum2:p201:19.4

Example 45.15 (Tutorial 6, Q3). Apply Taylor method of order four to obtain a finite difference equation for the IVP

$$\dot{x} = tx^{1/3}, \quad x(1) = 1$$

and use it to find the approximation for $x(1.1)$.

Solution.

$$x(t+h) = x(t) + hx'(t) + \frac{h^2}{2} x''(t) + \frac{h^3}{6} x'''(t) + \frac{h^4}{24} x^{(4)}(t)$$

with a local truncation error $E = |\frac{h^5}{120} x^{(5)}(\xi)|$ and

$$x''(t) = \frac{1}{3} t^2 x^{-1/3} + x^{1/3},$$

$$x'''(t) = -\frac{1}{9} t^3 x^{-1} + tx^{-1/3},$$

$$x^{(4)}(t) = \frac{1}{9} t^4 x^{-5/3} - \frac{2}{3} t^2 x^{-1} + x^{-1/3}$$

The finite difference equation is

$$\begin{aligned} x_{i+1} &= x_i + ht_i x_i^{1/3} + \frac{h^2}{2} \left(\frac{1}{3} t_i^2 x_i^{-1/3} + x_i^{1/3} \right) + \frac{h^3}{6} \left(-\frac{1}{9} t_i^3 x_i^{-1} + t_i x_i^{-1/3} \right) \\ &\quad + \frac{h^4}{24} \left(\frac{1}{9} t_i^4 x_i^{-5/3} - \frac{2}{3} t_i^2 x_i^{-1} + x_i^{-1/3} \right) \end{aligned}$$

and let $t_0 = 1$, $x_0 = 1$ and $h = 0.1$, we obtain an approximation for

$$x(1.1) \approx x_1 = 1 + 0.1(1)(1) + \frac{0.01}{2} \left(\frac{1}{3} + 1 \right) + \frac{0.001}{6} \left(-\frac{1}{9} + 1 \right) + \frac{0.0001}{24} \left(\frac{1}{9} - \frac{2}{3} + 1 \right) = 1.10681667$$

Example 45.16. Write down the difference equations by applying Taylor method of order two for Lotka-Volterra model (45.8) in Example 45.4.

Solution:

$$\begin{cases} \dot{x} = x(1-y) \\ \dot{y} = y(x-1) \end{cases} \Rightarrow \begin{cases} \ddot{x} = \dot{x}(1-y) + x(-\dot{y}) = x(1-y)^2 + xy(x-1) \\ \ddot{y} = \dot{y}(x-1) + y(\dot{x}) = y(x-1)^2 + xy(1-y) \end{cases}$$

Therefore, Taylor of order two is

$$\begin{aligned} x_{i+1} &= x_i + h\dot{x}_i + \frac{h^2}{2!}\ddot{x}_i = x_i + hx_i(1-y_i) + \frac{h^2}{2}[x_i(1-y_i)^2 + x_iy_i(x_i-1)] \\ y_{i+1} &= y_i + h\dot{y}_i + \frac{h^2}{2!}\ddot{y}_i = y_i + hy_i(x_i-1) + \frac{h^2}{2}[y_i(x_i-1)^2 + x_iy_i(1-y_i)] \end{aligned}$$

Remark 45.17. We will find that standard numerical methods for solving (45.2) do not use Taylor's method. Taylor's method is used in *validated* (or *interval*) methods for producing bounds for the IVP (45.2) that guarantees to contain the solution and is useful in certain mathematical proofs based on numerical methods. A software implementation of validated ODE in C++ is the VNODE package.

Taylor's method is inconvenient due to the need for differentiation. However, with suitable derivation, it is possible to use it to derived fast and high-order methods for second order ODEs such as Numerov's method for $\ddot{x} = U(t) + V(t)x$ and Leap-frog method for $\ddot{x} = f(x)$.

45.3 Runge-Kutta Methods

sec:rk

The Taylor methods have the desirable property of high-order local truncation error, but the disadvantage of requiring the computation and evaluation of the derivatives of $f(t, \mathbf{x})$. This is similar to the Newton method in Topic 3, where we need to find the Jacobian. The computation of derivatives is a complicated and time-consuming procedure for many problems, so the Taylor methods are seldom used in practice.

Carl David Tolmé Runge (we knew him in Topic 4 for the famous Runge phenomena) and Martin Kutta developed a family of methods called the *Runge-Kutta Methods*, which modify the Taylor methods so that the high-order error bounds are preserved, while eliminating the need to determine and evaluate the high-order partial derivative.

Definition 45.5. Let the *step size* $h > 0$ be a positive real number and *number of stages* s be a positive integer. Define $t_i := t_0 + ih$ and let $\mathbf{x}_i$ be the approximation of $\mathbf{x}(t_i)$ where $i \geq 0$. An *s-stage (explicit) Runge-Kutta method* is of the form

$$\mathbf{x}_{i+1} = \mathbf{x}_i + h \left[b_1 \mathbf{k}_1 + \cdots + b_s \mathbf{k}_s \right] \quad (45.16)$$

e:rk

where

$$\begin{aligned} \mathbf{k}_1 &= \mathbf{f}(t_i, \mathbf{x}_i) \\ \mathbf{k}_2 &= \mathbf{f}(t_i + c_2h, \mathbf{x}_i + ha_{21}\mathbf{k}_1) \\ \mathbf{k}_3 &= \mathbf{f}(t_i + c_3h, \mathbf{x}_i + h(a_{31}\mathbf{k}_1 + a_{32}\mathbf{k}_2)) \\ &\vdots \\ \mathbf{k}_s &= \mathbf{f}(t_i + c_sh, \mathbf{x}_i + h(a_{s1}\mathbf{k}_1 + \cdots + a_{s,s-1}\mathbf{k}_{s-1})) \end{aligned}$$

where $a_{ij} \in \mathbb{R}$, $i = 2, \dots, s$, $j < i$; $b_i, c_i \in \mathbb{R}$, $i = 1, \dots, s$.

These methods makes use of the following theorem.

urden-faires:5.13

Theorem 45.18 (Taylor's Theorem for Two Variables). *Suppose that $f(t, x)$ and all its partial derivative of order $\leq (n + 1)$ are continuous on $D = \{(t, x) | a \leq t \leq b, a \leq x \leq d\}$ and let $(t, x) \in D$ and $(t + \alpha, x + \beta) \in D$. Then*

$$f(t + \alpha, x + \beta) = f(t, x) + \left[\alpha \frac{\partial f}{\partial t}(t, x) + \beta \frac{\partial f}{\partial x}(t, x) \right] + \left[\frac{\alpha^2}{2} \frac{\partial^2 f}{\partial t^2}(t, x) + \alpha \beta \frac{\partial^2 f}{\partial t \partial x}(t, x) + \frac{\beta^2}{2} \frac{\partial^2 f}{\partial x^2}(t, x) \right] + \cdots + \frac{1}{n!} \sum_{j=1}^n \binom{n}{j} \alpha^{n-j} \beta^j \frac{\partial^n f}{\partial t^{n-j} \partial x^j}(t, x) + R_n(t, x)$$

where

$$R_n(t, x) = \frac{1}{(n+1)!} \sum_{j=0}^{n+1} \binom{n+1}{j} \alpha^{n+1-j} \beta^j \frac{\partial^{n+1} f}{\partial t^{n+1-j} \partial x^j}(\xi, \mu).$$

Note that Euler method can also be regarded as a Runge-Kutta method of order one. Therefore our discussion will begin with Runge-Kutta methods of order two.

45.3.1 Runge-Kutta Methods of Order Two

Let's look at the Taylor method (45.15) of order 2 (for ODE of dimensional 1):

$$x_{i+1} = x_i + h T^{(2)}(t_i, x_i) = x_i + h \left[f(t_i, x_i) + \frac{h}{2} f'(t_i, x_i) \right].$$

The idea of Runge-Kutta method is to approximate $T^{(2)}(t, x)$ by

$$a_1 f(t + \alpha_1, x + \beta_1) \quad (45.17) \quad \text{e:t2approx}$$

with appropriate values of a_1 , α_1 , and β_1 so that the error is no greater than $O(h^2)$. Since

$$f'(t, x) = \frac{\partial f}{\partial t}(t, x) + \frac{\partial f}{\partial x}(t, x) \cdot x'(t) \quad \text{and} \quad x'(t) = f(t, x),$$

this implies

$$T^{(2)}(t, x) = f(t, x) + \frac{h}{2} \frac{\partial f}{\partial t}(t, x) + \frac{h}{2} \frac{\partial f}{\partial x}(t, x) \cdot f(t, x). \quad (45.18) \quad \text{eq:5.18}$$

On the other hand, expanding (45.17) in its Taylor polynomial of degree one about (t, y) gives

$$a_1 f(t + \alpha_1, x + \beta_1) = a_1 f(t, x) + a_1 \alpha_1 \frac{\partial f}{\partial t}(t, x) + a_1 \beta_1 \frac{\partial f}{\partial x}(t, x) + a_1 R_1(t + \alpha_1, x + \beta_1) \quad (45.19) \quad \text{eq:5.19}$$

where

$$R_1(t + \alpha_1, x + \beta_1) = \frac{\alpha_1^2}{2} \frac{\partial^2 f}{\partial t^2}(\xi, \mu) + \alpha_1 \beta_1 \frac{\partial^2 f}{\partial t \partial x}(\xi, \mu) + \frac{\beta_1^2}{2} \frac{\partial^2 f}{\partial x^2}(\xi, \mu) \quad (45.20) \quad \text{eq:5.20}$$

for some $\xi \in (t, t + \alpha_1)$ and $\mu \in (x, x + \beta_1)$ according to Theorem 45.18.

Matching the coefficients of f and its derivatives in Equations (45.18) and (45.19) gives

Term	$f(t, x)$	$\frac{\partial f}{\partial t}(t, x)$	$\frac{\partial f}{\partial x}(t, x)$
Coefficient	$a_1 = 1$	$a_1 \alpha_1 = \frac{h}{2}$	$a_1 \beta_1 = \frac{h}{2} f(t, x)$

in which the parameters a_1 , α_1 , and β_1 are uniquely determined to be $a_1 = 1$, $\alpha_1 = \frac{h}{2}$ and $\beta_1 = \frac{h}{2} f(t, x)$. So

$$T^{(2)}(t, x) = f\left(t + \frac{h}{2}, x + \frac{h}{2} f(t, x)\right) - R_1\left(t + \frac{h}{2}, x + \frac{h}{2} f(t, x)\right),$$

where Equation (45.20) implies

$$R_1 \left(t + \frac{h}{2}, x + \frac{h}{2} f(t, x) \right) = \frac{h^2}{8} \frac{\partial^2 f}{\partial t^2}(\xi, \mu) + \frac{h^2}{4} f(t, x) \frac{\partial^2 f}{\partial t \partial x}(\xi, \mu) + \frac{h^2}{8} (f(t, x))^2 \frac{\partial^2 f}{\partial x^2}(\xi, \mu).$$

If all the second-order partial derivatives of f are bounded, then $R_1 \left(t + \frac{h}{2}, x + \frac{h}{2} f(t, x) \right)$ is $O(h^2)$, the order of the local truncation error of Taylor method of order two. As a consequence, using the new procedure in place of the Taylor method of order two might add error, but it does not increase the order of the error.

The difference-equation method resulting from replacing $T^{(2)}(t, x)$ in Taylor method of order two by $f(t+h/2, x+h f(t, x)/2)$ is a specific Runge-Kutta method known as the *midpoint method*.

Method 45.2 (Midpoint Method). For $i = 0, 1, \dots, N-1$,

$$\begin{aligned} \mathbf{k}_1^{(i)} &= \mathbf{f}(t_i, \mathbf{x}_i) \\ \mathbf{x}_{i+1} &= \mathbf{x}_i + h \mathbf{f} \left(t_i + \frac{h}{2}, \mathbf{x}_i + \frac{h}{2} \mathbf{k}_1^{(i)} \right) \end{aligned} \quad (45.21)$$

e:midpt

Let us study an example.

Example 45.19. By applying the midpoint method (45.21) with $h = 0.2$ to solve (45.7) numerically for the range $0 \leq t \leq 2$.

1. First, calculate the approximation x_2 ;

Solution. Note that $x_0 = 0.5$. The difference equation is

$$\begin{aligned} k_1^{(i)} &= f(t_i, x_i), \quad t_i = 0 + ih = 0.2i, \\ x_{i+1} &= x_i + h f \left(t_i + \frac{0.2}{2}, x_i + \frac{0.2}{2} k_1^{(i)} \right) \\ &= x_i + 0.2 \left[(x_i + 0.1 k_1) - t_i^2 + 1 \right], \quad \text{for } i = 0, 1, \dots, 9. \end{aligned}$$

$$x_0 = 0.5, \quad t_0 = 0$$

$$k_1^{(0)} = x_0 - t_0^2 + 1 = 1.5$$

$$\begin{aligned} x_1 &= x_0 + 0.2 f(t_0 + 0.1, x_0 + 0.1 k_1^{(0)}) = x_0 + 0.2 f(0.1, 0.65) \\ &= 0.5 + 0.2(0.65 - 0.1^2 + 1) = 0.828 \end{aligned}$$

□

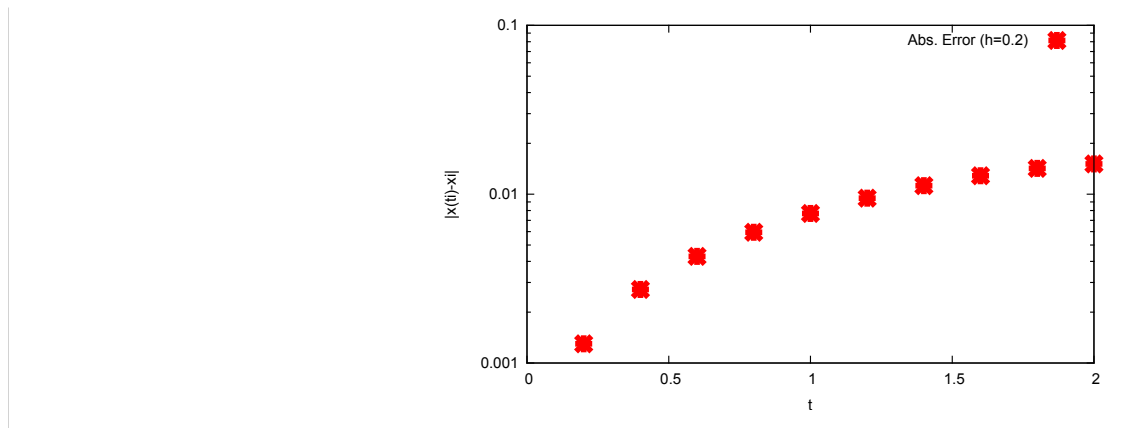
$$k_1^{(1)} = x_1 - t_1^2 + 1 = 0.828 - 0.2^2 + 1 = 1.788$$

$$\begin{aligned} x_2 &= x_1 + 0.2 f(t_1 + 0.1, x_1 + 0.1 k_1^{(1)}) = 0.828 + 0.2 f(0.3, 0.828 + 0.1788) \\ &= 0.828 + 0.2[1.0068 - 0.3^2 + 1] = 1.21136 \end{aligned}$$

2. Then plot the (global) absolute error against the time using $h = 0.2$.

Solution: The numerical solution and the absolute error are shown below.

t_i	x_i
0.0	0.500000000000
0.2	0.828000000000
0.4	1.211360000000
0.6	1.644659200000
0.8	2.121284224000
1.0	2.633166753280
1.2	3.170463439002
1.4	3.721165395582
1.6	4.270621782610
1.8	4.800958574784
2.0	5.290369461237



eg:midpoint1

Example 45.20. The implementation of Midpoint method (45.21) in Fortran is listed below:

```

1  procedure(funcn) :: f
2  integer, intent(in) :: neq
3  double precision, intent(in) :: x(neq), t, tout
4  double precision :: midpoint(neq), h, k1(neq)
5  h = tout-t
6  k1 = f(neq, t, x)
7  midpoint = x + h*f(neq, t+h/2d0, x + h*k1/2d0)
8  end function

```

Use this function with $h = 0.1$ to find the numerical solution $\mathbf{x}_1$ for (45.8) by hand.

Solution. Note that $\mathbf{f}(t, \mathbf{x}) = [x(1 - y), y(x - 1)]^T$.

$$\begin{aligned}
 \mathbf{x}_0 &= \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad t_0 = 0 \\
 \mathbf{k}_1^{(0)} &= \begin{bmatrix} x_0(1 - y_0) \\ y_0(x_0 - 1) \end{bmatrix} = \begin{bmatrix} 2(1 - 1) \\ 1(2 - 1) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
 \mathbf{x}_1 &= \mathbf{x}_0 + 0.1\mathbf{f}\left(t_0 + \frac{0.1}{2}, \mathbf{x}_0 + \frac{0.1}{2}\mathbf{k}_1^{(0)}\right) = \begin{bmatrix} 2 \\ 1 \end{bmatrix} + 0.1\mathbf{f}(0.05, \begin{bmatrix} 2 \\ 1.05 \end{bmatrix}) \\
 &= \begin{bmatrix} 2 \\ 1 \end{bmatrix} + 0.1 \begin{bmatrix} 2(1 - 1.05) \\ 1.05(2 - 1) \end{bmatrix} = \begin{bmatrix} 1.99 \\ 1.105 \end{bmatrix}
 \end{aligned}$$

□

Example 45.21 (April 2018 Final Exam Question 5(a)). The circular motion θ of an idealised pendulum with a string of negligible mass, length L , a bob of mass m and a damping proportional to the angular velocity θ' can be modelled by

$$m\theta'' + b\theta' + \frac{mg}{L} \sin \theta = 0.$$

where $\theta'' = \frac{d^2\theta}{dt^2}$ and $\theta' = \frac{d\theta}{dt}$. Suppose $m = 0.05$, $b = 0.001$, $L = 1$, $g = 9.8$, and at time $t = 0$, $\theta = 0.4$ and $\theta' = 0$. You are planning to use the following Fortran program to perform numerical calculation for this initial value problem.

```

1  program exam_q5_midpt
2  use iso_c_binding, only : c_double
3  use exam_func, only : f
4  interface
5  function funcn(neq,t,x)
6  use iso_c_binding, only : c_double
7  integer, intent(in) :: neq
8  real(c_double) :: t, x(neq), funcn(neq)
9  end function funcn

```

```

10     end interface
11     integer, parameter :: neq=2
12     real(c_double) :: t, tnext, x(neq)
13     t = 0d0; x = (/0.4d0,0d0/)
14     tnext = t + 0.1d0
15     x = midpoint(neq,f,x,t,tnext)
16     print *, "t=", tnext, "x=", x
17 contains
18 function midpoint(neq, f, x, t, tout)
19     integer, intent(in) :: neq
20     procedure(funcn) :: f
21     real(c_double), intent(in) :: x(neq), t, tout
22     real(c_double) :: midpoint(neq), h, k1(neq)
23     h = tout-t
24     k1 = f(neq, t, x)
25     midpoint = x + h*f(neq, t+h/2.0, x + h*k1/2.0)
26 end function
27 end program exam_q5_midpt

```

- (i) Implement the Fortran function f (line 3) for solving the initial value problem. (3 marks)

Solution. One possible implementation is

```

1 function f(neq, t, x)
2     real(c_double), parameter :: m=5d-2, b=1d-3, L=1d0, g=9.8d0
3     integer, intent(in) :: neq
4     real(c_double) :: t, x(neq), f(neq)
5     f(1) = x(2)
6     f(2) = -(b*x(2)+m*g/L*sin(x(1)))/m
7 end function

```

Declaration of constants [1 mark]

Correct implementation of the return array [2 marks] ☐

- (ii) By perform hand calculation using scientific calculator, verify the output of the Fortran program. All calculations are required to have at least 6 decimal places. (7 marks)

Solution. Let $\omega = \theta'$. $h = 0.1$.

$$\mathbf{k}_1 = \begin{bmatrix} \omega \\ -\frac{1}{0.05}(0.001\omega + 0.05 \times 9.8 \sin \theta) \end{bmatrix} \quad [1 \text{ mark}]$$

$$= \begin{bmatrix} 0 \\ -20.0 \times (0.001 \times 0 + 0.49 \sin(0.4)) \end{bmatrix} = \begin{bmatrix} 0 \\ -3.816300 \end{bmatrix} \quad [2 \text{ marks}]$$

$$\mathbf{f}(0.05, \begin{bmatrix} 0.4 \\ 0 \end{bmatrix} + 0.05 \begin{bmatrix} 0 \\ -3.816300 \end{bmatrix}) = \mathbf{f}(0.05, \begin{bmatrix} 0.4 \\ -0.190815 \end{bmatrix})$$

$$= \begin{bmatrix} -0.190815 \\ -20.0 \times (0.001 \times (-0.190815) + 0.49 \sin(0.4)) \end{bmatrix} = \begin{bmatrix} -0.190815 \\ -3.812483 \end{bmatrix} \quad [2 \text{ marks}]$$

The output is $t = 0.1$ and

$$\mathbf{x} = \begin{bmatrix} 0.4 \\ 0 \end{bmatrix} + 0.1 \begin{bmatrix} -0.190815 \\ -3.812483 \end{bmatrix} = \begin{bmatrix} 0.3809185 \\ -0.3812483 \end{bmatrix} \quad [2 \text{ marks}]$$

☐

- (iii) Is it possible for us to use higher order Taylor method

$$\mathbf{x}(t_{i+1}) = \mathbf{x}(t_i) + h\mathbf{x}'(t_i) + \frac{h^2}{2}\mathbf{x}''(t_i) + \cdots + \frac{h^n}{n!}\mathbf{x}^{(n)}(t_i)$$

to obtain a more accurate estimate of $\theta(0.1)$? Justify your answer. What is the problem with higher order Taylor method? (2 marks)

Solution. Yes, since the ODE has analytic expression. [1 mark]

The problem with higher order Taylor method is the lengthy analytic expression for higher order n [1 mark] $\square$

Now, let's look at the Taylor method (45.15) of order 3 (for ODE of dimensional 1):

$$x_{i+1} = x_i + h \underbrace{\left[f(t_i, x_i) + \frac{h}{2} f'(t_i, x_i) + \frac{h^2}{6} f''(t_i, x_i) \right]}_{T^{(3)}(t_i, x_i)}.$$

With appropriate values of a_1 , α_1 , and β_1 , can we use

$$a_1 f(t, x) + a_2 f(t + \alpha_2, x + \delta_2 f(t, x)) \quad (45.22)$$

e:t3approx

to approximate Taylor method to order 3? The answer is negative and according to Burden and Faires [2015], when $a_1 = a_2 = \frac{1}{2}$ and $\alpha_2 = \delta_2 = h$, we obtain a different Runge-Kutta method of order 2 called the *modified Euler method* or *Heun's (Second Order) Method*.

Method 45.3 (Modified Euler Method). For $i = 0, 1, \dots, N - 1$,

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \frac{h}{2} [\mathbf{k}_1^{(i)} + \mathbf{k}_2^{(i)}], \quad \mathbf{k}_1^{(i)} = \mathbf{f}(t_i, \mathbf{x}_i), \quad \mathbf{k}_2^{(i)} = \mathbf{f}(t_{i+1}, \mathbf{x}_i + h\mathbf{k}_1^{(i)}). \quad (45.23)$$

e:mod-euler

Example 45.22. The Fortran implementation for modified Euler method (45.23) of step-size h is given below.

```
1  procedure(funcn) :: f
2  integer, intent(in) :: neq
3  double precision, intent(in) :: x(neq), t, tout
4  double precision :: modeuler(neq), h, k1(neq), k2(neq)
5  h = tout - t
6  k1 = f(neq, t, x)
7  k2 = f(neq, t+h, x+h*k1)
8  modeuler = x + h*(k1+k2)/2d0
9  end function
```

Use this function with $h = 0.2$ to estimate $x(0.2)$ for Equation (45.7).

e:burden-faires

Solution: The data to be passed into the modified Euler method modeuler are

$$f(t, x) = x - t^2 + 1, \quad x_0 = 0.5, \quad t_0 = 0, \quad t_1 = t_0 + h = 0.2.$$

By substituting these values to lines 7 to 9, we are able to estimate $x(0.2)$ with x_1 :

$$\begin{aligned} k_1^{(0)} &= 0.5 - 0^2 + 1 = 1.5 \\ k_2^{(0)} &= f(t_1, x_0 + 0.2k_1^{(0)}) = f(0.2, 0.5 + 0.3) = 0.8 - 0.2^2 + 1 = 1.76 \\ x_1 &= x_0 + \frac{0.2}{2} [k_1^{(0)} + k_2^{(0)}] = 0.5 + 0.1(1.5 + 1.76) = 0.826 \end{aligned}$$

Example 45.23 (April 2014 Final Exam Question 5(b)). Let $x(t)$ and $y(t)$ denote the population of rabbits and foxes, respectively, at time t . The Predator-Prey model asserts that x and y satisfies

$$\dot{x} = 2x - 0.02xy, \quad \dot{y} = 0.0002xy - 0.8y$$

Given the initial condition $x(0) = 3000$ rabbits and $y(0) = 120$ foxes. Find an approximation for $x(0.1)$ and $y(0.1)$ using modified Euler method:

$$\mathbf{w}_{i+1} = \mathbf{w}_i + \frac{h}{2}[\mathbf{f}(t_i, \mathbf{w}_i) + \mathbf{f}(t_{i+1}, \mathbf{w}_i + h\mathbf{f}(t_i, \mathbf{w}_i))], \quad i = 0, 1, \dots$$

where h is a suitable time step. (6 marks)

Solution. Given $t_0 = 0$, $\mathbf{w}_0 = (x_0, y_0)^T = (3000, 120)^T$, by choosing $h = 0.1$ (since 0.1 is small),

$$\mathbf{f}(t_0, \mathbf{w}_0) = \begin{bmatrix} 2x_0 - 0.02x_0y_0 \\ 0.0002x_0y_0 - 0.8y_0 \end{bmatrix} = \begin{bmatrix} -1200 \\ -24 \end{bmatrix} \quad [2 \text{ marks}]$$

$$\mathbf{f}(t_0, \mathbf{w}_0 + h\mathbf{f}(t_0, \mathbf{w}_0)) = \begin{bmatrix} 2(2880) - 0.2(2880)(117.6) \\ 0.0002(2880)(117.6) - 0.8(117.6) \end{bmatrix} = \begin{bmatrix} -1013.76 \\ -26.3424 \end{bmatrix} \quad [2 \text{ marks}]$$

$$x(0.1) \approx x_1 = 3000 + \frac{0.1}{2}[-1200 - 1013.76] = 2889.312 \quad [1 \text{ mark}]$$

$$y(0.1) \approx y_1 = 120 + \frac{0.1}{2}[-24 - 26.3424] = 117.4829 \quad [1 \text{ mark}]$$

□

45.3.2 Runge-Kutta Method of Order Three

In the previous subsection, we failed to obtain order 3 with the approximation (45.22). But what about the approximation

$$f(t + \alpha_1, x + \delta_1 f(t + \alpha_2, x + \delta_2 f(t, y)))? \quad (45.24)$$

A German mathematician Karl Heun (1859–1929) obtained the *Heun's third order method* after tedious calculation.

Method 45.4 (Heun's Third Order Method). For $i = 0, 1, \dots, N - 1$:

$$\begin{aligned} \mathbf{k}_1 &= \mathbf{f}(t_i, \mathbf{x}_i), \\ \mathbf{k}_2 &= \mathbf{f}(t_i + \frac{h}{3}, \mathbf{x}_i + \frac{h}{3}\mathbf{k}_1), \\ \mathbf{k}_3 &= \mathbf{f}(t_i + \frac{2h}{3}, \mathbf{x}_i + \frac{2h}{3}\mathbf{k}_2), \\ \mathbf{x}_{i+1} &= \mathbf{x}_i + \frac{h}{4}(\mathbf{k}_1 + 3\mathbf{k}_3). \end{aligned} \quad (45.25)$$

Example 45.24. The Heun's Third Order method (45.25) can be implemented in Fortran as follow.

```
1  procedure(funcn) :: f
2  integer, intent(in) :: neq
3  double precision, intent(in) :: x(neq), t, tout
4  double precision :: heun3(neq), h, k1(neq), k2(neq), k3(neq)
5  h = tout - t
6  k1 = f(neq, t, x)
7  k2 = f(neq, t+h/3d0, x+h*k1/3d0)
8  k3 = f(neq, t+2d0*h/3d0, x+2d0*h*k2/3d0)
9  heun3 = x + h*(k1+3d0*k3)/4d0
10 end function
```

Use the Fortran function with $h = 0.2$ to calculate x_1 for the system (45.7) and generate a table for the time domain $0 \leq t \leq 2$. Write a script to plot the absolute error against the time.

Solution. $x_0 = 0.5$, $t_0 = 0$

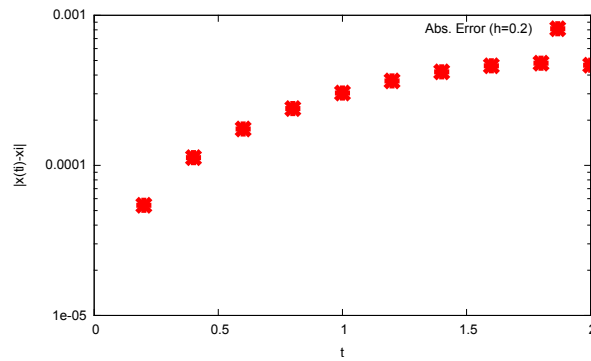
$$k_1 = x_0 - t_0^2 + 1 = 1.5$$

$$k_2 = f(\frac{0.2}{3}, x_0 + \frac{0.2}{3}k_1) = f(\frac{0.2}{3}, 0.6) = 0.6 - (\frac{0.2}{3})^2 + 1 = 1.59555556$$

$$k_3 = f\left(\frac{0.4}{3}, 0.5 + \frac{2 \times 0.2}{3}(1.59555556)\right) = f(0.13333333, 0.71274074) = 1.69496296$$

$$x_1 = x_0 + \frac{0.2}{4}(k_1 + 3k_3) = 0.5 + \frac{0.2}{4}(1.5 + 3(1.69496296)) = 0.82924444$$

t_i	x_i
0.0	0.500000000000
0.2	0.829244444444
0.4	1.213974992593
0.6	1.648765902064
0.8	2.126990532832
1.0	2.640555548543
1.2	3.179576287732
1.4	3.731980283861
1.6	4.283023031134
1.8	4.814696573136
2.0	5.305007192434



eg:heun3

Example 45.25. Write a program to apply the Heun's third order method with $h = 0.1$ to solve (45.8) for $0 \leq t \leq 10$.

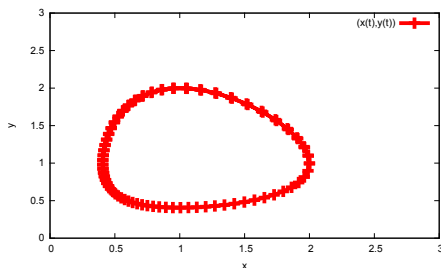
Solution:

```

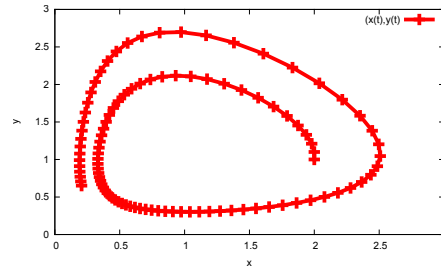
1 program eg_heun2
2   use iso_c_binding, only : c_double
3   use textbook_odeivp, only : heun3, lotka_volterra
4   implicit none
5   integer, parameter :: neq = 2
6   real(c_double), parameter :: h=0.1d0, tend=10d0
7   real(c_double) :: t, tnext, x(neq)
8   integer :: i
9   t = 0d0; x = [2d0,1d0]
10  print "(a5, a34)", "t", "x"
11  do i=1,int((tend-t)/h)
12    print 100, t, x
13    tnext = t+h
14    x = heun3(neq, lotka_volterra, x, t, tnext)
15    t = tnext
16  end do
17  print 100, t, x
18  100 format(f5.2, 2f17.12)
19 end program eg_heun2

```

If we plot the graph of Heun's third order method and compare it to the graph of Euler method, we will see how poorly Euler method performed.



Huen's Third Order



Euler method (Example 45.11)

45.3.3 Runge-Kutta Methods of Order Four

Although $T^{(3)}(t, y)$ can be approximated with error $O(h^3)$ by an expression of the form (45.24), however, (variations of) Runge-Kutta of order four, with higher order compare to the previous methods, is often used in practise.

m:rk4

Method 45.5 (Runge-Kutta Order Four Method). For $i = 0, 1, \dots, N - 1$,

$$\begin{aligned} \mathbf{k}_1 &= \mathbf{f}(t_i, \mathbf{x}_i), \\ \mathbf{k}_2 &= \mathbf{f}(t_i + \frac{h}{2}, \mathbf{x}_i + \frac{h}{2}\mathbf{k}_1), \\ \mathbf{k}_3 &= \mathbf{f}(t_i + \frac{h}{2}, \mathbf{x}_i + \frac{h}{2}\mathbf{k}_2), \\ \mathbf{k}_4 &= \mathbf{f}(t_{i+1}, \mathbf{x}_i + h\mathbf{k}_3), \\ \mathbf{x}_{i+1} &= \mathbf{x}_i + \frac{h}{6}(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4). \end{aligned} \quad (45.26) \quad \text{e:rk4}$$

This method has local truncation error $O(h^4)$, provided the solution $\mathbf{x}(t)$ has five continuous derivatives.

burden-faires:2

Example 45.26. By applying the fourth order Runge-Kutta method with $h = 0.2$ to (45.7),

1. find x_1 ;

Solution. $x_0 = 0.5, \quad t_0 = 0, \quad t_i = 0.2i$
 $k_1 = f(t_0, x_0) = 0.5 - 0^2 + 1 = 1.5$
 $k_2 = f(t_0 + \frac{0.2}{2}, x_0 + \frac{0.2}{2}k_1) = f(0.1, 0.65) = 0.65 - 0.1^2 + 1 = 1.64$
 $k_3 = f(t_0 + \frac{0.2}{2}, x_0 + \frac{0.2}{2}k_2) = f(0.1, 0.664) = 0.664 - 0.1^2 + 1 = 1.654$
 $k_4 = f(t_1, x_0 + 0.2k_3) = f(0.2, 0.5 + 0.2(1.654)) = 1.7908$
 $x_1 = x_0 + \frac{0.2}{6}(1.5 + 2(1.64 + 1.654) + 1.7908) = 0.82929333$

□

2. Write a Fortran function for Runge-Kutta 4 method (45.26) based on Example 45.20.

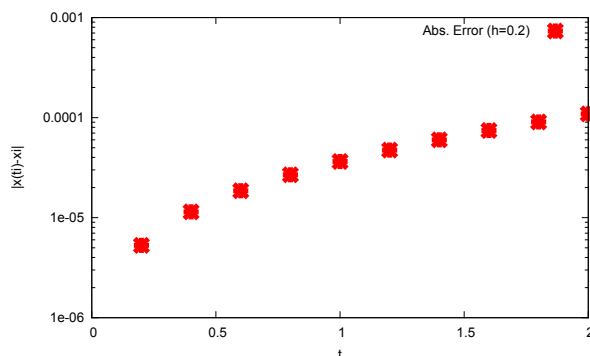
Solution:

```
1  procedure (funcn) :: ff
2  integer, intent(in) :: neq
3  double precision, intent(in) :: x(neq), t, tout
4  double precision, dimension(neq) :: rungekutta4, k1, k2, k3, k4
5  double precision :: h
6  h = tout - t
7  k1 = ff(neq, t, x)
8  k2 = ff(neq, t+h/2d0, x+h*k1/2d0)
9  k3 = ff(neq, t+h/2d0, x+h*k2/2d0)
10 k4 = ff(neq, t+h, x+h*k3)
11 rungekutta4 = x + h*(k1+2d0*(k2+k3)+k4)/6d0
12 end function
```

3. Use the Fortran to perform the calculation for $0 \leq t \leq 2$ and plot the (global) absolute error against the time.

Solution: By applying Runge-Kutta order 4 method 45.5, we obtain

t_i	x_i
0.00	0.500000000000
0.20	0.829293333333
0.40	1.214076210667
0.60	1.648922017042
0.80	2.127202684948
1.00	2.640822692729
1.20	3.179894170232
1.40	3.732340072855
1.60	4.283409498318
1.80	4.815085694579
2.00	5.305363000693



Example 45.27 (April 2014 Final Exam Question 2(c)). Use the fourth order Runge-Kutta method

$$\begin{cases} w_0 = x_0, \\ k_1 = hf(t_i, w_i), \\ k_2 = hf(t_i + \frac{h}{2}, w_i + \frac{1}{2}k_1), \\ k_3 = hf(t_i + \frac{h}{2}, w_i + \frac{1}{2}k_2), \\ k_4 = hf(t_{i+1}, w_i + k_3), \\ w_{i+1} = w_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4), \quad i = 0, 1, \dots, N-1. \end{cases}$$

with $h = 0.1$ to approximate the solution of the following IVP at $t = 1.1$ to at least 5 decimal places:

$$t^2 x' = x^2 + tx, \quad 1 \leq t \leq 1.1, \quad x(1) = 2. \quad (6 \text{ marks})$$

Solution: First, we need to transform the IVP to the form

$$x' = f(t, x) = \frac{x^2 + tx}{t^2} = \left(\frac{x}{t}\right)^2 + \frac{x}{t}. \quad [1 \text{ mark}]$$

$$w_0 = x_0 = 2$$

$$k_1 = (0.1)f(t_0, w_0) = 0.6 \quad [1 \text{ mark}]$$

$$k_2 = (0.1)f\left(t_0 + \frac{0.1}{2}, w_0 + \frac{1}{2}k_1\right) = 0.69887 \quad [1 \text{ mark}]$$

$$k_3 = (0.1)f\left(t_0 + \frac{0.1}{2}, w_0 + \frac{1}{2}k_2\right) = 0.72442 \quad [1 \text{ mark}]$$

$$k_4 = (0.1)f(t_1, w_0 + k_3) = 0.86110 \quad [1 \text{ mark}]$$

$$w_1 = w_0 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) = 2.7179467 \quad [1 \text{ mark}]$$

Example 45.28 (Tutorial 6, Q4). Consider the initial value problem

$$x' = -5x + 5t^2 + 2t, \quad 0 \leq t \leq 1, \quad x(0) = 1.$$

With a step size of $h = 0.1$, solve the IVP for $0 \leq t \leq 1$ by using the following numerical methods. Calculate x_1 by hand and the rest by computer program.

1. Midpoint method (Runge-Kutta's method of order 2)

Solution.

$$k_1^{(0)} = f(t_0, x_0) = f(0, 1) = -5$$

$$x_1 = x_0 + 0.1f\left(t_0 + \frac{0.1}{2}, x_0 + \frac{0.1}{2}k_1^{(0)}\right) = 1 + 0.1f(0.05, 0.75)$$

$$= 1 + 0.1(-5(0.75) + 5(0.05^2) + 2(0.05)) = 1 - 0.36375 = 0.63625$$

$$t = 0.00 \quad x_1 = 1.000000000000$$

$$t = 0.10 \quad x_1 = 0.636250000000$$

$$t = 0.20 \quad x_1 = 0.432656250000$$

$$t = 0.50 \quad x_1 = 0.348382873535$$

$$t = 0.60 \quad x_1 = 0.422739295959$$

$$t = 0.90 \quad x_1 = 0.827836742178$$

$$t = 1.00 \quad x_1 = 1.012397963861$$

□

2. Modified Euler method (Runge-Kutta's method of order 2)

Solution.

$$k_1^{(0)} = f(t_0, x_0) = -5$$

$$k_2^{(0)} = f(t_1, x_0 + hk_1^{(0)}) = f(0.1, 1 + 0.1(-5)) = -2.25$$

$$x_1 = x_0 + \frac{0.1}{2}(k_1^{(0)} + k_2^{(0)}) = 1 + 0.05(-5 - 2.25) = 0.6375$$

```
t= 0.00 x1= 1.00000000000000
t= 0.10 x1= 0.63750000000000
t= 0.20 x1= 0.43468750000000
t= 0.50 x1= 0.351398315430
t= 0.60 x1= 0.425873947144
t= 0.90 x1= 0.831121569127
t= 1.00 x1= 1.015700980704
```

□

3. Heun's third order method (Runge-Kutta's method of order 3)

Solution.

$$k_1 = f(t_0, x_0) = -5$$

$$k_2 = f(t_0 + \frac{0.1}{3}, x_0 + \frac{0.1}{3}k_1) = f(0.03333333, 0.83333333) = -4.09444444$$

$$k_3 = f(t_0 + \frac{2(0.1)}{3}, x_0 + \frac{2(0.1)}{3}k_2) = f(0.06666666, 0.72703704) = -3.47962966$$

$$x_1 = x_0 + \frac{h}{4}(k_1 + 3k_3) = 1 + \frac{0.1}{4}(-5 + 3(-3.47962966)) = 0.61402778$$

```
t= 0.00 x1= 1.00000000000000
t= 0.10 x1= 0.61402777777778
t= 0.20 x1= 0.404794560185
t= 0.50 x1= 0.330175128958
t= 0.60 x1= 0.408300251523
t= 0.90 x1= 0.820378220556
t= 1.00 x1= 1.006131286030
```

□

4. Runge-Kutta's method of order 4

Solution.

$$k_1 = f(t_0, x_0) = -5$$

$$k_2 = f(t_0 + \frac{0.1}{2}, x_0 + \frac{0.1}{2}k_1) = f(0.05, 0.75) = -3.6375$$

$$k_3 = f(t_0 + \frac{0.1}{2}, x_0 + \frac{0.1}{2}k_2) = f(0.05, 0.818125) = -3.978125$$

$$k_4 = f(t_1, x_0 + 0.1k_3) = f(0.1, 0.6021875) = -2.7609375$$

$$x_1 = x_0 + \frac{0.1}{6}(k_1 + 2k_2 + 2k_3 + k_4) = 1 + \frac{0.1}{6}(-5 + 2(-3.6375 - 3.978125) - 2.7609375) = 0.616796875$$

```
t= 0.00 x1= 1.00000000000000
t= 0.10 x1= 0.616796875000
t= 0.20 x1= 0.408212687174
t= 0.50 x1= 0.332308425510
t= 0.60 x1= 0.409968393604
t= 0.90 x1= 0.821214136548
t= 1.00 x1= 1.006830452645
```

Example 45.29. Write a program to apply the Runge-Kutta's method of order 4 with $h = 0.1$ to solve the Lotka-Volterra system (45.8) for $0 \leq t \leq 10$.

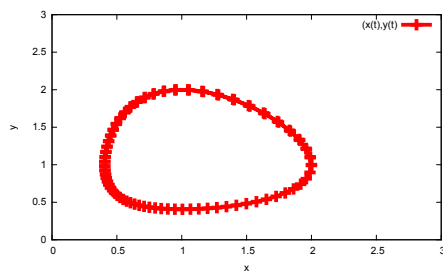
Solution:

```

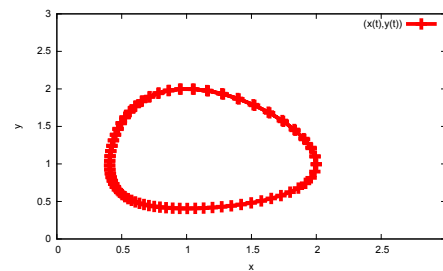
1 program eg_rk4_2
2   use iso_c_binding, only : c_double
3   use textbook_odeivp, only : rungekutta4, lotka_volterra
4   implicit none
5   integer, parameter :: neq = 2
6   real(c_double), parameter :: h=0.1d0, tend=10d0
7   real(c_double) :: t, tend, tnext, x(neq)
8   integer :: i
9   t = 0d0; tend=10d0; x = [2d0,1d0]
10  print "(a5, a34)", "t", "x"
11  do i=1,int((tend-t)/h)
12    print 100, t, x
13    tnext = t+h
14    x = rungekutta4(neq, lotka_volterra, x, t, tnext)
15    t = tnext
16  end do
17  print 100, t, x
18  100 format(f5.2, 2f17.12)
19 end program eg_rk4_2

```

If we plot the graph of Runge-Kutta's method of order 4 and compare it to the graph of Heun's third order method. We can see that both perform well with small h and for a short period.



Runge-Kutta's method of order 4



Heun's Method of Order 3 (Example 45.25)

Example 45.30 (<http://www.solvingproblems.inf.ethz.ch/>). The Weidlich and Haag's Population Migration Model is described by a 2-dimensional ODE-IVP below:

$$\begin{cases} \dot{x} = 2A(\sinh u - x \cosh u), \\ \dot{y} = 2A(\sinh v - y \cosh v), \end{cases} \quad x(0) = 0, \quad y(0) = 0.05, \quad (45.27)$$

where $u = a_m + b_m m_0 x + c_m n_0 y$, $v = a_n + b_n m_0 x + c_n n_0 y$, $A = 1/2$, $a_m = 0$, $b_m = 1.2 \times 10^{-4}$, $c_m = 0.5 \times 10^{-3}$, $m_0 = 10000$, $a_n = 0$, $b_n = - \times 10^{-4}$, $c_n = 1.2 \times 10^{-3}$, $n_0 = 1000$.

1. Write a script in either Python, Octave or R to apply the 4th order Runge-Kutta method with $h = 0.1$ to solve the IVP (45.27) for $0 \leq t \leq 60$.

Solution: A possible implementation in R is given below.

```

1 library(deSolve) # deSolve::..., rkMethod()
2
3 ### Population Migration Model
4 param = c(A=0.5, m0=10000, n0=1000, am=0, bm=1.2e-4,
5           cm=0.5e-3, an=0, bn=-1e-4, cn=1.2e-3)
6 state = c(x=0., y=0.05)
7 popul = function(t, state, parameters) {
8   with(as.list(c(state, parameters)), {

```

```

9      u = am + bm*m0*x + cm*n0*y
10     v = an + bn*m0*x + cn*n0*y
11     dx = 2.*A*(sinh(u) - x*cosh(u))
12     dy = 2.*A*(sinh(v) - y*cosh(v))
13     list(c(dx, dy))
14   })
15 }
16 times = seq(0,60,by=0.1)
17 out = ode(y=state,times=times,func=popul,parms=param,method="rk4")
18 diagnostics(out)
19 setEPS()
20 postscript("populationR.eps")
21 plot(out[, "x"], out[, "y"], pch=".", xlab="x", ylab="y")

```

Running it generates the following output

```

1 The return code : 0
2 The number of steps taken for the problem so far: 6000
3 The number of function evaluations for the problem so far: 24001
13 The number of error test failures of the integrator so far: 0
18 The order (or maximum order) of the method: 4

```

2. Write a Fortran program using the 4th order Runge-Kutta method to solve the IVP (45.27) ^{eg: population} for $0 \leq t \leq 60$.

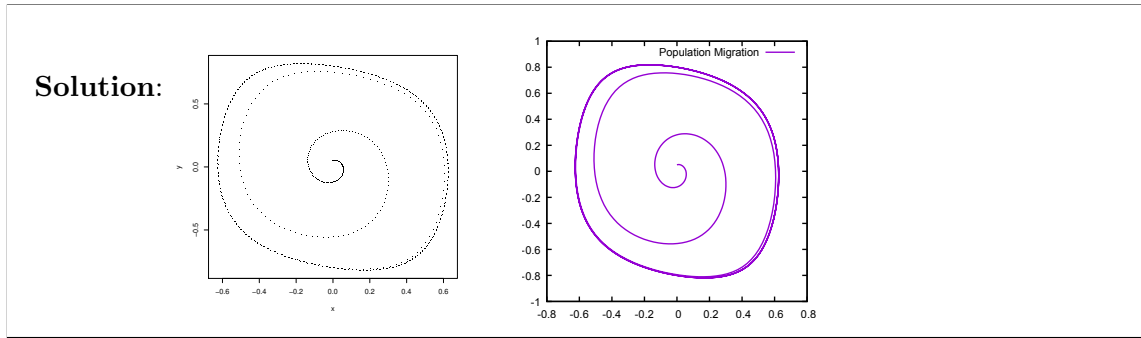
Solution: A implementation in Fortran 2003 linking against the function rungekutta4 in Example 45.26 ^{eg: Burden-faires:2} is given below.

```

1 program RK4_PopulationModel
2   use textbook_odeivp
3   implicit none
4   integer, parameter :: odedim = 2
5   double precision :: x(odedim) = (/ 0.0d0, 0.05d0 /)
6   double precision :: t=0d0, h=1d-1, tend=60d0, tnext
7   integer :: i, N
8   procedure(funcn) :: f
9   N = int((tend-t)/h)
10  do i=1,N
11    print 200, t, x(1), x(2)
12    tnext = t+h
13    x = rungekutta4(odedim,f,x,t,tnext)
14    t = tnext
15  end do
16  print 200, t, x(1), x(2)
17  200 format(f18.12,f18.12,f18.12)
18 end program RK4_PopulationModel
19 function f(neq, t, x)      ! Population Migration Model
20   integer, intent(in) :: neq
21   real(kind=8), intent(in) :: x(neq), t
22   real(kind=8) :: f(neq), u, v
23   real(kind=8), parameter :: A=0.5, m0=10000, n0=1000, am=0, bm=1.2e-4
24   real(kind=8), parameter :: cm=0.5e-3, an=0, bn=-1e-4, cn=1.2e-3
25   u = am + bm*m0*x(1) + cm*n0*x(2)
26   v = an + bn*m0*x(1) + cn*n0*x(2)
27   f(1) = 2.*A*(sinh(u) - x(1)*cosh(u))
28   f(2) = 2.*A*(sinh(v) - x(2)*cosh(v))
29 end function f

```

3. Compare the computational result / plot of R (or Python or octave) against Fortran.



dr.gohyk:t6q6

Example 45.31 (Tutorial 6, Q5). Consider the initial value problem

$$x'' - 3x' + 2x = t, \quad 0 \leq t \leq 1, \quad x(0) = 0, \quad x'(0) = 0.5.$$

With a step size of $h = 0.25$, solve the IVP for $0 \leq t \leq 1$ by using the following algorithms. Calculate $\mathbf{x}_1 = (x_1, y_1)$ by hand and the rest by computer program. Here y_1 is the approximation of $x'(0.25)$.

1. Midpoint method (Runge-Kutta's method of order 2)

Solution. First, we need to write the IVP as $\begin{cases} x' = y \\ y' = 3y - 2x + t \end{cases}$

$$\mathbf{k}_1 = \mathbf{f}(t_0, \mathbf{x}_0) = \mathbf{f}\left(0, \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}\right) = \begin{bmatrix} y_0 \\ -2x_0 + 3y_0 + t_0 \end{bmatrix} \Big|_{t_0=0, x_0=0, y_0=0.5} = \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix}$$

$$\mathbf{x}_1 = \mathbf{x}_0 + 0.25\mathbf{f}\left(t_0 + \frac{0.25}{2}, \mathbf{x}_0 + \frac{0.25}{2}\mathbf{k}_1\right) = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} + 0.25\mathbf{f}\left(0.125, \begin{bmatrix} 0.0625 \\ 0.6875 \end{bmatrix}\right)$$

$$= \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} + 0.25 \begin{bmatrix} y \\ -2x + 3y + t \end{bmatrix} \Big|_{t=0.125, x=0.0625, y=0.6875} = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} + 0.25 \begin{bmatrix} 0.6875 \\ 2.0625 \end{bmatrix} = \begin{bmatrix} 0.171875 \\ 1.015625 \end{bmatrix}$$

t	x	y
0.00	0.000000000000	0.500000000000
0.25	0.171875000000	1.015625000000
0.50	0.518066406250	1.998535156250
0.75	1.188308715820	3.781570434570
1.00	2.437391757965	6.917067050934

2. Modified Euler method (Runge-Kutta's method of order 2)

Solution.

$$\mathbf{k}_1 = \mathbf{f}(t_0, \mathbf{x}_0) = \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix}$$

$$\mathbf{k}_2 = \mathbf{f}\left(t_1, \mathbf{x}_0 + h\mathbf{k}_1\right) = \mathbf{f}\left(0.25, \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} + \frac{1}{4} \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix}\right)$$

$$= \begin{bmatrix} y \\ -2x + 3y + t \end{bmatrix} \Big|_{t=0.25, x=0.125, y=0.875} = \begin{bmatrix} 0.875 \\ 2.625 \end{bmatrix}$$

$$\mathbf{x}_1 = \mathbf{x}_0 + \frac{0.25}{2}[\mathbf{k}_1 + \mathbf{k}_2] = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} + \frac{0.25}{2} \left(\begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix} + \begin{bmatrix} 0.875 \\ 2.625 \end{bmatrix} \right) = \begin{bmatrix} 0.171875 \\ 1.015625 \end{bmatrix}$$

t	x	y
0.00	0.000000000000	0.500000000000
0.25	0.171875000000	1.015625000000

0.50	0.518066406250	1.998535156250
0.75	1.188308715820	3.781570434570
1.00	2.437391757965	6.917067050934

□

3. Heun's third order method (Runge-Kutta's method of order 3)

Solution.

$$\mathbf{k}_1 = \mathbf{f}(t_0, \mathbf{x}_0) = \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix}$$

$$\mathbf{k}_2 = \mathbf{f}\left(t_0 + \frac{0.25}{3}, \mathbf{x}_0 + \frac{0.25}{3}\mathbf{k}_1\right) = \mathbf{f}\left(\frac{1}{12}, \begin{bmatrix} 0.0416667 \\ 0.625 \end{bmatrix}\right) = \begin{bmatrix} 0.625 \\ 1.875 \end{bmatrix}$$

$$\mathbf{k}_3 = \mathbf{f}\left(t_0 + \frac{2 \times 0.25}{3}, \mathbf{x}_0 + \frac{2 \times 0.25}{3}\mathbf{k}_2\right) = \mathbf{f}\left(\frac{1}{6}, \begin{bmatrix} 0.104167 \\ 0.8125 \end{bmatrix}\right) = \begin{bmatrix} 0.8125 \\ 2.39583333 \end{bmatrix}$$

$$\mathbf{x}_1 = \mathbf{x}_0 + \frac{0.25}{4}(\mathbf{k}_1 + 3\mathbf{k}_3) \approx \begin{bmatrix} 0.183594 \\ 1.04297 \end{bmatrix}$$

t		x
0.00	0.000000000000	0.500000000000
0.25	0.183593750000	1.042968750000
0.50	0.559153238932	2.090728759766
0.75	1.294405062993	4.013039774365
1.00	2.677817502438	7.430882964904

□

4. Runge-Kutta's method of order 4

Solution. $\mathbf{k}_1 = \mathbf{f}(t_0, \mathbf{x}_0) = \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix}$

$$\mathbf{k}_2 = \mathbf{f}\left(t_0 + \frac{0.25}{2}, \mathbf{x}_0 + \frac{0.25}{2}\mathbf{k}_1\right) = \mathbf{f}(0.125, \begin{bmatrix} 0.0625 \\ 0.6875 \end{bmatrix}) = \begin{bmatrix} 0.6875 \\ 2.0625 \end{bmatrix}$$

$$\mathbf{k}_3 = \mathbf{f}\left(t_0 + \frac{0.25}{2}, \mathbf{x}_0 + \frac{0.25}{2}\mathbf{k}_2\right) = \mathbf{f}(0.125, \begin{bmatrix} 0.0859375 \\ 0.7578125 \end{bmatrix}) = \begin{bmatrix} 0.7578125 \\ 2.2265625 \end{bmatrix}$$

$$\mathbf{k}_4 = \mathbf{f}(t_1, \mathbf{x}_0 + 0.25\mathbf{k}_3) = \mathbf{f}(0.25, \begin{bmatrix} 0.1894531 \\ 1.0566406 \end{bmatrix}) = \begin{bmatrix} 1.0566406 \\ 3.0410156 \end{bmatrix}$$

$$\mathbf{x}_1 = \mathbf{x}_0 + \frac{0.25}{6}(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4) = \begin{bmatrix} 0.1853027 \\ 1.0466309 \end{bmatrix}$$

0.00	0.000000000000	0.500000000000
0.25	0.185302734375	1.046630859375
0.50	0.564960440000	2.102970083555
0.75	1.309094482923	4.043626004720
1.00	2.710662834161	7.498640577123

□

45.4 Error Control and Runge-Kutta-Fehlberg Method

In practice, adaptive methods are used in complement to the Runge-Kutta methods mentioned in the previous section. We have seen that adaptive methods for numerical quadrature in Topic 5 makes appropriate use of varying step sizes for integral approximations. They incorporate in the step-size procedure an estimate of the truncation error that does not require the approximation of the higher derivatives of the function. They adapt the number and position of the nodes used

in the approximation to ensure that the truncation error is kept within a specified bound.

There is a close connection between the problem of approximating the value of a definite integral and that of approximating the solution to an IVP (45.2) as follows:

$$(45.2) \Leftrightarrow \mathbf{x}(t) - \mathbf{x}(t_0) = \int_{t_0}^t \mathbf{f}(\tau, \mathbf{x}(\tau)) d\tau.$$

So it may not be surprising that there are adaptive methods for approximating the solutions to IVPs and that these methods are not only efficient, but also incorporate the control of error.

According to the main reference Burden and Faires [2015], give a tolerance ϵ , one popular technique that uses the error control

$$q \leq \left[\frac{\epsilon h}{|\tilde{x}_{i+1} - x_{i+1}|} \right]^{1/n}$$

is the Runge-Kutta-Fehlberg method. This technique uses a Runge-Kutta method with local truncation error of order five,

$$\tilde{\mathbf{x}}_{i+1} = \mathbf{x}_i + h \left(\frac{16}{135} \mathbf{k}_1 + \frac{6656}{12825} \mathbf{k}_3 + \frac{28561}{56430} \mathbf{k}_4 - \frac{9}{50} \mathbf{k}_5 + \frac{2}{55} \mathbf{k}_6 \right), \quad (45.28)$$

to estimate the local error in a Runge-Kutta method of order four given by

$$\mathbf{x}_{i+1} = \mathbf{x}_i + h \left(\frac{25}{216} \mathbf{k}_1 + \frac{1408}{2565} \mathbf{k}_3 + \frac{2197}{4104} \mathbf{k}_4 - \frac{1}{5} \mathbf{k}_5 \right), \quad (45.29)$$

where the coefficient equations are

$$\begin{aligned} \mathbf{k}_1 &= \mathbf{f}(t_i, \mathbf{x}_i), \\ \mathbf{k}_2 &= \mathbf{f}\left(t_i + \frac{h}{4}, \mathbf{x}_i + \frac{h}{4} \mathbf{k}_1\right), \\ \mathbf{k}_3 &= \mathbf{f}\left(t_i + \frac{3h}{8}, \mathbf{x}_i + \frac{3}{32} h \mathbf{k}_1 + \frac{9}{32} h \mathbf{k}_2\right), \\ \mathbf{k}_4 &= \mathbf{f}\left(t_i + \frac{12h}{13}, \mathbf{x}_i + \frac{1932}{2197} h \mathbf{k}_1 - \frac{7200}{2197} h \mathbf{k}_2 + \frac{7296}{2197} h \mathbf{k}_3\right), \\ \mathbf{k}_5 &= \mathbf{f}\left(t_i + h, \mathbf{x}_i + \frac{439}{216} h \mathbf{k}_1 - 8h \mathbf{k}_2 + \frac{3680}{513} h \mathbf{k}_3 - \frac{845}{4104} h \mathbf{k}_4\right), \\ \mathbf{k}_6 &= \mathbf{f}\left(t_i + \frac{h}{2}, \mathbf{x}_i - \frac{8}{27} h \mathbf{k}_1 + 2\mathbf{k}_2 - \frac{3544}{2565} h \mathbf{k}_3 + \frac{1859}{4104} h \mathbf{k}_4 - \frac{11}{40} h \mathbf{k}_5\right). \end{aligned}$$

Both the ode45 and ode23 in Matlab implements the Runge-Kutta-Fehlberg method. An implementation of the pseudo-code of the Runge-Kutta-Fehlberg method in Burden and Faires [2015, p.297] in Fortran is given below.

```

1      integer, intent(in) :: neq ! neq = length(x0)
2      procedure(funcn) :: f
3      double precision :: t0, tN, hmin, hmax, t, h, R, delta
4      double precision, dimension(neq) :: x0, x, k1, k2, k3, k4, k5, k6
5      double precision, external :: dnrn2 ! BLAS Level 1 (Topic 2)
6      integer :: i
7      t = t0
8      h = hmax
9      x = x0
10     call disp(neq,t,x)
11     do i=1,maxIter
12         k1 = f(neq,t, x)
13         k2 = f(neq,t+h/4., x+h*k1/4.)
14         k3 = f(neq,t+3d0/8*h, x+3d0/32*h*k1+9d0/32*h*k2)
15         k4 = f(neq,t+12d0/13*h, x+1932d0/2197*h*k1 -
16                7200d0/2197*h*k2 + 7296d0/2197*h*k3)
17         k5 = f(neq,t+h, x+439d0/216*h*k1 - 8d0*h*k2 +
18                3680d0/513*h*k3 - 845d0/4104*h*k4)

```

```

19      k6 = f(neq,t+0.5*h, x-8d0/27*h*k1 + 2d0*h*k2 -      &
20              3544d0/2565*h*k3 + 1859d0/4104*h*k4 - 11d0/40*h*k5)
21      R = dnrms2(neq, k1/360d0 - 128d0/4275d0*k3 - 2197d0/75240d0*k4 + &
22              k5/50d0 + 2d0/55d0*k6,1) ! norm(rhs(6.28) - rhs(6.29))
23      if (R .le. tol) then
24          t = t + h
25          x = x + h*(25d0/216d0*k1 + 1408d0/2565d0*k3 + &
26              2197d0/4104*k4 - k5/5d0)
27          call disp(neq,t,x)
28      end if
29      delta = 0.84*(tol/R)**(1d0/4d0)
30      ! The approximation is bad, reduced step size
31      if (delta.le.0.1) then
32          h = 0.1*h
33      else if (delta .ge. 4) then
34          h = 4*h
35      else
36          h = delta*h
37      end if
38      if (h .gt. hmax) h = hmax
39      if (t>=tN) then
40          exit
41      else if ((t+h)>tN) then
42          h = tN-t
43      else if (h<hmin) then
44          print *, "Minimum h exceeded!"
45          exit
46      end if
47  end do
48 end subroutine

```

An advantage to this method is that only six evaluations of $\mathbf{f}$ are required per step. Arbitrary Runge-Kutta methods of order four and five used together require at least four evaluations of $\mathbf{f}$ for the fourth-order method and an additional six for the fifth-order method, for a total of at least ten function evaluations. So the Runge-Kutta-Fehlberg method has at least a 40% decrease in the number of function evaluations over the use of a pair of arbitrary fourth- and fifth-order methods.

In the error-control theory, an initial value of h at the i th step is used to find the first values of $\mathbf{x}_{i+1}$ and $\tilde{\mathbf{x}}_{i+1}$, which leads to the determination of q for that step, and then the calculations were repeated. This procedure requires twice the number of function evaluations per step as without the error control. In practice, the value of q to be used is chosen somewhat differently in order to make the increased function-evaluation cost worthwhile. The value of q determined at the i th step is used for two purposes:

- When $q < 1$: to reject the initial choice of h at the i th step and repeat the calculations using qh ;
- When $q \geq 1$: to accept the computed value at the i th step using the step size h , but change the step size to qh for the $(i + 1)$ st step.

Example 45.32. Study $\dot{x} = 1 + x^2$, $x(0) = 0$ for $0 \leq t \leq 2$ for all the numerical integration methods using $h = 0.2$ discussed so far and compare them with Runge-Kutta-Fehlberg method with $h_{\min} = 10^{-6}$ and $h_{\max} = 0.2$.

Solution: This is an interesting ODE where we can find the exact using the knowledge from Calculus I and II:

$$\frac{dx}{dt} = 1 + x^2 \Rightarrow \int_{x_0}^x \frac{1}{1 + x^2} dx = \int_{t_0}^t dt \Rightarrow \tan^{-1} x - \tan^{-1} x_0 = t - t_0.$$

Since $t_0 = x_0 = 0$, we have $\tan^{-1} x = t \Rightarrow \boxed{x = \tan t}$, $0 \leq x < \frac{\pi}{2}$.

A Fortran program to compute all the numerical integration methods discussed so far

with $h = 0.2$ as well as the Runge-Kutta-Fehlberg method is listed below.

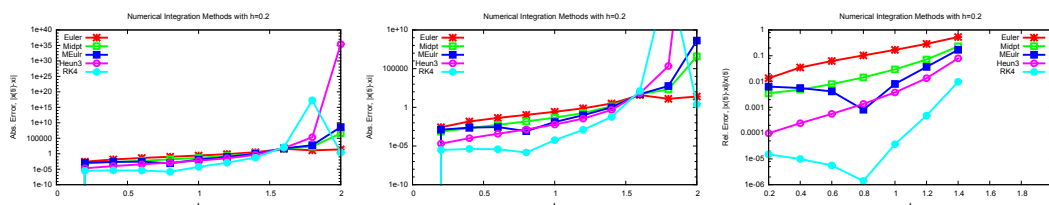
```

1 program eg_tangent
2   use iso_c_binding, only : c_double
3   use textbook_odeivp
4   implicit none
5   procedure(funcn) :: tangentode
6   integer, parameter :: neq=1
7   integer :: i
8   real(c_double) :: h, tend, t, tnext
9   real(c_double), dimension(neq) :: x1, x2, x3, x4, x5
10  t = 0d0; tend=2d0; h=2d-1
11  x1=[0d0]; x2=[0d0]; x3=[0d0]; x4=[0d0]; x5=[0d0]
12  print 200, "t", "euler", "midpoint", "modified euler", &
13    "heun (3rd)", "runge-kutta"
14  do i=1,int((tend-t)/h)
15    print 100, t, x1, x2, x3, x4, x5
16    tnext = t+h
17    x1 = euler(neq,tangentode,x1,t,tnext)
18    x2 = midpoint(neq,tangentode,x2,t,tnext)
19    x3 = modeuler(neq,tangentode,x3,t,tnext)
20    x4 = heun3(neq,tangentode,x4,t,tnext)
21    x5 = rungekutta4(neq,tangentode,x5,t,tnext)
22    t = tnext
23  end do
24  print 100, t, x1, x2, x3, x4, x5
25  print "(a29)", "# Runge-Kutta-Fehlberg method"
26  call setMaxIter(1000)
27  call rk_fehlberg(neq,tangentode,[0d0],0d0,tend,1d-6,0.2d0)
28  100 format(f5.2, d15.6, d15.6, d15.6, d15.6, d15.6)
29  200 format('#',a4, a15, a15, a15, a15, a15)
30 end program eg_tangent
31 function tangentode(neq,t,x)
32   use iso_c_binding, only : c_double
33   integer, intent(in) :: neq
34   real(c_double) :: tangentode(neq), t, x(neq)
35   tangentode = 1d0*x*x
36 end function

```

We find that at $t = 1.57$, i.e. a value close to $t = \frac{\pi}{2}$, the Runge-Kutta-Fehlberg failed to converge!

Plotting the error of various methods (excluding the Runge-Kutta-Fehlberg because it will comply with the error control), we have the absolute error (left and middle figures) and relative error (right figure) shown below.



Note that from the the absolute error plots we can see that Euler method performed the worst and Runge-Kutta order 4 method performed better but the error curves are all cluttered together. By re-scaling using relative error, we can see that Runge-Kutta order 4 method is clearly the best. The relative error for $t = 1.6$ to 2.0 were not plotted probably due to not a number being generated. The script to generate the curves are shown below.

```

1 y(x) = tan(x)
2 set border linewidth 0.2
3 set terminal pdfcairo font "Arial,12" linewidth 8
4 set key top left
5 set output "eg-tan.pdf"

```

```

6  set datafile fortran
7  set logscale y
8  set xlabel "t"
9  set ylabel "Abs. Error, |x(ti)-xi|"
10 set title "Numerical Integration Methods with h=0.2"
11 plot "eg-tan.dat" using 1:(abs(y($1)-$2)) t "Euler" w lp pt 3, \
12      "" using 1:(abs(y($1)-$3)) t "Midpt" w lp pt 4, \
13      "" using 1:(abs(y($1)-$4)) t "MEulr" w lp pt 5, \
14      "" using 1:(abs(y($1)-$5)) t "Heun3" w lp pt 6, \
15      "" using 1:(abs(y($1)-$6)) t "RK4" w lp pt 7
16 set output # close file
17 set key top right
18 set output "eg-tan3.pdf"
19 set ylabel "Rel. Error, |x(ti)-xi|/x(ti)"
20 plot "eg-tan.dat" using 1:(abs(y($1)-$2)/y($1)) t "Euler" w lp pt 3, \
21      "" using 1:(abs(y($1)-$3)/y($1)) t "Midpt" w lp pt 4, \
22      "" using 1:(abs(y($1)-$4)/y($1)) t "MEulr" w lp pt 5, \
23      "" using 1:(abs(y($1)-$5)/y($1)) t "Heun3" w lp pt 6, \
24      "" using 1:(abs(y($1)-$6)/y($1)) t "RK4" w lp pt 7
25 set output "eg-tan2.pdf"
26 set key top left
27 set yrange [1e-10:1e10]
28 set ylabel "Abs. Error, |x(ti)-xi|"
29 plot "eg-tan.dat" using 1:(abs(y($1)-$2)) t "Euler" w lp pt 3, \
30      "" using 1:(abs(y($1)-$3)) t "Midpt" w lp pt 4, \
31      "" using 1:(abs(y($1)-$4)) t "MEulr" w lp pt 5, \
32      "" using 1:(abs(y($1)-$5)) t "Heun3" w lp pt 6, \
33      "" using 1:(abs(y($1)-$6)) t "RK4" w lp pt 7

```

45.5 Multi-Step Formulae

sec:multistep

Explicit Runge-Kutta methods are single-step methods. They approximate the next mesh point t_{i+1} by using information from only one of the current mesh points, t_i and make use functional evaluation information at points between t_i and t_{i+1} . However, they do not retain the information at points in the interval (t_i, t_{i+1}) for direct use in future approximations. All the information used by these methods is obtained within the subinterval over which the solution is being approximated.

Since the approximation solution is available at each of the mesh points $t_0, t_1, \dots, t_i$ before the approximation at t_{i+1} is obtained, and because the error $\|\mathbf{x}_j - \mathbf{x}(t_j)\|$ tends to increase with j , it seems reasonable to develop methods that use these more accurate previous data when approximating the solution at t_{i+1} .

Methods using the approximation at more than one previous mesh point to determine the approximation at the next point are called *multi-step methods*. There are many multi-step IVP solvers using the multi-step formulae, such as ODEPACK (using lsode algorithm by A. C. Hindmarsh), vode [Brown et al., 1989] and desk [Brenan et al., 1996].

Definition 45.6. An m -step multi-step method (e:multi-step (45.5)) for solving the ODE-IVP (e:ode-ivp (45.2)) has a difference equation for finding the approximation $\mathbf{x}_{i+1}$ at the mesh point t_{i+1} represented by the following equation, where m is an integer greater than 1:

$$\begin{aligned} \mathbf{x}_{i+1} = & a_{m-1}\mathbf{x}_i + a_{m-2}\mathbf{x}_{i-1} + \dots + a_0\mathbf{x}_{i+1-m} + h[b_m\mathbf{f}(t_{i+1}, \mathbf{x}_{i+1}) + b_{m-1}\mathbf{f}(t_i, \mathbf{x}_i) + \dots \\ & + b_0\mathbf{f}(t_{i+1-m}, \mathbf{x}_{i+1-m})], \quad i = m-1, m, \dots, N-1, \end{aligned} \quad (45.30)$$

where $h = (b-a)/N$, the $a_0, a_1, \dots, a_{m-1}$ and $b_0, \dots, b_m$ are constants, and the starting values $\mathbf{x}_0 = \mathbf{x}(t_0), \mathbf{x}_1 = \mathbf{x}(t_1), \dots, \mathbf{x}_{m-1} = \mathbf{x}(t_{m-1})$ are specified.

When $b_m = 0$ the method is called *explicit*, or *open*, since (e:burden-faires:5.23 (45.30)) then gives $\mathbf{x}_{i+1}$ explicitly in terms of previously determined values. When $b_m \neq 0$ the method is called *implicit*, or *closed*, since $\mathbf{x}_{i+1}$ occurs on both sides of (e:burden-faires:5.23 (45.30)) and is specified only implicitly.

e:burden-faires:

The starting values in a multi-step method must be specified, generally by assuming $\mathbf{x}_0 = \mathbf{x}(t_0)$ and generating the remaining values by a Runge-Kutta method.

To begin the derivation of a multi-step method, note that the solution to IVP (45.2) if integrated over the interval $[t_i, t_{i+1}]$ becomes

$$\int_{t_i}^{t_{i+1}} \dot{\mathbf{x}}(t) dt = \int_{t_i}^{t_{i+1}} \mathbf{f}(t, \mathbf{x}(t)) dt \Rightarrow \mathbf{x}(t_{i+1}) = \mathbf{x}(t_i) + \int_{t_i}^{t_{i+1}} \mathbf{f}(t, \mathbf{x}(t)) dt. \quad (45.31) \quad \text{eq:5.26}$$

Since we cannot integrate $\mathbf{f}(t, \mathbf{x}(t))$ without knowing $\mathbf{x}(t)$, the solution to the problem, we instead integrate an interpolating polynomial $\mathbf{P}(t)$ to $\mathbf{f}(t, \mathbf{x}(t))$ that is determined by some of the previously obtained data points $(t_0, \mathbf{x}_0), \dots, (t_i, \mathbf{x}_i)$. When we assume, in addition, that $\mathbf{x}(t_i) \approx \mathbf{x}_i$, (45.31) becomes

$$\mathbf{x}(t_{i+1}) \approx \mathbf{x}_i + \int_{t_i}^{t_{i+1}} \mathbf{P}(t) dt. \quad (45.32) \quad \text{eq:5.27}$$

Although Taylor series or any form of the interpolating polynomial can be used for the derivation, it is most convenient to use the Newton backward-difference formula.

Some commonly used multi-step methods which can be derived from (45.32) are

- *Adams-Bashforth* (explicit m -step) technique; and
- *Adams-Moulton* (implicit m -step) technique (used in ODEPACK).

Let $\mathbf{f}_i = \mathbf{f}(t_i, \mathbf{x}(t_i))$. To derive an Adams-Bashforth explicit m -step method, we form the backward difference polynomial $\mathbf{P}_{m-1}(t)$ through $(t_i, \mathbf{f}_i), (t_i, \mathbf{f}_{i-1}), \dots, (t_{i+1-m}, \mathbf{f}_{i+1-m})$. Since $\mathbf{P}_{m-1}(t)$ is an interpolatory polynomial of degree $m-1$, some value $\xi_i \in (t_{i+1-m}, t_i)$ exists with

$$\mathbf{f}(t, \mathbf{x}(t)) = \mathbf{P}_{m-1}(t) + \frac{1}{m!} \frac{d^m}{dt^m} \mathbf{f}(\xi_i, \mathbf{x}(\xi_i)) \underbrace{(t - t_i)(t - t_{i-1}) \cdots (t - t_{i+1-m})}_{\prod_{j=0}^{m-1} (t - t_{i-j})}.$$

Introducing the substitution $t = t_i + sh$, with $dt = hds$ into the above expression and integrate it gives

$$\int_{t_i}^{t_{i+1}} \mathbf{f}(t, \mathbf{x}(t)) dt = \sum_{k=0}^{m-1} \nabla^k \mathbf{f}(t_i, \mathbf{x}(t_i)) h (-1)^k \int_0^1 \binom{-s}{k} ds + \underbrace{\frac{h^{m+1}}{m!} \int_0^1 \mathbf{f}^{(m)}(\xi_i, \mathbf{x}(\xi_i)) \prod_{j=0}^{m-1} (s + j) ds}_{O(h^{m+1})}. \quad (45.33) \quad \text{e:adams-bashfo}$$

To derive an Adams-Moulton implicit $(m-1)$ -step method, we form the backward difference polynomial $\tilde{\mathbf{P}}_{m-2}(t)$ through $(t_{i+1}, \mathbf{f}_{i+1}), (t_i, \mathbf{f}_i), \dots, (t_{i+2-m}, \mathbf{f}_{i+2-m})$ and integrate $\tilde{\mathbf{P}}_{m-2}(t)$ over the interval $[t_i, t_{i+1}]$.

For any m , it is not difficult to derive the coefficients of the m -step Adams-Bashforth formula and $(m-1)$ -step Adams-Moulton formula, both having order of accuracy $p = m$ (see Def. 45.3). For Runge-Kutta methods, by contrast, there is no simple relation between the number of “stages” (i.e. function evaluations per step) and the attainable order of accuracy. The classical method with $s = 1, 2, 3, 4$ were known to Kutta in 1901 and have order $p = s$, but it not until 1963 that it was proved that $s = 6$ stages are required to achieve order $p = 5$ [Trefethen, 2008].

Example 45.33. Derive the 3-step Adams-Moulton formula for Equation (45.2). le:ode-ivp

Solution: By definition, $t_{i+1} - t_i = h$, $t_i - t_{i-1} = h$, $t_{i-1} - t_{i-2} = h$. Then (45.31) implies

$$\begin{aligned} & \mathbf{x}(t_{i+1}) - \mathbf{x}(t_i) \\ & \approx \int_{t_i}^{t_{i+1}} \left[\mathbf{f}_{i+1} + \frac{\mathbf{f}_{i+1} - \mathbf{f}_i}{h} (t - t_{i+1}) + \frac{\mathbf{f}_{i+1} - 2\mathbf{f}_i + \mathbf{f}_{i-1}}{2h^2} (t - t_{i+1})(t - t_i) \right. \\ & \quad \left. + \frac{\mathbf{f}_{i+1} - 3\mathbf{f}_i + 3\mathbf{f}_{i-1} - \mathbf{f}_{i-2}}{6h^3} (t - t_{i+1})(t - t_i)(t - t_{i-1}) \right] dt \\ & = \mathbf{f}_i \int_{t_i}^{t_{i+1}} dt + \frac{\mathbf{f}_i - \mathbf{f}_{i-1}}{h} \int_{t_i}^{t_{i+1}} (t - t_{i+1}) dt + \frac{\mathbf{f}_i - 2\mathbf{f}_{i-1} + \mathbf{f}_{i-2}}{2h^2} \int_{t_i}^{t_{i+1}} (t - t_{i+1})(t - t_i) dt \end{aligned}$$

$$\begin{aligned}
& + \frac{\mathbf{f}_{i+1} - 3\mathbf{f}_i + 3\mathbf{f}_{i-1} - \mathbf{f}_{i-2}}{6h^3} \int_{t_i}^{t_{i+1}} (t - t_{i+1})(t - t_i)(t - t_{i-1})dt \\
& = \mathbf{f}_i \left[t \right]_{t_{i-1}}^{t_i} + \frac{\mathbf{f}_{i+1} - \mathbf{f}_i}{h} \left[\frac{(t - t_{i+1})^2}{2} \right]_{t_i}^{t_{i+1}} + \frac{\mathbf{f}_{i+1} - 2\mathbf{f}_i + \mathbf{f}_{i-1}}{2h^2} \left[\frac{(t_i - t_{i+1})^3}{6} \right] + \frac{\mathbf{f}_{i+1} - 3\mathbf{f}_i + 3\mathbf{f}_{i-1} - \mathbf{f}_{i-2}}{6h^3} \\
& \quad \times \left[-2t_i^3 t_{i-1} + 6t_{i+1} t_i^2 t_{i-1} - 6t_{i+1}^2 t_i t_{i-1} + 2t_{i+1}^3 t_{i-1} + t_i^4 - 2t_{i+1} t_i^3 + 2t_{i+1}^3 t_i - t_{i+1}^4 \right] \\
& = h\mathbf{f}_{i+1} + \frac{\mathbf{f}_{i+1} - \mathbf{f}_i}{h} \left[-\frac{h^2}{2} \right] + \frac{\mathbf{f}_{i+1} - 2\mathbf{f}_i + \mathbf{f}_{i-1}}{2h^2} \left[-\frac{h^3}{6} \right] + \frac{\mathbf{f}_{i+1} - 3\mathbf{f}_i + 3\mathbf{f}_{i-1} - \mathbf{f}_{i-2}}{6h^3} \left[-\frac{3h^4}{12} \right] \\
& = h \left[\mathbf{f}_{i+1} - \frac{\mathbf{f}_{i+1}}{2} + \frac{\mathbf{f}_i}{2} - \frac{\mathbf{f}_{i+1}}{12} + \frac{\mathbf{f}_i}{6} - \frac{\mathbf{f}_{i-1}}{12} - \frac{\mathbf{f}_{i+1}}{24} + \frac{\mathbf{f}_i}{8} - \frac{\mathbf{f}_{i-1}}{8} + \frac{\mathbf{f}_{i-2}}{24} \right] \\
& = h \left[\frac{3}{8}\mathbf{f}_{i+1} + \frac{19}{24}\mathbf{f}_i - \frac{5}{24}\mathbf{f}_{i-1} + \frac{1}{24}\mathbf{f}_{i-2} \right].
\end{aligned}$$

Example 45.34 (Tutorial 6, Q6). Derive the three-step Adams-Bashforth formula for an ODE-IVP using the interpolating polynomial of Newton form from Topic 4.

Solution. Note that $t_i - t_{i-1} = t_{i-1} - t_{i-2} = h$ and $\mathbf{f}_i = \mathbf{f}(t_i, \mathbf{x}(t_i))$.

$$\begin{aligned}
\mathbf{x}_{i+1} & \approx \mathbf{x}(t_{i+1}) = \mathbf{x}(t_i) + \int_{t_i}^{t_{i+1}} \mathbf{f}(t, \mathbf{x}(t))dt \approx \mathbf{x}_i + \int_{t_i}^{t_{i+1}} \mathbf{P}[t_i, t_{i-1}, t_{i-2}](t)dt \\
& = \mathbf{x}_i + \int_{t_i}^{t_{i+1}} \left[\mathbf{f}(t_i, \mathbf{x}(t_i)) + \frac{\mathbf{f}(t_i, \mathbf{x}(t_i)) - \mathbf{f}(t_{i-1}, \mathbf{x}(t_{i-1}))}{t_i - t_{i-1}}(t - t_i) \right. \\
& \quad \left. + \frac{(t_{i-1} - t_{i-2})(\mathbf{f}_i - \mathbf{f}_{i-1}) - (t_i - t_{i-1})(\mathbf{f}_{i-1} - \mathbf{f}_{i-2})}{(t_i - t_{i-2})(t_i - t_{i-1})(t_{i-1} - t_{i-2})}(t - t_i)(t - t_{i-1}) \right] dt \\
& = \mathbf{x}_i + \int_{t_i}^{t_{i+1}} \left[\mathbf{f}_i + \frac{\mathbf{f}_i - \mathbf{f}_{i-1}}{h}(t - t_i) + \frac{(h(\mathbf{f}_i - \mathbf{f}_{i-1}) - h(\mathbf{f}_{i-1} - \mathbf{f}_{i-2}))}{2h \times h^2}(t - t_i)(t - t_{i-1}) \right] dt \\
& = \mathbf{x}_i + h\mathbf{f}_i + \frac{\mathbf{f}_i - \mathbf{f}_{i-1}}{h} \int_{t_i}^{t_{i+1}} (t - t_i)dt + \frac{(\mathbf{f}_i - 2\mathbf{f}_{i-1} + \mathbf{f}_{i-2})}{2h^2} \int_{t_i}^{t_{i+1}} (t - t_i)(t - t_{i-1})dt \\
& = \mathbf{x}_i + h\mathbf{f}_i + \frac{\mathbf{f}_i - \mathbf{f}_{i-1}}{h} \left[\frac{(t - t_i)^2}{2} \right]_{t_i}^{t_{i+1}} + \frac{(\mathbf{f}_i - 2\mathbf{f}_{i-1} + \mathbf{f}_{i-2})}{2h^2} \left[\frac{2h^3 + 3h^2(t_i - t_{i-1})}{6} \right] \\
& = \mathbf{x}_i + h\mathbf{f}_i + \frac{h}{2}(\mathbf{f}_i - \mathbf{f}_{i-1}) + \frac{5h}{12}(\mathbf{f}_i - 2\mathbf{f}_{i-1} + \mathbf{f}_{i-2}) = \mathbf{x}_i + \frac{h}{12}[23\mathbf{f}_i - 16\mathbf{f}_{i-1} + 5\mathbf{f}_{i-2}]
\end{aligned}$$

□

ires:p326:5.37

Example 45.35 (Tutorial 6, Q7). Derive the finite difference equation corresponding to the Adams-Moulton Two-Step Implicit Method.

Solution. Note that $t_{i+1} - t_i = h$. So $x_0 = x(t_0)$, $x_1 = x(t_1)$ and

$$\begin{aligned}
x(t_{i+1}) & \approx x(t_i) + \int_{t_i}^{t_{i+1}} \left\{ f(t_{i+1}, x_{i+1}) + \frac{f(t_{i+1}, x_{i+1}) - f(t_i, x_i)}{h}(t - t_{i+1}) \right. \\
& \quad \left. + \frac{f(t_{i+1}, x_{i+1}) - 2f(t_i, x_i) + f(t_{i-1}, x_{i-1})}{2h^2}(t - t_{i+1})(t - t_i) \right\} dt \\
& = x(t_i) + hf(t_{i+1}, x_{i+1}) + \frac{f(t_{i+1}, x_{i+1}) - f(t_i, x_i)}{h} \left[-\frac{h^2}{2} \right] \\
& \quad + \frac{f(t_{i+1}, x_{i+1}) - 2f(t_i, x_i) + f(t_{i-1}, x_{i-1})}{2h^2} \left[-\frac{h^3}{6} \right] \\
& \Rightarrow x_{i+1} = x_i + h[5f(t_{i+1}, x_{i+1}) + 8f(t_i, x_i) - f(t_{i-1}, x_{i-1})]
\end{aligned}$$

□

45.6 Predictor-Corrector Methods

In practice, implicit multi-step methods are used to improve approximations obtained by explicit methods. The combination of explicit and implicit technique is called a *predictor-corrector method*. The explicit method *predicts* an approximation, and the implicit method *corrects* this prediction.

Consider the following fourth-order method for solving IVP. The first step is to calculate the starting values $\mathbf{x}_0$, $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ for the four-step explicit Adams-Bashforth method. To do this, we use a fourth-order one-step method, the Runge-Kutta method of order four. Next, calculate an approximation, $\mathbf{x}_4^{(0)}$, to $\mathbf{x}(t_4)$ using the explicit Adams-Bashforth methods (refer to Equation (45.33)) as predictor:

```

25          59d0*f(neq,t-2d0*h,xp(:,3)) +      &
26          37d0*f(neq,t-3d0*h,xp(:,2)) -      &
27          9d0*f(neq,t-4d0*h,xp(:,1)))/24d0
28      ! Correct x_i
29      x = xp(:,4) + h*( 9d0*f(neq,t, x) +      &
30          19d0*f(neq,t-h, xp(:,4)) -      &
31          5d0*f(neq,t-2d0*h, xp(:,3)) +      &
32          f(neq,t-3d0*h, xp(:,2)))/24d0
33      call show(neq,t,x) ! or disp
34      do j=1,3
35          xp(:,j) = xp(:,j+1)
36      end do
37      xp(:,4) = x
38  end do
39  end subroutine
40
41  subroutine rk_fehlberg(neq, f, x0, t0, tN, hmin, hmax)

```

Example 45.36. Use the Adams-Bashforth/Adams-Moulton based predictor-corrector method with $h = 0.2$ mentioned above to obtain an approximation to $x(0.8)$ for the solution of the IVP (45.7). You should re-use the results from Example 45.26.

Solution: With $h = 0.2$, $x(0.8) \approx x_4$. From Example 45.26, we know that the Runge-Kutta order 4 method gives

- $t_0 = 0.0$, $x_0 = 0.5$
- $t_1 = 0.2$, $x_1 = 0.829293333333$
- $t_2 = 0.4$, $x_2 = 1.214076210667$
- $t_3 = 0.6$, $x_3 = 1.648922017042$

The predictor is calculated as follows

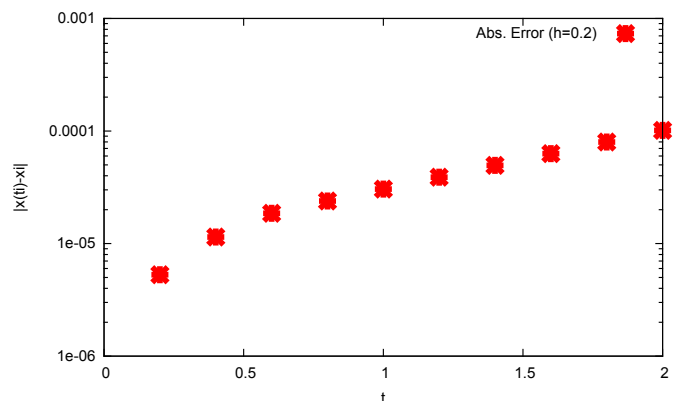
- $f_3 = f(t_3, x_3) = x_3 - t_3^2 + 1 = 2.28892202$
- $f_2 = f(t_2, x_2) = x_2 - t_2^2 + 1 = 2.05407621$
- $f_1 = f(t_1, x_1) = x_1 - t_1^2 + 1 = 1.78929333$
- $f_0 = f(t_0, x_0) = 0.5 - 0^2 + 1 = 1.5$
- $x_4^{(0)} = x_3 + \frac{0.2}{24} [55f_3 - 59f_2 + 37f_1 - 9f_0] \approx 1.648922017042 + 0.47836723 \approx 2.12728925$

The corrector is calculated as follows

- $f_4 = f(t_4, x_4^{(0)}) = f(0.8, 2.12728925) = 2.48728925$
- $x(0.8) \approx x_4^{(1)} = x_3 + \frac{0.2}{24} [9f_4 + 19f_3 - 5f_2 + f_1] \approx 1.64892202 + 0.47828362 \approx 2.12720564$

Remark 45.37. In Example 45.36 If we use a numerical computing software with double precision arithmetic, we will obtain the following solutions and the absolute error:

t_i	x_i
0.0	0.500000000000
0.2	0.829293333333
0.4	1.214076210667
0.6	1.648922017042
0.8	2.127205632419
1.0	2.640828595970
⋮	⋮
1.8	4.815096355330
2.0	5.305370671516



Example 45.38 (Tutorial 6, Q8). Write a program using the Adams Fourth-Order Predictor-Corrector Method `adams4(f, x0, a, b, N)` to solve $x'' - 3x' + 2x = t$, $0 \leq t \leq 1$, $x(0) = 0$, $x'(0) = 0.5$ [Burden and Faires, 2015].

Solution: A Fortran program will do the trick:

```

1  program tut_adams_2
2      use iso_c_binding, only : c_double
3      use textbook_odeivp, only : adams4, funcn
4      implicit none
5      integer, parameter :: neq = 2
6      real(c_double) :: x0(neq) = [0d0, 0.5d0]
7      procedure(funcn) :: ff
8      call adams4(neq, ff, x0, 0d0, 1d0, 10)
9  end program tut_adams_2
10 function ff(neq,t,x)
11     use iso_c_binding, only : c_double
12     integer, intent(in) :: neq
13     real(c_double) :: t, x(neq), ff(neq)
14     ff = [x(2), 3d0*x(2) - 2d0*x(1) + t]
15 end function ff

```

Other multi-step methods can be derived using integration of interpolating polynomials over intervals $[t_j, t_{i+1}]$, for $j \leq i - 1$, to obtain an approximation to $x(t_{i+1})$.

According to Cleve Moler, Matlab `ode113` uses a variable-order Adams-Bashforth-Moulton predictor-corrector algorithm. It is often more efficient than `ode45` at stringent tolerances and if the ODE is expensive to evaluate.

In practice, adaptive method with suitable error control is combined with predictor-corrector method. For example, in Nicholas B. Tufillaro's `ode` (in `plotutils`), various of the above numerical integration methods are implemented:

- `rk.c`: Runge-Kutta with fixed step size;
- `rka.c`: Runge-Kutta with adaptive step size;
- `am.c`: Adams-Moulton with fixed step size;
- `ama.c`: Adams-Moulton with adaptive step size.

In `ode.h`, the minimum step size is defined to be 10^{-36} and the maximum step size is defined as 0.5 for the adaptive methods.

45.7 Convergence, Stability and Stiffness

So far, a number of methods have been presented for approximating the solution to an IVP (45.2). Although numerous other techniques are available, the above methods were chosen because they generally satisfied three criteria [Burden and Faires, 2015]:

1. Their development is clear enough so that one can understand how and why they work.
2. One or more of the methods will give satisfactory results for most of the problems in science and engineering.
3. Most of the more advanced and complex techniques are based on one or a combination of the procedures described above.

In this section, we discuss why these methods give satisfactory results when some similar methods do not. Three central concepts that characterise that a numerical method for solving (45.2) is satisfactory are

- convergence: the method approximates the solution;

- order: how well the solution can be approximated if the step size is small enough.
- stability: errors should not propagate (recall from Topic 1, since we are using floating point representation for numbers, rounding errors are unavoidable).

Since single-step methods are special cases of multi-step methods, the definitions of convergence, consistency, and stability will be defined for explicit multi-step methods [Mattheij and Molenaar, 2002, Chapter VII].

Definition 45.7. A numerical method is *convergent* if the solution to the difference equation approaches the solution of (45.2) as the step size approaches zero. This means that

$$\lim_{h \rightarrow 0} \max_{0 \leq i \leq N} \|\mathbf{x}_i - \mathbf{x}(t_i)\| = 0.$$

A general representation of the numerical method for IVP (45.2) is a special form of (45.5).

Definition 45.8. The *local truncation error* (LTE) for an explicit multi-step method is

$$\tau_{i+1}(h) = \Phi_h(t_i; \mathbf{x}(t_i), \dots, \mathbf{x}(t_{i+1-m})) - \mathbf{x}(t_{i+1}). \quad (45.34)$$

Here, $\mathbf{x}(t_{i+1}), \mathbf{x}(t_i), \dots, \mathbf{x}(t_{i+1-m})$ are the exact solutions of (45.2) at time $t_{i+1}, t_i, \dots, t_{i+1-m}$ respectively.

An explicit multistep method is said to be *consistent* if

$$\lim_{h \rightarrow 0} \frac{\|\tau_{i+1}(h)\|}{h} = 0.$$

The method has order p if

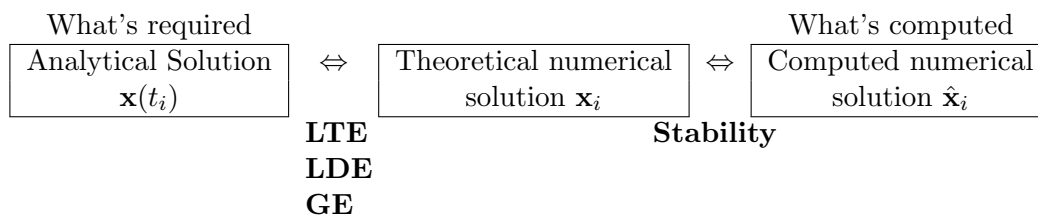
$$\|\tau_{i+1}(h)\| = O(h^{p+1}) \quad \text{as } h \rightarrow 0.$$

Therefore, a numerical method for solving (45.2) is consistent if it has an order greater than zero. The Euler method has an order one, so it is consistent. Most methods discussed previously has higher order.

Theorem 45.39. A Runge-Kutta method (45.16) is consistent if $c_i = a_{i1} + \dots + a_{i,i-1}$, $2 \leq i \leq s$.

Definition 45.9. A numerical method is said to be (*zero*) *stable* if the solution remains bounded as $h \rightarrow 0$ for finite final time T .

The link between these concepts is summarised in the following diagram:



and the following theorem.

Theorem 45.40 (Dahlquist Equivalence Theorem). *For a multi-step method,*

$$\text{convergent} \Leftrightarrow \text{consistent} + \text{stable}.$$

To have a reliable numerical method for solving ODE it is essential that accuracy (LTE, LDE, GE) and insensitivity to errors (A-stability and L-stability) are established so that the computed numerical solution $\hat{\mathbf{x}}_i$ is “close to” the required analytical solution $\mathbf{x}(t_i)$.

The following special ODE related to the concept stability (and stiffness):

$$\dot{x} = \lambda x, \quad x(0) = C. \quad (45.35)$$

It is obvious that this ODE has a solution $x(t) = Ce^t$. What happens if we study the difference equation of an ODE-IVP method corresponding to Equation (45.35)? In the case of Taylor method (45.15), $f(t, x) = \lambda x \Rightarrow f'(t, x) = \lambda x' = \lambda f(t, x) = \lambda^2 x$ and in general $f^{(k)}(t, x) = \lambda^{(k+1)} x$ and we have

$$x_{i+1} = x_i + h\lambda x_i + \frac{h^2}{2}\lambda^2 x_i + \cdots + \frac{h^n}{n!}\lambda^n x_i = Q(h\lambda)x_i$$

where $Q(h\lambda) = 1 + h\lambda + \cdots + \frac{h^n}{n!}\lambda^n$ is called the absolute stability function of a given ODE-IVP numerical method.

Definition 45.10. The region R of absolute stability for a one-step method is $R = \{h\lambda \in \mathbb{C} : |Q(h\lambda)| < 1\}$.

Example 45.41 (April 2013 Final Exam Question 5(b)). Consider the Runge-Kutta method below for the ODE $y' = -\lambda y$:

$$\begin{aligned} k_1 &= hf(t_i, y_i) \\ k_2 &= hf(t_i + h, y_i + k_1) \\ y_{i+1} &= y_i + (k_1 + k_2)/2 \end{aligned}$$

1. Show that the absolute stability function $Q(h\lambda)$ for the method is $Q(h\lambda) = 1 + h\lambda + (h\lambda)^2/2$.

Solution:

$$\begin{aligned} k_1 &= h\lambda y_i \\ k_2 &= h\lambda(y_i + k_1) = h\lambda y_i + h^2\lambda^2 y_i \\ y_{i+1} &= y_i + (h\lambda y_i + h\lambda y_i + h^2\lambda^2 y_i)/2 = y_i + \frac{2h\lambda + h^2\lambda^2}{2} y_i = \underbrace{\left(1 + h\lambda + \frac{h^2\lambda^2}{2}\right)}_{Q(h\lambda)} y_i. \end{aligned}$$

2. If the method is used to solve the equation $y' = -1000y$, determine the largest stable time step h .

Solution: We need step size $h > 0$, so

$$\begin{aligned} |Q(-1000h)| &= |1 - 1000h + (1000h)^2/2| < 1 \\ \Rightarrow -1 &< 1 - 1000h + (1000h)^2/2 < 1 \\ \Rightarrow -2 &< -1000h + 500000h^2 < 0 \Rightarrow h < \frac{1000}{500000} = \boxed{0.002}. \end{aligned}$$

All the methods for approximating the solution to IVP have error terms that involve a higher derivative of the solution of the equation. If the derivative can be reasonably bounded, then the method will have a predictable error bound that can be used to estimate the accuracy of the approximation. Even if the derivative grows as the steps increase, the error can be kept

in relative control, provided that the solution also grows in magnitude. However, problems frequently arise when the magnitude of the derivative increases but the solution does not. In this situation, the error can grow so large that it dominates the calculations. IVP for which this is likely to occur are called *stiff equations* [Mattheij and Molenaar, 2002, Chapter VIII] and are quite common, particularly in the study of vibrations, chemical reactions, and electrical circuits. The word *stiff* comes from an analogy to a spring system where the natural frequency of vibration is very great if the spring constant is large.

When the solution of a DE (say, of second order) has a general solution that involves the sum or difference of terms of the form ae^{-ct} and be^{-dt} where $0 < c \ll d$, the numerical solution can be very unstable even with a very small step size. The be^{-dt} part is usually the *transient* part of the solution, called the transient solution. The more important part is the *steady-state solution*. The transient part will rapidly decay to zero as t increases, but since n th derivative of this term has magnitude $(-1)^n b d^n e^{-dt}$, the derivative does not decay as quickly. In fact, since the derivative in the error term is evaluated not at t , but at a number between 0 and t , the derivative terms can increase as t increases — and very rapidly indeed.

The difficulty with stiff problems is the prohibitive amounts of computer time required for their solution by classical ODE solution methods, such as the popular explicit Runge-Kutta (Section 45.3) and Adams methods (Section 45.5). The reason is the excessively small step sizes that these methods must use to satisfy stability requirements. Because of the approximate nature of the solutions generated by numerical integration to be stable, errors introduced at any one step should not grow unbounded as the calculation proceeds. To maintain numerical stability, classical ODE solution methods must use small step size of order $1/\max\{-Re(\lambda_i)\}$ even after the rapidly decaying components have decreased to negligible levels.

The system

$$\begin{aligned}x' &= 1195x - 1995y, & x(0) &= 2, \\y' &= 1197x - 1997y, & y(0) &= -2\end{aligned}$$

has an analytical solution of

$$x(t) = 10e^{-2t} - 8e^{-800t}, \quad -y(t) = 6e^{-2t} - 8e^{-800t}$$

The transient part e^{-800t} in the solution causes this system to be stiff. Applying Runge-Kutta Fourth-Order Method for Systems will show that for not small enough step-size h , the result is disastrous.

Since the region of absolute stability of a method is generally the critical factor in producing accurate approximations for stiff systems, numerical methods are sought with as large a region of absolute stability as possible. A numerical method is said to be *A-stable* if its region R of absolute stability contains the entire left half-plane. The *implicit trapezoidal method*, given by

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \frac{h}{2}[f(t_i, \mathbf{x}_i) + f(t_{i+1}, \mathbf{x}_{i+1})], \quad i = 0, 1, 2, \dots$$

is an A-stable method and is the only A-stable multi-step method. Although the trapezoidal method does not give accurate approximations for large step sizes, its error will not grow exponentially.

Apart from all the theory we discussed so far, real-world numerical integration software has many aspects that are of concerned to us. A comparison of various of numerical integration software for (45.2) (Figure 45.1) is given in <http://www.stochasticlifestyle.com/comparison-differential-equation-solver-suites-matlab-r-julia-python-c-fortran/>.

Comparison Of Differential Equation Solver Software														
Subject/Item	MATLAB	SciPy	deSolve	DifferentialEquations.jl	Sundials	Halper	ODEPACK and Netlib	JICODE	PyDSTool	FATODE	GSL	BOOST	Mathematica	Maple
Language	MATLAB	Python	R	Julia	C++ and Fortran	Fortran	Fortran	Python	Python	Fortran	C	C++	Mathematica	Maple
Selection of Methods for ODEs	Fair	Poor	Poor	Excellent	Good (Including ARKCODE)	Good	Good	Poor	Poor	Good	Poor	Fair	Good	Fair
Efficiency*	Poor	Poor	Poor	Excellent	Good	Good	Good	Good	Good	Good	Fair	Fair	Fair	Fair
Tweakability	Fair	Poor	Poor	Excellent	Good	Good	Good	Fair	Fair	Fair	Fair	Fair	Fair	Fair
Event Handling	Good	None	Fair	Excellent	Good**	None	Good**	None	Fair	None	None	None	Good	Fair
Symbolic Calculation of Jacobians and Autodifferentiation	None	None	None	Good	None	None	None	None	None	None	None	None	Excellent	Excellent
Complex Numbers	Excellent	Fair	None	Good	Good	None	None	None	None	None	None	Good	Fair	Fair
Arbitrary Precision Numbers	None	None	None	Excellent	None	None	None	None	None	None	None	None	Fair	Fair
Control Over Linear/Nonlinear Solvers	None	None	None	Excellent	Excellent	Good	Depends on the solver	None	None	None	None	None	None	None
Built-in Parallelism	None	None	None	Excellent	Excellent	None	None	None	None	None	None	Fair	None	None
Differential-Algebraic Equation (DAE) Solvers	Good	None	None	Excellent	Excellent	Excellent	Excellent	None	Fair	Good	None	None	Good	Fair
Constant-Lag Delay Differential Equation (DDE) Solvers	Fair	None	None	Excellent	None	Good	None	Fair	Fair	None	None	None	Good	None
State-Dependent DDE Solvers	Poor	None	None	Excellent	None	Excellent	None	None	Fair	None	None	None	None	None
Stochastic Differential Equation (SDE) Solvers	Poor	None	None	Excellent	None	None	None	Good	None	None	None	None	Fair	None
Specialized Methods for 2nd Order ODEs and Hamiltonians (and Symplectic Integrators)	None	None	None	Excellent	None	Good	None	None	None	None	None	None	Good	None
Boundary Value Problem (BVP) Solvers	Good	Fair	None	Good	None	None	Good	None	None	None	None	None	Good	None
GPU Compatibility	None	None	None	Excellent	None	None	None	None	None	None	None	Excellent	None	None
Analysis Addons (Sensitivity Analysis, Parameter Estimation, etc.)	None	None	None	Excellent	Good	None	None	None	None	Good	None	None	None	None

*Efficiency takes into account not only the efficiency of the implementation, but the features of the implemented methods (advanced timestepping controls, existence of methods which are known to be more efficient, Jacobian handling)

**Event handling needs to be implemented yourself using basic rootfinding functionality

For more detailed explanations and comparisons, see the following blog post:
<http://www.stochasticlifestyle.com/a-comparison-between-differential-equation-solver-suites-in-matlab-r-julia-python-c-and-fortran>

Scale	None	Poor	Fair	Good	Excellent
Explanation	Functionality does not exist	Functionality exists, but is incomplete	The basic features exist	The basic features exist and some extra tweakability exists. May include extra methods for efficiency.	Has all of the basic features and more features for flexibility and efficiency.

Figure 45.1: Comparison of ODE Solvers

fig:comparison

Chapter 46

ODE-BVPs

ch:ode-bvp

A general *two-point boundary-value problem (BVP)* is given by [Betts, 2010, Chapter 3], [Ascher et al., 1994]:

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}), \quad \psi(\mathbf{y}(t_0), \mathbf{y}(t_F)) = \mathbf{0}, \quad \mathbf{y} \in \mathbb{R}^n, \quad t_0 \leq t \leq t_F. \quad (46.1)$$

e:bvp-gen

When $\psi(\mathbf{y}(t_0), \mathbf{y}(t_F)) = B_0 \mathbf{y}(t_0) + B_F \mathbf{y}(t_F) - \mathbf{b}$, the boundary conditions of the BVP (46.1) are said to be *linear*. Here B_0 and B_F and $\mathbf{b}$ are constants.

This topic will only investigate a special *two-point boundary-value problem (BVP) (of the first type)* of a second-order ordinary differential equation (ODE-BVP) which can be regarded as a special case of (46.1):

$$y'' = f(x, y, y'), \quad y(a) = A, \quad y(b) = B. \quad (46.2)$$

e:ode-bvp:1

We will limit ourselves to the less complicated form above and refers Bailey et al. [1968] for the study of general boundary conditions:

$$y'' = f(x, y, y'), \quad r_1(y(a), y'(a)) = r_2(y(b), y'(b)) = 0. \quad (46.3)$$

e:ode-bvp

Two-point BVPs normally aroused from the steady-state problems of heat equations (Topic 8) and the motion controlled problems. Many examples are given in Ascher et al. [1994, Chapter 1].

There are three fundamentally different types of methods for solving ODE-BVP (46.2), the *shooting method*, the *finite-difference method* and *Galerkin method* (which includes finite element method as special case) [Hoffman, 1992, 445].

A (single) shooting method for obtaining a solution to (46.2) is to try to find an initial slope μ such that the solution of the IVP $y(a) = A, y'(a) = \mu$ for the ODE satisfies (46.2). This will be investigated in Section 46.3. Linear BVP is a special case of nonlinear BVP, however, since the solution is much simpler, it will be introduced in Section 46.2. *Multiple shooting* method divides the interval into N subintervals and choose an initial value at each of the intervals. The continuity requirement leads to $N - 1$ matching conditions as well as a nonlinear system which can be solved using Newton method (see Deuffhard and Bornemann [2002, Chapter 8] and Mattheij and Molenaar [2002, Chapter X]).

Section 46.4 introduces the linear finite difference method, which is also used in PDE (Topic 8). The non-linear BVP, which is very complicated, is mentioned in Section 46.5. Galerkin method is simply too complicated and we will just illustrate it with a simple example with linear finite element approximation in Section 46.4.

46.1 Existence and Uniqueness Theorem

The following theorem gives the general conditions that ensure BVP (46.2) has a solution and is unique.

thm:bvp:existence

Theorem 46.1 (BVP Existence Theorem). Suppose that $f(x, u, v)$ in the second order ODE in (46.3) is **uniformly Lipschitz continuous**, i.e. there exist constants K and L such that

$$|f(x, u_1, v_1) - f(x, u_2, v_2)| \leq K|u_1 - u_2| + L|v_1 - v_2|, \quad \forall (x, u_1, v_1), (x, u_2, v_2) \in D$$

in the domain

$$D = \{(x, u, v) | a \leq x \leq b, -\infty < u, v < \infty\},$$

for a fixed x and suppose that the partial derivatives f_u and f_v are continuous on D . Let the boundary conditions in (46.3) be linear, i.e.

$$r_1(y(a), y'(a)) = a_0 y(a) - a_1 y'(a) - A, \quad \text{and} \quad r_2(y(b), y'(b)) = b_0 y(b) + b_1 y'(b) - B \quad (46.4)$$

where the constants satisfy $a_0 a_1 \geq 0$, $b_0 b_1 \geq 0$, $|a_0| + |a_1| \neq 0$ and $|b_0| + |b_1| \neq 0$. If

1. $f_u(x, u, v) > 0$, and
2. there is a constant M such that $|f_v(x, u, v)| \leq M$

for all $(x, u, v) \in D$, then the BVP (46.2) has a unique solution $y(x)$.

e:ode-bvp:bc

If we are dealing with a constant coefficient linear ODE, the best way to determine whether the BVP has a solution is to find the solution instead of applying the BVP Existence Theorem.

j.cash:test11

Example 46.2 (J. Cash Test Problem 11). Show that the BVP below

$$\epsilon y'' - y = -(\epsilon \pi^2 + 1) \cos(\pi x), \quad -1 \leq x \leq 1, \quad y(-1) = y(1) = -1$$

has a solution (and find the analytic solution).

Solution: Using the knowledge from Calculus 2 (second order non-homogeneous linear ODE of constant coefficient), we can find the solution $y(x) = \cos(\pi x)$.

The reason we want to try to find an analytic solution first is because sometimes a solution may exist, but it does not satisfy the conditions of the BVP Existence Theorem.

Example 46.3. Show that the BVP below has a solution:

$$y'' + y = 0, \quad 0 \leq x \leq 1, \quad y(0) = 0, \quad y(1) = 1.$$

Solution: Using the knowledge from Calculus 2, we can find the general solution $y = C_1 \sin x + C_2 \cos x$. From the BCs,

$$y(0) = 0 \Rightarrow C_2 = 0,$$

$$y(1) = 1 \Rightarrow C_1 \sin 1 = 1 \Rightarrow C_1 = \frac{1}{\sin 1}.$$

the unique solution is obtained: $y = \frac{\sin x}{\sin 1}$.

However, in this example, $f(x, y, y') = -y$ and $f_y(x, y, y') = -1 \not> 0$. This means that the BVP Existence Theorem cannot be applied.

Most of the time, it is impossible for us to find a solution using the techniques from *Calculus 2* and *UECM2023 Ordinary Differential Equations*. In this case, we may have to seek the help of the BVP Existence Theorem, although it cannot tell us the existence of solution for many examples which do not satisfy the sufficient conditions.

Example 46.4. Show that the BVP $y'' - y \cos x = 0$, $0 \leq x \leq 1$, $y(0) = y(1) = 0$ has a unique solution.

Proof: Let $D = \{(x, u, v) \mid 0 \leq x \leq 1, -\infty < u, v < \infty\}$. Rewriting the equation, we have

$$f(x, u, v) = u \cos x, \quad r_1(y(0), y'(0)) = y(0) = 0, \quad r_2(y(1), y'(1)) = y(1) = 0.$$

where $a_0 = 1, a_1 = 0, b_0 = 1, b_1 = 0, A = 0, B = 0$. Thus,

$$a_0 a_1 = 0 \geq 0, \quad |a_0| + |a_1| = 1 \neq 0, \quad b_0 b_1 = 0 \geq 0, \quad |b_0| + |b_1| = 1 \neq 0.$$

1. Since f is continuous on D and for any $(x, u, v), (x, \tilde{u}, \tilde{v}) \in D$,

$$|f(x, u, v) - f(x, \tilde{u}, \tilde{v})| = |u \cos x - \tilde{u} \cos x| = |u - \tilde{u}| |\cos x| \leq |u - \tilde{u}|,$$

$f(x, u, v)$ is uniformly Lipschitz continuous on D .

2. $f_u(x, u, v) = \cos x$ is continuous and > 0 for $0 \leq x \leq 1$, i.e. $(x, u, v) \in D$;
3. $f_v(x, u, v) = 0$ is continuous on D and there exists a constant $M = 1$ such that $|f_v| \leq M$ for all $(x, u, v) \in D$.

By Theorem [tthm:bvp:existence](#) 46.1, the BVP has a unique solution.

A more general BVP existence and uniqueness theorem is given in Bailey et al. [1968, Theorem 2.2].

46.2 Linear Shooting Method

When f is linear, i.e. $f(x, y, y') = p(x)y' + q(x)y + r(x)$, [\(46.2\)](#) becomes a linear BVP:

$$y'' = p(x)y' + q(x)y + r(x), \quad y(a) = A, \quad y(b) = B. \quad (46.5)$$

Theorem [tthm:bvp:existence](#) 46.1 implies the following corollary.

Corollary 46.5. *If the linear BVP [\(46.5\)](#) satisfies*

1. $q(x) > 0$ on $[a, b]$, and
2. $p(x), q(x)$ and $r(x)$ are continuous on $[a, b]$

then the BVP [\(46.5\)](#) has a unique solution.

Proof: Let $D = \{(x, u, v) \mid a \leq x \leq b, -\infty < u, v < \infty\}$. Comparing [\(46.3\)](#) with linear boundary conditions [\(46.4\)](#) against [\(46.5\)](#), we have

$$a_0 = 1, \quad a_1 = 0, \quad b_0 = 1, \quad b_1 = 0 \Rightarrow \begin{cases} a_0 a_1 = 0 \geq 0, & |a_0| + |a_1| = 1 \neq 0, \\ b_0 b_1 = 0 \geq 0, & |b_0| + |b_1| = 1 \neq 0. \end{cases}$$

1. Since for any $(x, u, v), (x, \tilde{u}, \tilde{v}) \in D$,

$$\begin{aligned} |f(x, u, v) - f(x, \tilde{u}, \tilde{v})| &= |p(x)v + q(x)u + r(x) - (p(x)\tilde{v} + q(x)\tilde{u} + r(x))| \\ &\leq |p(x)||v - \tilde{v}| + |q(x)||u - \tilde{u}| \end{aligned}$$

$f(x, u, v)$ is uniformly Lipschitz continuous on D since p and q are continuous on $[a, b]$.

2. $f_u(x, u, v) = q(x)$ is continuous and > 0 for $a \leq x \leq b$, i.e. $(x, u, v) \in D$;
3. $f_v(x, u, v) = p(x)$ is continuous on D and there exists a constant M such that $|p(x)| \leq M$ for all $(x, u, v) \in D$.

By Theorem [tthm:bvp:existence](#) 46.1, the BVP has a unique solution.

Linear shooting method is a numerical method to solved (46.5) by replacing it with two IVPs:

$$y'' = p(x)y' + q(x)y + r(x), \quad y(a) = A, \quad y'(a) = 0, \quad (46.6)$$

$$y'' = p(x)y' + q(x)y, \quad y(a) = 0, \quad y'(a) = 1. \quad (46.7)$$

for $a \leq x \leq b$ based on the following theorem.

Theorem 46.6. If the solutions to Equations (46.5), (46.6) and (46.7) are y , y_n and y_h respectively, then they are related by the formula

$$y(x) = y_n(x) + \frac{B - y_n(b)}{y_h(b)} y_h(x). \quad (46.8)$$

Proof: Assuming y_n and y_h are the solutions of (46.6) and (46.7). By substituting (46.8) into LHS – RHS of (46.5), we have

$$y_n''(x) + \frac{B - y_n(b)}{y_h(b)} y_h''(x) - p(x)y_n'(x) + p(x) \frac{B - y_n(b)}{y_h(b)} y_h'(x) - q(x)y_n(x) + q(x) \frac{B - y_n(b)}{y_h(b)} y_h(x) - r(x).$$

After rearrangement, we have

$$\underbrace{y_n''(x) - p(x)y_n'(x) - q(x)y_n(x) - r(x)}_{=0, \text{ by definition of } y_n} + \frac{B - y_n(b)}{y_h(b)} \underbrace{\left[y_h''(x) - p(x)y_h'(x) + q(x)y_h(x) \right]}_{=0, \text{ by definition of } y_h} = 0$$

This shows that y satisfies (46.5).

When $x = a$, $y(a) = y_n(a) + \frac{B - y_n(b)}{y_h(b)} y_h(a) = A + 0 = A$

When $x = b$, $y(b) = y_n(b) + \frac{B - y_n(b)}{y_h(b)} y_h(b) = y_n(b) + B - y_n(b) = B$.

These proved that (46.8) is indeed a solution to the BVP (46.5).

An advantage of the shooting method is that the existing ODE-IVP programs from Topic 6 may be used to solve (46.5).

Method 46.1 (Linear Shooting Method). Let $z = y'$.

$$1. \text{ Solve the IVP } \begin{cases} y' = z, & y(a) = A, \\ z' = p(x)y' + q(x)y + r(x), & z(a) = 0 \end{cases}$$

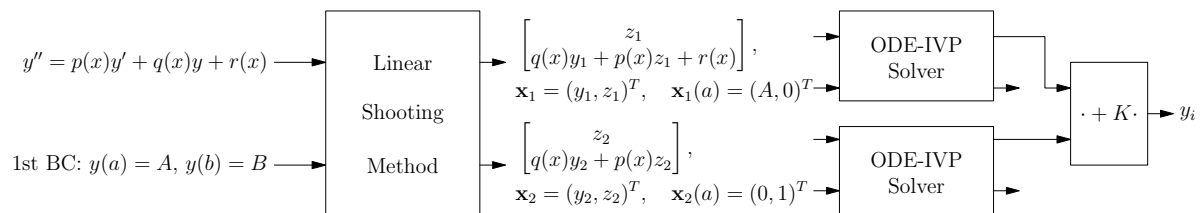
for the solution $y_n(x_i)$, $i = 0, \dots, N$.

$$2. \text{ Solve the IVP } \begin{cases} y' = z, & y(a) = 0, \\ z' = p(x)y' + q(x)y, & z(a) = 1 \end{cases}$$

for the solution $y_h(x_i)$, $i = 0, \dots, N$.

$$3. \text{ Let } K = \frac{B - y_n(x_N)}{y_h(x_N)}. \text{ Output the solution } y_i \approx y_n(x_i) + K y_h(x_i) \text{ for } i = 0, \dots, N.$$

A block diagram of linear shooting method will look like



An implementation in Fortran 2008 using the fourth order Runge-Kutta method in Topic 6 is listed as follows.

```

1 ! Using Fortran2008 feature: passing internal function to called function
2 subroutine linearshooting(npt,p,q,r,a,b,alpha,beta)
3   ! f(x,y,y') = p(x)y' + q(x)y + r(x); Endpoints: a,b;
4   ! BC: alpha, beta (Fortran is case insensitive)
5   use textbook_odeivp, only: rungekutta4
6   integer, intent(in) :: npt
7   procedure(linc) :: p, q, r
8   double precision :: a,b,alpha,beta, xi,y1(2),y2(2),coeff,h
9   double precision, dimension(npt) :: y1l,y2l
10  integer :: i
11  h = (b-a)/real(npt,8)
12  xi = a; y1 = (/alpha, 0.0d0/); y2 = (/0.0d0, 1.0d0/)
13  print 100, xi, y1(1), y2(1), y1(1)
14  do i=1,npt
15    y1 = rungekutta4(2,f1,y1,xi,xi+h) ! Using Internal Proc
16    y1l(i) = y1(1)
17    y2 = rungekutta4(2,f2,y2,xi,xi+h) ! Fortran 2008 Feature
18    y2l(i) = y2(1)
19    xi = xi+h
20  end do
21  coeff = (beta-y1l(npt))/y2l(npt)
22  do i=1,npt
23    xi = a+i*h
24    print 100, xi, y1l(i), y2l(i), y1l(i) + coeff*y2l(i)
25  end do
26 100 format(f10.6,f18.12,f18.12,f18.12)
27 contains
28   function f1(neq,x,y)
29     integer, intent(in) :: neq
30     double precision :: x, y(neq), f1(neq)
31     f1(1) = y(2)
32     f1(2) = p(x)*y(2)+q(x)*y(1)+r(x)
33   end function f1
34   function f2(neq,x,y)
35     integer, intent(in) :: neq
36     double precision :: x, y(neq), f2(neq)
37     f2(1) = y(2)
38     f2(2) = p(x)*y(2)+q(x)*y(1)
39   end function f2
40 end subroutine linearshooting

```

Note that we are using our naïve Runge-Kutta implementation which gives poor accuracy compare to commercial implementation such as Matlab's ode45.

Example 46.7. Use the linear shooting method together with the (naïve) fourth-order Runge-Kutta method of step-size $h = 1/2$ to solve the J. Cash Test Problem 11 in Example 46.2 with $\epsilon = 0.1$.

Solution. J. Cash Test Problem 11 becomes $y'' = 10y - (\pi^2 + 10) \cos \pi x$, $-1 \leq x \leq 1$, $y(-1) = y(1) = -1$.

$$1. \begin{cases} y'' = 10y - (\pi^2 + 10) \cos \pi x, \\ y(-1) = -1, y'(1) = 0 \end{cases} \Rightarrow \begin{cases} y'_n = z_n, \\ z'_n = 10y_n - (\pi^2 + 10) \cos \pi x, \end{cases} \quad \begin{matrix} y_n(0) = -1, \\ z_n(0) = 0. \end{matrix} \quad (*)$$

Using naïve RK4 to solve (*), we obtain the following numerically

$$\begin{aligned} y_n(-0.5) &\approx 0.005748798949, \\ y_n(0.0) &\approx 0.819190325460, \\ y_n(0.5) &\approx -0.998142298247, \\ y_n(1.0) &\approx -5.569522860539. \end{aligned}$$

$$2. \begin{cases} y'' = 10y, \\ y(-1) = 0, \quad y'(1) = 1 \end{cases} \Rightarrow \begin{cases} y'_h = z_h, & y_h(0) = 0, \\ z'_h = 10y_h, & z_h(0) = 1. \end{cases} \quad (**)$$

Use RK4 to solve (**): we obtain

$$\begin{aligned} y_h(-0.5) &\approx 0.708333333333, \\ y_h(0.0) &\approx 3.556423611111, \\ y_h(0.5) &\approx 16.946121780961, \\ y_h(1.0) &\approx 80.514250767084. \end{aligned}$$

$$3. \text{ Since } y(x) = y_n(x) + \frac{2 - y_n(1)}{y_h(1)} y_h(x) = y_n(x) + \underbrace{\frac{-1 - (-5.569522860539)}{80.514250767084}}_{= 0.056754212} y_h(x),$$

$$\begin{aligned} y(-0.5) &\approx 0.045949699146, \\ y(0.0) &\approx 1.021032345201, \\ y(0.5) &\approx -0.036378509380, \\ y(1.0) &\approx -1. \end{aligned}$$

□

Remark 46.8. It is too tedious to calculate Example 46.7 and is wise to use a computer program to perform the calculation:

```
1 program eg_linshoot
2   use textbook_odebvp, only : linearshooting
3   implicit none
4   double precision, parameter :: eps = 0.1d0, pi = 4*datan(1d0)
5   call linearshooting(4, jcl1p, jcl1q, jcl1r, -1d0, 1d0, -1d0, -1d0)
6 contains
7 function jcl1p(x)
8   intent(in) :: x
9   double precision :: x, jcl1p
10  jcl1p = 0.0d0
11 end function jcl1p
12 function jcl1q(x)
13   intent(in) :: x
14   double precision :: x, jcl1q
15   jcl1q = 1d0/eps
16 end function jcl1q
17 function jcl1r(x)
18   intent(in) :: x
19   double precision :: x, jcl1r
20   jcl1r = -(eps*pi**2+1)*cos(pi*x)/eps
21 end function jcl1r
22 end program eg_linshoot
```

Example 46.9 (Tutorial 7, Q1). Given the following BVP

$$(1 + x^2)y''(x) = 2xy'(x) + 3y(x) + 1 + x^2, \quad 0 \leq x \leq 4, \quad y(0) = 1.25, \quad y(4) = -0.95.$$

1. Determine whether a unique solution exists by using the BVP Existence Theorem.

Solution. The function

$$f(x, y, y') = \frac{2x}{1 + x^2} y' + \frac{2}{1 + x^2} y + 1$$

is continuous on $D = [0, 4] \times \mathbb{R}^2$. For all $x \in [0, 4]$,

$$f_y(x, y, y') = \frac{2}{1 + x^2} > 0$$

and

$$|f_{y'}(x, y, y')| = \left| \frac{2x}{1+x^2} \right| \leq 1.$$

By BVP Existence Theorem, the problem has a unique solution. $\square$

2. Write down the steps to solve the BVP using linear shooting method.

Solution. Let $z = y'$.

(a) Solve the IVP
$$\begin{cases} y' = z, & y(0) = 1.25, \\ z' = \frac{2x}{1+x^2}z + \frac{2}{1+x^2}y + 1, & z(0) = 0 \end{cases}$$
 for the solution $y_n(x_i)$, $i = 0, \dots, N$.

(b) Solve the IVP
$$\begin{cases} y' = z, & y(0) = 0, \\ z' = \frac{2x}{1+x^2}z + \frac{2}{1+x^2}y, & z(0) = 1 \end{cases}$$
 for the solution $y_h(x_i)$, $i = 0, \dots, N$.

(c) Output the approximation of solution y_i .

$$y(x_i) = y_n(x_i) + \frac{-0.95 - y_n(x_N)}{y_h(x_N)} y_h(x_i), \text{ for } i = 0, \dots, N$$

$\square$

3. Find the numerical solution using naïve Runge-Kutta order 4 method with $N = 10$ (i.e. 10 intervals) by writing a computer program.

0.000000	1.25000000000000	0.00000000000000	1.25000000000000
0.400000	1.656863905325	0.453254437870	0.571938238889
0.800000	3.196172579481	1.259467261804	0.181467715720
1.200000	6.796494651909	2.886986595602	-0.113897380758
1.600000	13.929861861049	5.976634561599	-0.376019544687
2.000000	26.557937188811	11.353403536056	-0.617966384393
2.400000	47.095588608504	20.022995493432	-0.832153327922
2.800000	78.385903289593	33.165513151773	-1.000228323367
3.200000	123.681827154496	52.129402136206	-1.096932032848
3.600000	186.631827005540	78.426171783824	-1.091799239586
4.000000	271.268187165749	113.725857188934	-0.950000000000

Solution. A Fortran program is shown below:

```

1 program tut_linshoot1
2   use textbook_odebvp, only : linearshooting
3   implicit none
4   call linearshooting(10,p,q,r,0d0,4d0,1.25d0,-0.95d0)
5 contains
6 function p(x)
7   double precision, intent(in) :: x
8   double precision :: p
9   p = 2d0*x/(1+x**2)
10 end function p
11 function q(x)
12   double precision, intent(in) :: x
13   double precision :: q
14   q = 3d0/(1+x**2)
15 end function q
16 function r(x)
17   double precision, intent(in) :: x
18   double precision :: r
19   r = 1d0
20 end function r
21 end program tut_linshoot1

```

$\square$

46.3 Nonlinear (Single) Shooting Method

When f is nonlinear, the steps in the linear shooting technique no longer work for (46.2) because the solution cannot be expressed as a linear combination of solutions to two IVPs. Instead, we have to approximate the solution to the BVP by using the solutions of a sequence of IVPs involving parameter t :

$$y'' = f(x, y, y'), \quad a \leq x \leq b, \quad y(a) = A, \quad y'(a) = t. \quad (46.9)$$

Suppose that the solution of (46.9) is $y(x; t)$. If we can find the right $t = t^*$ such that $y(b; t^*) = B$, then $y(x; t^*)$ is the solution of (46.3). Finding t^* is equivalent to solving a nonlinear problem below for t^* (Topic 3):

$$F(t) := y(b, t) - B = 0. \quad (46.10)$$

As mentioned in Topic 3, when solving a nonlinear problem, we should always start with the secant method and try other methods if secant method fails. According to Deuffhard and Bornemann [2002, Chapter 8], the nonlinear (single) shooting method has limited use because in many cases, the estimated initial value at $t = a$ cannot continue over the entire interval $[a, b]$. This difficulty could be overcome by the multiple shooting method.

The numerical nonlinear shooting method can be stated as follows.

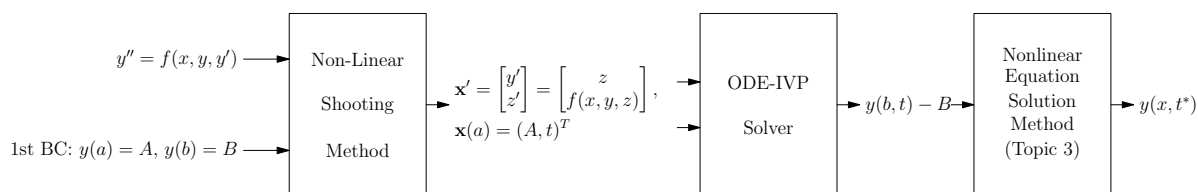
Method 46.2 (Nonlinear Shooting Method based on Secant Method).

1. Let t_0 and t_1 be the initial approximations to the parameter t that solves the nonlinear equation (46.10).
2. Solve the IVP $y'' = f(x, y, y')$, $y(a) = A$, $y'(a) = t_{k-2}$ for $a \leq x \leq b$ to find $y(b, t_{k-2})$.
3. Solve the IVP $y'' = f(x, y, y')$, $y(a) = A$, $y'(a) = t_{k-1}$ for $a \leq x \leq b$ to find $y(b, t_{k-1})$.
4. Obtain a improved estimate of t using the secant method (see Topic 3):

$$t_k = t_{k-1} - \frac{t_{k-1} - t_{k-2}}{y(b, t_{k-1}) - y(b, t_{k-2})} (y(b, t_{k-1}) - B).$$

5. If $|y(b, t_{k-1}) - B|$ is sufficiently small, then t_{k-1} is a good approximation for t^* . Otherwise, increase k by one and continue with step 2.

In block diagram form, the nonlinear shooting method seems simpler, but it is actually more complicated because we have a family of ODE-IVP with parameter t (compare to just two ODE-IVP in the linear shooting method).



One can observe its complexity based on a naïve implementation of the above algorithm using the Runge-Kutta method of order 4 in Fortran 2008 is given below.

```

1 subroutine shooting(npt,f,a,b,alpha,beta,t1_,t2_)
2   ! f(x,y,y') with BC: y(a)=alpha, y(b)=beta
3   ! Initial Guess of Time: t1, t2
4   use textbook_odeivp, only: rungekutta4
5   procedure(twoptbvfpf) :: f
6   intent(in) :: npt, a, alph, b, beta, t1_, t2_
7   integer :: npt, trial, i
8   double precision :: a, alph, b, beta, t1_, t2_, t1,t2,tmp,err1,err2,xi

```

```

9      double precision :: h, ivp1(npt), ivp2(npt), y1(2), y2(2)
10     t1 = t1_; t2 = t2_
11     h = (b-a)/real(npt)
12     do trial=0,120 ! Max iteration for Secant Method
13         xi = a; y1 = (/alph,t1/); y2 = (/alph,t2/)
14         do i=1,npt
15             y1 = rungekutta4(2, myf, y1, xi, xi+h) ! Using Internal Proc
16             ivp1(i) = y1(1)
17             y2 = rungekutta4(2, myf, y2, xi, xi+h) ! Fortran 2008 Feature
18             ivp2(i) = y2(1)
19             xi = xi+h
20         end do
21         err1=beta-y1(1); err2=beta-y2(1)
22         if (abs(err1)<tol) then
23             call printvector(npt, [(a+i*h, i=1,npt)], ivp1)
24             return
25         else if (abs(err2)<tol) then
26             call printvector(npt, [(a+i*h, i=1,npt)], ivp2)
27             return
28         else
29             tmp = t2
30             t2 = t2 - (t2-t1)/(err2-err1)*err2
31             t1 = tmp
32         end if
33     end do
34     print *, "Maximum iteration (120) reach"
35 contains
36     function myf(neq, x, y)
37         integer, intent(in) :: neq
38         double precision :: x, y(neq), myf(neq)
39         myf = [y(2), f(x,y(1),y(2))]
40     end function myf
41 end subroutine shooting

```

The following are some mathematical examples of nonlinear two-point BVPs found in textbooks.

Example 46.10. $y'' + \cos y = 0$, $y(0) = y(1) = 0$ [Bailey et al., 1968, p. 37].

Example 46.11. $y'' + e^y = 0$, $y(0) = y(1) = 0$ has exactly two solutions [Bailey et al., 1968, p. 54].

Example 46.12. $y'' = \frac{3}{2}y^2$, $y(0) = 4$, $y(1) = 1$ has exactly two solutions [Stoer and Bulirsch, 2002, p. 544].

Example 46.13. $y'' = e^y - e^{-y} - c(x)$, $y(0) = \varphi_D(0) + u_0$, $y(L) = \varphi_D(L)$. Here $c(x) = \tanh(20(\frac{x}{L} - \frac{1}{2}))$ and $\varphi_D(x) = \sinh^{-1}(\frac{c(x)}{2})$. [Gander and Hřebíček, 1997, Chapter 5].

Example 46.14 (April 2014 Final Exam Question 5(c)). Consider the following nonlinear boundary value problem

$$y'' = 2 - 2y + y^2, \quad y(0) = y(1) = 0.$$

Explain how the shooting method and secant method can be used to solve the above problem numerically by writing a pseudo-code assuming you have a function `odeint(f, x0, t)` to integrate ODE with given initial conditions. (8 marks)

Solution: The steps in shooting methods to solve the BVP below are listed in (i) to (v):

$$y'' = f(x, y, y'), \quad a \leq x \leq b, \quad y(a) = A, \quad y(b) = B$$

- (i) Let t_0 and t_1 be the initial approximations to the parameter t that solves the nonlinear equation $y(b, t) - B = 0$[1 mark]
- (ii) Solve the IVP $y'' = f(x, y, y')$, $y(a) = A$, $y'(a) = t_{k-2}$ for $a \leq x \leq b$ to find $y(b, t_{k-2})$.

- [2 marks]
- (iii) Solve the IVP $y'' = f(x, y, y')$, $y(a) = A$, $y'(a) = t_{k-1}$ for $a \leq x \leq b$ to find $y(b, t_{k-1})$
 [1 mark]
- (iv) Improve the estimate, t : $t_k = t_{k-1} - \frac{t_{k-1} - t_{k-2}}{y(b, t_{k-1}) - y(b, t_{k-2})}(y(b, t_{k-1}) - B)$. [2 marks]
- (v) If $|y(b, t_{k-1}) - B|$ is sufficiently small, then $y(x; t_{k-1})$ is a good approximation for $y(x; t^*)$. Otherwise, increase k by one and continue with step (ii). [2 marks]

Example 46.15. The main reference book Burden and Faires [2015, p. 682] applied the shooting method with Newton's Method to solve the following BVP:

$$y'' = \frac{32 + 2x^3 - yy'}{8}, \quad 1 \leq x \leq 3, \quad y(1) = 17, \quad y(3) = \frac{43}{3}. \quad (46.11)$$

e:burden-nlbvp

By using the shooting method with Secant method, write a computer program to solve the BVP and compare with the result in Burden and Faires [2015].

Solution: A Fortran program calling the shooting subroutine is shown below.

```

1 program eg_odebvp_nlsh ! Nonlinear Shooting Example from Burden & Faires, p700
2   use textbook_odebvp, only : shooting
3   implicit none
4   call shooting(20, f, 1d0, 3d0, 17d0, 43/3d0, 0d0, 0.5d0)
5 contains
6 function f(x, y, yp)
7   intent(in) :: x, y, yp
8   double precision :: x, y, yp, f
9   f = (32d0 + 2d0*x**3 - y*yp)/8d0
10 end function f
11 end program eg_odebvp_nlsh

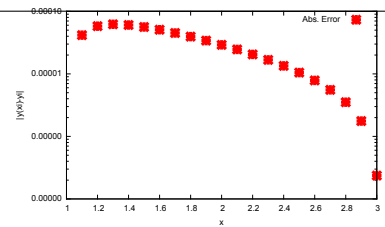
```

The result obtained from the program is very close to the result in Burden and Faires [2015], the absolute error (plot using gnuplot, a software mentioned in Topic 4) is less than 10^{-4} with the guess $t_0 = 0$ and $t_1 = 0.5$ ($N = 20$, $h = \frac{3-1}{20} = 0.1$):

```

1 y(x) = x**2 + 16/x
2 set border linewidth 0.2
3 set terminal pdfcairo font "Arial,12" linewidth 12
4 set output "eg-odebvp_nlsh.pdf"
5 set logscale y
6 set xlabel "x"
7 set ylabel "|y(xi)-yi|"
8 plot "eg-odebvp_nlsh.dat" using 1:(abs(y($1)-$2)) t "Abs. Error" w p pt 3

```



Example 46.16 (Tutorial 7, Q2). Given a nonlinear BVP

$$y'' + (y')^2 + y = \ln x, \quad 1 \leq x \leq 2, \quad y(1) = 0, \quad y(2) = \ln 2.$$

- Does the ODE-BVP satisfy the BVP Existence Theorem?

Solution. No. We are given $f(x, y, y') = \ln x - y - (y')^2$ which is continuous on $D = [1, 2] \times \mathbb{R}^2$. But

$$f_y(x, y, y') = -1 \not\equiv 0, \quad x \in [1, 2].$$

□

- Show that $y = \ln x$ is a solution of the ODE-BVP.

p702:ex11.2:q1

Solution. By Substitution,

$$y' = \frac{1}{x}, \quad y'' = -\frac{1}{x^2} \Rightarrow y'' + (y')^2 + y = -\frac{1}{x^2} + \frac{1}{x^2} + \ln x = \ln x.$$

□

3. Use the nonlinear shooting method with $h = 0.5$ to approximate the solution to the BVP by writing down the steps.

Solution. Secant Method or Newton Method [Burden and Faires, 2015].

- (a) Let t_0 and t_1 be the initial approximations to the parameter t that solves the nonlinear equation $y(2, t) - \ln 2 = 0$.

- (b) Solve the IVP

$$y'' = -(y')^2 - y + \ln x, \quad y(1) = 0, \quad y'(1) = t_{k-2}$$

for $1 \leq x \leq 2$ to find $y(2, t_{k-2})$.

- (c) Solve the IVP

$$y'' = -(y')^2 - y + \ln x, \quad y(1) = 0, \quad y'(1) = t_{k-1}$$

for $1 \leq x \leq 2$ to find $y(2, t_{k-1})$.

- (d) Use secant method to obtain a better estimate of t :

$$t_k = t_{k-1} - \frac{t_{k-1} - t_{k-2}}{y(2, t_{k-1}) - y(2, t_{k-2})} (y(2, t_{k-1}) - \ln 2).$$

- (e) If $|y(2, t_{k-1}) - \ln 2|$ is sufficiently small, then $y(x, t_{k-1})$ is a good approximation. Otherwise, increase k by one and continue with step ii.

□

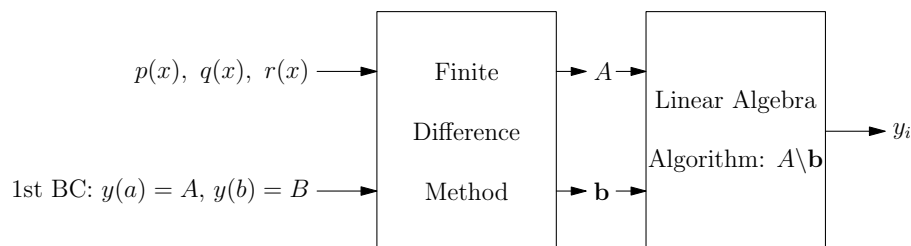
4. Use the nonlinear shooting method with $h = 0.5$ and secant method with initial guesses $t_1 = 0$, $t_2 = 0.5$ to approximate $y(1.5)$. Compare your results to the actual solution $y = \ln x$.

Solution. The nonlinear shooting algorithm gives $x_1 = 0.40549863$ which approximates $\ln 1.5 = 0.405465$.

□

46.4 Linear Finite Difference Method

The linear shooting method for BVPs can present problems of instability due to round-off errors. The *finite difference method* (FDM) in this section have better numerical stability characteristics, but generally requires more computation to obtain a certain accuracy. FDM makes use of the “finite difference approximation” that was introduced in Topic 3 and the block diagram below outlines what FDM does:



Consider (46.2) with $f(x, y, y') = -p(x)y' - q(x)y + r(x)$ so that we have (instead of (46.5)):

$$y''(x) + p(x)y'(x) + q(x)y(x) = r(x). \quad (46.12)$$

At the mesh points $x_i = a + ih$, $i = 1, \dots, N-1$, the values of (46.12) are

$$y''(x_i) + p(x_i)y'(x_i) + q(x_i)y(x_i) = r(x_i).$$

By using the Taylor polynomial approximation of $y(x_{i+1})$ and $y(x_{i-1})$, we obtain

$$\begin{aligned} y(x_{i+1}) &= y(x_i + h) = y(x_i) + hy'(x_i) + \frac{h^2}{2}y''(x_i) + \frac{h^3}{6}y'''(x_i) + \frac{h^4}{24}y^{(4)}(\xi_i^+), \quad x_i < \xi_i^+ < x_{i+1}, \\ y(x_{i-1}) &= y(x_i - h) = y(x_i) - hy'(x_i) + \frac{h^2}{2}y''(x_i) - \frac{h^3}{6}y'''(x_i) + \frac{h^4}{24}y^{(4)}(\xi_i^-), \quad x_{i-1} < \xi_i^- < x_i. \end{aligned}$$

After some algebraic manipulation, we can obtain the *centre-difference approximations*:

$$\begin{aligned} y''(x_i) &= \frac{1}{h^2}(y(x_{i+1}) - 2y(x_i) + y(x_{i-1})) - \frac{h^2}{12}y^{(4)}(\xi_i), \\ y'(x_i) &= \frac{1}{2h}(y(x_{i+1}) - y(x_{i-1})) - \frac{h^2}{6}y'''(\eta_i) \end{aligned}$$

where $\frac{h^2}{12}y^{(4)}(\xi_i) = \frac{h^2}{24}(y^{(4)}(\xi_i^+) + y^{(4)}(\xi_i^-))$ for some $\xi_i, \eta_i \in (x_{i-1}, x_{i+1})$ according to Mean Value Theorem.

Let $y_i \approx y(x_i)$. Then by replacing y' and y'' in (46.12) with their centre-difference approximations in the BVP and ignoring the higher order terms $O(h^2)$, we get

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + p(x_i)\frac{y_{i+1} - y_{i-1}}{2h} + q(x_i)y_i = r(x_i), \quad i = 1, 2, \dots, N-1.$$

After simplifying gives

$$\left(1 - \frac{h}{2}p(x_i)\right)y_{i-1} + (h^2q(x_i) - 2)y_i + \left(1 + \frac{h}{2}p(x_i)\right)y_{i+1} = h^2r(x_i).$$

This is known as a *finite difference equation* and it is an approximation to the ODE (46.12). It enables us to approximate the solution at the mesh points $x_1, \dots, x_{N-1}$.

Example 46.17. A subroutine to solve (46.12) using finite difference method using dense matrix linear algebra solver from LAPACK (Topic 2) is

```

1  subroutine linear_fd_dgesv(p,q,r,a,b,alph,beta,N,y,info)
2      ! f(x,y,y') = p(x)y' + q(x)y + r(x); BC: y(a)=alph,y(b)=beta
3      use textbook_odeivp, only: rungekutta4
4      procedure(linc) :: p, q, r
5      integer :: N, i, info, PP(N+1)
6      double precision :: a,b,alph,beta,h, x
7      double precision, allocatable :: M(:,,:), y(:)
8      h = (b-a)/real(N)
9      if (N > 1000) then
10         print *, "FAIL: N =", N, "is too large"
11         return
12     end if
13     allocate(M(N+1,N+1))
14     if(.not. allocated(y)) allocate(y(N+1))
15     x = a
16     M = 0d0 ! Initialise M to zero matrix
17     M(1,1) = 1.0d0; y(1) = alph
18     M(N+1,N+1) = 1.0d0; y(N+1) = beta
19     do i=2,N
20         x = x+h
21         M(i,i-1) = -p(x)*h/2.0d0-1.0d0
22         M(i,i) = 2.0d0 + h**2*q(x)
23         M(i,i+1) = p(x)*h/2.0d0 - 1.0d0
24         y(i) = -h**2*r(x)
25     end do
26     call printmatrix(M,N+1)
27     print *, ""

```

```

28     print '(a,10f18.12)', "f=", y
29     print *, ""
30     ! DGESV( N, NRHS, A, LDA, IPIV, B, LDB, INFO ) solves A * X = B,
31     ! A is an N-by-N matrix and X and B are N-by-NRHS matrices.
32     call dgesv(N+1,1,M,N+1,PP,y,N+1,info)
33     deallocate(M)
34 end subroutine linear_fd_dgesv

```

Explain how this could be improved.

Solution: In view that the matrix is tridiagonal, we could improve it using the subroutine dgtsv that specifically handles tridiagonal linear algebra problem.

```

1  subroutine linear_fd(npt,p,q,r,a,b,alph,beta,y,info)
2      ! f(x,y,y') = p(x)y' + q(x)y + r(x); BC: y(a)=alph,y(b)=beta
3      use textbook_odeivp, only: rungekutta4
4      procedure(linc) :: p, q, r
5      integer :: npt, i, info
6      double precision :: a,b,alph,beta,h, x
7      double precision, allocatable, intent(out) :: y(:)
8      double precision :: D(npt+1), DU(npt), DL(npt)
9      h = (b-a)/real(npt,8) ! Convert integer 8 to double
10     if(.not. allocated(y)) allocate(y(npt+1))
11     x = a
12     DU( 1) = 0.0d0; D( 1) = 1.0d0; y( 1) = alph
13     DL(npt) = 0.0d0; D(npt+1) = 1.0d0; y(npt+1) = beta
14     do i=2,npt
15         x = x+h
16         D(i) = 2.0d0 + h**2*q(x)
17         DU(i) = p(x)*h/2.0d0-1.0d0
18         DL(i-1) = -p(x)*h/2.0d0-1.0d0
19         y(i) = -h**2*r(x)
20     end do
21     ! DGTSV( N, NRHS, DL, D, DU, B, LDB, INFO )
22     call dgtsv(npt+1,1,DL,D,DU,y,npt+1,info)
23 end subroutine linear_fd

```

Example 46.18 (April 2014 Final Exam Question 3(b)). Consider the following boundary value problem:

$$y'' + x(1-x)y' - 2y = 0, \quad y(0) = 3, \quad y(1) = 3.$$

- Derive the following linear equation

$$(2 + hx_i(1 - x_i))y_{i+1} - 4(1 + h^2)y_i + (2 - hx_i(1 - x_i))y_{i-1} = 0$$

by using suitable finite-difference approximations to the derivatives. (6 marks)

Solution: By using the approximation $y'' \approx \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2}$ and $y' \approx \frac{y_{i+1} - y_{i-1}}{2h}$, we have

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + x_i(1 - x_i)\frac{y_{i+1} - y_{i-1}}{2h} - 2y_i = 0 \quad \dots \dots \dots [3 \text{ marks}]$$

$$\text{Rearranging leads to } 2y_{i+1} - 4y_i + 2y_{i-1} + hx_i(1 - x_i)(y_{i+1} - y_{i-1}) - 4h^2y_i = 0 \quad [2 \text{ marks}]$$

$$\text{and } (2 + hx_i(1 - x_i))y_{i+1} - 4(1 + h^2)y_i + 2y_{i-1} - hx_i(1 - x_i)y_{i-1} = 0 \quad \dots \dots [1 \text{ mark}]$$

- Explain which of the LAPACK subroutine DGESV, DGTSV or DPOSV is most appropriate for solving the type of linear problem in (i) above. (2 marks)

Solution. DGTSV is the most appropriate since it handles tridiagonal matrix. DGESV can be used but will be slower if n is large. $\dots \dots \dots$ [2 marks] □

Example 46.19. Use the finite difference method with grid $N = 4$ to solve the J. Cash Test Problem 11 in Example 46.2 with $\epsilon = 0.1$ by writing down (a) the corresponding finite difference approximation; (b) the corresponding linear algebra problem; (c) the approximation y_i ; (d) compare your results with the linear shooting method in Example 46.7 and the exact solution.

Solution. Here $h = \frac{1 - (-1)}{4} = 0.5$, $x_i = -1 + 0.5i$, $i = 0, 1, 2, 3, 4$.

(a) Hence, the difference equation is

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} = 10y_i - (\pi^2 + 10) \cos \pi x_i$$

$$\Rightarrow y_{i-1} - (2 + 0.5^2 \cdot 10)y_i + y_{i+1} = -0.5^2(\pi^2 + 10) \cos \pi x_i \quad i = 1, 2, 3.$$

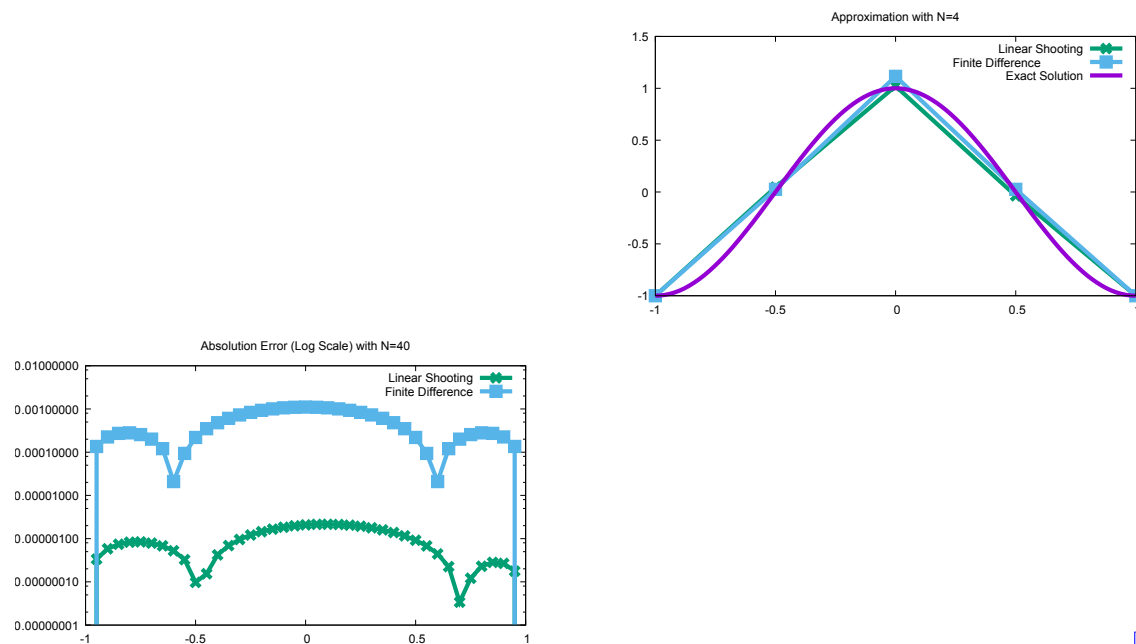
(b) Together with the BCs, we obtain the linear algebra problem

$$\begin{array}{l} i = 0 : \\ i = 1 : \\ i = 2 : \\ i = 3 : \\ i = 4 : \end{array} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & -4.5 & 1 & 0 & 0 \\ 0 & 1 & -4.5 & 1 & 0 \\ 0 & 0 & 1 & -4.5 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ -4.96740110 \\ 0 \\ -1 \end{bmatrix}$$

(c) Solving the linear algebra problem $A\mathbf{x} = \mathbf{b}$ above gives

$$y_1 = 0.02561102, \quad y_2 = 1.11524959, \quad y_3 = 0.02561102.$$

(d) By comparing the values of y_i , we find that FDM preserves symmetry while linear shooting method may not. FDM has a larger absolute error around $x = 0$. Both of them are quite close to the original solution:



Remark 46.20. We can improve the accuracy by using smaller h . But for that we have to pay a price, i.e. we have to solve a larger system of equations.

Example 46.21 (April 2015 Final Exam Question 6(a)). Consider the following boundary value problem (BVP):

$$\frac{d^2 y}{dx^2} + y = 1, \quad y(0) = 0, \quad \frac{dy}{dx}(\pi) = 0.$$

- (i) By using centre-difference approximation for the second order derivative and backward-difference approximation for the first order derivative, and assuming the x -space is equally spaced by $h = \pi/8$, derive the linear algebra problem for the approximate solution y_i , $i = 0, 1, \dots, 8$. (8 marks)

Solution. In the interval $[0, \pi]$, the requirement of $h = \frac{\pi}{8}$ splits the interval to 9 points: $x_i, i = 0, 1, \dots, 8$. By apply the principle of finite-difference approximation, we obtain

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + y_i = 1 \Rightarrow y_{i+1} + (h^2 - 2)y_i + y_{i-1} = h^2, \quad i = 1, 2, \dots, 7. \quad [3 \text{ marks}]$$

and

$$y_0 = 0, \quad \frac{y_8 - y_7}{h} = 0 \quad [2 \text{ marks}]$$

Let $d = \frac{\pi^2}{64} - 2$. We can rearrange them to a linear algebra problem:

$$\begin{bmatrix} 1 & & & & & & & & \\ 1 & d & 1 & & & & & & \\ & & 1 & d & 1 & & & & \\ & & & 1 & d & 1 & & & \\ & & & & 1 & d & 1 & & \\ & & & & & 1 & d & 1 & \\ & & & & & & 1 & d & 1 \\ & & & & & & & 1 & d & 1 \\ & & & & & & & & -1 & 1 \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{bmatrix} = \begin{bmatrix} 0 \\ h^2 \\ h^2 \\ h^2 \\ h^2 \\ h^2 \\ h^2 \\ h^2 \\ 0 \end{bmatrix} \quad [3 \text{ marks}]$$

□

(ii) The BVP can be solved using the following subroutine:

```

1  subroutine linear_fd(p,q,r,a,b,alpha,betap,N,y,info)
2      use iso_c_binding, only : c_double
3      procedure(real(c_double)) :: p, q, r
4      integer, intent(in) :: N
5      real(c_double), intent(in) :: a,b,alpha,betap
6      real(c_double), intent(out) :: y(N+1)
7      integer :: i, info
8      real(c_double) :: h, x, D(N+1), DU(N), DL(N)
9      h = (b-a)/real(N,8)
10     x = a
11     D(1) = 1d0; y(1) = alpha; DU(1) = 0d0
12     D(N+1) = 1d0; y(N+1) = betap*h; DL(N) = -1d0
13     do i=2,N
14         x = x+h
15         D(i) = h**2*q(x) - 2d0
16         DU(i) = 1d0+p(x)*h/2d0
17         DL(i-1) = 1d0-p(x)*h/2d0
18         y(i) = h**2*r(x)
19     end do
20     ! DGTSV( N, NRHS, DL, D, DU, B, LDB, INFO )
21     call dgtsv(N+1,1,DL,D,DU,y,N+1,info)
22 end subroutine linear_fd

```

Explain what is the subroutine dgtsv in line 21 for. Will iterative methods perform better for large N? Justify your answer. (4 marks)

Solution. The subroutine dgtsv solves a linear algebra problem $Ax = b$ with A a tridiagonal matrix using efficient PLU factorisation. [1+1=2 marks]

No. When N is large, dgtsv takes $O(N)$ steps where as iterative methods such as Jacobi method or Gauss-Seidel method may take a longer time to converge. [2 marks] □

$$x^2 y'' = xy' + 3y + x \ln x - x^2, \quad 1 \leq x \leq 2, \quad y(1) = y(2) = 0$$

by finite-difference method with a step size $h = 0.2$. Make sure you **reduce the BVP to a linear algebra problem**.

Solution. By finite-difference approximation, we have

$$x_i^2 \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} = x_i \frac{y_{i+1} - y_{i-1}}{2h} + 3y_i + x_i \ln x_i - x_i^2, \quad i = 1, 2, 3, 4.$$

After simplification, we have

$$(x_i^2 + \frac{hx_i}{2})y_{i-1} - (3h^2 + 2x_i^2)y_i + (x_i^2 - \frac{hx_i}{2})y_{i+1} = h^2(x_i \ln x_i - x_i^2)$$

and reduce the BVP to linear algebra problem

$$\begin{bmatrix} 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 1.56 & -3.00 & 1.32 & 0.00 & 0.00 & 0.00 \\ 0.00 & 2.10 & -4.04 & 1.82 & 0.00 & 0.00 \\ 0.00 & 0.00 & 2.72 & -5.24 & 2.40 & 0.00 \\ 0.00 & 0.00 & 0.00 & 3.42 & -6.60 & 3.06 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} 0 \\ -0.04884857 \\ -0.05955755 \\ -0.07231977 \\ -0.08727936 \\ 0 \end{bmatrix}.$$

$$\mathbf{x} = 1 : 0.2 : 2; \mathbf{y} = \begin{bmatrix} 0.0000000000 \\ 0.0493637994 \\ 0.0751839643 \\ 0.0772097156 \\ 0.0532328163 \\ 0.0000000000 \end{bmatrix}$$

□

ers:p72:ex1.C.1

Example 46.23 (Tutorial 7, Q4). Consider the following boundary value problem involving two coupled fields governed by ODE:

$$-\frac{d^2\varphi}{dx^2} = 1 + \psi(x), \quad -\frac{d^2\psi}{dx^2} = \varphi(x), \quad 0 \leq x \leq 1.$$

with boundary conditions $\varphi(0) = \varphi(1) = \psi(0) = \psi(1) = 0$. We wish to solve this problem numerically using the method of finite differences for a uniform grid of points $x_j, j = 1, 2, \dots, N$, in the interior of the domain.

1. Derive the corresponding linear algebra problem.

$$\text{Solution. } \begin{cases} \varphi_{i-1} - 2\varphi_i + \varphi_{i+1} + h^2\psi_i = -h^2, \\ h^2\varphi_i + \psi_{i-1} - 2\psi_i + \psi_{i+1} = 0. \end{cases}$$

$$\varphi_{i+1} - 2\varphi_i + \varphi_{i-1} = -h^2(1 + \psi_i),$$

$$\psi_{i+1} - 2\psi_i + \psi_{i-1} = -h^2\varphi_i.$$

□

2. Which of the following ways of stacking the unknowns into a single vector do you recommend using? Explain your reasoning.

$$\mathbf{x} = \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_N \\ \psi_1 \\ \psi_2 \\ \vdots \\ \psi_N \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} \varphi_1 \\ \psi_1 \\ \varphi_2 \\ \psi_2 \\ \vdots \\ \vdots \\ \varphi_N \\ \psi_N \end{bmatrix}.$$

This ensures, by Theorem [46.1](#), that a unique solution exists.

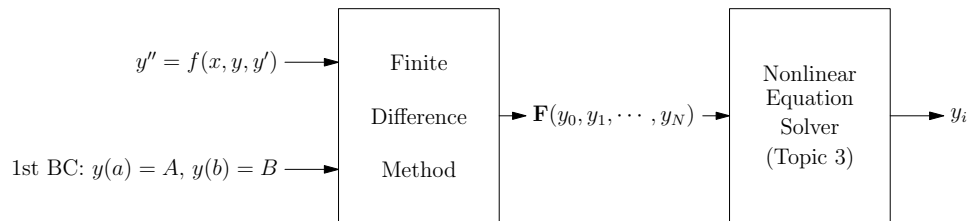
As in the linear case, we divide $[a, b]$ into N equal subintervals whose endpoints are at $x_i = a + ih$, for $i = 0, 1, \dots, N$. Assuming that the exact solution has a bounded fourth derivative allows us to replace $y(x_i)$ and $y'(x_i)$ in [\(46.2\)](#) by the appropriate centre-difference approximations, we obtain

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} = f(x_i, y_i, \frac{y_{i+1} - y_{i-1}}{2h}), \quad i = 1, \dots, N-1$$

where $y_0 = A$, $y_N = B$. From this we can obtain an $(N-1) \times (N-1)$ nonlinear equation:

$$\begin{aligned} -A + 2y_1 - y_2 + h^2 f\left(x_1, y_1, \frac{y_2 - A}{2h}\right) &= 0 \\ -y_1 + 2y_2 - y_3 + h^2 f\left(x_2, y_2, \frac{y_3 - y_1}{2h}\right) &= 0 \\ &\vdots \\ -y_{N-3} + 2y_{N-2} - y_{N-1} + h^2 f\left(x_{N-2}, y_{N-2}, \frac{y_{N-1} - y_{N-3}}{2h}\right) &= 0 \\ -y_{N-2} + 2y_{N-1} - B + h^2 f\left(x_{N-1}, y_{N-1}, \frac{B - y_{N-2}}{2h}\right) &= 0 \end{aligned}$$

which can be solved by Newton method. The block diagram view of this method is shown below.



A Fortran 2008 implementation using secant method is stated below.

```

1  subroutine nlinear_fd(npt,f,a,b,alph,beta,y,info)
2    ! f(x,y,y'); BC: y(a)=alph,y(b)=beta
3    use textbook_rootn, only : newton
4    procedure(twoptbvfpf) :: f
5    integer :: npt, i, info, nsteps
6    double precision :: a,b,alph,beta,h,x0(npt-1)
7    double precision, intent(out) :: y(npt-1)
8    h = (b-a)/real(npt,8) ! Convert integer 8 to double
9    x0 = ((alph + i*(beta-alph)/npt, i=1,npt-1)/)
10   call newton(fdm_f, npt-1, x0, y, info, nsteps)
11   print *, "y=", y
12   contains
13   function fdm_f(np, y)
14     intent(in) :: np, y
15     integer :: np ! Number of Pieces
16     double precision :: y(np), fdm_f(np)
17     fdm_f( 1) = -alph + 2d0*y( 1)-y(2) + h**2* &
18               f(a+ 1*h, y( 1), (y(2) - alph)/2d0/h)
19     fdm_f(np) = -y(np-1)+2d0*y(np)-beta + h**2* &
20               f(a+np*h, y(np), (beta-y(np-1))/2d0/h)
21     do i=2,np-1
22       fdm_f(i) = -y(i-1)+2d0*y(i)-y(i+1) + h**2* &
23               f(a+ i*h, y( i), (y(i+1)-y(i-1))/2d0/h)
24     end do
25   end function fdm_f
26 end subroutine nlinear_fd
  
```

It is interesting to compare this to the algorithm in Burden and Faires [2015, page 693].

Example 46.24. Write a computer program to solve the BVP [\(46.11\)](#).

Solution: A sample Fortran code is listed below.

```

1 program eg_odebvp_nlfed ! Nonlinear Shooting Example from Burden & Faires, p700
2   use iso_c_binding, only : c_double
3   use textbook_odebvp, only : nlinear_fd
4   implicit none
5   integer, parameter :: npt=20
6   integer :: info
7   real(c_double) :: y(npt-1)
8   call nlinear_fd(20,f,1d0,3d0,17d0,43/3d0,y,info)
9 contains
10  function f(x,y,yp)
11    use iso_c_binding, only : c_double
12    intent(in) :: x, y, yp
13    real(c_double) :: x, y, yp, f
14    f = (32d0+2d0*x**3-y*yp)/8d0
15  end function f
16 end program eg_odebvp_nlfed

```

Remark: It is an interesting assignment for students to compare this with the nonlinear shooting as well the algorithms used in Burden and Faires [2015].

46.6 Finite Element Method (FEM)

sec:fem

The shooting method for approximating the solution to a BVP (46.2) replaced the BVP with pair of ODE-IVP. The finite-difference approach replaces the continuous operation of differentiation with the discrete operation of finite differences.

The third approach — *finite element methods* (FEM), based on the *Galerkin method*, which includes *Rayleigh-Ritz (variational) method* as a special case, are extremely important class of numerical methods for solving BVPs in PDEs [Ascher et al., 1994, 266]. The solution of a BVP also satisfy an equivalent variational formulation. In this approach, BVP is first reformulated as a problem of choosing, from the set of all sufficiently differentiable functions satisfying the boundary conditions, the function to minimise a certain integral. Then the set of feasible functions is reduced in size, and an approximation is found from this set to minimize the integral. This gives our approximation to the solution of the BVP (46.2).

Consider approximating the solution to a linear two-point BVP from beam-stress analysis. This BVP is described by the ODE

$$-\frac{d}{dx}(p(x)\frac{dy}{dx}) + q(x)y = f(x), \quad \text{for } 0 \leq x \leq 1,$$

with the boundary conditions

$$y(0) = y(1) = 0.$$

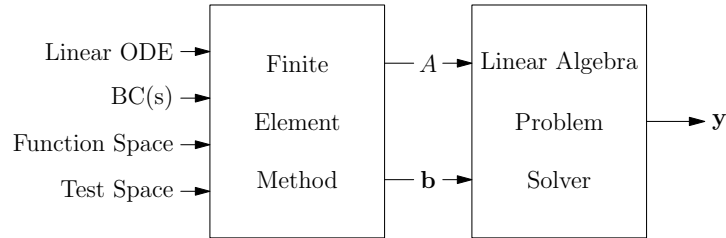
The ODE describes the deflection $y(x)$ of a beam of length 1 with variable cross section represented by $q(x)$. The deflection is due to the added stresses $p(x)$ and $f(x)$.

For specific $p(x)$ and $q(x)$, the above equation is equivalent to minimising the integral (or functional)

$$I[u] = \int_0^1 \{p(x)[u'(x)]^2 + q(x)[u(x)]^2 - 2f(x)u(x)\} dx$$

where u are second order differentiable and continuous functions on $[0, 1]$. With the right basis, it is possible to transform the minimisation problem to linear algebra problem. This whole process is called the *Rayleigh-Ritz method* [Burden and Faires, 2015].

The most difficult step in FEM is to rewrite the linear ODE in *weak form* and to choose an appropriate *function space* and *test space*. The block diagram is shown below.



The following example demonstrates this process roughly.

lu-jin-pu:p238

Example 46.25. Consider the following BVP:

$$\begin{aligned} -u'' &= 2, & 0 < x < 1, \\ u(0) &= 0, & u'(1) = 0. \end{aligned}$$

1. Find the exact solution.

Solution: By integration, we have $u = -x^2 + Ax + C$. $u(0) = 0 \Rightarrow C = 0$ and $u'(1) = 0 \Rightarrow -2 + A = 0 \Rightarrow A = 2$. The exact solution is $u = 2x - x^2$.

2. Use finite element method with linear approximation over grid with 4 equal cells to estimate u .

Solution: The original BVP will be multiplied with a test function v and re-written as functional form:

$$\begin{aligned} \int_0^1 (-u'')v dx &= \int_0^1 2v dx \Rightarrow \left[-u'v \right]_0^1 - \int_0^1 (-u')v' dx = \int_0^1 2v dx \\ &\Rightarrow -u'(1)v(1) + u'(0)v(0) + \int_0^1 u'v' dx = \int_0^1 2v dx \end{aligned}$$

We will choose test functions such that $v(0) = 0$ so that the functional form reduces to

$$\int_0^1 u'v' dx = \int_0^1 2v dx. \quad (46.13)$$

e:fem:1

Let

$$u \approx u_0\phi_0(x) + u_1\phi_1(x) + u_2\phi_2(x) + u_3\phi_3(x) + u_4\phi_4(x)$$

and choose a basis for the space of test functions (closely related the notion of spline in Topic 4):

$$v_0(x) = \begin{cases} 4x & 0 \leq x < \frac{1}{4}, \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad v_i(x) = \phi_i(x), \quad i = 1, 2, 3, 4,$$

where

$$\begin{aligned} \phi_0(x) &= \begin{cases} 1 - 4x & 0 \leq x < \frac{1}{4}, \\ 0 & \text{otherwise} \end{cases}, & \phi_4(x) &= \begin{cases} 4(x - \frac{3}{4}) & \frac{3}{4} \leq x \leq 1, \\ 0 & \text{otherwise} \end{cases}, \\ \phi_i(x) &= \begin{cases} 4(x - \frac{i-1}{4}) & \frac{i-1}{4} \leq x < \frac{i}{4}, \\ 1 - 4(x - \frac{i}{4}) & \frac{i}{4} \leq x < \frac{i+1}{4}, \\ 0 & \text{otherwise.} \end{cases} & i &= 1, 2, 3. \end{aligned}$$

Equation [46.13](#) becomes

$$\begin{aligned}\int_0^1 (u_0 \phi'_0(x) + \dots + u_4 \phi'_4(x)) v'_0(x) dx &= \int_0^1 2v_0(x) dx \Rightarrow \frac{16}{4}u_0 + \frac{16}{4}u_1 = \frac{2}{8} \\ \int_0^1 (u_0 \phi'_0(x) + \dots + u_4 \phi'_4(x)) v'_1(x) dx &= \int_0^1 2v_1(x) dx \Rightarrow \frac{16}{4}u_0 + \frac{32}{4}u_1 + \frac{-16}{4}u_2 = \frac{2}{4} \\ \int_0^1 (u_0 \phi'_0(x) + \dots + u_4 \phi'_4(x)) v'_2(x) dx &= \int_0^1 2v_2(x) dx \Rightarrow -\frac{16}{4}u_1 + \frac{32}{4}u_2 + \frac{-16}{4}u_3 = \frac{2}{4} \\ \int_0^1 (u_0 \phi'_0(x) + \dots + u_4 \phi'_4(x)) v'_3(x) dx &= \int_0^1 2v_3(x) dx \Rightarrow -\frac{16}{4}u_2 + \frac{32}{4}u_3 + \frac{-16}{4}u_4 = \frac{2}{4} \\ \int_0^1 (u_0 \phi'_0(x) + \dots + u_4 \phi'_4(x)) v'_4(x) dx &= \int_0^1 2v_4(x) dx \Rightarrow -\frac{16}{4}u_3 + \frac{16}{4}u_4 = \frac{2}{8}\end{aligned}$$

Since for each cell, $\phi'_i(x)$ and $v'_j(x)$ are just polynomial of degree 0, therefore, Gaussian integral could be applied to calculate the exact value of the left-hand-side of [\(46.13\)](#) and approximate the right-hand-side of [\(46.13\)](#).

We obtained a linear algebra problem below

$$\begin{bmatrix} 4 & 4 & 0 & 0 & 0 \\ 4 & 8 & -4 & 0 & 0 \\ 0 & -4 & 8 & -4 & 0 \\ 0 & 0 & -4 & 8 & -4 \\ 0 & 0 & 0 & -4 & 4 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} 1/4 \\ 1/2 \\ 1/2 \\ 1/2 \\ 1/4 \end{bmatrix}.$$

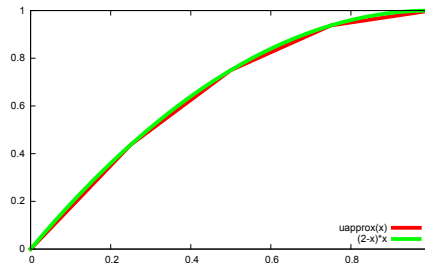
Our $v_0(x)$ was not properly chosen and is linearly dependent on $\{v_1, v_2, v_3, v_4\}$. Hence the matrix above is singular. We need to apply the boundary condition $v(0) = 0 \Rightarrow v_0 = 0$ to obtain

$$\begin{bmatrix} 8 & -4 & 0 & 0 \\ -4 & 8 & -4 & 0 \\ 0 & -4 & 8 & -4 \\ 0 & 0 & -4 & 4 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/4 \end{bmatrix}.$$

The solution of the above linear algebra problem is $u_1 = 0.4375$, $u_2 = 0.75$, $u_3 = 0.9375$ and $u_4 = 1.0$.

Hence, $u \approx 0.4375\phi_1(x) + 0.75\phi_2(x) + 0.9375\phi_3(x) + \phi_4(x)$.

Plotting it against the exact solution we obtain



Example 46.26 (Tutorial 7, Q5). Given

$$y''(x) = f(x), \quad 0 < x < 1, \quad y(0) = y(1) = 0.$$

Let $y = \sum_i y_i B_i(x)$ where $B_i(x)$ is a Lagrange interpolation basis such that $B_i(0) = B_i(1) = 0$ (Topic 4). Derive a linear algebra problem using the FEM. Explain how Gauss-Legendre quadrature rule can be applied.

Solution. By using $B_i(x)$ as the basis of test functions, we have the weak formulation

$$\begin{aligned}\int_0^1 y''(x)B_i(x)dx &= \int_0^1 f(x)B_i(x)dx \\ \Rightarrow \left[y'(x)B_i(x) \right]_{x=0}^{x=1} - \int_0^1 y'(x)B'_i(x)dx &= \int_0^1 f(x)B_i(x)dx \\ \Rightarrow \sum_j^\infty y_j \int_0^1 B'_j(x)B'_i(x)dx &= - \int_0^1 f(x)B_i(x)dx\end{aligned}$$

Here the integral $\int_0^1 B'_j(x)B'_i(x)dx$ could be evaluated using Gauss-Legendre quadrature rule. $\square$

Chapter 47

Linear Partial Differential Equations

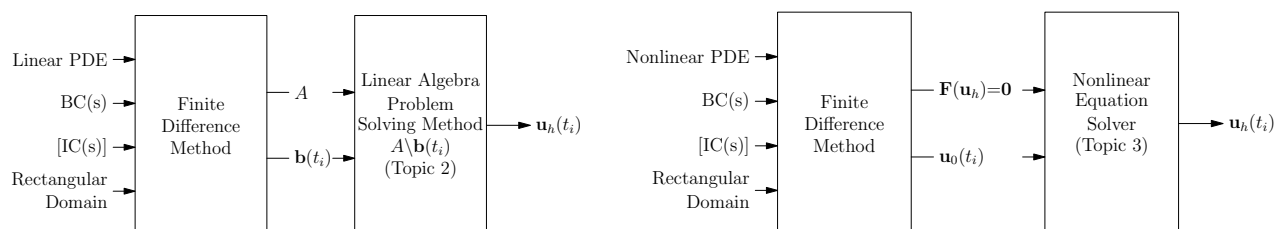
ch:pde-linear

Partial differential equation (PDE) is such a huge branch in mathematics, it is doubtful whether there is a single person who knows everything. Many PDEs in physics and engineering can be classified into *elliptic equation*, *parabolic equation* and *hyperbolic equation*. Instead of studying the general theory of PDE, the decision to limit ourselves to these equations is more practical.

The numerical methods for solving PDE are so complicated that what will be studied in this topic are just a tiny portion of PDE solution methods. Conceptually, the numerical methods for PDE in this topic is a generalisation of the numerical methods of ODE-IVP in Topic 6 and ODE-BVP in Topic 7. ODE-IVP describes the dynamic behaviour of the state, which is denoted as a point, in “system state space”. A simple *time-dependent PDE*, to be investigated in Sections 47.3 and 47.4, describe the dynamic behaviour of a “field” characterised by one “space” variable, e.g. temperature field, pressure field, etc., which changes with respect to time. The “steady state” behaviour of a time-dependent PDE is characterised by a time-independent PDE with two “space” variables, as in Section 47.2.

All the knowledge we learned in Topic 2 to Topic 5 are required in the design of numerical methods for PDE. When finite difference approximation (Topics 3 and 4) is applied to a linear PDE problem, we will normally obtain a sparse linear algebra problem. Interpolation (Topic 4), and Gaussian quadrature (and cubature) from Topic 5, are the foundations of *finite element method (FEM)*. Despite its importance in the numerical solution of PDE, FEM is not part of our syllabus.

When we apply *finite difference method (FDM)* (based on finite difference approximation) and FEM to the *linear PDE* on a simple rectangular domain, we will obtain numerical linear algebra problems $\mathbf{Ax} = \mathbf{b}$ and the numerical methods in Topic 2 can be used to obtain numerical solutions. The FDM framework can be extended to solve *nonlinear PDE* over a finite rectangular domain. Similar to Topic 7, a systems of nonlinear equations will be produced and the solution process is complicated and will not be discussed in this topic. These “discretisation” processes can be visualised as follows.



FDM is only suitable for linear PDE on a finite rectangular domain with equally spaced mesh. When the domain is *irregular* or the mesh spacing is not equally spaced, FDM may work badly. In these situations, FVM and FEM are comparatively simpler because of the geometric flexibility with the finite elements. One of the advantage of FVM and FEM over FDM is the relative ease with which the *boundary conditions (BCs)* of the problem are handled. Many physical problems have BCs involving derivatives and of irregular shape. BCs of this type are

difficult to handle using FDM since each BC involving a derivative must be approximated by a difference quotient at the grid points.

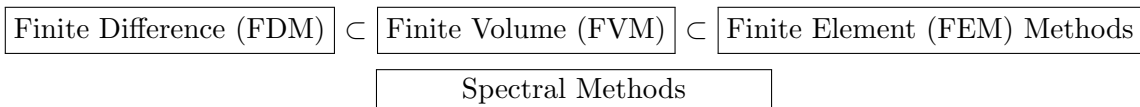
47.1 Overview

We must be aware of *limitations* of numerical methods. Many beginners and engineers tend to think that when a certain numerical method is employed, they will always get a good approximation of the solution. However, there are many problems where the numerical methods will either be too slow or will produce wrong results. Therefore, a careful mathematical analysis (involving complex mathematics [Dautray and Lions, 1988]) is necessary to *make sure* that not only that the *solution exists* for the mathematical model but the numerical method is suitable to give a good approximate solution based on the mathematical properties of the problem.

In calculus, we learn that $x^2 + 1 = 0$ does not have a real solution. In PDE theory, mathematicians told us that only nicely constructed PDE will have solutions. Solving a nonlinear equation $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ (Topic 3) normally gives us a solution vector $\mathbf{x}^*$. But solving a PDE gives us many functions over a domain Ω in $\mathbb{R}^n$ depending on whether the PDE is time-varying (if the PDE is time-invariant, remove the t):

$$u = u(t, \mathbf{x}), \quad (t, \mathbf{x}) \in [0, t_{\max}) \times \Omega, \quad \Omega \subset \mathbb{R}^n.$$

To make sure that a PDE problem has a *unique* solution, we need to add conditions until it becomes *well-posed*. In general, the solution of a well-posed PDE problem is so complicated that more than hundreds of books are written. Grossmann and Roos [2007], White [2004], Quarteroni and Valli [2008] and Shapira [2006] are just a few books on the introduction to the numerical methods of PDE. The numerical methods for solving well-posed PDE problems can be classified roughly into four classes [Mattiussi, 1997]:



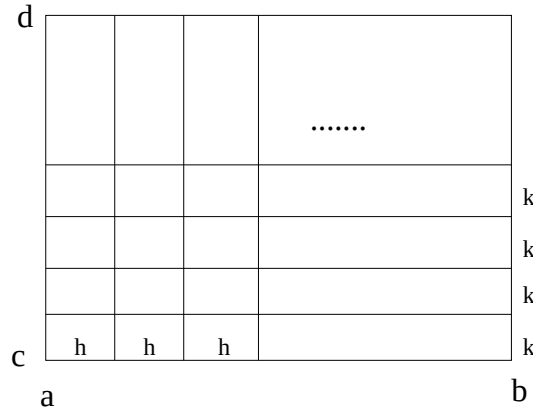
FDM is the least complicated class. According to Grossmann and Roos [2007, 34], the FDM is easily set up and on a coarse mesh it generally supplies, after a relatively short calculation, an overview of the solution function that is often sufficient in practice. Therefore, we are limited to FDMs in this topic. Leveque [2007], Smith [1985], Duffy [2006], Achdou and Pironneau [2005], and Zhu et al. [2013] are some books on FDMs.

For FVM, FEM and spectral methods, which are not discussed in this topic, read Croft [1998], Patankar [1980], Versteeg and Malalasekera [2007], Hughes [1987], Brenner and Scott [2008] Topper [2005], Akin [1982], Nikishkov [2010], Kikuchi [1986], Fish and Belytschko [2007], Löhner [2008], Lewis et al. [2004], Zienkiewicz et al. [2005], Monk [2003], and Canuto et al. [2006]. Note that engineers called the process of applying FEM to analyse an engineering problem by *Finite Element Analysis* (FEA).

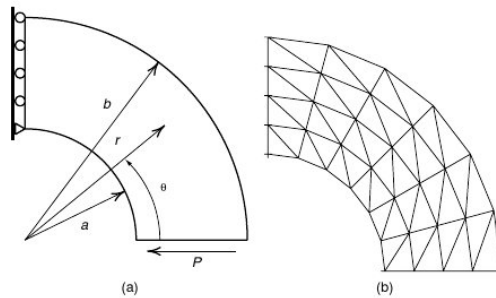
For various other techniques, such as method of lines, spectral methods, specialised FEM such as discontinuous Galerkin method, higher order methods and adaptive techniques, read Schiesser and Griffiths [2009], Kopriva [2009], Hesthaven and Warburton [2008], Deuffhard and Weiser [2012], and Deville et al. [2004].

FEM is mentioned or taught in UEMM3243 Engineering Analysis Topic 3 (only a brief introduction), UEMX4323 Bridge Engineering Topic 6, UEME2163 Solid Mechanics II Topic 8, UEME4343 Computational Methods in Solid Mechanics Topics 4–6, and UEMX4293 Finite Element Method in Structural Engineering. FVM is mentioned in UEME4213 Computational Fluid Dynamics (CFD) and practised with the commercial software Gambit and Fluent. Gambit is a meshing software package typically purchased with Fluent. Fluent is a popular commercial CFD package sold by ANSYS. A CFD course using Python is available at <https://github.com/barbagroup/CFDPython>.

The domain Ω in which the PDE is defined may not be rectangular in engineering problems. However, for simple mathematical analysis, it is assumed to be rectangular (or there is a way to transform it to rectangular). When $n = 2$, $\Omega = [a, b] \times [c, d]$ and a rectangular grid is formed:



Examples of regular and irrregular non-rectangular meshes can be found in mesh generation books and PDE books such as George and Frey [2008]. A 2-D non-rectangular and irrregular mesh was given in Zienkiewicz et al. [2005, 43] as follows:



There are many people who ask the question of the usefulness of the subjects taught in universities, complaining that they are too theoretical. It is the hope that one at least understand why mathematical theories such as Calculus are required — they are the foundations for the development of numerical methods for continuous models (Topic 1). Numerical methods are not only the core behind computer-aided design, analysis and manufacturing, but also applied in finance, statistics and machine learning (one can find related the necessary mathematical knowledge to master machine learning from the Internet such as <https://canvas.cmu.edu/courses/603/assignments/syllabus>, <https://mml-book.github.io/>, <https://hplgit.github.io/fdm-book/doc/web/>, and <https://www.youtube.com/playlist?list=PL7y-1rk2cCsAqRtWoZ95z-GMcecVG5mzA>.)

According to Wikipedia, some of the applications of numerical methods in real-world engineering domains are listed below.

- Material Engineering PDE (Topic 8)
The modelling of biomaterials, carbon, ceramics, composite materials, glass, metals, nano-materials, polymers, refractory, semiconductors, thin films, functionally graded materials require very complicated partial differential equations which is beyond this subject.
- Mechanical Engineering ODE/DAE (Topics 6 & 7)
Mechanical engineering is a discipline of engineering that applies the principles of physics for the analysis, design, manufacturing, and maintenance of mechanical systems. It is the branch of engineering that involves the production and usage of heat and mechanical power for the design, production, and operation of machines and tools. Mechanical engineers design and analyse manufacturing plants, industrial equipment and machinery, heating and cooling systems, transport systems, aircraft, watercraft, robotics, weapons, medical devices, and others. Robotics models involve systems of ordinary differential equations and differential algebraic equations (DAE):

$$F(t, \mathbf{x}, \dot{\mathbf{x}}) = \mathbf{0}.$$

- Agricultural engineering Statistics (Design and Experiment)

It is the engineering discipline that applies engineering science and technology to agricultural production and processing. Agricultural engineering combines the disciplines of animal biology, plant biology, and mechanical, civil, electrical and chemical engineering principles with a knowledge of agricultural principles.

- Power Systems and Electrical Engineering ODE (Topic 6)
Power systems engineering is an electrical engineering that concerns with the generation (e.g. model with second-order ODE), transmission, distribution and utilisation of electric power as well as the electrical devices connected to such systems including generators, motors and transformers. A significant fraction of the field is concerned with the conversion between AC and DC power as well as the development of specialised power systems such as those used in aircraft or for electric railway networks.
- Chemical Engineering ODE/DAE (Topic 6)
Advancements in computer science found applications designing and managing plants, simplifying calculations and drawings that previously had to be done manually. Chemical engineering principles were used to produce DNA sequences in large quantities, making it possible for biotechnology to generate medicines, catalysts, etc. Currently, environmental-friendly chemical process design is of important concern. Therefore, the modelling of chemical pollution is also an important task.
- Electronic Engineering Matrix/ODE/DAE (Topics 4 & 6)
It is an engineering discipline where non-linear and active electrical components such as semiconductor devices are utilised to design electronic circuits, devices (e.g. signal processing devices for camera) and systems (e.g. digital computers, control systems, etc.) Electronic components are mostly model by nonlinear functions and ODE.
- Telecommunications Engineering ODE/PDE (Topics 6 & 8)
This is an engineering discipline that brings together electrical engineering with computer science to enhance telecommunication systems. The work ranges from basic circuit design to strategic mass developments. A telecommunication engineer is responsible for designing and overseeing the installation of telecommunications equipment and facilities, such as complex electronic switching systems, copper wire telephone facilities, and fibre optics.
- Financial engineering . Time Series, Stochastic DE (UECM3763 Computational Finance)
It is a multidisciplinary field involving financial theory, the methods of engineering, the tools of mathematics and the practice of programming. Financial engineering draws on tools from applied mathematics, computer science, statistics and economic theory. In broadest definition, anyone who uses technical tools in finance could be called a financial engineer, for example any computer programmer in a bank or any statistician in a government economic bureau. However, most practitioners restrict the term to someone educated in the full range of tools of modern finance and whose work is informed by financial theory.

The various numerical methods to be studied are summarised in the Figure [Fig:1](#) 47.1.

It is also useful to see how finance use numerical methods. A good place to start is probably to have a look at QuantLib, a C++ library for quantitative finance. There are a lot comments on the Internet on why some lecturers are still using Fortran, which is regarded as an “old” language. My argument is simply because it is “less complex”. For example, to solve a nonlinear equation (Topic 3), one can write a C++ program below:

```

1 #include <ql/types.hpp>
2 #include <ql/math/solversld/secant.hpp>
3 #include <ql/math/solversld/newton.hpp>
4 #include <ql/math/solversld/bisection.hpp>
5 #include <cmath>
6 #include <iostream>
7
8 using namespace QuantLib;
9
10 class F1 {
```

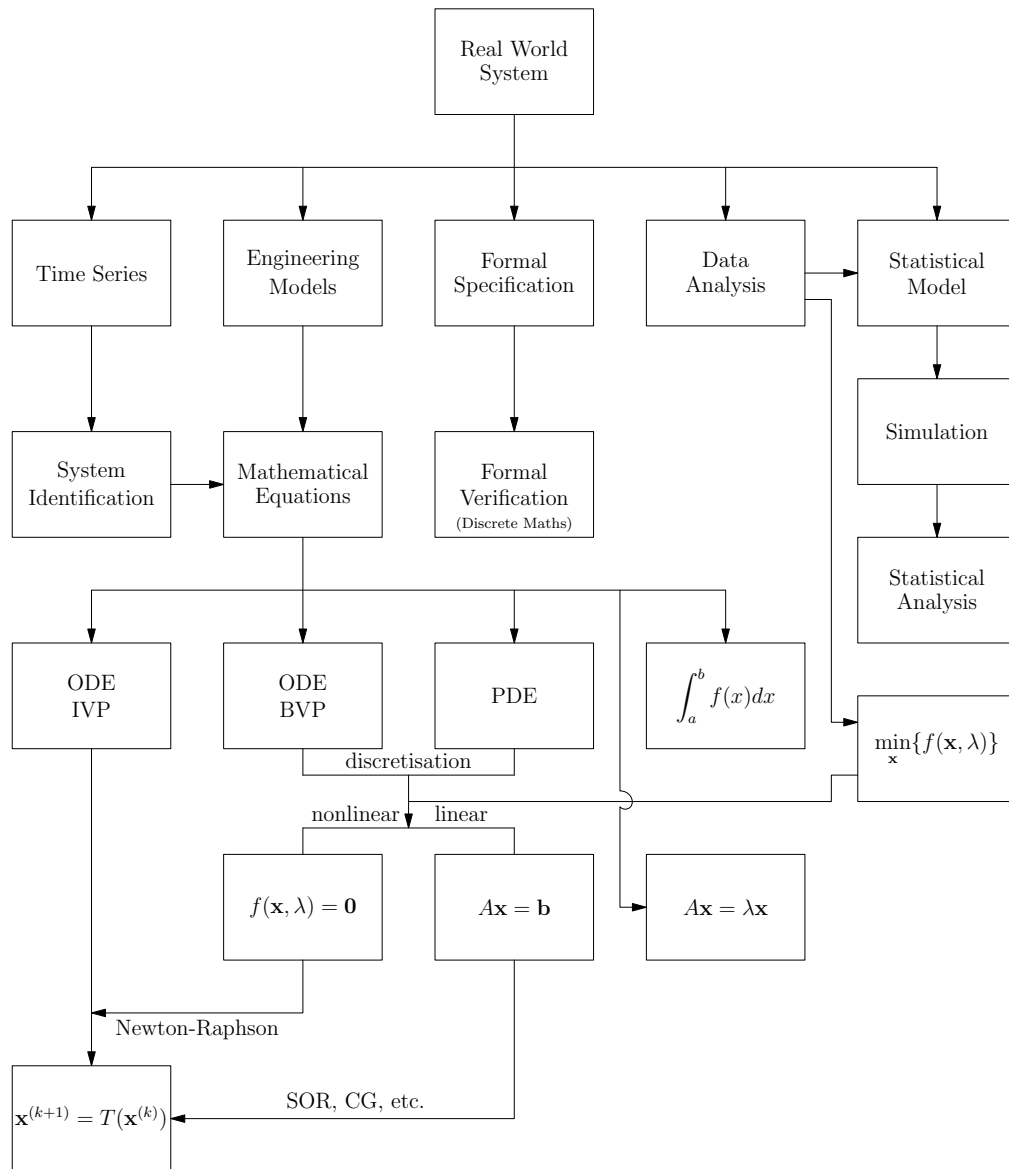


Figure 47.1: Real-world problems and numerical methods

fig:1

```

11     public:
12     Real operator()(Real x) const { return pow(x,5) - 3.0*x + 7.0; }
13     Real derivative(Real x) const { return 5.0*pow(x,4) - 3.0; }
14 } f;
15
16 /*
17 // A simple function below works for Secant and Bisection
18 Real f(Real x) {
19     return pow(x,5) - 3.0*x + 7.0;
20 }
21 */
22
23 int main() {
24     Secant solver1;
25     Newton solver2;
26     Bisection solver3;
27     const Real a = -2.0, b = 1.0, guess = 0.0, prec = 1e-8;
28     Real root;
29     root = solver1.solve(f, prec, guess, a, b);
30     std::cout << "Secant(f)      = " << root << "\n";
31     root = solver2.solve(f, prec, guess, a, b);
32     std::cout << "Newton(f)     = " << root << "\n";
33     root = solver3.solve(f, prec, guess, a, b);
34     std::cout << "Bisection(f) = " << root << "\n";
35 }

```

With the undergraduate C++ knowledge, it is impossible to explain the implementation of Secant, Newton and Bisection classes. On the other hand, the Fortran implementations in Topic 3 are much easier to understand.

According to <http://www.quantlib.org/slides/qlws14/duffy.pdf> (one can also see how complex C++ is from these slides), quite a few financial models are PDEs:

- Equity models: Black-Scholes, Heston, Asian;
- Interest rate models: CIR, HW, Cheyette;
- Fixed income models: caps, floors, swap, swaptions;
- UVM, Anchoring, etc.
- Hamilton-Jacobi-Bellman model
- Fokker-Planck model, Kolmogorov model

Numerical methods for PDE are also implemented in QuantLib, but only for parabolic equation (Section 47.3). ^{sec:pde:parabolic}To illustrate the complexity of C++ language, a header file `ql/methods/finitedifferences/boundarycondition.hpp` (slightly modified) is listed below.

```

1  /*
2  Copyright (C) 2000, 2001, 2002, 2003 RiskMap srl
3
4  This file is part of QuantLib, a free-software/open-source library
5  for financial quantitative analysts and developers - http://quantlib.org/
6
7  QuantLib is free software: you can redistribute it and/or modify it
8  under the terms of the QuantLib license (mostly a modified BSD license)
9  */
10
11 /*! \file boundarycondition.hpp
12     \brief boundary conditions for differential operators
13 */
14
15 #ifndef quantlib_boundary_condition_hpp
16 #define quantlib_boundary_condition_hpp
17
18 #include <ql/utilities/null.hpp>

```

```

19 #include <ql/methods/finitedifferences/tridiagonaloperator.hpp>
20
21 namespace QuantLib {
22
23     //! Abstract boundary condition class for finite difference problems
24     /*! \ingroup findiff */
25     template <class Operator>
26     class BoundaryCondition {
27     public:
28         // types and enumerations
29         typedef Operator operator_type;
30         typedef typename Operator::array_type array_type;
31         //! \todo Generalize for n-dimensional conditions
32         enum Side { None, Upper, Lower };
33         // destructor
34         virtual ~BoundaryCondition() {}
35         // interface
36         /*! This method modifies an operator \f$ L \f$ before it is
37             applied to an array \f$ u \f$ so that \f$ v = Lu \f$ will
38             satisfy the given condition. */
39         virtual void applyBeforeApplying(operator_type&) const = 0;
40         /*! This method modifies an array \f$ u \f$ so that it satisfies
41             the given condition. */
42         virtual void applyAfterApplying(array_type&) const = 0;
43         /*! This method modifies an operator \f$ L \f$ before the linear
44             system \f$ Lu' = u \f$ is solved so that \f$ u' \f$ will
45             satisfy the given condition. */
46         virtual void applyBeforeSolving(operator_type&,
47                                         array_type& rhs) const = 0;
48         /*! This method modifies an array \f$ u \f$ so that it satisfies
49             the given condition. */
50         virtual void applyAfterSolving(array_type&) const = 0;
51         /*! This method sets the current time for time-dependent
52             boundary conditions. */
53         virtual void setTime(Time t) = 0;
54     };
55
56     // Time-independent boundary conditions for tridiagonal operators
57
58     //! Neumann boundary condition (i.e., constant derivative)
59     /*! \warning The value passed must not be the value of the derivative.
60         Instead, it must be comprehensive of the grid step
61         between the first two points--i.e., it must be the
62         difference between f[0] and f[1].
63         \todo generalize to time-dependent conditions.
64
65         \ingroup findiff
66     */
67     class NeumannBC : public BoundaryCondition<TridiagonalOperator> {
68     public:
69         NeumannBC(Real value, Side side);
70         // interface
71         void applyBeforeApplying(TridiagonalOperator&) const;
72         void applyAfterApplying(Array&) const;
73         void applyBeforeSolving(TridiagonalOperator&, Array& rhs) const;
74         void applyAfterSolving(Array&) const;
75         void setTime(Time) {}
76     private:
77         Real value_;
78         Side side_;
79     };
80
81     //! Neumann boundary condition (i.e., constant value)
82     /*! \todo generalize to time-dependent conditions.
83         \ingroup findiff
84     */
85     class DirichletBC : public BoundaryCondition<TridiagonalOperator> {

```

```

86     public:
87         DirichletBC(Real value, Side side);
88         // interface
89         void applyBeforeApplying(TridiagonalOperator&) const;
90         void applyAfterApplying(Array&) const;
91         void applyBeforeSolving(TridiagonalOperator&, Array& rhs) const;
92         void applyAfterSolving(Array&) const;
93         void setTime(Time) {}
94     private:
95         Real value_;
96         Side side_;
97 };
98
99 }
100 #endif

```

The boundary condition is defined as an “interface”. Two kinds of boundary conditions (BCs) are implemented, one is the Type 1 BC (which was mentioned in Topic 7, called `DirichletBC` in QuantLib) and the other is the Type 2 BC (called `NeumannBC` in QuantLib). The `TridiagonalOperator` is responsible for the construction of tri-diagonal matrix (Topic 2).

47.2 Elliptic Problems

Elliptic problems are pure PDE-BVP, many of them arise from steady state solutions of transient problems of PDE-BVP-IVP. For example, if we consider the general transient heat flow problem for diffusion plus a source term, we would arrive at the time-dependent PDE

$$\frac{\partial u}{\partial t} = \Delta u - f(\mathbf{x}) \quad (47.1)$$

and an initial condition (plus appropriate boundary conditions). Section 47.3 discusses a simple case of the above heat flow problem where $f(\mathbf{x}) = 0$. Starting with the initial condition and/or appropriate boundary conditions, the state would tend to a steady state where $\frac{\partial u}{\partial t} = 0$ and we would obtain a *Poisson equation* (a type of elliptic problem) from (47.1):

$$\Delta u = f(\mathbf{x}) \quad \text{or} \quad \frac{\partial^2 u}{\partial x_1^2} + \cdots + \frac{\partial^2 u}{\partial x_n^2} = f(x_1, \cdots, x_n), \quad \mathbf{x} = (x_1, \cdots, x_n) \in \Omega. \quad (47.2)$$

When $n = 1$ and considering a 1-D interval $\Omega = [a, b]$, (47.2) becomes a 1-D *Poisson equation*:

$$\frac{d^2 u}{dx^2} = f(x), \quad a < x < b. \quad (47.3)$$

In addition to the given boundary conditions, it is a special case of the ODE-BVP we encountered in Topic 7.

When $n = 2$ and considering a 2-D rectangle region $\Omega = (a, b) \times (c, d)$, (47.2) becomes a 2-D *Poisson equation*:

$$u_{xx} + u_{yy} = f(x, y), \quad (x, y) \in R := (a, b) \times (c, d). \quad (47.4)$$

In this section, we will only consider the type 1 boundary conditions:

$$\begin{aligned} u|_{x=a} &= f(a, y) = f_1(y), & u|_{x=b} &= f(b, y) = f_2(y), \\ u|_{y=c} &= f(x, c) = g_1(x), & u|_{y=d} &= f(x, d) = g_2(x). \end{aligned} \quad (47.5)$$

More specifically, we will study how to derive a linear algebra problem associated with the 2-D elliptic PDE-BVP, which consists (47.4) and (47.5), using finite difference approximation. Type 2 and type 3 boundary conditions will not be discussed.

The u and f in a Poisson equation has the following meaning depending on the physical context.

Application	u	f
Steady state temperature in a solid	Temperature	Heat source/volume
Static deflection of a thin membrane in two dimensions	Deflection	Pressure
Electrostatics	Electrostatic potential	Charge/unit volume
Incompressible irrotational flow in 2-D or 3-D	Velocity potential	Mass source strength/volume
2-D incompressible Steady flow	Stream function	–Vorticity
Newtonian gravitation	Gravitational potential	Mass density

47.2.1 Finite Difference Approximation

The theory of finite difference is based on Taylor's approximation (Topic 3), i.e. the value of function f at $x + h$ can be estimated from the values the function f and its derivatives at x :

$$f(x + h) = f(x) + f'(x)h + f''(x)\frac{h^2}{2!} + O(h^3).$$

Here, we require h to be sufficiently small. To derive more general finite difference formula, the interpolation technique (Topic 4) may be used to obtain [Mathews and Fink, 2004] the centred difference approximations with order $O(h^2)$:

- $f'(x) = \frac{f(x+h)-f(x-h)}{2h} - h^2 \frac{f^{(3)}(\xi)}{3!} \approx (f_1 - f_{-1})/(2h);$
- $f''(x) \approx (f_1 - 2f_0 + f_{-1})/h^2;$
- $f'''(x) \approx (f_2 - 2f_1 + 2f_{-1} - f_{-2})/(2h^3);$
- $f^{(4)}(x) \approx (f_2 - 4f_1 + 6f_0 - 4f_{-1} + f_{-2})/h^4;$

the centred difference approximations with order $O(h^4)$:

- $f'(x) = \frac{-f(x+2h)+8f(x+h)-8f(x-h)+f(x-2h)}{12h} + h^4 \frac{f^{(5)}(\xi)}{30} \approx (-f_2 + 8f_1 - 8f_{-1} + f_{-2})/(12h)$
- $f''(x) \approx (-f_2 + 16f_1 - 30f_0 + 16f_{-1} - f_{-2})/(12h^2)$
- $f'''(x) \approx (-f_3 + 8f_2 - 13f_1 + 13f_{-1} - 8f_{-2} + f_{-3})/(8h^3)$
- $f^{(4)}(x) \approx (-f_3 + 12f_2 - 39f_1 + 56f_0 - 39f_{-1} + 12f_{-2} - f_{-3})/(6h^4)$

the forward and backward difference approximations with order $O(h^2)$:

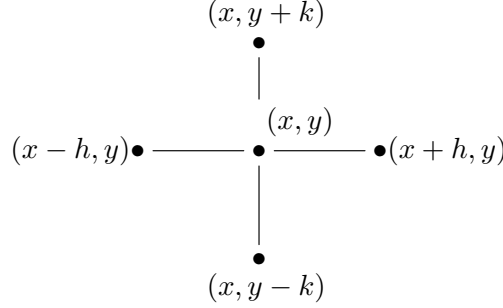
- $f'(x) \approx (-3f_0 + 4f_1 - f_2)/(2h)$
- $f'(x) \approx (3f_0 - 4f_{-1} + f_{-2})/(2h)$
- $f''(x) \approx (2f_0 - 5f_1 + 4f_2 - f_3)/h^2$
- $f''(x) \approx (2f_0 - 5f_{-1} + 4f_{-2} - f_{-3})/h^2$
- $f'''(x) \approx (-5f_0 + 18f_1 - 24f_2 + 14f_3 - 3f_4)/(2h^3)$
- $f'''(x) \approx (5f_0 - 18f_{-1} + 24f_{-2} - 14f_{-3} + 3f_{-4})/(2h^3)$
- $f^{(4)}(x) \approx (3f_0 - 14f_1 + 26f_2 - 24f_3 + 11f_4 - 2f_5)/h^4$
- $f^{(4)}(x) \approx (3f_0 - 14f_{-1} + 26f_{-2} - 24f_{-3} + 11f_{-4} - 2f_{-5})/h^4$

In this topic, we will limit ourselves to a two-variable function:

$$u = u(x, y) \quad \text{or} \quad u = u(x, t)$$

and ask: What is the “approximation” for the first order and second order partial derivatives at the point (x, y) (or (x, t) , which we will be using in the next two sections)?

To answer the above question, let us sketch the neighbourhood of u on a rectangular grid at (x, y) :



the finite difference approximation of u_x , u_y , u_{xx} , etc. based on the Taylor's approximation are given below.

- The *forward difference (FD)* formula for u_x at (x, y) is

$$\frac{\partial u}{\partial x}(x, y) = \frac{u(x+h, y) - u(x, y)}{h} + O(h).$$

- The *backward difference (BD)* formula for u_y at (x, y) is

$$\frac{\partial u}{\partial y}(x, y) = \frac{u(x, y) - u(x, y-k)}{k} + O(k).$$

- The *centre difference (CD)* formula for u_x at (x, y) is

$$\frac{\partial u}{\partial x}(x, y) = \frac{u(x+h, y) - u(x-h, y)}{2h} + O(h^2).$$

- The *centre difference* formula for u_{xx} at (x, y) is

$$\frac{\partial^2 u}{\partial x^2}(x, y) = \frac{u(x+h, y) - 2u(x, y) + u(x-h, y)}{h^2} + O(h^2).$$

In the case when u is a function of (x, t) , the finite difference approximation is similar to the above definitions.

47.2.2 Finite Difference Method on Rectangular Grid

By employing the finite difference approximation from Section 47.2.1, a numerical method to approximate the solution of the PDE-BVP (47.4) and (47.5) can be described as follows.

Method 47.1 (Three-Point Centre Difference Approximations to Elliptic Problem (47.4), (47.5)).

1. **SELECTING A grid** $\{(x_i, y_j) : 0 \leq i \leq M, 0 \leq j \leq N\}$: Choose appropriate integers M and N to define step sizes $h = \frac{b-a}{M}$ and $k = \frac{d-c}{N}$. Partition the interval $[a, b]$ into M equal parts of width h and the interval $[c, d]$ into N equal parts of width k . Let $x_i = a + ih$, $i = 0, 1, \dots, M$ and $y_j = c + jk$, $j = 0, 1, \dots, N$.
2. **APPLY AN APPROPRIATE FINITE DIFFERENCE APPROXIMATION** for u_{xx} and u_{yy} to the 2-D Poisson equation (47.4):

$$\frac{u(x_{i+1}, y_j) - 2u(x_i, y_j) + u(x_{i-1}, y_j))}{h^2} + \frac{u(x_i, y_{j+1}) - 2u(x_i, y_j) + u(x_i, y_{j-1}))}{k^2} + O(h^2 + k^2) = f(x_i, y_j).$$

Assume $u_{i,j} \approx u(x_i, y_j)$ and $f_{i,j} = f(x_i, y_i)$, we can obtain a finite difference equation:

$$\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{k^2} = f_{i,j}.$$

3. **SETUP A SYSTEM OF ALGEBRAIC EQUATIONS:** If we rewrite and rearrange the finite difference equation, we will obtain a linear algebra problem

$$2 \left[1 + \left(\frac{h}{k} \right)^2 \right] u_{i,j} - (u_{i+1,j} + u_{i-1,j}) - \left(\frac{h}{k} \right)^2 (u_{i,j+1} + u_{i,j-1}) = -h^2 f_{i,j}, \quad (47.6)$$

for $i = 1, \dots, M-1$ and $j = 1, \dots, N-1$. The boundary condition (47.5) becomes

$$\begin{aligned} u_{0,j} &= f_1(y_j), & u_{M,j} &= f_2(y_j), & j &= 0, 1, \dots, N, \\ u_{i,0} &= g_1(x_i), & u_{i,N} &= g_2(x_i), & i &= 0, 1, \dots, M. \end{aligned} \quad (47.7)$$

4. **SOLVE THE ALGEBRAIC EQUATIONS USING LINEAR ALGEBRA METHODS:** The matrix is large and normally diagonally dominant, therefore iteration methods are most appropriate.

By applying finite-difference approximation (47.4) and (47.5), we derived the linear algebra approximation (47.6) and (47.7). By using Taylor's polynomial approximation analysis, the approximation is found to have a local truncation error of order $O(h^2 + k^2)$. We learn from Topic 1 that rounding error happens with floating point numbers. How can we be sure that the errors will not be magnified?

When we look at (47.6), we discover that $2 \left[1 + \left(\frac{h}{k} \right)^2 \right] \geq |-1| + |-1| + |-\left(\frac{h}{k} \right)^2| + |-\left(\frac{h}{k} \right)^2|$ which means (47.6) is diagonally dominant. Therefore, the linear algebra problem has a unique solution and it is possible for us to use either the direct method for a block matrix or the iteration methods for sparse matrix.

A simple implementation of the FDM for solving (47.6) and (47.7) in Fortran is as follows:

code/nm/textbook_pdesolvers.f03

```
1 subroutine poisson2d_cs(M, N, Lx, Ly, lhs, rhs, top, bot, f, verbose)
2   ! Three-Point Centre-Difference for Poisson Eqn. u_xx + u_yy = f(x,y)
3   integer, intent(in) :: M, N
4   procedure(one_var_func) :: lhs, rhs, top, bot
5   procedure(two_var_func) :: f
6   logical, intent(in) :: verbose
7   double precision :: Lx, Ly, h, k, r, x, y
8   double precision, allocatable :: A(:, :), b(:)
9   integer, allocatable :: P(:)
10  integer :: i, j, idx, Asize, info
11
12  Asize = (M-1)*(N-1)
13  if (Asize > 1000) then
14    print *, "FAIL: Matrix size is larger than 1000"
15    return
16  end if
17  allocate(A(Asize, Asize)); allocate(b(Asize)); allocate(P(Asize))
18  h = Lx/M; k = Ly/N; r = (h/k)**2
19  do i=1, M-1          ! x-axis
20    do j=1, N-1        ! y-axis
21      idx = (j-1)*(M-1)+i
22      A(idx, idx) = 2d0*(1+r)
23      x = i*h; y = j*k
24      b(idx) = -h**2*f(x, y)
25      if (1 < i .and. i < M-1) then
26        A(idx, idx+1) = -1.0d0      ! East
27        A(idx, idx-1) = -1.0d0      ! West
28      else if (i == 1) then
```

```

29         A(idx,idx+1) = -1.0d0
30         b(idx) = b(idx) + lhs(y)
31     else if (i == M-1) then
32         A(idx,idx-1) = -1.0d0
33         b(idx) = b(idx) + rhs(y)
34     end if
35     if (1 < j .and. j < N-1) then
36         A(idx,j*(M-1)+i) = -r          ! North
37         A(idx,(j-2)*(M-1)+i) = -r      ! South
38     else if (j == 1) then
39         A(idx,j*(M-1)+i) = -r
40         b(idx) = b(idx) + r*bot(x)
41     else if (j == N-1) then
42         A(idx,(j-2)*(M-1)+i) = -r
43         b(idx) = b(idx) + r*top(x)
44     end if
45 end do
46 end do
47 if (verbose) call print_mat_vec(Asize, A, b)
48 call dgesv(Asize,1,A,Asize,P,b,Asize,info)
49 do i=0,M
50     write (*,100,advance="no") top(i*h)
51 end do
52 print *
53 do j=N-1,1,-1      ! y-axis
54     write (*,100,advance="no") lhs(j*k)
55     do i=1,M-1      ! x-axis
56         idx = (j-1)*(M-1)+i
57         write (*,100,advance="no") b(idx)
58     end do
59     write (*,100) rhs(j*k)
60 end do
61 do i=0,M
62     write (*,100,advance="no") bot(i*h)
63 end do
64 print *
65 100 format(f14.8)
66 end subroutine poisson2d_cs

```

When the 2-D domain is a **square** and the space grid is uniform with $h = k$ and $M = N$, the equation (47.6) becomes

$$4u_{i,j} - (u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}) = -h^2 f_{i,j}, \quad 1 \leq i, j \leq M-1. \quad (47.8)$$

By applying the index $(j-1) \times (M-1) + i$ to $u_{i,j}$ to (47.8) and treating $u_{i,j}$ as a vector, the coefficient matrix of the linear algebra problem (47.8) is a sparse block matrix:

$$\begin{bmatrix} T_{N-1} + 2I_{N-1} & -I_{N-1} & & \\ -I_{N-1} & T_{N-1} + 2I_{N-1} & -I_{N-1} & \\ \ddots & \ddots & \ddots & \\ & -I_{N-1} & T_{N-1} + 2I_{N-1} & -I_{N-1} \\ & & -I_{N-1} & T_{N-1} + 2I_{N-1} \end{bmatrix} \begin{bmatrix} u_{1,1} \\ \vdots \\ u_{1,N-1} \\ u_{2,1} \\ \vdots \\ u_{N-1,N-1} \end{bmatrix} = -h^2 \begin{bmatrix} f_{1,1} \\ \vdots \\ f_{1,N-1} \\ f_{2,1} \\ \vdots \\ f_{N-1,N-1} \end{bmatrix}$$

where I_{N-1} is an $(N-1) \times (N-1)$ identity matrix and $T_{N-1} = \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}$.

Note that the matrix is a $(N-1)^2 \times (N-1)^2$ matrix. Therefore, if N is 101, the matrix storage requirement for double precision is $100^4 \times 8 = 8 \times 10^8$ bytes = 0.8G of ram, which is very large!

Example 47.1 (April 2014 Final Exam Question 5(c)). Explain, first by writing out the finite-difference approximation formula, then by analysing the formula, why the Jacobi method could be used to solve the finite-difference formula of the steady-state heat distribution in a thin square metal plate problem described by

$$\frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial y^2}(x, y) = 0, \quad 0 < x, y < 1,$$

and $u(0, y) = u(x, 0) = 0$, $u(x, 1) = 200x$, $u(1, y) = 200y$. Furthermore, explain why SOR method and conjugate gradient method are preferable. [5 marks]

Solution: Consider the case $m = n = 4$. Then $h = \frac{1}{m}$, $k = \frac{1}{n}$. Let $w_{i,j} = u(x_i, y_j)$, by using the finite-difference approximation, we obtain a linear equation

$$-4w_{i,j} + w_{i+1,j} + w_{i-1,j} + w_{i,j+1} + w_{i,j-1} = 0, \quad i, j = 1, 2, 3. \quad [2 \text{ marks}]$$

Writing it in matrix form by arranging $w_{i,j}$ appropriately as a vector $\mathbf{w}$, we obtain a *diagonally dominant matrix* which can be solved by Jacobi method (which is guaranteed theoretically) [1 mark]

$$A\mathbf{w} = \begin{bmatrix} -4 & 1 & 0 & 1 & 0 & \cdots & 0 & 0 \\ 1 & -4 & 1 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -4 & 1 & 0 & 1 & \ddots & 0 \\ 1 & 0 & 1 & -4 & 1 & 0 & 1 & \cdots \\ \vdots & \ddots & \vdots & \ddots & \ddots & \ddots & \ddots & \cdots \\ 0 & \cdots & 0 & 0 & 1 & 0 & 1 & 4 \end{bmatrix} \mathbf{w} = \mathbf{b} \quad [1 \text{ mark}]$$

where $\mathbf{b}$ is determined by the boundary conditions.

SOR method and conjugate gradient method ($-A$ is symmetric positive definite) are preferable because they maintain low memory storage requirement and they converge faster than Jacobi method in computation. [1 mark]

The following is an unrealistic example, but it illustrates how to perform calculation with Equations (47.6) and (47.7) (as well as Equation (47.8)).

Example 47.2. [Burden and Faires, 2015, Exercise 12.1, Q3d] Approximate the solutions to the following elliptic PDE problem

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} &= \frac{x}{y} + \frac{y}{x}, \quad 1 < x < 2, \quad 1 < y < 2; \\ u(x, 1) &= x \ln x, \quad u(x, 2) = 2x \ln(2x), \quad 1 \leq x \leq 2; \\ u(1, y) &= y \ln y, \quad u(2, y) = 2y \ln(2y), \quad 1 \leq y \leq 2 \end{aligned}$$

using the finite difference approximation above with $h = k = 0.1$. Compare the results with the actual solution $u(x, y) = xy \ln(xy)$.

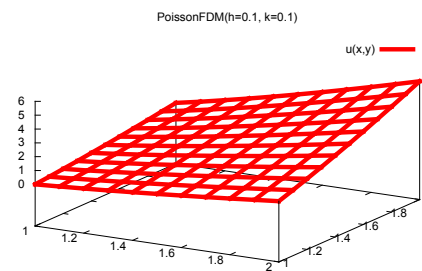
Solution: It is impossible to solve this problem by hand because we need to solve a linear algebra problem with $9 \times 9 = 81$ unknown variables. Before using the FDM above, we need to translate the domain from $[1, 2] \times [1, 2]$ to $[0, 1] \times [0, 1]$. Therefore, the functions f_1 , f_2 , g_2 in (47.7) are shifted in the computer program below:

```
1 program eg_poisson1 ! Burden & Faires, Exercise 12.1, Q3d
2   use textbook_pdesolvers
3   implicit none
4   double precision, parameter :: h=0.1d0, k=0.1d0, xa=1d0, xb=2d0, ya=1d0, yb=2d0
5   integer, parameter :: M=idint((xb-xa)/h), N=idint((yb-ya)/k)
6   integer :: i, j
```

```

7      double precision :: x, y
8      call poisson2d_cs(M, N, xb-xa, yb-ya, lhs, rhs, top, bot, f, .false.)
9      print *
10     do j=N,0,-1
11         y = ya + j*k
12         do i=0,M
13             x = xa + i*h
14             write (*,"(f14.8)",advance="no") x*y*dlog(x*y) ! Exact Soln
15         end do
16         print *
17     end do
18 contains
19 function f(x,y)
20     double precision :: x, y, f
21     f = (x+1d0)/(y+1d0) + (y+1d0)/(x+1d0)
22 end function f
23 function lhs(y)
24     double precision :: y, lhs
25     lhs = (y+1d0)*dlog(y+1d0)
26 end function lhs
27 function rhs(y)
28     double precision :: y, rhs
29     rhs = 2d0*(y+1d0)*dlog(2d0*(y+1d0))
30 end function rhs
31 function bot(x)
32     double precision :: x, bot
33     bot = (x+1d0)*dlog(x+1d0)
34 end function bot
35 function top(x)
36     double precision :: x, top
37     top = 2*(x+1d0)*dlog(2d0*(x+1d0))
38 end function top
39 end program eg_poisson1

```



If we plot the data from the output (after formatting it appropriately), we obtain the figure on the right.

Example 47.3. The four sides of a square plate of side 12 cm made of uniform material are kept at constant temperature such that

$$\text{LHS} = \text{RHS} = \text{Bottom} = 100^\circ \text{ and } \text{Top} = 0^\circ.$$

By using a mesh size of 4 cm, calculate the temperature $u(x, y)$ at the internal mesh points if $u(x, y)$ satisfies the Laplace equation $u_{xx} + u_{yy} = 0$ (a special case of Poisson equation where $f(x, y) = 0$).

Solution. Let $A = (4, 4)$, $B = (8, 4)$, $C = (8, 8)$ and $D = (4, 8)$. Applying the equation $u_{i,j} = \frac{1}{4}[u_E + u_N + u_W + u_S]$ to the 4 internal mesh points, we obtain

$$\begin{aligned}
 4u_A - u_B - u_D &= 200 \\
 -u_A + 4u_B - u_C &= 200 \\
 -u_A + 4u_D - u_C &= 100 \\
 -u_B - u_D + 4u_C &= 100
 \end{aligned}$$

This is a linear algebra problem with 4 variables, it is not too difficult to solve it to find $u_A = u_B = 87.5$, $u_C = u_D = 62.5$. $\square$

Example 47.4 (April 2018 Final Exam Question 5(b)). Consider the Laplace equation

$$\frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial y^2}(x, y) = 0, \quad 0 < x < \pi, \quad 0 < y < \ln 5,$$

with the following boundary conditions

$$u(0, y) = u(\pi, y) = 0, \quad u(x, 0) = \sin x, \quad u(x, \ln 5) = 5 \sin x.$$

(b) APPLY AN CENTRE DIFFERENCE APPROXIMATION, we obtain

$$\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{(\frac{1}{3})^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{(\frac{1}{2})^2} = 4.$$

(c) SETUP A SYSTEM OF ALGEBRAIC EQUATIONS: Let $f_{i,j} = f(x_i, y_j)$, the above equation becomes

$$u_{i+1,j} - 2u_{i,j} + u_{i-1,j} + \frac{4}{9}(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}) = (\frac{1}{3})^2 f_{i,j} = \frac{4}{9}.$$

The corresponding linear algebra problem is

$$\begin{bmatrix} -\frac{26}{9} & 1 & \frac{4}{9} & 0 & 0 & 0 \\ 1 & -\frac{26}{9} & 0 & \frac{4}{9} & 0 & 0 \\ \frac{4}{9} & 0 & -\frac{26}{9} & 1 & \frac{4}{9} & 0 \\ 0 & \frac{4}{9} & 1 & -\frac{26}{9} & 0 & \frac{4}{9} \\ 0 & 0 & \frac{4}{9} & 0 & -\frac{26}{9} & 1 \\ 0 & 0 & 0 & \frac{4}{9} & 1 & -\frac{26}{9} \end{bmatrix} \begin{bmatrix} u_{1,1} \\ u_{2,1} \\ u_{1,2} \\ u_{2,2} \\ u_{1,3} \\ u_{2,3} \end{bmatrix} = \begin{bmatrix} \frac{4}{9} - \frac{4}{9}u_{1,0} - u_{0,1} \\ \frac{4}{9} - \frac{4}{9}u_{2,0} - u_{3,1} \\ \frac{4}{9} - u_{0,2} \\ \frac{4}{9} - u_{3,2} \\ \frac{4}{9} - u_{0,3} - \frac{4}{9}u_{1,4} \\ \frac{4}{9} - u_{3,3} - \frac{4}{9}u_{2,4} \end{bmatrix}$$

$$RHS = \begin{bmatrix} \frac{4}{9} - \frac{4}{9}(\frac{1}{3})^2 - (\frac{1}{2})^2 \\ \frac{4}{9} - \frac{4}{9}(\frac{2}{3})^2 - (\frac{1}{2} - 1)^2 \\ \frac{4}{9} - 1^2 \\ \frac{4}{9} - (1 - 1)^2 \\ \frac{4}{9} - (\frac{3}{2})^2 - \frac{4}{9}(\frac{1}{3} - 2)^2 \\ \frac{4}{9} - (\frac{3}{2} - 1)^2 - \frac{4}{9}(\frac{2}{3} - 2)^2 \end{bmatrix} = \begin{bmatrix} \frac{47}{324} \\ -\frac{1}{324} \\ -\frac{5}{9} \\ \frac{4}{9} \\ -\frac{985}{324} \\ -\frac{193}{324} \end{bmatrix}$$

where the boundary conditions are

$$\begin{aligned} u_{i,0} &= x_i^2, & u_{i,4} &= (x_i - 2)^2, & i &= 0, 1, 2, 3; \\ u_{0,j} &= y_j^2, & u_{3,j} &= (y_j - 1)^2, & j &= 0, 1, 2, 3, 4. \end{aligned}$$

In summary,

$$\begin{bmatrix} -2.888889 & 1.000000 & 0.444444 & 0.000000 & 0.000000 & 0.000000 \\ 1.000000 & -2.888889 & 0.000000 & 0.444444 & 0.000000 & 0.000000 \\ 0.444444 & 0.000000 & -2.888889 & 1.000000 & 0.444444 & 0.000000 \\ 0.000000 & 0.444444 & 1.000000 & -2.888889 & 0.000000 & 0.444444 \\ 0.000000 & 0.000000 & 0.444444 & 0.000000 & -2.888889 & 1.000000 \\ 0.000000 & 0.000000 & 0.000000 & 0.444444 & 1.000000 & -2.888889 \end{bmatrix} \mathbf{u} = \begin{bmatrix} 0.145062 \\ -0.003086 \\ -0.555556 \\ 0.444444 \\ -3.040123 \\ -0.595679 \end{bmatrix}$$

□

(b) What numerical methods from Topic 2 do you think is appropriate for solve the linear algebra problem in part (a)? Justify your answer.

Solution. Direct Method or SOR for Sparse Matrix. Sparse matrix: save memory, speed; SOR because it is fast when suitable ω can be found, like in this specific matrix. □

(c) Solve the linear algebra problem using a computer program and compare the results obtained with the actual solution $u(x, y) = (x - y)^2$.

Solution. Writing the solution in “grid” or matrix form:

$$\mathbf{u} = [u_{i,j}]_{\substack{i=0..3; \\ j=0..4}} = \begin{bmatrix} 4.00 & 2.77777778 & 1.77777778 & 1.00 \\ 2.25 & 1.36111111 & 0.69444444 & 0.25 \\ 1.00 & 0.44444444 & 0.11111111 & 0.00 \\ 0.25 & 0.02777778 & 0.02777778 & 0.25 \\ 0.00 & 0.11111111 & 0.44444444 & 1.00 \end{bmatrix}$$

$$\text{In summary, } \begin{bmatrix} u_{1,3} & u_{2,3} \\ u_{1,2} & u_{2,2} \\ u_{1,1} & u_{2,1} \end{bmatrix} = \begin{bmatrix} 1.36111111 & 0.69444444 \\ 0.44444444 & 0.11111111 \\ 0.02777778 & 0.02777778 \end{bmatrix} \quad \square$$

47.3 Parabolic Problems

Conceptually, the general form of *nonlinear parabolic PDE*, a type of *time-dependent PDE*, is given by

$$u_t = -Lu \quad (47.9)$$

where L is a *second-order elliptic operator* (this implies that L must be “positive”). The solution of (47.9) normally exists for a short time and may explode in a singularity in a finite amount of time. Hence, the difficulty of the research in Equation (47.9) is in determining solutions for all time, or more generally studying the singularities that arise. This is a difficult research, as seen in the solution of the *Poincaré conjecture* via *Ricci flow*.

In engineering, one normally studies a more specific *heat equation* or *advection-diffusion* (or *diffusion-convection*) equation [Hundsdorfer and Verwer, 2003]:

$$\frac{\partial u}{\partial t} + \underbrace{(\mathbf{v} \cdot \nabla)u}_{\text{advection}} = \underbrace{\nabla \cdot (D \nabla u)}_{\text{diffusion}} + \frac{f}{\rho} \quad (47.10)$$

where u stands for temperature or concentration, f is a *source* or *reaction*, etc.

With the knowledge of undergraduate calculus and numerical methods we learned so far, it is still difficult for us to explore the two and three dimensional space heat equation. The study of two- or three- dimensional heat equation will be carried out in *UECM3003 Partial Differential Equations*.

If we limit ourselves to one-dimensional space, by suitable assumptions and transforms, mathematicians managed to reduce Equation (47.10) into a 1-D *advection-diffusion-reaction equation* [Soetaert and Herman, 2009]:

$$\frac{\partial u}{\partial t} = - \underbrace{\frac{1}{A\xi} \frac{\partial(v \cdot A \cdot \xi \cdot u)}{\partial x}}_{\text{advection}} + \underbrace{\frac{1}{A\xi} \frac{\partial}{\partial x} \left(A \cdot D \frac{\partial(\xi \cdot u)}{\partial x} \right)}_{\text{diffusion}} + \underbrace{f(x)}_{\text{reaction}}. \quad (47.11)$$

Here,

- D is the “diffusion coefficient”, v is the “advection rate”, and
- $A = A(x)$ and $\xi = \xi(x)$ are the surface area and volume fraction associated with the “shape” of the object. For cylindrical shape $A = 2\pi rL$, for spherical shape, $A = 4\pi r^2$.

In this section, we will analyse the 1-dimensional linear homogeneous parabolic equation in standard form using an explicit scheme and an implicit scheme:

$$\begin{aligned} \text{PDE : } & \frac{\partial u}{\partial t} = \alpha^2 \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < L, \quad t > 0, \\ \text{BCs : } & u(0, t) = g_1(t), \quad u(L, t) = g_2(t), \\ \text{IC : } & u(x, 0) = f(x), \end{aligned} \quad (47.12)$$

where $u = u(x, t)$ is a function of time t and space x .

The numerical methods for solving parabolic equation is more complicated than the elliptic case due to the “time variable” t . In this section, we will be investigating FTCS, BTCS and Crank-Nicolson methods. These methods are implemented in QuantLib’s `expliciteuler.hpp`, `impliciteuler.hpp` and `cranknicolson.hpp` as special cases of the `mixedscheme.hpp`.

47.3.1 FTCS (Explicit) Scheme

If we apply the forward difference (FD) approximation to $\frac{\partial u}{\partial t}$ and centre difference (CD) approximation to $\frac{\partial^2 u}{\partial x^2}$ in (47.12) we will obtain the *FTCS* (or *explicit Euler*) scheme below, which stands for “*forward in time and centre in space*”.

Method 47.2 (FTCS Explicit Finite Difference Scheme for (47.12)).

1. GENERATE A GRID for $[0, L] \times [0, T_{end}]$: Let $u_{i,j} \approx u(x_i = ih, t_j = jk)$, $i = 0, \dots, M$, $j = 0, \dots, N$.
2. REPRESENT THE DERIVATIVE BY SOME FINITE DIFFERENCE FORMULAE: Using FD for u_t and CD for u_{xx} to obtain a finite difference equation from (47.12):

$$\frac{u_{i,j+1} - u_{i,j}}{k} = \alpha^2 \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2}, \quad i = 1, \dots, M, \quad j = 0, \dots, N-1.$$

3. TOGETHER WITH THE BOUNDARY CONDITIONS:

$$u_{0,j} = g_1(t_j), \quad u_{M,j} = g_2(t_j), \quad j = 0, \dots, N-1, \quad (47.13)$$

REARRANGE THE FD APPROXIMATION ABOVE INTO AN ITERATION

$$u_{i,j+1} = (1 - 2r)u_{i,j} + r(u_{i+1,j} + u_{i-1,j}), \quad r = \frac{\alpha^2 k}{h^2} \quad (47.14)$$

4. SOLVE THE ITERATION (47.14) AND (47.13) FOR A GIVEN INITIAL CONDITION

$$u_{i,0} = u(x_i, 0) = f(x_i). \quad (47.15)$$

A Fortran 2003 implementation is given below.

```

1  subroutine parabolic_ftcs(M, N, f, a, b, Tend, alph, filename)
2      procedure(one_var_func) :: f
3      intent(in) :: M, N, a, b, Tend, alph
4      integer :: M, N, i, j, unitnum
5      double precision :: a, b, Tend, alph, u(M+1), upre(M+1), x, h, k, t, r
6      character(len=*), optional :: filename
7      if (present(filename)) then
8          unitnum=10
9          open(unit=unitnum, file=filename, action="write", status="replace")
10     else
11         unitnum=6 ! Standard Output
12     end if
13     h = (b-a)/M; k = Tend/N; t = 0d0
14     r = alph**2*k/(h**2) ! Condition number
15     !if (r >= 0.5d0) then
16     !    write (unitnum,*), "Condition Number, r =", r
17     !    return
18     !end if
19     write (unitnum,"(a,f8.4,a,f8.4,a)") "# set title "FTCS(h=", h, ", k=", k, ")"
20     write (unitnum,"(a,i5,i5)") "# set dgrid3d", M+1, N+1
21     write (unitnum,"(a)") '# splot "test.dat" u 1:2:3 w l'
22     do i=1,M+1
23         x = a+h*(i-1)
24         upre(i) = f(x)

```

```

25         write (unitnum,100) t, x, upre(i)
26     end do
27     write (unitnum,*)
28     do j=1,N
29         t = t + k
30         write (unitnum,100) t, a, 0d0
31         do i=2,M
32             u(i) = (1.0d0-2.0d0*r)*upre(i) + r*(upre(i+1)+upre(i-1))
33             write (unitnum,100) t, a+h*(i-1), u(i)
34         end do
35         write (unitnum,100) t, b, 0d0
36         write (unitnum,*)
37         upre(:) = u(:)
38     end do
39     if (present(filename)) close(unitnum)
40 100 format(es16.8,es16.8,es16.8)
41 end subroutine parabolic_ftcs

```

There are two things of concern for the numerical methods of PDE-IVP, one is the numerical stability, the other is the convergence of solution, i.e. $u_h \rightarrow u$ as $h \rightarrow 0^+$.

Definition 47.1 (Stability of Numerical Methods). A method is (*numerically*) *unstable* if round-off errors or any other errors grow too rapidly as the computations proceed. Otherwise, it is said to be (*numerically*) *stable*.

The FTCS scheme is stable for a given space step h provided the time step k is restricted by

$$0 < \boxed{r = \alpha^2 \frac{k}{h^2}} \leq \frac{1}{2}.$$

where r is called the (*Courant-Friedrichs-Lewy*) *condition number*. The restriction recommends that we do not move too fast in the t -direction.

Example 47.6 (April 2014 Final Exam Question 3(c)). Given a heat equation

$$\frac{\partial u}{\partial t}(x, t) = a^2 \frac{\partial^2 u}{\partial x^2}(x, t), \quad 0 < x < L, \quad t > 0$$

subject to the initial and boundary conditions below

$$u(0, t) = u(L, t) = 0, \quad t > 0, \quad u(x, 0) = f(x), \quad 0 \leq x \leq L.$$

Derive a forward-difference iteration for this equation and explain the conditions for the numerical solution of the forward-difference to be close to the real solution of the heat equation. (7 marks)

Solution: Consider the following approximations

$$\frac{\partial^2 u}{\partial x^2}(x_i, t_i) \approx \frac{u(x_i + h, t_i) - 2u(x_i, t_i) + u(x_i - h, t_i)}{h^2}, \quad i = 1, \dots, N-1 \quad [1 \text{ mark}]$$

$$\frac{\partial u}{\partial t}(x_i, t_i) \approx \frac{u(x_i, t_i + k) - u(x_i, t_i)}{k}, \quad j = 0, 1, \dots, i = 0, 1, \dots, N \quad [1 \text{ mark}]$$

Let $w_{i,j} = u(x_i, t_i)$, $w_{i,j+1} = u(x_i, t_i + k)$ and $w_{i+1,j} = u(x_i + h, t_i)$. Then

$$\begin{aligned} \frac{w_{i,j+1} - w_{i,j}}{k} &= a^2 \frac{w_{i+1,j} - 2w_{i,j} + w_{i-1,j}}{h^2} \\ \Rightarrow w_{i,j+1} &= w_{i,j} + \frac{a^2 k}{h^2} [w_{i+1,j} - 2w_{i,j} + w_{i-1,j}] = \frac{a^2 k}{h^2} w_{i+1,j} + \left(1 - \frac{2a^2 k}{h^2}\right) w_{i,j} + \frac{a^2 k}{h^2} w_{i-1,j} \end{aligned} \quad [2 \text{ marks}]$$

Let $\mathbf{w}^{(j)} = (w_{1,j}, \dots, w_{N-1,j})^T$. Note that from the boundary condition, $w_{0,j} = u(0, t_i) =$

$w_{N,j} = u(L, t_i) = 0$ such that

$$\mathbf{w}^{(j+1)} = \underbrace{\begin{bmatrix} 1 - \frac{2a^2k}{h^2} & \frac{a^2k}{h^2} & 0 & 0 & \cdots & 0 \\ \frac{a^2k}{h^2} & 1 - \frac{2a^2k}{h^2} & \frac{a^2k}{h^2} & 0 & \cdots & 0 \\ 0 & \frac{a^2k}{h^2} & 1 - \frac{2a^2k}{h^2} & \frac{a^2k}{h^2} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ 0 & \cdots & 0 & \frac{a^2k}{h^2} & 1 - \frac{2a^2k}{h^2} & 0 \end{bmatrix}}_{(N-1) \times (N-1) \text{ matrix}} \mathbf{w}^{(j)} \quad [1 \text{ mark}]$$

For the square matrix in the forward-difference iteration, we need to consider the *numerical stability* and *convergence*. Numerical stability is required to make sure that small errors are not magnify while convergence make sure the values tends to the real solution. [2 marks]

Example 47.7 (April 2018 Final Exam Question 3(c)). Given a heat equation

$$\frac{\partial u}{\partial t}(x, t) = 4 \frac{\partial^2 u}{\partial x^2}(x, t), \quad 0 < x < 1, \quad t > 0$$

subject to the initial and boundary conditions below

$$u(x, 0) = \sin 2\pi x, \quad 0 \leq x \leq 1;$$

$$u(0, t) = u(1, t) = 0, \quad t > 0.$$

By applying the FTCS scheme with $\Delta x = 0.1$ and $\Delta t = 0.05$, estimate the value of $u(0.7, 0.05)$. Write your answer accurate to 6 decimal places. (8 marks)

Solution. By applying FTCS scheme, we obtain the following equation.

$$\frac{u_i^{(j+1)} - u_i^{(j)}}{\Delta t} = 4 \frac{u_{i+1}^{(j)} - 2u_i^{(j)} + u_{i-1}^{(j)}}{(\Delta x)^2}, \quad i = 1, \dots, 9, \quad j = 0, 1, 2, \dots \quad [2 \text{ marks}]$$

When $j = 1$ and $i = 7$, the equation simplifies to

$$u_7^{(1)} = u_7^{(0)} + \frac{4\Delta t}{(\Delta x)^2}(u_8^{(0)} - 2u_7^{(0)} + u_6^{(0)}) = u_7^{(0)} + 20(u_8^{(0)} - 2u_7^{(0)} + u_6^{(0)}) \quad [2 \text{ marks}]$$

From the boundary condition, we know that

$$u_i^{(0)} = u(0.1i, 0) = \sin 2\pi(0.1i), \quad i = 0, 1, \dots, 10. \quad [1 \text{ mark}]$$

Therefore,

$$u(0.7, 0.05) \approx u_7^{(1)} = \sin 1.4\pi + 20(\sin 1.6\pi - 2\sin 1.4\pi + \sin 1.2\pi) = 6.314369 \quad [3 \text{ marks}]$$

□

Let us study the “numerical stability” with the following example.

Example 47.8. Let $h = 0.1$ and (a) $k = 0.001$ and (b) $k = \frac{0.5}{50} = 0.01$. Solve the problem (47.12) numerically for the time interval $[0, 0.5]$ using a computer program assuming that $\alpha = 1$, the boundaries at $x = 0$ and $x = 1$ are homogeneous and the initial condition $u(x, 0) = \sin(\pi x)$. Compare the two cases (a) and (b).

Solution: With the following computer program:

```
1 program eg_parab_ftcs
2   use textbook_pdesolvers
3   implicit none
4   call parabolic_ftcs(10, 500, f, 0d0, 1d0, 0.5d0, 1d0, "ftcs_ok_f.dat")
5   call parabolic_ftcs(10, 50, f, 0d0, 1d0, 0.5d0, 1d0, "ftcs_ko_f.dat")
```

```

6 contains
7 function f(x)
8     double precision :: x, f
9     f = sin(4d0*datan(1d0)*x)
10 end function f
11 end program eg_parab_ftcs

```

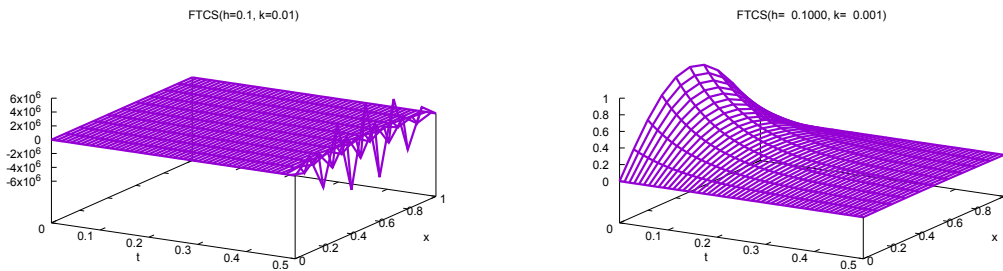
we can find the numerical solutions of the given problem and by using the visualisation tool gnuplot mentioned in Topic 4, we can write a script

```

1 set xlabel "t"
2 set ylabel "x"
3 set border linewidth 0.2
4 set terminal pdfcairo font "Arial,12" linewidth 3
5 set title "FTCS(h=0.1, k=0.01)"
6 set dgrid3d 11,51
7 set output "ftcs_ko.pdf"
8 splot "ftcs_ko_f.dat" u 1:2:3 w l t ""
9 set title "FTCS(h= 0.1000, k= 0.001)"
10 set output "ftcs_ok.pdf"
11 splot "ftcs_ok_f.dat" u 1:2:3 w l t ""

```

to plot the data generated by the computer program for visualisation. The figure on the left is the result of $k = 0.01$, where $r = \frac{0.01}{0.1^2} = 1 > \frac{1}{2}$. The numerical error is magnified and hence the solution is wrong! The figure on the right is the result of $k = 0.001$ where $r = \frac{0.001}{0.1^2} = 0.1 \leq \frac{1}{2}$. Hence, the numerical method is stable and generates a good approximation to the real solution.



Finally, let us investigate a hand-calculation example to learn how FTCS works.

eg:heat:1d:ftcs

Example 47.9. Consider the one-dimensional heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad t > 0$$

subject to the initial condition $u(x, 0) = \sin \pi x$, $0 \leq x \leq 1$ and boundary conditions $u(0, t) = u(1, t) = 0$, $t \geq 0$. Using an x -step of $h = 0.1$ and a t -step of $k = 0.0005$, use the FTCS scheme to estimate $u(0.1, 0.001)$.

Solution. Let $r = \frac{k}{h^2} = \frac{0.0005}{(0.1)^2} = 0.05$ and note that $\alpha = 1$. The grid is

$$(x_i, t_j) = (ih, kj) = (0.1i, 0.0005j), \quad i = 1, \dots, 9 \quad j = 1, 2, \dots$$

1. $u(0, t) = u(1, t) = 0 \Rightarrow u_{0,j} = u_{10,j} = 0$.
2. $u(x, 0) = \sin \pi x \Rightarrow u_{i,0} = \sin 0.1\pi i$, $i = 1, \dots, 9$.
3. Applying FTCS scheme, we obtain

$$u_{i,j+1} = (1 - 2r)u_{i,j} + r(u_{i+1,j} + u_{i-1,j}) = 0.9u_{i,j} + 0.05(u_{i+1,j} + u_{i-1,j})$$

4. $j = 0$: $u_{i,1} = 0.9u_{i,0} + 0.05(u_{i+1,0} + u_{i-1,0})$

- (a) $i = 1$: Since $u_{i,0} = \sin 0.1\pi i$, $u_{0,0} = 0$, $u_{1,0} = \sin 0.1\pi = 0.30901699$ and $u_{2,0} = \sin 0.2\pi = 0.58778525$,

$$u(0.1, 0.0005) \approx u_{1,1} = 0.9u_{1,0} + 0.05(u_{2,0} + u_{0,0}) = 0.30750455.$$

(b) $i = 2$: Since $u_{3,0} = \sin 0.3\pi = 0.80901699$,

$$u(0.2, 0.0005) \approx u_{2,1} = 0.9u_{2,0} + 0.05(u_{3,0} + u_{1,0}) = 0.58490842.$$

5. $j = 1$: $u_{i,2} = 0.9u_{i,1} + 0.05(u_{i+1,1} + u_{i-1,1})$

(a) $i = 1$: $u(0.1, 0.001) \approx u_{1,2} = 0.9u_{1,1} + 0.05(u_{2,1} + u_{0,1}) = 0.30599952.$

□

es:p754:12.2q5a

Example 47.10 (Tutorial 8, Q2). Use the FTCS method to approximate the solution of the following parabolic PDE:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 2, \quad t > 0, \quad (47.16)$$

e:ftcs1

subject to the initial and boundary conditions

$$u(0, t) = u(2, t) = 0, \quad t > 0, \quad \text{and} \quad u(x, 0) = \sin 2\pi x, \quad 0 \leq x \leq 2. \quad (47.17)$$

e:ftcs1:bc+ic

(a) Write down the “iterative linear algebra problem” $\mathbf{u}_{j+1} = A\mathbf{u}_j$ by applying the FTCS method with $h = 0.4$ and $k = 0.1$ to the parabolic problem (47.16) and (47.17).

Solution. Given $h = 0.4$, $k = 0.1$. So $r = \frac{0.1}{0.4^2} = 0.625 = \frac{5}{8} > 0.5$ and the FTCS is numerical unstable. The difference equation is

$$u_{i,j+1} = (1 - 2r)u_{i,j} + r(u_{i+1,j} + u_{i-1,j}) = -0.25u_{i,j} + 0.625(u_{i+1,j} + u_{i-1,j})$$

where $i = 1, 2, 3, 4$ and $j = 0, 1, \dots, 5$.

$$\begin{bmatrix} u_{0,j+1} \\ u_{1,j+1} \\ u_{2,j+1} \\ u_{3,j+1} \\ u_{4,j+1} \\ u_{5,j+1} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{5}{8} & -\frac{1}{4} & \frac{5}{8} & 0 & 0 & 0 \\ 0 & \frac{5}{8} & -\frac{1}{4} & \frac{5}{8} & 0 & 0 \\ 0 & 0 & \frac{5}{8} & -\frac{1}{4} & \frac{5}{8} & 0 \\ 0 & 0 & 0 & \frac{5}{8} & -\frac{1}{4} & \frac{5}{8} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} u_{0,j} \\ u_{1,j} \\ u_{2,j} \\ u_{3,j} \\ u_{4,j} \\ u_{5,j} \end{bmatrix}, \quad \begin{bmatrix} u_{0,0} \\ u_{1,0} \\ u_{2,0} \\ u_{3,0} \\ u_{4,0} \\ u_{5,0} \end{bmatrix} = \begin{bmatrix} 0 \\ \sin 0.8\pi \\ \sin 1.6\pi \\ \sin 2.4\pi \\ \sin 3.2\pi \\ 0 \end{bmatrix} \quad \square$$

(b) Compare your results at $t = 0.5$ to the exact solution $u(x, t) = \exp(-4\pi^2 t) \sin 2\pi x$.

Solution. At $t = 0.5$, we find $\mathbf{x}(x_i, 0.5) \approx \mathbf{x}_5 = A^5 \mathbf{x}_0$ which approximates the exact solution extremely poorly due to numerical instability.

x_i	$u_{i,j}$	$u(x_i, 0.5)$
0.0	0.00000000	0
0.4	-1.87612234	1.572495×10^{-9}
0.8	3.03562971	-2.544350×10^{-9}
1.2	-3.03562971	2.544350×10^{-9}
1.6	1.87612234	-1.572495×10^{-9}
2.0	0.00000000	0

□

(c) Write down the “iterative linear algebra problem” $\mathbf{u}_{i+1} = A\mathbf{u}_i$ by applying the FTCS method with an x step size of $h = 0.4$ and a t step size of $k = 0.05$ to the parabolic problem (47.16) and (47.17). By using the knowledge of eigenvalue analysis from Topic 2, compare the difference between this and the “iterative linear algebra problem” in part (a).

Solution. When $h = 0.4$, $k = 0.05$, $r = \frac{0.05}{0.4^2} = 0.3125 = \frac{5}{16} < 0.5$ and the difference equation is

$$u_{i,j+1} = (1 - 2r)u_{i,j} + r(u_{i+1,j} + u_{i-1,j}) = 0.375u_{i,j} + 0.3125(u_{i+1,j} + u_{i-1,j}).$$

By analysing the eigenvalues of the matrix “A” using

```

1 import numpy as np
2 N = 6
3 A = np.eye(N)*(-1.0/4)
4 for i in range(N-1): A[i,i+1] = A[i+1,i] = 5.0/8
5 A[0,:] = A[5,:] = 0
6 eigvals, eigvecs = np.linalg.eig(A)
7 print("eigvals for (a) = ", eigvals)
8 for i in range(N): A[i,i] = 3.0/8
9 for i in range(N-1): A[i,i+1] = A[i+1,i] = 5.0/16
10 A[0,:] = A[5,:] = 0
11 eigvals, eigvecs = np.linalg.eig(A);
12 print("eigvals for (c) = ", eigvals)

```

we obtain

- For (a), $\lambda = -1.26127124, -0.63627124, 0.76127124, 0.13627124, 0, 0$. We have eigenvalues with absolute value > 1 .
- Since $\rho(A) = \max_i |\lambda_i| = \max\{-0.13063562, 0.18186438, 0.88063562, 0.56813562, 0, 0\} = 0.88063562 \leq 1$. A is stable.

$$\begin{bmatrix} u_{0,j+1} \\ u_{1,j+1} \\ u_{2,j+1} \\ u_{3,j+1} \\ u_{4,j+1} \\ u_{5,j+1} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{5}{16} & \frac{3}{8} & \frac{5}{16} & 0 & 0 & 0 \\ 0 & \frac{5}{16} & \frac{3}{8} & \frac{5}{16} & 0 & 0 \\ 0 & 0 & \frac{5}{16} & \frac{3}{8} & \frac{5}{16} & 0 \\ 0 & 0 & 0 & \frac{5}{16} & \frac{3}{8} & \frac{5}{16} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} u_{0,j} \\ u_{1,j} \\ u_{2,j} \\ u_{3,j} \\ u_{4,j} \\ u_{5,j} \end{bmatrix}, \quad \begin{bmatrix} u_{0,0} \\ u_{1,0} \\ u_{2,0} \\ u_{3,0} \\ u_{4,0} \\ u_{5,0} \end{bmatrix} = \begin{bmatrix} 0 \\ \sin 0.8\pi \\ \sin 1.6\pi \\ \sin 2.4\pi \\ \sin 3.2\pi \\ 0 \end{bmatrix} \quad \square$$

47.3.2 BTCS (Implicit) Scheme

If we apply the backward difference (BD) approximation to $\frac{\partial u}{\partial t}$ and CD approximation to $\frac{\partial^2 u}{\partial x^2}$ in (47.12) we will obtain the *BTCS* (or *implicit Euler*) scheme below, which stands for “backward in time and centre in space”.

Method 47.3 (BTCS Implicit Finite Difference Scheme for (47.12)).

1. GENERATE A GRID for $[0, L] \times [0, T_{end}]$: Let $u_{i,j} \approx u(x_i = ih, t_j = jk)$, $i = 0, \dots, M$, $j = 0, \dots, N$.
2. REPRESENT THE DERIVATIVE BY SOME FINITE DIFFERENCE FORMULAE: Using BD for u_t and CD for u_{xx} to obtain a finite difference equation from (47.12):

$$\frac{u_{i,j} - u_{i,j-1}}{k} = \alpha^2 \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2}, \quad i = 1, \dots, M-1, \quad j = 1, \dots, N.$$

3. TOGETHER WITH THE BOUNDARY CONDITIONS:

$$u_{0,j} = g_1(t_j), \quad u_{M,j} = g_2(t_j), \quad j = 1, \dots, N, \quad (47.18)$$

REARRANGE THE FD APPROXIMATION ABOVE INTO A LINEAR ALGEBRA PROBLEM

$$-ru_{i-1,j} + (1 + 2r)u_{i,j} - ru_{i+1,j} = u_{i,j-1}, \quad r = \frac{\alpha^2 k}{h^2} \quad (47.19)$$

4. SOLVE THE LINEAR ALGEBRA PROBLEM (47.19) AND (47.18) FOR A GIVEN INITIAL CONDITION

$$u_{i,0} = u(x_i, 0) = f(x_i). \quad (47.20)$$

e:pde:btcs:ic

According to Burden and Faires [2015], the BTCS method has no stability issue. However, mathematicians was found that the Crank-Nicolson method, which is an averaging of FTCS and BTCS methods, gives better accuracy. The software implementation of BTCS is similar to the Crank-Nicolson method shown in the next subsection.

Example 47.11 (Tutorial 8, Q3). Given the PDE-BVP-IVP

$$\begin{aligned} \text{PDE : } u_t &= u_{xx} + (x - t^2)e^{xt}, \\ \text{BC : } u(0, t) &= 1, \quad u(1, t) = e^t, \\ \text{IC : } u(x, 0) &= 1 \end{aligned}$$

Use BTCS with $h = 0.1$ and $k = 1$ to find the solution numerically by plotting the figure.

Solution. By using FDM it is not difficult to obtain

$$\frac{u_{i,j} - u_{i,j-1}}{k} = \alpha^2 \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} + S_{i,j}.$$

```

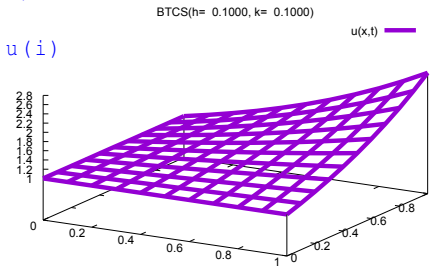
1 program tut_parab_inhomog_btcs
2   use textbook_pdesolvers
3   call parabolic_btcs(10, 10, one, 0d0, 1d0, 1d0, 1d0, one, g2, S)
4   !call parabolic_btcs(10,50,f2,0d0,1d0,0.5d0,1d0,zero,zero,zero2) ! Textbook Eg
5 contains
6 function zero(x)
7   double precision :: x, zero
8   zero = 0d0
9 end function zero
10 function one(x)
11   double precision :: x, one
12   one = 1d0
13 end function one
14 function f2(x)
15   double precision :: x, f2
16   f2 = sin(4d0*datan(1d0)*x)
17 end function f2
18 function g2(x)
19   double precision :: x, g2
20   g2 = dexp(x)
21 end function g2
22 function S(x,t)
23   double precision :: x, t, S
24   S = (x-t**2)*exp(x*t)
25 end function S
26 function zero2(x,t)
27   double precision :: x, t, zero2
28   zero2 = 0d0
29 end function zero2
30 subroutine parabolic_btcs(M, N, f, a, b, Tend, alph, g1, g2, S, filename)
31   intent(in) :: M, N, a, b, Tend, alph
32   procedure(one_var_func) :: f, g1, g2
33   procedure(two_var_func) :: S
34   integer :: M, N, unitnum, i, j, info
35   double precision :: a, b, Tend, alph, h, k, r, t, x
36   double precision, dimension(M+1) :: u, upre, CNMDU, CNMDL, CNMD
37   character(len=*), optional :: filename
38   if (present(filename)) then
39     unitnum=10
40     open(unit=unitnum,file=filename,action="write",status="replace")
41   else
42     unitnum=6 ! Standard Output
43   end if
44   ! Discretise Space ; Discretise Time ; Condition number

```

```

45 h = (b-a)/real(M,8); k = Tend/real(N,8); r = alph**2*k/(h**2)
46 t = 0d0 ! Starting Time = 0
47 write (unitnum,"(a,f8.4,a,f8.4,a)") "# set title "BTCS(h=", h, ", k=", k, ")""
48 write (unitnum,"(a,i5,i5)") "# set dgrid3d", M+1, N+1
49 write (unitnum,"(a)") '# splot "test.dat" u 1:2:3 w l'
50 do i=1,M+1
51     x = a+h*(i-1)
52     upre(i) = f(x)
53     write (unitnum,100) t, x, upre(i)
54 end do
55 write (unitnum,*) ""
56 do j=1,N
57     CNMD(1) = 1d0; CNMD(M+1) = 1d0; CNMDL(M) = 0d0; CNMDU(1) = 0d0
58     do i=2,M
59         CNMD(i) = 1+2d0*r; CNMDL(i-1) = -r; CNMDU(i) = -r
60     end do
61     ! dgtsv will change CNMD, CNMDL, CNMDU, so reinitialise them everytime
62     t = t + k ! or t = j*k
63     u(1) = g1(t); u(M+1) = g2(t) ! Nonhomogeneous BC
64     do i=2,M
65         u(i) = upre(i) + k*S(a+h*(i-1),t)
66     end do
67     ! DGTSV( N, NRHS, DL, D, DU, B, LDB, INFO )
68     call dgtsv(M+1,1,CNMDL,CNMD,CNMDU,u,M+1,info)
69     do i=1,M+1
70         write (unitnum,100) t, a+h*(i-1), u(i)
71     end do
72     write (unitnum,*) ""
73     upre(:) = u(:)
74 end do
75 if (present(filename)) close(unitnum)
76 100 format(f13.8,1x,f13.8,1x,f13.8)
77 end subroutine parabolic_btcs
78 end program tut_parab_inhomog_btcs

```



47.3.3 Crank-Nicolson (Implicit) Method

The averaging of FTCS and BTCS leads to the *Crank-Nicolson Implicit (CNI) Scheme* for the heat equation (47.12), which replaces u_{xx} by an average of two centre difference quotients, one at the time level j and another at $j+1$:

$$\frac{u(x_i, t_{j+1}) - u(x_i, t_j)}{k} + O(k) = \alpha^2 \left[\frac{u(x_{i+1}, t_j) - 2u(x_i, t_j) + u(x_{i-1}, t_j)}{h^2} + O(h^2) \right]$$

$$\frac{u(x_i, t_{j+1}) - u(x_i, t_j)}{k} + O(k) = \alpha^2 \left[\frac{u(x_{i+1}, t_{j+1}) - 2u(x_i, t_{j+1}) + u(x_{i-1}, t_{j+1})}{h^2} + O(h^2) \right].$$

Assume that $u_{i,j} \approx u(x_i, t_j)$. Then summing the above two equations and rearranging it gives

$$u_{i,j+1} - u_{i,j} = \frac{k\alpha^2}{2h^2} [u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1} + u_{i+1,j} - 2u_{i,j} + u_{i-1,j}] + O(k^2, kh^2).$$

Let $r = \alpha^2 \frac{k}{h^2}$. Ignoring the error terms and we obtain the finite difference equation of the CNI scheme:

$$-\frac{r}{2}u_{i+1,j+1} + (1+r)u_{i,j+1} - \frac{r}{2}u_{i-1,j+1} = \frac{r}{2}u_{i+1,j} + (1-r)u_{i,j} + \frac{r}{2}u_{i-1,j}. \quad (47.21)$$

For each time level j , we obtain an $(M-1) \times (M-1)$ tridiagonal system which can be solved using iterative methods. The CNI scheme has no stability restriction and it is more accurate than the FTCS scheme. The implementation of the CNI scheme (47.21) is shown below:

```

1 subroutine parabolic_cni(M, N, f, a, b, Tend, alph)
2   procedure(one_var_func) :: f ! cni=Crank Nicholson Implicit
3   intent(in) :: M, N, a, b, Tend, alph
4   integer :: M, N, i, j, info
5   double precision :: a, b, Tend, alph, x, h, k, t, r
6   double precision :: CNMD(M+1), CNMDL(M), CNMDU(M), u(M+1), upre(M+1)
7   h = (b-a)/M ! Discretise Space
8   k = Tend/N ! Discretise Time
9   r = alph**2*k/(h**2) ! Condition number
10  t = 0d0
11  print "(a,f8.4,a,f8.4,a)", "# set title "CNI(h=", h, ", k=", k, ")"
12  print "(a,i5,i5)", "# set dgrid3d", M+1, N+1
13  print "(a)", "# splot "test.dat" u 1:2:3 w l'
14  do i=1,M+1
15     x = a+h*(i-1)
16     upre(i) = f(x)
17     print 100, t, x, upre(i)
18  end do
19  print *
20  do j=1,N
21     CNMD(1) = 1.0d0; CNMD(M+1) = 1.0d0
22     do i=2,M
23        CNMD(i) = 1+r
24     end do
25     CNMDL(M) = 0d0; CNMDU(1) = 0d0
26     do i=1,M-1
27        CNMDL(i) = -r/2d0
28        CNMDU(i+1) = -r/2d0
29     end do
30     ! dgtsv will change CNMD, CNMDL, CNMDU, so reinitialise them everytime
31     u(1) = 0d0; u(M+1) = 0d0 ! Homogeneous BC
32     do i=2,M
33        u(i) = (1-r)*upre(i) + r/2d0*(upre(i-1)+upre(i+1))
34     end do
35     ! DGTSV( N, NRHS, DL, D, DU, B, LDB, INFO )
36     call dgtsv(M+1,1,CNMDL,CNMD,CNMDU,u,M+1,info)
37     upre(:) = u(:)
38     t = j*k
39     print 100, t, a, 0d0
40     do i=2,M
41        print 100, t, a+h*(i-1), u(i)
42     end do
43     print 100, t, b, 0d0
44     print *
45  end do
46  100 format(f13.8,f13.8,f13.8)
47 end subroutine parabolic_cni

```

Example 47.12. Consider a one-dimensional heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad t > 0.$$

subject to the initial condition

$$u(x, 0) = \sin \pi x, \quad 0 \leq x \leq 1$$

and boundary conditions

$$u(0, t) = u(1, t) = 0, \quad t \geq 0.$$

Use the Crank-Nicolson method with an x -step of $h = 0.2$ and a t -step of $k = 0.001$ to estimate $u(0.4, 0.001)$.

Solution: Since the initial temperature distribution is symmetric with respect to $x = 0.5$, we only need to consider the grid points over $0 \leq x \leq 0.5$.

In this case $\alpha = 1$, $r = \frac{k}{h^2} = \frac{0.001}{(0.2)^2} = 0.025$.

Since $u(0.4, 0.001) \approx u_{2,1}$, we will perform the calculations as follows:

1. $u(0, t) = u(1, t) = 0 \Rightarrow u_{0,j} = u_{5,j} = 0, j = 0, 1, \dots$
2. $u(x, 0) = \sin \pi x \Rightarrow u_{i,0} = \sin 0.2\pi i$.
3. Substituting into (47.21) and ignoring higher order terms, we obtain

$$-0.025u_{i-1,j+1} + 2.05u_{i,j+1} - 0.025u_{i+1,j+1} = 0.025u_{i-1,j} + 1.95u_{i,j} + 0.025u_{i+1,j}. (*)$$

4. When $j = 0$, $(*) \Rightarrow -0.025u_{i-1,1} + 2.05u_{i,1} - 0.025u_{i+1,1} = 0.025u_{i-1,0} + 1.95u_{i,0} + 0.025u_{i+1,0}$.

- (a) $i = 1$: Since $u_{0,j} = 0$ and $u_{i,0} = \sin 0.2\pi i$,

$$u_{0,0} = 0, \quad u_{1,0} = \sin 0.2\pi = 0.58778525, \quad u_{2,0} = \sin 0.4\pi = 0.95105652,$$

we obtain the equation

$$-0.025u_{0,1} + 2.05u_{1,1} - 0.025u_{2,1} = 0.025u_{0,0} + 1.95u_{1,0} + 0.025u_{2,0}.$$

which simplifies to

$$2.05u_{1,1} - 0.025u_{2,1} = 1.95(0.58778525) + 0.025(0.95105652) = 1.1699576505 \quad (A)$$

- (b) $i = 2$: Since $u_{3,0} = \sin 0.6\pi = 0.95105652$,

$$-0.025u_{1,1} + 2.05u_{2,1} - 0.025u_{3,1} = 0.025u_{1,0} + 1.95u_{2,0} + 0.025u_{3,0}$$

which simplifies to

$$\begin{aligned} -0.025u_{1,1} + 2.05u_{2,1} - 0.025u_{3,1} &= 0.025(0.58778525) + 1.95(0.95105652) \\ &+ 0.025(0.95105652) = 1.89303126. \end{aligned} \quad (B)$$

Since $u_{3,1} = u_{2,1}$ by symmetry, we can further simplify it to

$$-0.025u_{1,1} + 2.025u_{2,1} = 1.89303126 \quad (B')$$

- (c) Solving (A) and (B') by Gauss elimination or Cramer's rule, we obtain

$$u_{1,1} = 0.58219907, \quad u_{2,1} = 0.94201789.$$

Therefore $u(0.4, 0.001) \approx u_{2,1} = 0.94201789$.

Example 47.13 (Tutorial 8, Q4). The exact solution of the 1-D heat equation

$$\begin{cases} u_t = u_{xx}, & 0 \leq x \leq 1, \quad t > 0 \\ u(0, t) = u(1, t) = 0, & t > 0, \\ u(x, 0) = f(x) = \sin \pi x, & 0 < x < 1 \end{cases}$$

is $u(x, t) = e^{-\pi^2 t} \sin(\pi x)$.

- (a) Write down the linear iteration problem associated with the Crank-Nicolson Scheme with an x step size of $h = 0.1$ and a t step size of $k = 0.01$.

Solution. $-\frac{1}{2}u_{i+1,j+1} + (1+1)u_{i,j+1} - \frac{1}{2}u_{i-1,j+1} = \frac{1}{2}u_{i+1,j} + (1-1)u_{i,j} + \frac{1}{2}u_{i-1,j}$.

By averaging FTCS and BTCS, we obtain

$$\begin{aligned} u_{i,j+1} - u_{i,j} &= \frac{k}{2h^2}[u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1} + u_{i+1,j} - 2u_{i,j} + u_{i-1,j}] \\ &= \frac{1}{2}[u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1} + u_{i+1,j} - 2u_{i,j} + u_{i-1,j}] \end{aligned}$$

which can be simplified to give

$$\begin{aligned} &\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.5 & 2.0 & -0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -0.5 & 2.0 & -0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.5 & 2.0 & -0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.5 & 2.0 & -0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.5 & 2.0 & -0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0.5 & 2.0 & -0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0.5 & 2.0 & -0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.5 & 2.0 & -0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.5 & 2.0 & -0.5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_{0,j+1} \\ u_{1,j+1} \\ u_{2,j+1} \\ u_{3,j+1} \\ u_{4,j+1} \\ u_{5,j+1} \\ u_{6,j+1} \\ u_{7,j+1} \\ u_{8,j+1} \\ u_{9,j+1} \\ u_{10,j+1} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0.0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0.0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.0 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0.0 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0.0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0.0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0.0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0.0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_{0,j} \\ u_{1,j} \\ u_{2,j} \\ u_{3,j} \\ u_{4,j} \\ u_{5,j} \\ u_{6,j} \\ u_{7,j} \\ u_{8,j} \\ u_{9,j} \\ u_{10,j} \end{bmatrix} \end{aligned}$$

Note that $M = \frac{1-0}{h} = 10$. Then $i = 1, \dots, 9$ and $j \geq 0$. □

- (b) Estimate $u(0.1, 0.02)$ using part (a). Note that a numerical computing environment is required to solve the linear algebra problem.

Solution. From $u(x, 0) = \sin \pi x$, it is easy to obtain

$$[u_{i,0}] = [0, 0.3090, 0.5878, 0.8090, 0.9511, 1.0000, 0.9511, 0.8090, 0.5878, 0.3090, 0]^T$$

Using a numerical computing environment, it is easy to find the next time step:

$$[u_{i,1}] = [0, 0.2802, 0.5329, 0.7335, 0.8623, 0.9067, 0.8623, 0.7335, 0.5329, 0.2802, 0]^T$$

and

$$[u_{i,2}] = [0, 0.2540, 0.4832, 0.6651, 0.7818, 0.8221, 0.7818, 0.6651, 0.4832, 0.2540, 0]^T$$

Thus, $u(0.1, 0.02) \approx u_{1,2} = 0.2540$. □

47.3.4 Method of Lines

The *method of lines (MOL)* is a general approach for the treatment of time-dependent PDEs. The standard MOL involves two steps. The first step is to discretize the spatial variables of a time-dependent PDE by finite differences to obtain an high-dimensional ODE-IVP. The second step is to integrate the ODE-IVP numerically using Runge-Kutta methods or multistep methods from Topic 6 [Schiesser and Griffiths, 2009], [Cellier and Kofman, 2006, Section 6.2].

Let us investigate how MOL works with a simple parabolic equation (47.12). If we discretize only the “spatial” component in (47.12), we will obtain an ODE below:

$$\frac{du_i}{dt} = \frac{\alpha^2}{h^2}(u_{i+1} - 2u_i + u_{i-1}), \quad i = 1, 2, \dots, N-1, \quad \text{BCs} \Rightarrow u_0 = u_N = 0. \quad (47.22)$$

From the initial condition, we have

$$u_i(x_i, 0) = \sin \pi x_i. \quad (47.23)$$

Both (47.22) and (47.23) form an ODE-IVP and we could use the numerical methods in Topic 6 to solve it.

In practise, more sophisticated MOLs are used. For example, the statistical software R has an add-on module called `ReacTran` which solves a advection-diffusion-reaction equation (47.11) based on MOL and it can be easily applied to solve (47.12).

Example 47.14. Use the R module `ReacTran` to solve the parabolic problem in Example 47.9.

Solution:

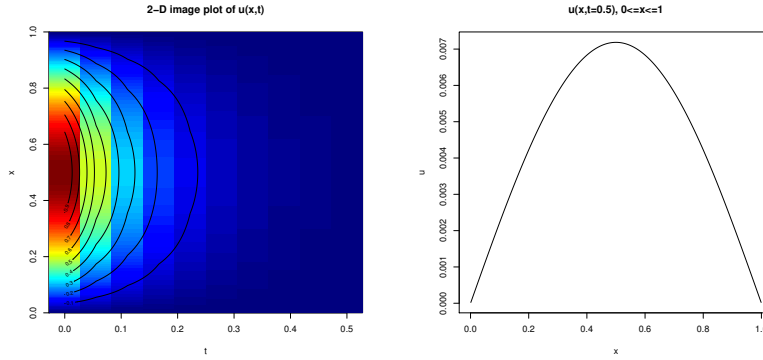
```

1 library("ReacTran")
2
3 N          = 500
4 L          = 1.0
5 diffusion  = 1.0
6 Tend       = 0.5
7 reaction   = 0.0
8
9 TimeGrid   = seq(0,Tend,length.out=10)
10 SpaceGrid  = setup.grid.1D(N=N,L=L)
11
12 # First, we discretise with respect to ``space`` variable
13 parabolic.rhs = function(t,x,parms){
14   # avection, v=0, by default
15   f = tran.1D(C=x,D=diffusion,dx=SpaceGrid,C.up=0,C.down=0)$dC
16   list(f + reaction)
17 }
18
19 # MOL: Perform numerical integration (Topic 6)
20 u0 = sin(pi*seq(0,L,length.out=N))
21 u = ode.1D(func=parabolic.rhs, y=u0, times=TimeGrid, parms=NULL, nspec=1)
22
23 setEPS()
24 postscript("eg_parab1.eps")
25 image(u, grid=SpaceGrid$x.mid, add.contour=TRUE, xlab="t", ylab="x",
26       main="2-D image plot of u(x,t)")
27 ### To see the ``temperature`` distribution at time t=0.5
28 #x11()
29 postscript("eg_parab2.eps")
30 plot(SpaceGrid$x.mid,u[10,2:(N+1)],type="l", xlab="x", ylab="u",
31      main="u(x,t=0.5), 0<=x<=1") #type="o","l","p"
32 ### To see how ``temperature`` u decay as time increases
33 #plot(TimeGrid,u[,1],type="o", xlab="x", ylab="u",
34      # main="u(any number between 1--N,t), 0<=t<=0.5")
35 #lines(c(0:endtime),u[,10])
36 dev.off()

```

The script generates the output of the approximation of solution as an image on the left.

Studying the “temperature” distribution u at time $t = 0.5$ gives us the right figure below.



47.3.5 Application: Financial Mathematics

The heat equation is fundamental to financial engineering for a number of reasons. First, it is a “component” (first order derivative in time and second order derivative in “space”) in Black-Scholes PDE (47.25). Second, the Black-Scholes PDE can be transformed to the heat equation by a change of variables $V(S, t) = e^{\alpha x + \beta \tau} u(x, t)$, where $\alpha = -\frac{1}{2}(2r/\sigma^2 - 1)$, $\beta = -\frac{1}{4}(2r/\sigma^2 + 1)^2$, $t = T - 2\tau/\sigma^2$ and $x = \ln S$ [Duffy, 2006, 39].

FDM is one popular numerical methods for solving option pricing problems. Application of FDM examples in finance can be found in White [2004, Section 5.5], Lyuu [2002, Section 18.1], Wilmott et al. [1994], Wilmott et al. [1995, Chapter 8] Schwartz [1977], Duffy [2006], Rebonato [1998, Chapter 9], Achdou and Pironneau [2005, Chapter 3], etc.

The *Black-Scholes model* is a continuous-time model involving an underlying asset whose price at time t is $S(t)$ and a risk-free asset whose price at time t is $S^0 = S^0(t)$: the evolution of S^0 is found by solving the ODE:

$$\frac{dS^0}{dt} = r(t)S^0$$

where $r(t)$ is the *force of interest* (refer to Topic 1 in UECM1403 Theory of Interest). For simplicity, $r(t)$ is normally assumed constant. The Black-Scholes model decomposes the return on a financial asset $\frac{dS}{S}$ as follows:

$$\frac{dS}{S} = \underbrace{\mu dt}_{\text{drift}} + \underbrace{\sigma(S, t) dB_t}_{\text{volatility}} \quad (47.24)$$

where μ is a measure of the average rate of growth of the asset price, B_t is a standard Brownian motion on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, i.e. a real-valued continuous stochastic process whose increments are independent and stationary, with $B_0 = 0$ $\mathbb{P}$ -almost surely, $\mathbb{E}(B_t) = 0$ and $\mathbb{E}(B_t^2) = t$.

Suppose the value of the derivative of S is $V(S, t)$. According to Black and Scholes [1973], by applying stochastic calculus (more precisely, Itô calculus) and by using a *no-arbitrage argument*, it is possible to derive the *Black-Scholes PDE* (formula (21) in Black and Scholes [1973]) from (47.24) [Brandimarte, 2006, 30]:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma(S, t)^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS \frac{\partial V}{\partial S} - rV = 0 \quad (47.25)$$

where

- $V(S(t), t)$ is the valuation of the option at time t if the current price of the underlying asset is $S(t)$;
- r is the risk-free interest rate, i.e. the rate of interest one can earn by investing her money in a safe account; and
- σ is a parameter related to the volatility of the price of the underlying asset S , which is risky.

The Black-Scholes equation is solved backwards from maturity at the expiration time $t = T$ to current time $t = 0$. Therefore, it is conventional to write (47.25) by change of time variable $\tau = T - t$:

$$\frac{\partial V}{\partial \tau} - \frac{1}{2}\sigma(S, \tau)^2 S^2 \frac{\partial^2 V}{\partial S^2} - rS \frac{\partial V}{\partial S} + rV = 0. \quad (47.26)$$

The Black-Scholes PDE (47.25) holds for any option (or portfolio of options) whose underlying asset satisfies the Black-Scholes assumptions. To determine a unique solution of (47.25), BCs and the solution at the expiry, $t = T$, needs to be given. This condition at $t = T$ is called *final condition* (or the *initial condition* for (47.26)). Different options satisfy the same PDE (47.25), but different BCs (and final conditions). Therefore, in order to determine the price of an option, we need to know the value of the option at time T [Zhu et al., 2013]. For example, the BCs for European Call are:

$$V(S, T) = \max(S - K, 0), \quad V(0, t) = 0, \quad V(S_{\max}, t) = Ke^{r(T-t)} \quad (47.27)$$

and the BCs for European Put are:

$$V(S, T) = \max(K - S, 0), \quad V(0, t) = Ke^{r(T-t)}, \quad V(S_{\max}, t) = 0. \quad (47.28)$$

So far, only a few types of options have analytic solutions [Wilmott et al., 1994]:

- European call option: $V(S, t) = SN(d_1) - Ke^{-r(T-t)}N(d_2)$
- European put option: $V(S, t) = Ke^{-r(T-t)}N(-d_2) - SN(-d_1)$

where

$$N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy, \quad d_1 = \frac{\log(S/K) + (r + \frac{\sigma^2}{2})(T-t)}{\sigma\sqrt{T-t}}, \quad d_2 = \frac{\log(S/K) + (r - \frac{\sigma^2}{2})(T-t)}{\sigma\sqrt{T-t}},$$

FDM is a popular numerical method for pricing option values because it is relatively simple and we can determine the option values at every grid points (S_i, t_j) . If the option value is not at a grid point, then interpolation from Topic 4 could be used to estimate the option price.

There are other two popular numerical methods for the valuation of options, such as Monte-Carlo simulation and binomial tree. However, one may face higher difficulty compare to FDM when applying them to solving exotic options.

Monte-Carlo method is a natural technique to price an option, but speeding up the method by variance reduction may be tricky. Tree method is another fast and intuitive technique to price an option but can become difficult as the complexity of the financial products grow.

It is possible solve the Black-Scholes PDE that describes European call and put options. However, since we have the analytic solution, it is a waste of time to calculate with PDE unless we are testing the effectiveness of the numerical method.

Instead, we will study the pricing of American put option found in Ødegaard [2014, Chapter 13], Brandimarte [2006, Chapter 9] and Seydel [2006, Chapter 4]. Note that American-style derivative securities usually have *free boundaries* and the numerical methods are far more complicated than any of the FDM mentioned above, for details, consult Zhu et al. [2013].

The BCs and final condition of the American put are given by

$$V(S, T) = \max(K - S, 0), \quad V(0, t) = Ke^{-r(T-t)}, \quad \underbrace{V(S, t) \rightarrow 0 \text{ as } S \rightarrow \infty}_{\text{free boundary}}. \quad (47.29)$$

Let us study the algorithm found in Ødegaard [2014, Chapter 13], which is written as Julia below:

eg-amerput.jl

```
1 function AmericanPutOption(S0, K, r, sigma, time, S_nsteps, t_nsteps)
2     const M = S_nsteps + rem(S_nsteps, 2) # We want M to be even
3     const N = t_nsteps
```

```

4      const sigma2 = sigma^2
5      const k = time/N
6      S_vals = 2*S0/float(M)*vec(float(collect(1:M+1))) # Assume S_max = 2*S0
7      A = spzeros(M+1,M+1)
8      A[1,1] = A[M+1,M+1] = 1.0
9      for i=2:M
10         A[i,i-1] = 0.5*i*k*( r-sigma2*i)
11         A[i,i ] = 1.0 + k*( r+sigma2*i^2)
12         A[i,i+1] = 0.5*i*k*(-r-sigma2*i)
13     end
14     B = max.(0, K-S_vals)
15     F = A\B
16     for t=N-1:-1:1
17         B = F
18         F = A\B
19         F = max.(F, K-S_vals) # Check for early exercise
20     end
21     P = F[div(M,2)] # Reason for requiring M to be even
22 end

```

This function prices of an American put option using “implicit difference formula”.

Example 47.15. Explain where do lines 10 to 12 come from in the Julia script `eg-amerput.jl`?

Solution: One of most annoying thing about Black-Scholes equation is that we normally have *final condition* and not initial condition. So, we don’t calculate from the “beginning of time”, we calculate from the “end of time” (termination date). This means the numerical method BTCS and FTCS are FTCS and BTCS respectively for Black-Scholes PDE! It can really confuse people.

The Julia script `eg-amerput.jl` uses implicit difference formula, therefore, we should be using FTCS formula for BTCS calculation (confusing, isn’t it?):

$$(47.25) \Rightarrow \frac{\partial V_{i,j}}{\partial t} + \frac{1}{2}\sigma^2 S_i^2 \frac{V_{i+1,j} - 2V_{i,j} + V_{i-1,j}}{h^2} + rS_i \frac{V_{i+1,j} - V_{i-1,j}}{2h} - rV_{i,j} = 0.$$

Here, we assume r and σ are constant.

Because, we are computing in the reverse of time, the arrangement of the above difference equation after multiplying by k should be

$$\left[\frac{\sigma^2}{2} S_i^2 \cdot \frac{k}{h^2} - rS_i \cdot \frac{k}{2h} \right] V_{i-1,j} + \left[1 + \sigma^2 S_i^2 \cdot \frac{k}{h^2} + rk \right] V_{i,j} + \left[-\frac{\sigma^2}{2} S_i^2 \cdot \frac{k}{h^2} - rS_i \cdot \frac{k}{2h} \right] V_{i+1,j} = V_{i,j+1} \quad (47.30)$$

Note that the underlying asset price S varies from 0 to current stock price “ $S(t)$ ” (we denote it S_0 in the Julia script). This is another very confusing thing about Black-Scholes PDE: $S(t)$ is actually the current price of the underlying asset, and after $T - t$, the derivative based on the underlying asset will terminate.

Therefore, assuming $S_i = ih$, where $h = \frac{S_0 - 0}{M}$. Then (47.30) becomes

$$\left[\frac{\sigma^2}{2} (ih)^2 \cdot \frac{k}{h^2} - rih \cdot \frac{k}{2h} \right] V_{i-1,j} + \left[1 + \sigma^2 (ih)^2 \cdot \frac{k}{h^2} + rk \right] V_{i,j} + \left[-\frac{\sigma^2}{2} (ih)^2 \cdot \frac{k}{h^2} - rih \cdot \frac{k}{2h} \right] V_{i+1,j} = V_{i,j+1}$$

or after simplification

$$\underbrace{\frac{1}{2} [\sigma^2 i^2 k - rik]}_{\text{line 10}} V_{i-1,j} + \underbrace{\left[1 + \sigma^2 i^2 k + rk \right]}_{\text{line 11}} V_{i,j} + \underbrace{\frac{1}{2} [-\sigma^2 i^2 k - rik]}_{\text{line 12}} V_{i+1,j} = V_{i,j+1} \quad (47.31)$$

In practise, solving option pricing problems using Crank-Nicolson method mentioned above is probably more accurate. Note however, that for non-smooth final conditions (which happen for most financial instruments), the Crank-Nicolson method is not satisfactory as numerical oscillations are not damped. For vanilla options, this results in oscillation in the gamma value

around the strike price. Therefore, special damping initialization steps are necessary (e.g., fully implicit finite difference method) Wikipedia [2015].

An alternative for is to transform the **primitive variable** S to $\ln S$, the **logarithmic price**. This will be illustrated in UECM3763 Computational Finance. The reason for this is because with a uniform grid in the primitive variable S , they are too many nodes for large values of S . This is not optimal because V tends to be flat for large S .

Example 47.16. Write a Julia script to price the value of an American put option at time $t = 0.5$ with the following data: $S = 50$, $K = 50$, $r = 0.1$, $\sigma = 0.4$. Use implicit difference formula with 200 steps in S and 200 steps in time to estimate the price.

Solution: Using the Julia function above, the price can be calculated as follows:

```
1 S0 = 50; K = 50; r = 0.1; sigma = 0.4; time = 0.5;
2 S_nsteps = 200; t_nsteps = 200;
3 price = AmericanPutOption(S0, K, r, sigma, time, S_nsteps, t_nsteps)
4 println("American Put price is approximately $price")
```

American Put price is approximately 4.6006397921238635

Remark 47.17. According to the article “The Epic Story of How a ‘Genius’ Hedge Fund Almost Caused a Global Financial Meltdown” from *Business Insider Malaysia*, around 1994, one bond-trading hedge fund (called the Long-Term Capital Management, LTCM for short) grew from launch (when?) to over \$100 billion ($= 10^{11}$, a large number indeed) in 3 years. The following are two aspects that made this fund famous:

- “yearly return of $\geq 40\%$ ”
- It was run by finance veterans, PhDs, professors and two Nobel Prize winners, Myron Scholes and Robert C. Merton.

But by 1998, LTCM lost approximately \$4.4 billion of its \$4.7 billion in capital and many banks have to bail it.

So finance is not maths or statistics! Statistical models only provide some kind of approximation when the financial market is complete (and continuous!). But remember, financial market is neither complete or continuous!

47.4 Hyperbolic Problems

Consider the following system of p first order partial differential equations for p unknown functions $\mathbf{u}(\mathbf{x}, t) = (u_1(\mathbf{x}, t), \dots, u_p(\mathbf{x}, t))$, where $\mathbf{x} \in \mathbb{R}^d$:

$$\frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^d \frac{\partial}{\partial x_j} \mathbf{f}^j(\mathbf{u}) = 0, \quad (47.32)$$

where $\mathbf{f}^j \in C^1(\mathbb{R}^p, \mathbb{R}^p)$, $j = 1, \dots, d$ are once continuously differentiable functions, nonlinear in general.

For each $\mathbf{f}^j$, $j = 1, \dots, d$, a $p \times p$ matrix A^j is defined as

$$A^j := \begin{bmatrix} \frac{\partial f_1^j}{\partial u_1} & \dots & \frac{\partial f_1^j}{\partial u_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_p^j}{\partial u_1} & \dots & \frac{\partial f_p^j}{\partial u_p} \end{bmatrix}.$$

We say that the system (47.32) is *hyperbolic* if for all $\alpha_1, \dots, \alpha_d \in \mathbb{R}$ the matrix $A := \alpha_1 A^1 + \dots + \alpha_d A^d$ has only real eigenvalues and is diagonalizable.

If the matrix A has distinct real eigenvalues, it follows that it's diagonalizable. In this case the system (47.32) is called *strictly hyperbolic*.

The interesting and really difficult fact of hyperbolic PDE (47.32) is that the solution does not have to be smooth even if the initial value is smooth in nonlinear problems.

Two special cases of hyperbolic equation (47.32) often studied in numerical methods are the first order wave equations

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0 \quad \text{and} \quad \frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f_1(u) + \frac{\partial}{\partial x} f_2(u) = 0$$

where $-\infty < x, y < \infty$ and $0 < t$. When f, f_1, f_2 are linear, we have

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = 0, \quad \text{and} \quad \frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} + b \frac{\partial u}{\partial y} = 0.$$

If you wonder what are the applications of hyperbolic PDE, you will find them in anything to do with “wave propagation”. For example, in wired and wireless telecommunication systems, in reservoir system (part of fluid mechanics), in seismology study, etc.

In wired and wireless telecommunication systems, the wave equation is used to determine the direction of signal power distribution and the effect of noise. Many people are more concerned with the health issue related to telecommunication signal stations. To conduct a scientific study on this, more complicated PDEs from quantum mechanics and chemistry are required.

Wave propagation in reservoir can be very dangerous, engineers must ensure that shock wave and resonance will not damage the reservoir, otherwise, a extremely serious flood may occur.

In seismology, we view earthquake and tsunami as the wave propagation cause by changes of earth plates. To determine the speed of how the sea wave travel is crucial in building tsunami warning system. Earthquake happens too fast and a warning system will never be useful. However, from the propagation of wave, it is possible for us to get a very rough picture about the earth beneath us and may be of little help for future prediction. However, the PDE problems in which we need to solve will be extremely complicated.

In this section, we will analyse the 1-dimensional homogeneous second order wave equation in standard form using an explicit scheme:

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} = \alpha^2 \frac{\partial^2 u}{\partial x^2}, & 0 < x < L, \quad 0 < t < T, \quad (\text{PDE}), \\ u(0, t) = u(1, t) = 0, & 0 < t < T, \quad (\text{boundary conditions}) \\ u(x, 0) = f(x), \quad u_t(x, 0) = g(x), & 0 \leq x \leq 1. \quad (\text{initial conditions}) \end{cases} \quad (47.33)$$

Here the PDE can be viewed as a “composite” of “hyperbolic equations” which are linear:

$$\left(\frac{\partial}{\partial t} - \alpha \frac{\partial}{\partial x} \right) \left(\frac{\partial}{\partial t} + \alpha \frac{\partial}{\partial x} \right) u = 0.$$

47.4.1 Finite Difference Method on Rectangular Grid

Using a 2-dimensional grid with x step size h and t step size k and by replacing both $\frac{\partial^2 u}{\partial t^2}$ and $\frac{\partial^2 u}{\partial x^2}$ with the centre difference approximations, we obtain a CTCS scheme as follows. A difference in sign makes 1-D wave equation very different from the Laplace equation (a special case of Poisson equation in Section 47.2.2).

Method 47.4 (CTCS Explicit Finite Difference Scheme).

1. GENERATE A GRID: Let $u_{i,j} = u(ih, jk) = u(x_i, t_j)$, $i = 0, \dots, M$, $j = 0, \dots, N$.
2. REPRESENT THE DERIVATIVE BY SOME FINITE DIFFERENCE FORMULAE: Applying centre difference approximation to (47.33), we obtain

$$\frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{k^2} = \alpha^2 \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2}, \quad i = 1, \dots, M-1, \quad j = 1, \dots, N-1.$$

3. REARRANGE THE CENTRE DIFFERENCE APPROXIMATION ABOVE INTO AN ITERATION:

$$u_{i,j+1} = ru_{i+1,j} + 2(1-r)u_{i,j} + ru_{i-1,j} - u_{i,j-1}, \quad r = \left(\frac{\alpha k}{h}\right)^2. \quad (47.34)$$

4. Compute the homogeneous boundary conditions:

$$u_{0,j} = u_{M,j} = 0, \quad j = 1, \dots, N.$$

and the numerical approximations of the initial conditions:

$$u_{i,0} = u(x_i, 0) = f(x_i), \quad i = 0, 1, \dots, M, \quad (47.35)$$

$$u_{i,1} = (1-r)f(x_i) + \frac{r}{2}(f(x_{i+1}) + f(x_{i-1})) + kg(x_i). \quad (47.36)$$

5. SOLVE THE ITERATION (47.34) FOR A GIVEN INITIAL CONDITION.

Remark 47.18. According to Courant-Friedrichs-Lewy (CFL) stability criterion, when the numerical method above is stable, the *Courant number* $\sigma = \frac{\alpha k}{h}$ must satisfy $\sigma \leq 1$. The

CTCS scheme is **stable** for $0 \leq r = \sigma^2 \leq 1$.

The higher-order approximation formula (47.36) is derived in Burden and Faires [2015, 741] with centre difference approximation and the use of a “ghost point” to the initial condition

$$u_t(x, 0) = g(x) \Rightarrow \frac{u_{i,1} - u_{i,-1}}{2k} = g_i \Rightarrow u_{i,-1} = u_{i,1} - 2kg_i$$

where $(i, -1)$ is a fictitious “ghost point”. Therefore, the calculation of $u_{i,1}$ becomes

$$\begin{aligned} (47.34) \wedge (j=1) &\Rightarrow u_{i,1} = 2(1-r)u_{i,0} + r(u_{i+1,0} + u_{i-1,0}) - u_{i,-1} \\ &\Rightarrow u_{i,1} = 2(1-r)u_{i,0} + r(u_{i+1,0} + u_{i-1,0}) - u_{i,1} + 2kg_i \\ &\Rightarrow 2u_{i,1} = 2(1-r)u_{i,0} + r(u_{i+1,0} + u_{i-1,0}) + 2kg_i \end{aligned}$$

which gives (47.36) after rearrangement of terms.

The implementation of the FDM for 1-D wave equation (47.33) is relatively straight-forward as follows.

```
1 subroutine hyperb_ctcs(M, N, f, g, a, b, Tend, alph, filename)
2 ! CTCS for Wave 1D Type Homogeneous 1 BC
3   procedure(one_var_func) :: f, g
4   intent(in) :: M, N, a, b, Tend, alph
5   character(len=*), optional :: filename
6   integer :: M, N, i, j, unitnum
7   double precision :: a, b, Tend, alph, x, h, k, t, r
8   double precision :: u(M+1), upre(M+1), upre2(M+1)
9   if (present(filename)) then
10     unitnum=10
11     open(unit=unitnum, file=filename, action="write", status="replace")
12   else
13     unitnum=6 ! Standard Output
14   end if
15   h = (b-a)/M ! Discretise Space
16   k = Tend/N ! Discretise Time
17   r = (alph*k/h)**2 ! Condition number
18
19   write (unitnum, "(a,f8.4,a,f8.4,a)") "# set title "CTCS(h=", h, ", k=", k, ")"
20   write (unitnum, "(a,i5,i5)") "# set dgrid3d", M+1, N+1
21   write (unitnum, "(a)") '# splot "test.dat" u 1:2:3 w l'
22
23   t = 0.0d0
24   do i=1,M+1
```

```

25      x = a+h*(i-1)
26      upre(i) = f(x)
27      write (unitnum,100) t, x, upre(i)
28  end do
29  write (unitnum,*)
30  t = k
31  upre2(1) = 0.0d0
32  write (unitnum,100) t, a, upre2(1)
33  do i=2,M
34      x = a+h*(i-1)
35      ! Better Approximation of u'(x,0). Read Burden & Faires, page 741
36      upre2(i) = (1d0-r)*upre(i) + 0.5d0*r*(upre(i+1)+upre(i-1)) + k*g(x)
37      write (unitnum,100) t, x, upre2(i)
38  end do
39  upre2(M+1) = 0.0d0
40  write (unitnum,100) t, b, upre2(1)
41  do j=2,N
42      t = t + k
43      write (unitnum,*)
44      write (unitnum,100) t, a, 0.0d0
45      do i=2,M
46          x = a+h*(i-1)
47          u(i) = 2d0*(1d0-r)*upre2(i)+r*(upre2(i+1)+upre2(i-1))-upre(i)
48          write (unitnum,100) t, x, u(i)
49      end do
50      write (unitnum,100) t, b, 0.0d0
51      upre(:) = upre2(:)
52      upre2(:) = u(:)
53  end do
54  if (present(filename)) then
55      close(unitnum)
56  end if
57  100 format (f10.6,f10.6,f15.10)
58  end subroutine hyperb_ctcs

```

Example 47.19. If a string with fixed ends at $x = 0$ and $x = 1$ is governed by the hyperbolic PDE

$$u_{tt} = u_{xx}, \quad 0 \leq x \leq 1, \quad t \geq 0$$

and it starts from an equilibrium position with initial velocity $g(x) = \sin \pi x$. (a) Find out what is its displacement u at time $t = 0.4$ and $x = 0.2, 0.4, 0.6, 0.8$ by using the CTCS explicit scheme with $h = k = 0.2$; (b) Plot the function approximation of $u(x, t)$ by using CTCS($h = k = 0.2$) numerical method.

Solution: (a)

$$1. \quad u(0, t) = u(1, t) = 0 \Rightarrow u_{0,j} = u_{5,j} = 0 \quad (\dagger)$$

$$2. \quad \boxed{j = 0}: u(x, 0) = f(x) = 0 \Rightarrow u_{i,0} = f_i = 0.$$

$$3. \quad g(x) = \sin \pi x \Rightarrow g_i = \sin 0.2i\pi.$$

4. With $\alpha = 1$, $h = k = 0.2 \Rightarrow r = (k/h)^2 = 1$, we have the following CTCS scheme:

$$u_{i,j+1} = u_{i+1,j} + u_{i-1,j} - u_{i,j-1} \quad (*)$$

5. $\boxed{j = 1}$: From the initial conditions, we have

$$u_{i,1} = (1 - r)f(x_i) + \frac{r}{2}(f(x_{i+1}) + f(x_{i-1})) + 0.2g_i = 0.2 \sin 0.2i\pi$$

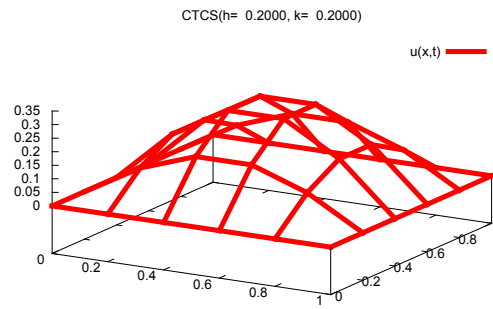
$$(a) \quad i = 0: u(0.0, 0.2) = u_{0,1} = 0$$

- (b) $i = 1$: $u(0.2, 0.2) \approx u_{1,1} = 0.2 \sin 0.2\pi = 0.1176$
(c) $i = 2$: $u(0.4, 0.2) \approx u_{2,1} = 0.2 \sin 0.4\pi = 0.1902$
(d) $i = 3$: $u(0.6, 0.2) \approx u_{3,1} = 0.2 \sin 0.6\pi = 0.1902$
(e) $i = 4$: $u(0.8, 0.2) \approx u_{4,1} = u_{1,1} = 0.1176$
(f) $i = 5$: $u(1.0, 0.2) = u_{5,1} = 0$

6. $j = 2$: Apply (*), we obtain:

- (a) $i = 1$: $u_{1,2} = u_{2,1} + u_{0,1} - u_{1,0} = 0.1902 + 0 - 0 = 0.1902$
(b) $i = 2$: $u_{2,2} = u_{3,1} + u_{1,1} - u_{2,0} = 0.1902 + 0.1176 - 0 = 0.3078$
(c) $i = 3$: $u_{3,2} = u_{4,1} + u_{2,1} - u_{3,0} = 0.1176 + 0.1902 - 0 = 0.3078$
(d) $i = 4$: $u_{4,2} = u_{5,1} + u_{3,1} - u_{4,0} = 0 + 0.1902 - 0 = 0.1902$

(b) If we use gnuplot to plot the data with $h = k = 0.2$, we obtain



Example 47.20 (Tutorial 8, Q5). Consider the 1-D wave equation together with the BCs and ICs

$$\begin{aligned} u_{tt} &= u_{xx}, & 0 \leq x \leq 1, \quad t > 0, \\ u(0, t) &= u(1, t) = 0, & t \geq 0, \\ u(x, 0) &= x(1 - x), & 0 \leq x \leq 1, \\ u_t(x, 0) &= \cos(\pi x), & 0 \leq x \leq 1. \end{aligned}$$

Use the CTCS explicit scheme with $\Delta x = 0.2$ and $\Delta t = 0.1$ to approximate the value of u at time $t = 0.2$ and $x = 0.4$.

Solution. Note that $\alpha = 1$, $r = \left(\frac{k}{h}\right)^2 = \left(\frac{0.1}{0.2}\right)^2 = \frac{1}{4} < 1$. The CTCS scheme is stable and the difference equation is

$$u_{i,j+1} - 2u_{i,j} + u_{i,j-1} = r(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) \Rightarrow u_{i,j+1} = \frac{3}{2}u_{i,j} + \frac{1}{4}(u_{i+1,j} + u_{i-1,j}) - u_{i,j-1}$$

From the ICs and when $j = 0, 1$, we have

$$\begin{aligned} u_{i,0} &= f(x_i) = x_i(1 - x_i), \quad i = 0, 1, \dots, 5, \\ u_{i,1} &= \frac{3}{2}u_{i,0} + \frac{1}{4}(u_{i+1,0} + u_{i-1,0}) - u_{i,-1} = \frac{3}{2}f(x_i) + \frac{f(x_{i+1}) + f(x_{i-1}))}{4} - (u_{i,1} - 2kg(x_i)) \\ &= \frac{3}{4}f(x_i) + \frac{f(x_{i+1}) + f(x_{i-1}))}{8} + 0.1g(x_i) \\ u_{i,2} &= \frac{3}{2}u_{i,1} + \frac{1}{4}(u_{i+1,1} + u_{i-1,1}) - u_{i,0} = \frac{3}{2}u_{i,1} + \frac{1}{4}(u_{i+1,1} + u_{i-1,1}) - f(x_i) \end{aligned}$$

To approximate $u(0.4, 0.2)$, we need to calculate $u_{2,2}$:

- $u_{2,2} = \frac{3}{2}u_{2,1} + \frac{u_{3,1} + u_{1,1}}{4} - f(x_2)$
- $u_{1,1} = 0.75x_1(1 - x_1) + 0.125(x_2(1 - x_2) + x_0(1 - x_0)) + 0.1 \cos(x_1\pi) = 0.2309016994$
- $u_{2,1} = 0.75x_2(1 - x_2) + 0.125(x_3(1 - x_3) + x_1(1 - x_1)) + 0.1 \cos(x_2\pi) = 0.2609016994$
- $u_{3,1} = 0.75x_3(1 - x_3) + 0.125(x_4(1 - x_4) + x_2(1 - x_2)) + 0.1 \cos(x_3\pi) = 0.1990983006$
- $\therefore u_{2,2} = \frac{3}{2}(0.2609016994) + \frac{0.2309016994 + 0.1990983006}{4} - 0.4(1 - 0.4) = 0.25885255$

□

47.4.2 Method of Lines

To apply the method of lines, we need to rewrite (47.33) as two coupled first order PDE in time so that we can apply the numerical integration method from Topic 6:

$$\begin{aligned}\frac{\partial u_1}{\partial t} &= u_2, \quad 0 \leq x \leq L, \quad 0 < t < T, \\ \frac{\partial u_2}{\partial t} &= \alpha^2 \frac{\partial^2 u_1}{\partial x^2} \\ u_1(0, t) &= u_1(1, t) = 0, \\ u_1(x, 0) &= f(x), \quad u_2(x, 0) = g(x)\end{aligned}$$

Let us study this method with the use of R module ReacTran. Personally, I find that R is a terrible programming environment. However, it has many additional package for computation.

Example 47.21. Consider a hyperbolic PDE $u_{tt} = u_{xx}$, $0 \leq x \leq 1$, $t \geq 0$ with boundary conditions $u(0, t) = u(1, t) = 0$ and initial conditions $u_t(x, 0) = 0$ and $u(x, 0) = \begin{cases} 4x - 1, & \frac{1}{4} \leq x \leq \frac{1}{2}, \\ 3 - 4x, & \frac{1}{2} \leq x \leq \frac{3}{4}, \\ 0, & \text{otherwise.} \end{cases}$

Plot the approximation of the solution by using ReacTran for the time interval $[0, 2]$.

Solution:

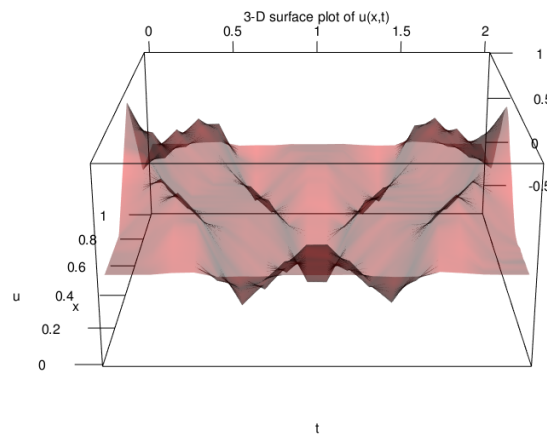
```
1 library("ReacTran")
2
3 N = 100; L = 1.0; diffusion = 1.0; Tend = 2.0; reaction = 0.0
4 TimeGrid = seq(0, Tend, length.out=20)
5 SpaceGrid = setup.grid.1D(N=N, L=L)
6
7 # First, we discretise with respect to ``space`` variable
8 hyperbolic.rhs = function(t, u, parms){
9   u1 = u[1:N]
10  u2 = u[-(1:N)]
11  du1 = u2
12  du2 = tran.1D(C=u1, D=diffusion, dx=SpaceGrid, C.up=0, C.down=0) $dC
13  return(list(c(du1, du2)))
14 }
15
16 # MOL: Perform numerical integration (Topic 6)
17 u1.ini = rep(0, N)
18 for(i in 1:N){
19   x = SpaceGrid$x.mid[i]
20   if (x > 0.25 && x <= 0.5) {
21     u1.ini[i] = 4*x-1
22   } else if (x > 0.5 && x < 0.75) {
23     u1.ini[i] = 3-4*x
24   }
25 }
26 u2.ini = rep(0, N)
27 u1and2.ini = c(u1.ini, u2.ini)
```

```

28 u1and2 = ode.1D(func=hyperbolic.rhs, y=u1and2.ini, times=TimeGrid,
29     parms=NULL, nspec=1)
30
31 library(rgl)
32 par3d("windowRect"= c(0,0,800,600),zoom=0.7)
33 #cat(length(TimeGrid), "x", length(SpaceGrid$x.mid), "=?", length(u1and2))
34 persp3d(x=TimeGrid, y=SpaceGrid$x.mid, z=u1and2[,2:(N+1)], xlab="t",
35     ylab="x", zlab="u", main="3-D surface plot of u(x,t)",
36     alpha=0.4, col="red", aspect=c(1,1,0.5))
37 #rgl.postscript("eg_hyper.eps" , fmt="eps") # core dump on Linux Mint
38 #rgl.postscript("eg_hyper.pdf" , fmt="pdf") # core dump on Linux Mint
39 rgl.snapshot("eg_hyper.png")

```

The plot does not reflect the exact solution properly probably due to the non-smooth initial condition.



Example 47.22. Consider a hyperbolic PDE

$$u_{tt} = 2000(u_{xx} + u_{yy}), \quad 0 \leq x, y \leq 600, \quad t \geq 0$$

with boundary conditions $u(0, y, t) = u(600, y, t) = u(x, 0, t) = u(x, 600, t) = 0$ and initial conditions

$$u(x, y, 0) = \begin{cases} 1, & (x, y) \in (200, 400) \times (250, 350), \\ 0, & \text{otherwise} \end{cases},$$

$$u_t(x, y, 0) = 0.$$

Write an R script to plot the approximation of the solution with x -step size 10 and y -step size 5 by using ReacTran for the time interval $[0, 5]$ with time step size 0.01.

Solution: We will be using the add-on module `animate` to generate a GIF file which “shows” the evolution of the wave:

```

1 library("ReacTran")
2
3 Nx = 60; Ny = 120; L = 600; diffusion = 2000; Tend = 5; reaction = 0.0
4 NT = Tend/0.01;
5 TimeGrid = seq(0, Tend, length.out=NT)
6 xGrid = setup.grid.1D(L=L, N=Nx)
7 yGrid = setup.grid.1D(L=L, N=Ny)
8 SpaceGrid = setup.grid.2D(xGrid, yGrid)
9
10 # First, we discretise with respect to ``space`` variable
11 hyperbolic.rhs = function(t, u, parms){
12     N = length(u)/2
13     u1 = matrix(nrow=Nx, ncol=Ny, u[1:N])
14     u2 = matrix(nrow=Nx, ncol=Ny, u[(N+1):(2*N)])

```

```

15     du1 = u2
16     du2 = tran.2D(C=u1,D.x=diffusion, D.y=diffusion, v.x=0, v.y=0,
17                 dx=xGrid$dx, dy=yGrid$dx,
18                 C.x.up=0,C.x.down=0,C.y.up=0,C.y.down=0)$dC
19     return(list(c(du1,du2)))
20 }
21
22 # MOL: Perform numerical integration (Topic 6)
23 u1.ini = matrix(nrow = Nx, ncol = Ny, data = 0, byrow = FALSE)
24 u2.ini = matrix(nrow = Nx, ncol = Ny, data = 0, byrow = FALSE)
25 u2.ini[((Nx+1)/2-10):((Nx+1)/2+10),((Ny+1)/2-10):((Ny+1)/2+10)] = 1
26 uand2.ini = c(u1.ini, u2.ini)
27 uand2 = ode.2D(func=hyperbolic.rhs, y=uand2.ini, times=TimeGrid,
28               dims=c(Nx,Ny), parms=NULL, method="rk4")
29
30 ###http://stackoverflow.com/questions/13023274/how-to-do-printf-in-r
31 #printf <- function(...) invisible(print(sprintf(...)))
32 N=Nx*Ny
33 library(animation)
34 saveGIF({
35     for(i in 1:(NT/20-1)) {
36         j = 20*i
37         u = matrix(nrow=Nx, ncol=Ny, data=uand2[j,2:(N+1)])
38         image(x=xGrid$x.mid, y=yGrid$y.mid, u, xlab="x", ylab="y",
39              main=sprintf("u(x, y, t=%6.3f)",TimeGrid[j]))
40     }
41 }, movie.name="eg_hyper2d.gif")

```

47.5 Open Source PDE Software

PDE models are used to solve many engineering problems such as:

- Structural analysis of civil engineering constructions such as skyscrapers, dams, etc.
- Underground gases, crude oils and minerals can be studied using PDE eigenvalue analysis.
- Disaster simulation related to fire, flood (this is especially important in city planning because of the sudden flood due to a heavy rain), earth-quake, tsunami, etc.

A general PDE software will be extremely complicated because the objects of interest are not rectangular as discussed above. The steps in solving a well-posed problem associated with a linear elliptic PDE or a general time-independent linear PDE goes like:

1. Subdivide the domain into appropriate meshes. This process is called *mesh generation*
2. Write the equation in its *weak form*, integrate it with the *test functions* V_i
3. Choose N simple *trial functions* ϕ_j and let $u = u_1\phi_1 + \cdots + u_N\phi_N$. Most of the time, $V_i = \phi_i$
4. Form a m -order finite element matrix B by using $(m+1)$ -point Gauss quadrature?
5. Assemble the *stiffness matrix* $K = \int B^T DB dV$ and the *load vector* F and process the boundary conditions.
6. Solve the large linear algebra problem $Ku = F$
7. Post-processing: improve the precision, posterior error estimate, data printout, data plotting, etc.

Therefore, implementing a PDE software is extremely difficult. In order to avoid the redundant efforts in building FDM/FVM/FEM, mathematicians not only write books but also write standard software for general use. These software can be categorised into:

1. Mesh/Grid generation software

- GTS (LGPL): GNU Triangulated Surface Library is an Open Source Free Software Library intended to provide a set of useful functions to deal with 3D surfaces meshed with interconnected triangles. <http://gts.sourceforge.net/>
- Gmsh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities <http://geuz.org/gmsh/>
- Jonathan Richard Shewchuk's Triangle: A Two-Dimensional Quality Mesh Generator and Delaunay Triangulator. <https://www.cs.cmu.edu/~quake/triangle.html>
- NETGEN: An automatic 3d tetrahedral mesh generator. It accepts input from *constructive solid geometry (CSG)* or boundary representation (BRep) from STL file format. The connection to a geometry kernel allows the handling of IGES and STEP files. NETGEN contains modules for mesh optimisation and hierarchical mesh refinement. <http://sourceforge.net/projects/netgen-mesher/>
- TetGen: A Quality Tetrahedral Mesh Generator and a 3D Delaunay Triangulator <http://wias-berlin.de/software/tetgen/>

2. General Programming Libraries

- deal.II
- FETK: The Finite Element ToolKit is a collaboratively developed, evolving collection of adaptive finite element method (AFEM) software libraries and tools for solving coupled systems of nonlinear geometric partial differential equations (PDE). The FETK libraries and tools are written in an object-oriented form of ANSI-C and in C++, and include a common portability layer (MALOC) for all of FETK, a collection of standard numerical libraries (PUNC), a stand-alone high-quality surface and volume simplex mesh generator (GAMer), a stand-alone networked polygon display tool (SG), a general nonlinear finite element modeling kernel (MC), and a MATLAB toolkit (MCLite) for prototyping finite element methods and examining simplex meshes using MATLAB. The entire FETK Suite of tools is highly portable (from iPhone to Blue Gene/L), thanks to use of a small abstraction layer (MALOC) and heavy use of the GNU Autoconf infrastructure. <http://www.fetk.org/>
- getfem++ (depends on Numpy)
- libMesh: Provides a framework for the numerical simulation of PDEs using arbitrary unstructured discretizations on serial and parallel platforms. A major goal of the library is to provide support for adaptive mesh refinement (AMR) computations in parallel while allowing a research scientist to focus on the physics they are modelling. <http://libmesh.github.io/>
- MODULEF (Fortran 77): A modular finite element library <https://www.rocq.inria.fr/modulef/english.html>
- DOUG (Fortran 95): Domain Decomposition on Unstructured Grids. <http://www.dougdevel.org/doug/docs/files.html>

3. Scripting PDE solvers:

- freefem++ [Hecht, 2012]: A PDE solver. It has its own language. freefem scripts can solve multiphysics non linear systems in 2D and 3D. Problems involving PDE (2d, 3d) from several branches of physics such as fluid-structure interactions require interpolations of data on several meshes and their manipulation within one program. FreeFem++ includes a fast 2^d -tree-based interpolation algorithm and a language for the manipulation of data on multiple meshes. FreeFem++ is written in C++ and the FreeFem++ language is a C++ idiom. <http://www.freefem.org/ff++/>
It's really nice! We can use it to do floating point arithmetic and string manipulation.:

```

real  x=3.14,y;
int   i,j;
complex c;
cout << " x = " << x << "\n";

x = 1;y=2;
x=y;

i=0;j=1;
cout << 1 + 3 << " " << 1/3 << "\n";
cout << 10 ^10 << "\n";
cout << 10 ^-10 << "\n";
cout << -10^-2+5 << "==" 4.99 \n";
cout << 10^-2+5 << "==" 5.01 \n";
cout << "----- complex ---- \n" ;
cout << 10-10i << " \n";
int prec=cout.precision(12);
cout.scientific << " scientific : \n";
cout << " -1^(1/3) = " << (-1+0i)^(1./3.) << " (precision=12)\n";
cout.precision(prec);
cout.fixed;
cout.showpos << " fixed and showpos :\n";
cout << " -1^(1/3) = " << (-1+0i)^(1./3.) << " (precision=" << prec << ")\n";
cout.noshowpos << "noshowpos\n";
cout << " 8^(1/3)= " << (8)^(1./3.) << " \n";
cout << " sqrt(-1) = " << sqrt(-1+0i) << " \n";
complex a=10+1i;
cout.default << "default \n";
cout << a << endl;
cout << " real(a) = " << real(a) << " conj(a) " << conj(a)
    << " arg(a) = " << arg(a) << endl;
cout << " ++i =" << ++i ;
cout << " i=" << i << "\n";
cout << " i++ = " << i++ << "\n";
cout << " i = " << i << "\n";
// ---- string concatenation -----
string str,str1;
str="abc+";
str1="+abcdddd+";
str=str + str1;
str = str + 2 ;
cout << "str= " << str << "==" abc++abcdddd+2;\n";

{ real x=0;
for (int i=0;i<10;i++)
    x += i*i;
cout << " x= " << x << endl;
// example of if arithmetic expression
real a = x == 0 ? x : -1;
real b = x != 0 ? x : -1;
cout << " a = " << a << " b = " << b << endl;
string ss="\z\a\b\f\-\-\\";
cout << "\"" << ss << "\"" << endl;
}

```

We can also use it to solve simple Laplace equation problem:

```

mesh Th=square(10,10);
fespace Vh(Th,P1); // P1 FE space
Vh uh,vh; // unkown and test function.
func f=1; // right hand side function
func g=0; // boundary condition function

problem laplace(uh,vh,solver=GMRES,tgv=1e5) = // Weak Form
    int2d(Th)( dx(uh)*dx(vh) + dy(uh)*dy(vh) ) // bilinear form
    + int1d(Th,1)( uh*vh)
    - int1d(Th,1)( vh)

```

```

- int2d(Th)( f*vh ) // linear form
+ on(2,3,4,u=h=g) ; // boundary condition

laplace; plot(uh); // solve the problem and view the result
//plot(uh,ps="LaplacePl.eps",value=true); // save the figure

```

Solving the following Poisson problem

$$-\Delta p = f \text{ on } \Omega, \quad \frac{dp}{d\mathbf{n}} = (g_1 d, g_2 d) \cdot \mathbf{n} \text{ on } \Gamma_{123}, \quad p = g d \text{ on } \Gamma_1$$

with de Mixte finite element formulation. Find $u \in L^2(\Omega)$ and $u \in H(\text{div})$ such that

$$u - \text{Grad} p = 0, \quad -\text{div} u = f, \quad u \cdot \mathbf{n} = (g_1 d, g_2 d) \cdot \mathbf{n} \text{ on } \Gamma_{123}, \quad p = g d \text{ on } \Gamma_1$$

the variationnel form is:

$$\forall v \in H(\text{div}) : \quad v \cdot \mathbf{n} = 0 \text{ on } \Gamma_4 : \int_{\Omega} uv + p \text{div} v - \int_{\Gamma_{123}} g d * v.n = 0$$

$$\forall q \in L^2 : \quad - \int_{\Omega} q \text{div} u = \int_{\Omega} f q \quad \text{and} \quad u \cdot \mathbf{n} = (g_1 n, g_2 n) \cdot \mathbf{n} \text{ on } \Gamma_4.$$

```

mesh Th=square(10,10);
fespace Vh(Th,RT0);
fespace Ph(Th,P0);

func gd = 1.;
func g1n = 1.;
func g2n = 1.;
func f = 1.;

Vh [u1,u2],[v1,v2];
Ph p,q;

problem laplaceMixte([u1,u2,p],[v1,v2,q],solver=UMFPACK,eps=1.0e-10,
    tgv=1e30,dimKrylov=150) =
    int2d(Th)( p*q*0e-10+ u1*v1 + u2*v2 + p*(dx(v1)+dy(v2))
        + (dx(u1)+dy(u2))*q )
+ int2d(Th)( f*q)
- int1d(Th,1,2,3)( gd*(v1*N.x +v2*N.y)) // int on gamma
+ on(4,u1=g1n,u2=g2n);

laplaceMixte;
plot([u1,u2],coef=0.1,wait=1,ps="lapRTuv.eps",value=true);
plot(p,fill=1,wait=1,ps="laRTp.eps",value=true);

```

- Fenics (Python2): a collection of free software with an extensive list of features for automated, efficient solution of PDE. <http://fenicsproject.org/>, <http://www.bodysoulmath.org/> (Old)

$$\text{Solving Poisson equation} \begin{cases} -\Delta u(x, y) = f(x, y), & (x, y) \in [0, 1] \times [0, 1] \\ u(0, y) = u(1, y) = 0, \\ \frac{du}{d\mathbf{n}}(x, y) = \sin(5x), \text{ for } y = 0 \text{ or } y = 1 \end{cases}$$

where

$$f(x, y) = 10 \exp\left(-\frac{(x - 0.5)^2 + (y - 0.5)^2}{0.02}\right)$$

goes like this:

```

# This file is part of DOLFIN. Copyright (C) 2007-2011 Anders Logg
from dolfin import *

# Create mesh and define function space

```

```

mesh = UnitSquareMesh(32, 32)
V = FunctionSpace(mesh, "Lagrange", 1)

# Define Dirichlet boundary (x = 0 or x = 1)
def boundary(x):
    return x[0] < DOLFIN_EPS or x[0] > 1.0 - DOLFIN_EPS

# Define boundary condition
u0 = Constant(0.0)
bc = DirichletBC(V, u0, boundary)

# Define variational problem
u = TrialFunction(V)
v = TestFunction(V)
f = Expression("10*exp(-(pow(x[0] - 0.5, 2) + pow(x[1] - 0.5, 2)) / 0.02)")
g = Expression("sin(5*x[0])")
a = inner(grad(u), grad(v))*dx
L = f*v*dx + g*v*ds

# Compute solution
u = Function(V)
solve(a == L, u, bc)

# Save solution in VTK format
file = File("poisson.pvd")
file << u

# Plot solution
plot(u, interactive=True)

```

- **GetDP:** An open source finite element solver using mixed elements to discretize de Rham-type complexes in one, two and three dimensions. The main feature of GetDP is the closeness between the input data defining discrete problems (written by the user in ASCII data files) and the symbolic mathematical expressions of these problems. <http://www.geuz.org/getdp/>

4. Application oriented PDE software:

- **FEM1D and FEM2D** in Reddy [2006]
- **Felt:** A package for introductory level Finite ELeMenT analysis. It is centred around a mathematical engine designed to simply, effectively, accurately, and flexibly solve most types of structural / mechanical problems that would be encountered in an introductory course in finite element analysis. It was developed in an overzealous fit of “we can do better than that” based on some antiquated Fortran code that we had been using for just such a course. <http://felt.sourceforge.net>
- **CalculiX:** ccx (for stress analysis job?) and cgx -b cylinder.fbd for editing mesh with GUI. If the file is not there, cgx will complain, and one has to create points and lines using commands. An example is given below:

```

PNT p0      -0.00000    -0.00000    0.00000
PNT py      -0.00000     1.00000    0.00000
PNT p1       1.00000    -0.00000    0.00000
PNT P001     1.00000     1.00000    0.00000
PNT P002    -0.00000    -0.00000   -1.00000
PNT P003    -0.00000     1.00000   -1.00000
PNT P006    -1.00000    -0.00000    0.00000
PNT P007    -1.00000     1.00000    0.00000
PNT P00A     0.00000    -0.00000    1.00000
PNT P00C     0.00000     1.00000    1.00000
LINE L001 p1 P001 2
LINE L002 P002 P003 2
LINE L003 p1 P002 p0 8
LINE L004 P001 P003 py 8

```

```

LINE L005 P006 P007 2
LINE L006 P002 P006 p0 8
LINE L007 P003 P007 py 8
LINE L008 P00A P00C 2
LINE L009 P006 P00A p0 8
LINE L00A P007 P00C py 8
LINE L00I P00A p1 p0 8
LINE L00J P00C P001 py 8
GSUR A001 + BLEND - L001 + L003 + L002 - L004
GSUR A002 + BLEND - L002 + L006 + L005 - L007
GSUR A003 + BLEND - L005 + L009 + L008 - L00A
GSUR A005 + BLEND - L008 + L00I + L001 - L00J
SETA p1 p p1
SETA p2 p P001
SETA 11 l L001
SETA 12 p P002
SETA 12 p P003
SETA 12 l L002
SETA 13 p P006
SETA 13 p P007
SETA 13 l L005
SETA 14 p P00A
SETA 14 p P00C
SETA 14 l L008
SETA 15 p p1
SETA 15 p P001
ELTY all QU4

```

To save data to file, type “send all fbd” to save the mesh data to “all.fbd” and to quit, type “quit” in the GUI window.

- Elmer:

- Heat transfer: for the modelling of conduction, radiation, ...

$$\rho c_p \left(\frac{\partial T}{\partial t} + (\mathbf{u} \cdot \nabla) T \right) = \nabla \cdot (k \nabla T) + \bar{\tau} : \bar{\varepsilon} + \rho h$$

where ρ = density, c_p = heat capacity at constant pressure, T = temperature, $\mathbf{u}$ = convection velocity, k = heat conductivity, h = source of heat, $\bar{\tau} : \bar{\varepsilon}$ = frictional viscous heating, which is negligible in most cases.

- Species transport (or convection-diffusion equation)

$$\rho \left(\frac{\partial c_i}{\partial t} + (\mathbf{v} \cdot \nabla) c_i \right) = \rho \nabla \cdot (D_i \nabla c_i) + S_i.$$

– ...

- Tochnog: A professional tochnog version is distributed by FEAT <http://www.feat.nl>. <http://sourceforge.net/projects/tochnog/>
- Impact: An explicit Finite Element Program Suite which simulates dynamic impact events. It has a range of elements, contact handling and different material laws. Models can be created, solved and analyzed with the included pre- and postprocessor. Impact has been designed to be easily extendible and modular to enable programmers a way to easy add features to the program without having to enter other parts of the code. Impact has been written in Java. This choice of language may seem strange at first, but with the recent development of Java engines, speed penalty is not that significant. On the other hand, the Object Oriented features and the high portability of Java is a clear advantage for the future.

Impact is a Finite Element Code which is based on an Explicit Time stepping algorithm. These kind of codes are used to simulate dynamic phenomena such as car crashes and similar, usually involving large deformations.

There are quite few explicit codes around which might seem strange since the other cousin (implicit finite element) are quite common. The implicit codes are used to

simulate static loads in structures. Something that explicit codes does not manage very well. <http://www.impact-fem.org/>

- OpenFOAM: A free, open source CFD software package, licensed and distributed by the OpenFOAM Foundation . It has a large user base across most areas of engineering and science, from both commercial and academic organisations. OpenFOAM has an extensive range of features to solve anything from complex fluid flows involving chemical reactions, turbulence and heat transfer, to solid dynamics and electromagnetics. <http://www.openfoam.org/>
- SU2 is a suite of open-source software tools for the numerical solution of Computational Fluid Dynamics problems (including design) <https://github.com/su2code>, <http://su2.stanford.edu/>

5. Post-Processing Software:

- IBM OpenDX
- Gnuplot
- Python's Matplotlib
- gl2ps: OpenGL to postscript
- POV-Ray
- <https://github.com/vispy/vispy>, <http://glumpy.github.io/index.html>
- VTK, Paraview, Enthought Mayavi2, Visit
- <https://github.com/Vaa3D/Vaa3D-Wiki/wiki/Vaa3D-Wiki>
- <https://github.com/OpenChemistry/tomviz>

Part VII

Signal Analysis and Control

Chapter 48

Signal and Filter

In signal processing, functions are normally changing with respect to time. Therefore, the variable t , a function $t \mapsto f(t)$ is used as an abstraction of a *physical signal*.

More work to be done:

RautR-SwamyMNS-ModernAnalogFilterAnalysisAndDesignAPracticalApproach-Wiley2010
D3E/UUEA3423 Control Systems

Laplace Transform (and Transfer Function for Linear-Time-Invariant (LTI) system, Definition 32.6) is used in linear digital signal processing and linear control.

Refer to control package LTI functions dssdata (dssdata.m), zpkdata (zpkdata.m) and octave-signal to find applications.

Relevant Info:

<https://github.com/hyumo/lti>

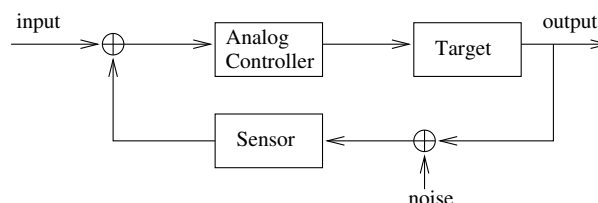
<https://github.com/TrymArve/Open-Loop-Optimization-of-LTI-Systems>

<https://python-control.readthedocs.io/en/latest/linear.html>

48.1 Systems and Linear Transforms

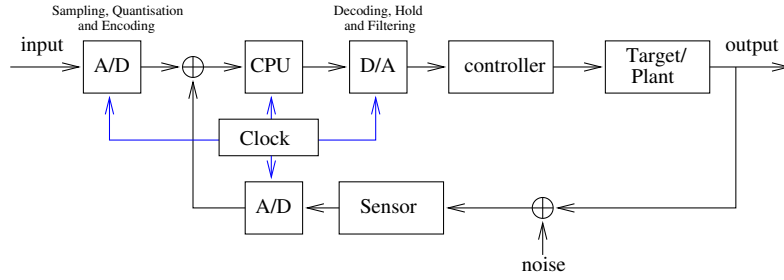
Engineering systems or simply *systems* can be categorised into analog systems, digital systems and hybrid systems. **Analog systems** have continuous states with respect to t ; **digital systems** have discrete states with respect to a discrete time t ; **hybrid systems** have a mix of discrete and continuous states. The *states* of a system are the signals of the system that can be measured/characterised by mathematical equations such as ODEs.

Analog systems can be approximated by the following block diagram.

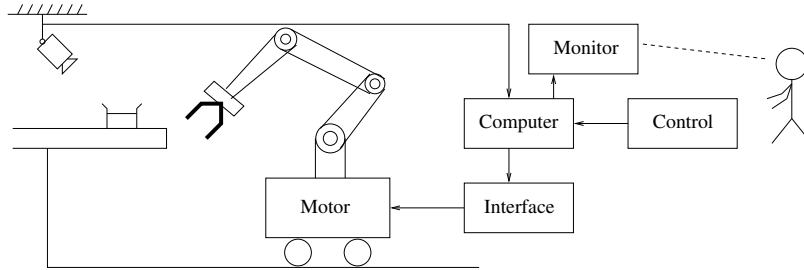


These systems are used in satellite control, satellite communication, electro-mechanical system, power system, (satellite) signal processing, feedback control system, image processing, radio astronomy, medical computer tomography, oil exploration, seismic research, etc.

Although analog controllers were cheaper, they were prone to noise and errors. When digital controllers, which are digital systems, were invented, they started to replace analog controllers because of digital controllers were resistant to noise and programmable (which were much more flexible compare to analog controllers). Hence we have a hybrid system with a digital controller (CPU) shown below.



Today, a mixture of analog and digital systems called hybrid systems are applied in digital filtering, industrial control, transport, agriculture, astronomy, home entertainment, etc. in the form of computer-aided manufacturing system, sensor network, radio telescope, robots (AIBO), etc. The following diagram shows a hybrid system that is used in chemical or medical industries to handle hazardous or delicate materials.



Note that the computer control is digital whereas the motor is analog.

Block diagrams are useful representations in engineering. However, in order to study the behaviour of an engineering system, a mathematical representation is required. This can be done by looking at the mathematical specification of the system or identifying the characteristic of the system using experiment. The specification or characterisation can be given in the form of a *state equation* or in the form of a *transfer function* below:

Type	Mathematical Model	State Equation	Transfer Function
analog system	differential equations	$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}, t)$ $\mathbf{y} = \mathbf{g}(\mathbf{x})$	$\mathbf{Y}(s) = \tilde{\mathbf{G}}(\mathbf{X}(s))$ $= \tilde{\mathbf{F}}(\mathbf{X}(s), \mathbf{U}(s), s)$
digital system	difference equations	$\mathbf{x}(k+1) = \mathbf{f}(\mathbf{x}(k), \mathbf{u}(k), k)$ $\mathbf{y}(k) = \mathbf{g}(\mathbf{x}(k))$	$\mathbf{Y}(z) = \tilde{\mathbf{G}}(\mathbf{X}(z))$ $= \tilde{\mathbf{F}}(\mathbf{X}(z), \mathbf{U}(z), z)$

The state equation characterises the system's behaviour in time-domain whereas the transfer function characterises the system's behaviour in frequency-domain. The state equation and transfer function are linked by a linear transform defined below.

Definition 48.1. A *linear transform* L of a set of functions over a subset of $\mathbb{R}$ satisfies the following property:

$$L\{\alpha f + \beta g\} = \alpha L\{f\} + \beta L\{g\}$$

for all constants α, β and for all functions over a subset of $\mathbb{R}$, f, g .

Fourier Transform and Laplace Transform are special linear transforms called integral transforms defined below.

Definition 48.2. An *integral transform* of a function f is given by

$$T\{f(x)\} = \int_{\gamma}^{\delta} K(x, s) f(s) ds$$

where $K(x, s)$ is the *kernel* of the transformation and $\gamma, \delta \in \mathbb{R} \cup \{\pm\infty\}$, $\gamma < \delta$. When $\gamma = -\infty$ or $\delta = \infty$ the integral is improper.

48.2 Fourier Integrals

The complex Fourier series representation of a $2L$ -periodic function f is

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{in\omega t}$$

where $c_n = \frac{1}{2L} \int_{-L}^L f(t) e^{-in\omega t} dt$. Substituting the formula for c_n into the Fourier series leads to the formula

$$f(t) = \sum_{n=-\infty}^{\infty} \left[\frac{1}{2L} \int_{-L}^L f(s) e^{-in\omega s} ds \right] e^{in\omega t}.$$

Let $\omega_n = n\omega = \frac{n\pi}{L}$, $\Delta\omega = \omega_{n+1} - \omega_n = \omega = \frac{\pi}{L}$ thus $\frac{1}{2L} = \frac{\omega}{2\pi}$ and the formula can be rewritten as

$$f(t) = \sum_{n=-\infty}^{\infty} \left[\int_{-L}^L f(s) e^{-i\omega_n s} ds \right] e^{i\omega_n t} \frac{\Delta\omega}{2\pi}.$$

When the period of the function becomes *very large*, i.e. when $L \rightarrow \infty$, the function becomes “non-periodic” and we obtain a *Fourier integral* representation of the “non-periodic” function f :

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f(s) e^{-i\omega s} ds \right] e^{i\omega t} d\omega \quad (48.1)$$

since $\lim_{L \rightarrow \infty} \sum_{n=-\infty}^{\infty} G(\omega_n) \Delta\omega = \lim_{\Delta\omega \rightarrow 0} \sum_{n=-\infty}^{\infty} G(\omega_n) \Delta\omega = \int_{-\infty}^{\infty} G(\omega) d\omega$.

Theorem 48.1 (Fourier Integral Theorem). *If f is a piecewise-continuous function defined over any finite closed interval on $\mathbb{R}$ and satisfies both the following conditions:*

1. *the left-hand and right-hand derivatives of f exist at every point,*
2. *the integral $\int_{-\infty}^{\infty} |f(t)| dt$ exists*

then the function f can be represented by the Fourier integral (48.1).

Note that the conditions in this theorem is similar to the Dirichlet’s conditions for Fourier series. At a discontinuous point t_d , the Fourier integral converges to the average value of the left-hand and right-hand limits of the function f , i.e. $\frac{1}{2}(f(t_d+) + f(t_d-))$.

Expanding equation (48.1), we obtain another form of Fourier integral representation of $f(t)$:

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\left(\int_{-\infty}^{\infty} f(s) \cos \omega s ds \right) \cos \omega t + \left(\int_{-\infty}^{\infty} f(s) \sin \omega s ds \right) \sin \omega t \right] dt.$$

Fourier Cosine Integral

When f is an even function, $\int_{-\infty}^{\infty} f(t) \sin \omega t dt = 0$ and $\int_{-\infty}^{\infty} f(t) \cos \omega t dt = 2 \int_0^{\infty} f(t) \cos \omega t dt$, hence we have Fourier Cosine integral:

$$f(t) = \frac{2}{\pi} \int_0^{\infty} \left[\int_0^{\infty} f(s) \cos \omega s ds \right] \cos \omega t dt.$$

Fourier Sine Integral

When f is an odd function, $\int_{-\infty}^{\infty} f(t) \cos \omega t dt = 0$ and $\int_{-\infty}^{\infty} f(t) \sin \omega t dt = 2 \int_0^{\infty} f(t) \sin \omega t dt$, hence we have Fourier Sine integral:

$$f(t) = \frac{2}{\pi} \int_0^{\infty} \left[\int_0^{\infty} f(s) \sin \omega s ds \right] \sin \omega t dt.$$

48.3 Fourier Transforms

In the Fourier integral representation of f :

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f(s) e^{-i\omega s} ds \right] e^{i\omega t} d\omega$$

The bracket part on the right acts like Fourier coefficients in the Fourier series. This inspires us to define Fourier transform as follows.

Definition 48.3. The *Fourier Transform* of a non-periodic function f is defined as

$$\mathcal{F}\{f(t)\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt = F(\omega). \quad (48.2)$$

Here the kernel $K(t, \omega) = \frac{e^{-i\omega t}}{\sqrt{2\pi}}$.

The *Inverse Fourier Transform* of a non-periodic function F is defined as

$$\mathcal{F}^{-1}\{F(\omega)\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega) e^{i\omega t} d\omega = f(t).$$

Note that some books use different notations and definitions for Fourier Transform. We will stick to this convention throughout.

eg: two-exp

Example 48.2. Find the Fourier Transform for $f(t) = e^{-|t|}$.

Solution.

$$\begin{aligned} \mathcal{F}\{e^{-|t|}\} &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-|t|} e^{-i\omega t} dt = \frac{1}{\sqrt{2\pi}} \left[\int_{-\infty}^0 e^{t-i\omega t} dt + \int_0^{\infty} e^{-t-i\omega t} dt \right] \\ &= \frac{1}{\sqrt{2\pi}} \left[\frac{1}{1-i\omega} + \frac{1}{1+i\omega} \right] = \frac{1}{\sqrt{2\pi}} \frac{2}{1+\omega^2}. \end{aligned}$$

□

Similarly, the Fourier Cosine and Sine integrals inspire us to introduce the definitions of Fourier Cosine Transform and Fourier Sine Transform as follows.

Definition 48.4. The *Fourier Cosine Transform* of f over $[0, \infty)$ is defined as

$$\mathcal{F}_c\{f(t)\} = \sqrt{\frac{2}{\pi}} \int_0^{\infty} f(t) \cos \omega t dt = F_c(\omega)$$

and the *inverse Fourier Cosine Transform* of F_c is defined as

$$\mathcal{F}_c^{-1}\{F_c(\omega)\} = \sqrt{\frac{2}{\pi}} \int_0^{\infty} F_c(\omega) \cos \omega t d\omega = f(t).$$

The *Fourier Sine Transform* of f over $[0, \infty)$ is defined as

$$\mathcal{F}_s\{f(t)\} = \sqrt{\frac{2}{\pi}} \int_0^{\infty} f(t) \sin \omega t dt = F_s(\omega).$$

and the *inverse Fourier Sine Transform* of F_s is defined as

$$\mathcal{F}_s^{-1}\{F_s(\omega)\} = \sqrt{\frac{2}{\pi}} \int_0^{\infty} F_s(\omega) \sin \omega t d\omega = f(t).$$

Example 48.3. Find the Fourier Cosine and Sine Transforms for $f(t) = e^{-at}$, $a > 0$.

Solution.

$$\begin{aligned}
 F_s(\omega) &= \sqrt{\frac{2}{\pi}} \int_0^\infty e^{-at} \sin \omega t dt = \sqrt{\frac{2}{\pi}} \left[\frac{e^{-at}}{a^2 + \omega^2} (-a \sin \omega t - \omega \cos \omega t) \right]_0^\infty \\
 &= \sqrt{\frac{2}{\pi}} \left[0 - \frac{e^0}{a^2 + \omega^2} (0 - \omega \cos 0) \right] = \sqrt{\frac{2}{\pi}} \frac{\omega}{a^2 + \omega^2} \\
 F_c(\omega) &= \sqrt{\frac{2}{\pi}} \int_0^\infty e^{-at} \cos \omega t dt = \sqrt{\frac{2}{\pi}} \left[\frac{e^{-at}}{a^2 + \omega^2} (-a \cos \omega t + \omega \sin \omega t) \right]_0^\infty \\
 &= \sqrt{\frac{2}{\pi}} \left[0 - \frac{e^0}{a^2 + \omega^2} (-a \cos 0 + 0) \right] = \sqrt{\frac{2}{\pi}} \frac{a}{a^2 + \omega^2}
 \end{aligned}$$

□

48.4 Step Function and Piecewise Continuous Function

Definition 48.5. A *Heaviside step function*, or a *(unit) step function*, is defined as

$$H(t - a) = \begin{cases} 0, & t < a, \\ 1, & t \geq a. \end{cases}$$

The product of arbitrary function f with Heaviside step function gives a function which is zero when $t < a$:

$$f(t)H(t - a) = \begin{cases} 0, & t < a, \\ f(t), & t \geq a. \end{cases}$$

Definition 48.6. A function f is called *piecewise continuous* on an interval $\alpha \leq t \leq \beta$ if the interval can be partitioned by a finite number of points $\alpha = t_0 < t_1 < \cdots < t_n = \beta$ so that

1. f is continuous on each open sub-interval $t_{i-1} < t < t_i$;
2. f approaches a finite limit as the endpoints of each sub-interval are approached from within the sub-interval.

The following example shows that piecewise continuous functions can be written in terms of Heaviside step function.

Example 48.4. Write the following function in terms of Heaviside step function:

(a) “On-off-switch” function $f(t) = \begin{cases} 1, & a \leq t < b, \\ 0, & t < a \text{ or } t \geq b. \end{cases}$

(b) $g(t) = \begin{cases} 7t^2, & 0 \leq t < 3, \\ t + 4, & 3 \leq t < 5, \\ 9, & t \geq 5. \end{cases}$

Solution. (a) $f(t) = H(t - a) - H(t - b)$.

(b) $g(t) = 7t^2H(t) + (t + 4 - 7t^2)H(t - 3) + (9 - t - 4)H(t - 5)$.

□

The following example calculates the Fourier transform of functions with Heaviside step function.

Example 48.5.

1. “Decay function”, $H(t)e^{-at}$, $a > 0$.

$$\begin{aligned}\mathcal{F}\{H(t)e^{-at}\} &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} H(t)e^{-at}e^{-i\omega t}dt = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-at-i\omega t}dt \\ &= \frac{1}{\sqrt{2\pi}} \left[\lim_{t \rightarrow +\infty} \frac{e^{(-a-i\omega)t}}{-a-i\omega} - \frac{e^0}{-a-i\omega} \right] = \frac{1}{\sqrt{2\pi}} \frac{1}{a+i\omega}.\end{aligned}$$

2. “On-off-switch” function, $\Pi_{[a,b]}(t) = H(t-a) - H(t-b)$.

$$\mathcal{F}\{\Pi_{[a,b]}(t)\} = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-i\omega t}dt = \frac{1}{\sqrt{2\pi}} \left[\frac{e^{-i\omega t}}{-i\omega} \right]_a^b = \frac{1}{\sqrt{2\pi}} \left[\frac{e^{-i\omega a} - e^{-i\omega b}}{i\omega} \right].$$

48.5 Properties of Fourier Transform

This section derives some general properties of Fourier Transform are useful in the calculation of Fourier Transform.

Theorem 48.6 (Linearity). *If $\mathcal{F}\{f(t)\} = F(\omega)$ and $\mathcal{F}\{g(t)\} = G(\omega)$, given α and β are constants, then*

$$\mathcal{F}\{\alpha f(t) + \beta g(t)\} = \alpha F(\omega) + \beta G(\omega).$$

Theorem 48.7 (Time-differentiation). *If $\mathcal{F}\{f(t)\} = F(\omega)$ then*

$$\mathcal{F}\left\{\frac{d^n}{dt^n}f(t)\right\} = (i\omega)^n F(\omega).$$

Example 48.8. If $\mathcal{F}\{y\} = Y(\omega)$ and $\mathcal{F}\{u\} = U(\omega)$, let $Y(\omega) = G(\omega)U(\omega)$ find $G(\omega)$ if $y(t)$ and $u(t)$ are related by the differential equation $y'' + 3y' + 7y = 3u'' + 2u$.

$$\begin{aligned}(i\omega)^2 Y(\omega) + 3(i\omega)Y(\omega) + 7Y(\omega) &= 3(i\omega)^2 U(\omega) + 2U(\omega) \\ \Rightarrow (-\omega^2 + 3i\omega + 7)Y(\omega) &= (-3\omega^2 + 2)U(\omega) \\ \Rightarrow G(\omega) = \frac{Y(\omega)}{U(\omega)} &= \frac{-3\omega^2 + 2}{-\omega^2 + 3i\omega + 7}.\end{aligned}$$

Theorem 48.9 (Time-shift). *If $\mathcal{F}\{f(t)\} = F(\omega)$ then*

$$\mathcal{F}\{f(t-\tau)\} = e^{-i\omega\tau} F(\omega).$$

Example 48.10. Find the Fourier transform of the following functions.

- (a) $H(t) - H(t-2a)$;
- (b) $H(t-1)e^{-2t}$;
- (c) $H(t-t_0)e^{-k(t-t_0)}$.

Solution.

$$\begin{aligned}\text{(a) } \mathcal{F}\{H(t) - H(t-2a)\} &= \mathcal{F}\{\Pi_{[0,2a]}(t)\} = \mathcal{F}\{\Pi_{[-a,a]}(t-a)\} \\ &= e^{-i\omega a} \frac{1}{\sqrt{2\pi}} \frac{e^{i\omega a} - e^{-i\omega a}}{i\omega} = \frac{1}{\sqrt{2\pi}} \frac{1 - e^{-i2a\omega}}{i\omega}; \\ \text{(b) } \mathcal{F}\{H(t-1)e^{-2t}\} &= e^{-2} \mathcal{F}\{H(t-1)e^{-2(t-1)}\} = e^{-2} e^{-i\omega} \mathcal{F}\{H(t)e^{-2t}\} \\ &= e^{-(2+i\omega)} \frac{1}{\sqrt{2\pi}} \frac{1}{2+i\omega} = \frac{e^{-(2+i\omega)}}{\sqrt{2\pi}(2+i\omega)};\end{aligned}$$

(c) In general we have

$$\mathcal{F}\{H(t-t_0)e^{-k(t-t_0)}\} = \frac{1}{\sqrt{2\pi}} \frac{e^{-i\omega t_0}}{k+i\omega}.$$

□

Theorem 48.11 (Frequency-shift). *If $\mathcal{F}\{f(t)\} = F(\omega)$ and α is a constant, then*

$$\mathcal{F}\{e^{i\alpha t} f(t)\} = F(\omega - \alpha).$$

Example 48.12. Find the Fourier transform of $g(t) = f(t) \cos \alpha t$.

$$\begin{aligned} \mathcal{F}\{g(t)\} &= \mathcal{F}\left\{f(t) \frac{e^{i\alpha t} + e^{-i\alpha t}}{2}\right\} = \frac{1}{2} \mathcal{F}\{e^{i\alpha t} f(t)\} + \frac{1}{2} \mathcal{F}\{e^{i(-\alpha)t} f(t)\} \\ &= \frac{1}{2} F(\omega - \alpha) + \frac{1}{2} F(\omega + \alpha). \end{aligned}$$

Theorem 48.13 (Symmetry). *If $\mathcal{F}\{f(t)\} = F(\omega)$, then*

$$\mathcal{F}\{F(t)\} = f(-\omega).$$

Example 48.14. Find the Fourier transform of $\frac{1}{1+t^2}$.

Solution. According to Example ^{eq:two-exp}48.2, $\mathcal{F}\{e^{-|t|}\} = \sqrt{\frac{2}{\pi}} \frac{1}{1+\omega^2}$, i.e. $\mathcal{F}\{\sqrt{\frac{\pi}{2}} e^{-|t|}\} = \frac{1}{1+\omega^2}$.

Using the symmetry property, $\mathcal{F}\{\frac{1}{1+t^2}\} = \sqrt{\frac{\pi}{2}} e^{-|\omega|} = \sqrt{\frac{\pi}{2}} e^{-|\omega|}$. □

Theorem 48.15 (Scaling). *If $\mathcal{F}\{f(t)\} = F(\omega)$, then*

$$\mathcal{F}\{f(at)\} = \frac{1}{a} F\left(\frac{\omega}{a}\right).$$

Example 48.16. Find the Fourier transform of $\frac{1}{1+9t^2}$.

$$\mathcal{F}\left\{\frac{1}{1+9t^2}\right\} = \mathcal{F}\left\{\frac{1}{1+(3t)^2}\right\} = \frac{1}{3} \sqrt{\frac{\pi}{2}} e^{-|\frac{\omega}{3}|}.$$

Definition 48.7. Given two piecewise-continuous function f and g , the *convolution* of the two functions is given by the operation

$$(f * g)(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(\tau) g(t - \tau) d\tau.$$

Properties of convolution:

1. $f * g = g * f$;
2. $f * (g + h) = f * g + f * h$.
3. $(f * g) * h = f * (g * h)$;

Example 48.17. Find the convolution of $f(t) = tH(t)$ and $g(t) = (\sin 2t)H(t)$.

$$\begin{aligned}
(f * g)(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sin 2\tau H(\tau) \cdot (t - \tau) H(t - \tau) d\tau \\
&= \frac{H(t)}{\sqrt{2\pi}} \int_0^t (t - \tau) \sin 2\tau d\tau = \frac{H(t)}{\sqrt{2\pi}} \left[-\frac{1}{2}(t - \tau) \cos 2\tau - \frac{1}{4} \sin 2\tau \right]_0^t \\
&= \frac{-1}{4\sqrt{2\pi}} (\sin 2t - 2t) H(t) = \frac{1}{\sqrt{2\pi}} \left(\frac{t}{2} - \frac{\sin 2t}{4} \right) H(t).
\end{aligned}$$

Theorem 48.18 (Convolution Theorem). *If $\mathcal{F}\{f(t)\} = F(\omega)$ and $\mathcal{F}\{g(t)\} = G(\omega)$ then*

1. *the Fourier transform for $f * g$ (convolution in time) is*

$$\mathcal{F}\{(f * g)(t)\} = F(\omega)G(\omega);$$

2. *the inverse Fourier transform for $F * G$ (convolution in frequency) is*

$$\mathcal{F}^{-1}\{(F * G)(\omega)\} = f(t)g(t).$$

Example 48.19. Find the inverse Fourier Transform for $\frac{1}{(1 + i\omega)(2 + i\omega)}$.

Solution. The answer is simple if we apply partial fraction below

$$\mathcal{F}^{-1}\left\{\frac{1}{(1 + i\omega)(2 + i\omega)}\right\} = \mathcal{F}^{-1}\left\{\frac{1}{1 + i\omega} - \frac{1}{2 + i\omega}\right\} = \sqrt{2\pi}e^{-t}H(t) - \sqrt{2\pi}e^{-2t}H(t).$$

However, we can also apply convolution as follows

$$\begin{aligned}
\mathcal{F}^{-1}\left\{\frac{1}{(1 + i\omega)(2 + i\omega)}\right\} &= \mathcal{F}^{-1}\left\{\frac{1}{1 + i\omega}\right\} * \mathcal{F}^{-1}\left\{\frac{1}{2 + i\omega}\right\} = \sqrt{2\pi}e^{-t}H(t) * \sqrt{2\pi}e^{-2t}H(t) \\
&= \frac{2\pi}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\tau}H(\tau) \cdot e^{-2(t-\tau)}H(t - \tau) d\tau = \sqrt{2\pi}H(t) \int_0^t e^{-\tau}e^{-2(t-\tau)} d\tau \\
&= \sqrt{2\pi}H(t)e^{-2t} \int_0^t e^{\tau} d\tau = \sqrt{2\pi}H(t)e^{-2t}(e^t - 1).
\end{aligned}$$

□

48.6 Impulse Function and Generalised Fourier Transforms

Physicists introduced the generalised function as the “derivative” of a Heaviside step function:

$$\delta(t - a) := \frac{d}{dt}H(t - a).$$

The generalised function $\delta(t - a)$ is called a *Dirac delta function* or an *impulse function*. It can be informally written as

$$\delta(t - a) = \begin{cases} \infty, & t = a, \\ 0, & t \neq a. \end{cases}$$

Dirac delta function is not a function in the traditional sense (which should not contain infinity). But we can usually treat it as a normal function as long as multiplication of two impulse function are avoided. The following are some basic properties of Dirac delta function:

1. Filtering: $\int_{-\infty}^{\infty} \delta(t - a)f(t)dt = f(a)$. Specifically $\int_{-\infty}^{\infty} \delta(t - a)dt = 1$;
2. Even: $\delta(-t) = \delta(t)$.

The following example calculates the Fourier transform of a non-absolutely-integrable Heaviside step function and the Dirac delta function when $a = 0$:

Example 48.20.

1. In order to calculate the Fourier transform of Heaviside step function $H(t)$, we need to treat it as a limit function of $e^{-\varepsilon|t|}H(t)$:

$$\begin{aligned}\mathcal{F}\{H(t)\} &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} H(t)e^{i\omega t} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \lim_{\varepsilon \rightarrow 0^+} e^{-\varepsilon|t|} H(t)e^{-i\omega t} dt \\ &= \frac{1}{\sqrt{2\pi}} \lim_{\varepsilon \rightarrow 0^+} \int_0^{\infty} e^{-\varepsilon t - i\omega t} dt = \frac{1}{\sqrt{2\pi}} \lim_{\varepsilon \rightarrow 0^+} \left[\frac{e^{(-\varepsilon - i\omega)t}}{-\varepsilon - i\omega} \right]_0^{\infty} \\ &= \frac{1}{\sqrt{2\pi}} \lim_{\varepsilon \rightarrow 0^+} \left[\lim_{t \rightarrow +\infty} \frac{e^{(-\varepsilon - i\omega)t}}{-\varepsilon - i\omega} - \frac{e^0}{-\varepsilon - i\omega} \right] = \frac{1}{\sqrt{2\pi}} \frac{1}{i\omega}\end{aligned}$$

2. Using the filtering property, we can calculate the Fourier transform of $\delta(t - a)$:

$$\mathcal{F}\{\delta(t)\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \delta(t)e^{i\omega t} dt = \frac{1}{\sqrt{2\pi}} e^{i\omega 0} = \frac{1}{\sqrt{2\pi}}.$$

From the above result, we observe that the “time-differentiation” property of δ and H are related by $\mathcal{F}\{\delta(t)\} = (i\omega)\mathcal{F}\{H(t)\}$.

In calculating the Fourier transform of Heaviside step function, we have changed the definition slightly in order to obtain an answer. We discover that it is impossible to obtain an answer using the definition 48.2. For example, the calculation of Fourier transform of $\cos$ leads to a result with no limit when $t \rightarrow \infty$. To tackle this problem, we will have to generalise Fourier transform based on the properties of Fourier transform introduced in the previous section. In this case, the generalised Fourier transform for functions like (a) $f(t) = 1$, (b) $f(t) = e^{i\alpha t}$, (c) $f(t) = \cos \alpha t$, (d) $f(t) = \sin \alpha t$ could be obtained as follows.

- (a) Since $\mathcal{F}\{\delta(t)\} = \frac{1}{\sqrt{2\pi}}$ and $\mathcal{F}\{\sqrt{2\pi}\delta(t)\} = 1$, therefore by the symmetry property of Fourier transform, we have $\mathcal{F}\{1\} = \sqrt{2\pi}\delta(-\omega) = \sqrt{2\pi}\delta(\omega)$.
- (b) By the frequency-shift property, $\mathcal{F}\{e^{i\alpha t}\} = \mathcal{F}\{e^{i\alpha t} \cdot 1\} = \sqrt{2\pi}\delta(\omega - \alpha)$.
- (c) $\mathcal{F}\{\cos \alpha t\} = \mathcal{F}\left\{\frac{e^{i\alpha t} + e^{-i\alpha t}}{2} \cdot 1\right\} = \frac{\sqrt{2\pi}\delta(\omega - \alpha) + \sqrt{2\pi}\delta(\omega + \alpha)}{2} = \sqrt{\frac{\pi}{2}}(\delta(\omega + \alpha) + \delta(\omega - \alpha))$
- (d) $\mathcal{F}\{\sin \alpha t\} = \mathcal{F}\left\{\frac{e^{i\alpha t} - e^{-i\alpha t}}{2i} \cdot 1\right\} = \frac{\sqrt{2\pi}\delta(\omega - \alpha) - \sqrt{2\pi}\delta(\omega + \alpha)}{2i} = i\sqrt{\frac{\pi}{2}}(\delta(\omega + \alpha) - \delta(\omega - \alpha))$

48.7 Plancherel's Theorem

In Fourier series, we learn that the sum of the power spectrum of a periodic signal is related to the mean square average of the signal by the Parseval's Theorem:

$$\frac{1}{2L} \int_{-L}^L |f(t)|^2 dt = \sum_{n=-\infty}^{\infty} |c_n|^2.$$

In Fourier Transform, an analogous theorem called Plancherel's Theorem exists.

Theorem 48.21 (Plancherel's Theorem or Parseval's Theorem for Fourier Transform).

$$\int_{-\infty}^{\infty} |f(t)|^2 dt = \int_{-\infty}^{\infty} |F(\omega)|^2 d\omega.$$

We call $|F(\omega)|^2 = F(\omega)\overline{F(\omega)}$ the *energy spectral density*, and the plot $|F(\omega)|^2$ against ω is called the *energy spectrum*.

Example 48.22. 1. Determine the energy spectrum of $H(t) \exp(-at)$, $a > 0$.

2. Evaluate $\int_{-\infty}^{\infty} \frac{1}{a^2 + \omega^2} d\omega$.

3. Find the Fourier sine transform of e^{-x} and hence evaluate $\int_0^{\infty} \frac{x^2}{(x^2 + 1)^2} dx$.

Solution.

1. Let $f(t) = H(t) \exp(-at)$, $a > 0$. Since $F(\omega) = \frac{1}{\sqrt{2\pi}} \frac{1}{a + i\omega}$, the energy spectrum density is

$$|F(\omega)|^2 = \frac{1}{2\pi} \frac{1}{a^2 + \omega^2}.$$

and the energy spectrum is the plot $|F(\omega)|^2$ against ω .

2. Using Parseval's Theorem, we know that

$$\begin{aligned} \int_{-\infty}^{\infty} |F(\omega)|^2 d\omega &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{a^2 + \omega^2} d\omega = \int_{-\infty}^{\infty} |f(t)|^2 dt \\ &= \int_{-\infty}^{\infty} \exp(-2at) H(t) dt = \left[\frac{\exp(-2at)}{-2a} \right]_0^{\infty} = \frac{1}{2a}. \end{aligned}$$

Hence

$$\int_{-\infty}^{\infty} \frac{1}{a^2 + \omega^2} d\omega = \frac{\pi}{a}.$$

3. Let $f(t) = e^{-t}$.

$$F_s(\omega) = \sqrt{\frac{2}{\pi}} \int_0^{\infty} e^{-t} \sin \omega t dt = \sqrt{\frac{2}{\pi}} \left[\frac{e^{-t}}{1 + \omega^2} (-\sin \omega t - \omega \cos \omega t) \right]_0^{\infty} = \sqrt{\frac{2}{\pi}} \frac{\omega}{1 + \omega^2}.$$

Using Parseval's Theorem, we have

$$\int_0^{\infty} |F_s(\omega)|^2 d\omega = \int_0^{\infty} \left| \sqrt{\frac{2}{\pi}} \frac{\omega}{1 + \omega^2} \right|^2 d\omega = \frac{2}{\pi} \int_0^{\infty} \frac{\omega^2}{(1 + \omega^2)^2} d\omega = \int_0^{\infty} |e^{-t}|^2 dt = \frac{1}{2}.$$

Hence

$$\int_0^{\infty} \frac{\omega^2}{(1 + \omega^2)^2} d\omega = \frac{\pi}{4}.$$

□

48.8 Application in Solving Ordinary Differential Equations

In system engineering, a single input single output linear system is usually described by an ODE:

$$a_n y^{(n)} + a_{n-1} y^{(n-1)} + \cdots + a_1 y' + a_0 y = b_m u^{(m)} + b_{m-1} u^{(m-1)} + \cdots + a_1 u' + a_0 u$$

where $y = y(t)$ is the output and $u = u(t)$ is the input. Assuming $y(0) = \cdots = y^{(n-1)}(0) = 0$ and $u(0) = \cdots = u^{(m-1)}(0) = 0$. Using the properties of Fourier Transform, it is possible to obtain

$$\begin{aligned} (a_n (i\omega)^n + a_{n-1} (i\omega)^{n-1} + \cdots + a_1 (i\omega) + a_0) Y(\omega) \\ = (b_m (i\omega)^m + b_{m-1} (i\omega)^{m-1} + \cdots + b_1 (i\omega) + b_0) U(\omega), \end{aligned}$$

in short,

$$Y(\omega) = \frac{B(i\omega)}{A(i\omega)} U(\omega).$$

The function $\frac{B(i\omega)}{A(i\omega)}$ is called the *frequency response* of the system. For a given $U(\omega)$, we can solve for the ODE by taking the inverse Fourier Transform to obtain the solution $y(t)$.

In general, however, engineers are more interest in knowing the frequency response based on particular inputs like an impulse or a Heaviside function. There are two ways to analyse the frequency response:

1. Bode analysis: Plotting magnitude of $\left| \frac{B(i\omega)}{A(i\omega)} \right|$ and the phase angle $\arg \frac{B(i\omega)}{A(i\omega)}$ against ω .
2. Nyquist analysis: Plotting the curve $\frac{B(i\omega)}{A(i\omega)}$, $\omega \in [0, \infty)$ on the complex plane.

48.9 DFT

sec:dft

Given a continuous function, we could represent the function either as a Fourier series or a Fourier transform (depending on whether it's periodic). However, if we have a sequence of data points, we will use discrete Fourier transform.

Fourier series is linear algebra in infinite dimensions. The “vectors” were functions f ; they were projected on the sines and cosines, which produces Fourier coefficients. From this infinite sequence of sines and cosines, multiplied by their Fourier coefficients, we can reconstruct f .

In actual calculations it is the *discrete Fourier transform* (DFT) that we compute. The theory is pure linear algebra, based on orthogonality. The input is a sequence of numbers $y_0, \dots, y_{n-1}$, instead of function f . The output is another sequence of the same length — a set of n Fourier coefficients $c_0, \dots, c_{n-1}$ instead of an infinite sequence.

Definition 48.8. Let $y = \{y_k = f(k\Delta t)\}_{k=0}^{n-1}$ (y_k is the sampling point) be a sequence of n data points. The n points DFT of y is defined by

$$c_k = \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} y_j \exp(-i(k\Delta\omega)(j\Delta t)), \quad \Delta\omega = \frac{2\pi}{n\Delta t}$$

where j is the index for the data points and k is the index of the DFT term or wave number.

An easier to remember form of DFT is

$$c_k = \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} \hat{w}^{jk} y_j, \quad \hat{w} = \exp(-i\Delta\omega\Delta t) = \exp\left(-\frac{2\pi i}{n}\right).$$

Example 48.23. When $y = \{c\}_{k=0}^{n-1}$, c is a constant, $\hat{w} = \exp\left(-\frac{2\pi i}{n}\right)$. The DFT is given by

$$c_k = \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} \left(\exp\left(-\frac{2\pi i}{n}\right) \right)^{jk} y_j = \frac{c}{\sqrt{n}} \sum_{j=0}^{n-1} \left(\exp\left(-\frac{2\pi i}{n}\right) \right)^j = \frac{c}{\sqrt{n}} \frac{1 - \left(\exp\left(-\frac{2\pi i}{n}\right) \right)^n}{1 - \exp\left(-\frac{2\pi i}{n}\right)} = 0, \quad k \neq 0,$$

$$c_0 = \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} \left(\exp\left(-\frac{2\pi i}{n}\right) \right)^0 y_j = \frac{nc}{\sqrt{n}} = c\sqrt{n}$$

since $e^{-2k\pi i} = 1$.

Similar to inverse Fourier transform, we can define

Definition 48.9. The *inverse DFT* of c_k is defined by

$$y_k = \frac{1}{\sqrt{n}} \sum_{j=0}^n \hat{w}^{-jk} c_j.$$

In practise, Fast Fourier Transform (FFT) is used to compute DFT and inverse DFT.

48.10 Laplace Transform

Laplace transform (Definition [32.3](#)^{[def:laplace](#)}) can analyse a larger class of functions/signals belonging to the exponential order class (Definition [32.4](#)^{[def:exp-order](#)}).

Chapter 49

Linear Difference Equations & Z-Transform

Z-Transform is used in linear digital signal processing and linear control.

Refer to octave-signal's zplane.m, etc. to find applications.

octave-signal requires control package LTI functions dssdata (dssdata.m), zpndata (zpndata.m).

Relevant Info:

<https://github.com/hyumo/liti>

<https://github.com/TrymArve/Open-Loop-Optimization-of-LTI-Systems>

<https://python-control.readthedocs.io/en/latest/linear.html>

Continuous **low pass Butterworth filter** has a gain function $B_c(s)$ satisfying

$$|B_c(i\omega)|^2 = \frac{1}{1 + (\frac{i\omega}{i\omega_c})^{2N}}, \quad B_c(s)B_c(-s) = \frac{1}{1 + (\frac{s}{i\omega_c})^{2N}}.$$

Bilinear transform for relating continuous-time filters to discrete-time filters.

$$H_c(s) \mapsto H_d(z) = H_c(s = \frac{2}{T} \frac{1 - z^{-1}}{1 + z^{-1}})$$
$$H_d(z) \mapsto H_c(s) = H_d(z = \frac{1 + (\frac{T}{2})s}{1 - (\frac{T}{2})s})$$

In particular, the imaginary line $\boxed{s = i\omega}$ for the continuous-time filter corresponds Fourier transform and it corresponds to the circle $\boxed{z = e^{i\Omega}}$ for the discrete-time filter, i.e.

$$s = i\omega = \frac{2}{T} \frac{1 - (e^{i\Omega})^{-1}}{1 + (e^{i\Omega})^{-1}} = \frac{2}{T} i \tan \frac{\Omega}{2}$$

which leads to the equality

$$\omega = \frac{2}{T} \tan \frac{\Omega}{2} \Leftrightarrow \Omega = 2 \arctan \frac{\omega T}{2}.$$

Digital Signal Processing References:

* ToC only: <https://dl.icdst.org/pdfs/files/a7887a54f98669046c23e789611d67c0.pdf>

* <https://cobb.ece.wisc.edu/courses/431Notes.pdf>

* https://warwick.ac.uk/fac/sci/eng/people/adam_noel/teaching/es3c5-2021.notes_v1.5.pdf

* <https://www.cl.cam.ac.uk/teaching/2425/L314/materials.html> : Digital Signal Processing

* <http://www.dspguide.com/>

The linear difference equation has a general form below:

$$x(k) + a_1x(k-1) + \cdots + a_nx(k) = b_0f(k) + b_1f(k-1) + \cdots + b_mf(k-m) \quad (49.1) \quad \text{e: difference}$$

It can be derived from the finite difference approximation of linear ODEs, ...

Similar to the application of Laplace transform in solving linear ODE-IVPs, z -transform is developed to solve linear difference initial valued problems.

In engineering, z -transform is used to model a simple feedback control loop of a digital process ...

49.1 Definition and Examples

A *sequence* of numbers $\{x_k\}_{k=-M}^N$ is a mapping of a set of integers $\{-M, \dots, N\}$ to real number set $\mathbb{R}$, where M, N can be finite positive integers or infinity, i.e. $k \mapsto x_k \in \mathbb{R}, k \in \{-M, \dots, N\}$. E.g. The sequence $\{1, 3, 5, 7, 9\}$ is a mapping of $k \mapsto y_k = 2k + 1$ where $k \in \{0, \dots, 4\}$. When M and N are finite, the sequence $\{x_k\}_{k=-M}^N$ is called a *finite sequence* and when M or N is ∞ the sequence $\{x_k\}_{k=-M}^N$ is called an *infinite sequence*.

A (*causal*) *sampling sequence* $\{y_k\}_{k=-\infty}^{\infty}$ is an infinite sequence generated by sampling a causal function $f(t)$ over $\mathbb{R}$ for every time T , i.e. $y_k = f(kT) = f(kT)H(kT)$, where $k = 0, 1, 2, \dots$. It is usually written as $\{y_k\}_{k=0}^{\infty}$ since $y_k = 0$ when $k < 0$.

Example 49.1. The sampling sequence for $f(t) = \sin 2t - e^t$ is $y_k = f(kT) = \sin 2kT - e^{kT}$, $k = 0, 1, 2, \dots$.

Definition 49.1 (z -Transform). Let $\{x_k\}_{k=0}^{\infty}$ be a causal sequence. The z -transform of x_k is

$$Z[x_k] = \sum_{k=0}^{\infty} \frac{x_k}{z^k} = X(z) \quad (49.2) \quad \text{e: ztransform}$$

where z belongs to an open subset of the complex set $\mathbb{C}$.

Example 49.2. Given the following sequences:

1. $\{y_k\} = \{1, 3, 5, 7, 9, 0, 0, \dots\}$;
2. $\{x_k\} = \{1, 2, 4, 8, 16, \dots\} = \{2^k\}_{k=0}^{\infty}$;
3. $x_k = a^k$;
4. $x_k = \left(-\frac{1}{4}\right)^k$;
5. $x_k = ka^{k-1}$;
6. $x_k = k$;
7. impulse sequence, $\{\delta_{0,k}\} = \{1, 0, 0, 0, \dots\}$;
8. unit step sequence, $\{h_k\} = \{1, 1, 1, \dots\}$;
9. $x_k = 3k$;
10. sampling sequence for $f(t) = e^{-t}$;

the z -transform can be obtained by using the definition.

$$1. Y(z) = Z[y_k] = 1 + \frac{3}{z} + \frac{5}{z^2} + \frac{7}{z^3} + \frac{9}{z^4}, |z| > 0;$$

$$2. X(z) = Z[2^k] = \sum_{k=0}^{\infty} \frac{2^k}{z^k} = \frac{1}{1 - \frac{2}{z}} = \frac{z}{z-2}, |z| > 2;$$

$$3. X(z) = Z[a^k] = \sum_{k=0}^{\infty} \frac{a^k}{z^k} = \frac{1}{1 - \frac{a}{z}} = \frac{z}{z-a}, |z| > |a|;$$

$$4. \text{ From the above, let } a = -\frac{1}{4}, \text{ we have } X(z) = Z[-\frac{1}{4}] = \frac{z}{z + \frac{1}{4}}, |z| > \frac{1}{4};$$

$$\begin{aligned} 5. X(z) = Z[ka^{k-1}] &= \sum_{k=0}^{\infty} \frac{ka^{k-1}}{z^k} = \frac{1}{z} \sum_{k=0}^{\infty} k \left(\frac{a}{z}\right)^{k-1} = \frac{1}{z} \left[\frac{d}{dy} \sum_{k=0}^{\infty} y^k \right]_{y \rightarrow \frac{a}{z}} \\ &= \frac{1}{z} \left[\frac{d}{dy} \frac{1}{1-y} \right]_{y \rightarrow \frac{a}{z}} = \frac{1}{z} \left[\frac{1}{(1-y)^2} \right]_{y \rightarrow \frac{a}{z}} = \frac{z}{(z-a)^2}, |z| > |a|; \end{aligned}$$

$$6. \text{ Let } a = 1, \text{ from the above we have } X(z) = Z[k] = \frac{z}{(z-1)^2}, |z| > 1;$$

$$7. \text{ From definition, } X(z) = 1, z \in \mathbb{C};$$

$$8. X(z) = \sum_{k=0}^{\infty} \frac{1}{z^k} = \frac{z}{z-1}, |z| > 1;$$

$$9. Z[3k] = \sum_{k=0}^{\infty} \frac{3k}{z^k} = 3 \sum_{k=0}^{\infty} \frac{k}{z^k} = \frac{3z}{(z-1)^2}, |z| > 1;$$

$$10. Z[e^{-kT}] = Z[(e^{-T})^k] = \frac{z}{z - e^{-T}}, |z| > e^{-T}.$$

49.2 Properties of Z-Transform

Finding z -transform using definition (49.2) is time-consuming. In order to find z -transform faster, its properties are investigated and a table of z -transform is build.

Theorem 49.3 (Linearity). *If α and β are constants, then*

$$Z[\alpha x_k + \beta y_k] = \alpha Z[x_k] + \beta Z[y_k].$$

Example 49.4. Using the linearity, we can find $Z[\cos k\omega T]$ and $Z[\sin k\omega T]$ as follows.

$$\begin{aligned} Z[\cos k\omega T] &= Z\left[\frac{1}{2}(e^{ik\omega T} + e^{-ik\omega T})\right] = \frac{1}{2}Z[e^{ik\omega T}] + \frac{1}{2}Z[e^{-ik\omega T}] \\ &= \frac{1}{2}Z[(e^{i\omega T})^k] + \frac{1}{2}Z[(e^{-i\omega T})^k] = \frac{1}{2} \frac{z}{z - e^{i\omega T}} + \frac{1}{2} \frac{z}{z - e^{-i\omega T}} \\ &= \frac{z}{2} \frac{z - e^{-i\omega T} + z - e^{i\omega T}}{z^2 - (e^{i\omega T} + e^{-i\omega T})z + 1} = \frac{z(z - \cos \omega T)}{z^2 - 2(\cos \omega T)z + 1} \end{aligned}$$

$$\begin{aligned} Z[\sin k\omega T] &= Z\left[\frac{1}{2i}(e^{ik\omega T} - e^{-ik\omega T})\right] = \frac{1}{2i}Z[e^{ik\omega T}] - \frac{1}{2i}Z[e^{-ik\omega T}] \\ &= \frac{1}{2i}Z[(e^{i\omega T})^k] - \frac{1}{2i}Z[(e^{-i\omega T})^k] = \frac{1}{2i} \frac{z}{z - e^{i\omega T}} - \frac{1}{2i} \frac{z}{z - e^{-i\omega T}} \\ &= \frac{z}{2i} \frac{z - e^{-i\omega T} - z + e^{i\omega T}}{z^2 - (e^{i\omega T} + e^{-i\omega T})z + 1} = \frac{z \sin \omega T}{z^2 - 2(\cos \omega T)z + 1} \end{aligned}$$

Theorem 49.5 (Time Delay Property). *If x_k is a causal sequence then*

$$Z[x_{k-k_0}] = \frac{1}{z^{k_0}} Z[x_k].$$

Example 49.6. We use this property to determine $Z[(1/2)^{k-2}]$.

$$Z[(\frac{1}{2})^{k-2}] = \frac{1}{z^2} Z[(\frac{1}{2})^k] = \frac{1}{z^2} \frac{z}{z - \frac{1}{2}}.$$

Theorem 49.7 (Time Advance Property). *If x_k is a causal sequence and $Z[x_k] = X(z)$ then*

$$Z[x_{k+k_0}] = z^{k_0} X(z) - z^{k_0} x_0 - z^{k_0-1} x_1 - \cdots - z x_{k_0-1}$$

In particular,

$$Z[x_{k+1}] = zX(z) - zx_0,$$

$$Z[x_{k+2}] = z^2 X(z) - z^2 x_0 - zx_1.$$

The time advance property are useful in solving difference equation using z -transform as demonstrated in Section [49.4](#).

Theorem 49.8 (Scaling Property). *If x_k is a causal sequence and $Z[x_k] = X(z)$ then*

$$Z[a^k x_k] = X\left(\frac{z}{a}\right).$$

Theorem 49.9 (Differentiation).

$$Z[k^n x_k] = \left(-z \frac{d}{dz}\right)^n Z[x_k].$$

Theorem 49.10 (Initial Value Theorem).

$$\lim_{z \rightarrow \infty} X(z) = x_0.$$

Theorem 49.11 (Final Value Theorem).

$$\lim_{k \rightarrow \infty} x_k = \lim_{z \rightarrow 1} \left(1 - \frac{1}{z}\right) X(z)$$

49.3 Inverse z -Transform

Strategies for finding inverse z -Transform involves a combination of the following techniques:

1. Utilise the properties of z -transform;
2. Using the z -Transform Table;
3. Algebraic manipulations, e.g. partial fraction, to a recognisable form.

Example 49.12. The above techniques are used to find the inverse z -transform for the following $X(z)$:

$$1. X(z) = \frac{z}{z+2} \text{ (Table of Transform)}$$

From the table $Z[a^k] = \frac{z}{z-a}$, $x_k = (-2)^k$, $k = 0, 1, 2, \dots$.

$$2. X(z) = \frac{z}{(z-1)(z-2)} \text{ (Partial Fraction)}$$

$X(z)$ is a rational function, ...

$$X(z) = z\left(\frac{A}{z-1} + \frac{B}{z-2}\right)$$

$$\frac{A}{z-1} + \frac{B}{z-2} = \frac{1}{(z-1)(z-2)} \Rightarrow A = -1, B = 1.$$

$$X(z) = \frac{-z}{z-1} + \frac{z}{z-2}.$$

Hence $x_k = -1 + 2^k = 2^k - 1$, $k = 0, 1, 2, \dots$.

$$3. X(z) = \frac{2z+1}{(z+1)(z-3)} \text{ (Delay Property)}$$

$$\frac{X(z)}{z} = \frac{2z+1}{z(z+1)(z-3)} = -\frac{1}{4(z+1)} - \frac{1}{3z} + \frac{7}{12(z-3)}$$

Hence $x_k = -\frac{1}{3}\delta_{k,0} - \frac{1}{4}(-1)^k + \frac{7}{12}(3^k)$, $k = 0, 1, 2, \dots$.

$$4. X(z) = \frac{z}{z^2+a^2} \text{ (Partial Fraction)}$$

$$\frac{X(z)}{z} = \frac{1}{2ai}\left(\frac{1}{z-ai} - \frac{1}{z+ai}\right)$$

Hence $x_k = \frac{1}{2ai}((ai)^k - (-ai)^k)$, $k = 0, 1, 2, \dots$.

$$5. X(z) = \frac{z^3+2z^2+1}{z^3} \text{ (Partial Fraction)}$$

$$X(z) = 1 + \frac{2}{z} + \frac{1}{z^3}$$

Hence $\{x_k\}_{k=0}^{\infty} = \{1, 2, 0, 1, 0, 0, \dots\}$.

$$6. X(z) = \frac{z(1-e^{-aT})}{(z-1)(z-e^{-aT})} \text{ (Partial Fraction)}$$

$$X(z) = z\left(\frac{1}{z-1} + \frac{1}{z-e^{-aT}}\right)$$

Hence $x_k = 1 + (e^{-aT})^k = 1 + e^{-kaT}$, $k = 0, 1, 2, \dots$.

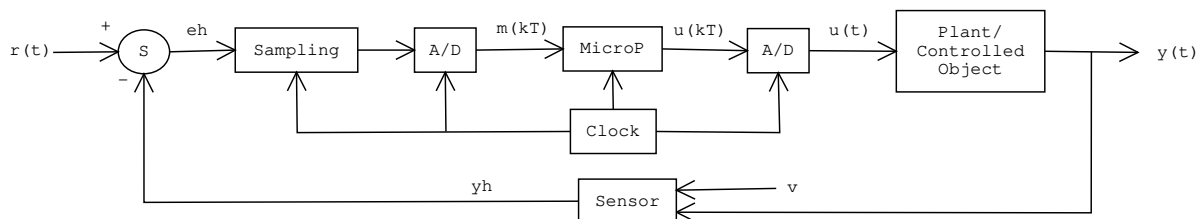
49.4 Applications in Solving Difference Equation

difference-eq

Discrete-Time System

In engineering applications, system with discrete signals can be categorised into *discrete-time system* (discrete only in the time, the amplitude can take any value), *data sampling system* (discrete in time and amplitude) and *digital system* (e.g. a microprocessor).

The following diagram (will be referred to as the FIGURE) shows a general diagram of a system with discrete signals:



where $r(t)$ is the reference input, $u(t)$ is the control signal, $y(t)$ is the output, $\hat{y}(t)$ is the observed signal, $\hat{e}(t)$ is the error of the observed signal and the reference input, $v(t)$ is the noise or interference in the sensor.

Design of Discrete-Time Systems

Once the controlled object (see FIGURE) and microcomputer have been setup, we will have a discrete-time system. In order to make the system *robust*, it is desirable to design a feedback loop. There are two ways to carry out the design:

- transform (frequency domain) method: Laplace, z transforms, *Delta Operator* and the *D-Transform*, etc.
- State-space method: Mentioned above.

Difference Equation

In general, discrete system can be transformed into difference equations. A general n -th order linear difference equation has the following form:

$$a_n y_{k+n} + a_{n-1} y_{k+n-1} + \cdots + a_1 y_{k+1} + a_0 y_k = u_k.$$

In particular, second order difference equations with constant coefficients can be roughly categorised into

- Homogeneous: $ay_{k+2} + by_{k+1} + cy_k = 0$
- Inhomogeneous: $ay_{k+2} + by_{k+1} + cy_k = u_k$

The solution of the difference equation is an expression of y_k that satisfies the difference equation with certain initial conditions.

Example 49.13. The following difference equations can be solved by applying z -transform:

1. $y_{k+2} + y_{k+1} - 2y_k = 1, y_0 = 0, y_1 = 1.$

$$z^2 Y(z) - z^2 y_0 - z y_1 + z Y(z) - z y_0 - 2Y(z) = \frac{z}{z-1}$$

$$\Rightarrow z^2 Y(z) - z + z Y(z) - 2Y(z) = \frac{z}{z-1}$$

$$\Rightarrow Y(z) = \frac{z^2}{z-1} \frac{1}{z^2 + z - 2}$$

$$\Rightarrow \frac{Y(z)}{z} = \frac{z}{(z-1)(z+2)(z-1)} = -\frac{2}{9} \frac{1}{z+2} + \frac{2}{9} \frac{1}{z-1} + \frac{1}{3(z-1)^2}$$

$$\text{Hence } y_k = -\frac{2}{9}(-2)^k + \frac{2}{9} + \frac{k}{3}.$$

2. $y_{k+2} + 2y_{k+1} = 0, y_0 = 1, y_1 = \sqrt{2}.$

$$z^2 Y - z^2 - \sqrt{2}z + 2(zY - z) = 0$$

$$\Rightarrow Y(z) = \frac{z^2 + \sqrt{2}z + 2z}{z^2 + 2z}$$

$$\Rightarrow \frac{Y(z)}{z} = \frac{z + \sqrt{2} + 2}{z(z+2)} = \frac{2 + \sqrt{2}}{2z} - \frac{\sqrt{2}}{2(z+2)}$$

$$\text{Hence } y_k = 1 + \frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}(-2)^k.$$

Relationship between Laplace and z Transforms

From the FIGURE, a signal $\hat{e}(t)$ coming out of the sampling can be described by

$$e(t) = \sum_{k=0}^{\infty} \hat{e}(kT) \delta(t - kT), \quad T \text{ is the period of the clock}$$

The Laplace transform gives

$$\mathcal{L}\{e(t)\} = \int_{-\infty}^{\infty} e(t)e^{-st} dt = \sum_{k=0}^{\infty} \int_{-\infty}^{\infty} \delta(t - kT) \hat{e}(t) e^{-st} dt = \sum_{k=0}^{\infty} \hat{e}(kT) e^{-skT}.$$

By letting $z = e^{sT}$, we have

$$\mathcal{L}\{e(t)\} = \sum_{k=0}^{\infty} \frac{\hat{e}(kT)}{z^k} = Z[\hat{e}(kT)].$$

So, the Laplace transform of sampling signal is related to the z -transform of the signal. Note that $z = e^{sT}$ can be read as a mapping between the z - and s - complex planes.

We can construct a table to compare Laplace transform and z -transform. Ref: Gene F. Franklin, J. David Powell, *Digital Control of Dynamical Systems*, Addison-Wesley 197?

49.5 Transfer Function and State-Space Model

Definition 49.2. The *transfer function* of the linear discrete system is the *ratio* of the z -transform of the *output* to the z -transform of the *input*, under a quiescent initial condition.

Example 49.14. The transfer function $\frac{Y}{U}$ for each of the following ODEs:

1. $y_{k+2} + 3y_{k+1} - y_k = u_k$
2. $y_{k+3} + 3y_{k+2} + 2y_k = u_{k+1} - u_k$

can be calculated as follows:

1. $z^2Y + 3zY - Y = U \Rightarrow \frac{Y}{U} = \frac{1}{z^2 + 3z - 1}$
2. $z^3Y + 3z^2Y + 2Y = zU - U \Rightarrow \frac{Y}{U} = \frac{z - 1}{z^3 + 3z^2 + 2}$

In engineering, state-space model is often used to solve discrete-time systems. In the following discrete-time systems will be solved by rewriting the system into a state-space model.

Stability

The stability of a transfer function is determined by the modulus of the poles:

- When all the poles have modulus less than 1, the transfer function is stable;
- When at least one pole has a modulus 1 and other poles have modulus less than or equal to 1, the transfer function is marginally stable;
- When at least one pole has a modulus greater than 1, the transfer function is unstable.

Eigenvalue Problem

Definition 49.3 (Eigenvalue Problem). A real or complex number λ is an eigenvalue of a matrix A , if there is a nonzero $n \times 1$ matrix E such that

$$AE = \lambda E$$

The nonzero vector E is called an *eigenvector*.

State-Space Model and Transition Matrix

Consider the n -th order discrete-time system given by the difference equation

$$y_{k+n} + a_{n-1}y_{k+n-1} + \cdots + a_1y_{k+1} + a_0y_k = b_0u_k. \quad (*)$$

By introducing a new set of state variables

$$\begin{cases} x_0(k) = y_k \\ x_1(k) = y_{k+1} \\ \cdots \\ x_{n-1}(k) = y_{k+n-1} \end{cases}$$

We could rewrite the above difference equation (*) into a matrix equation

$$\mathbf{x}(k+1) = \begin{bmatrix} x_0(k+1) \\ x_1(k+1) \\ \vdots \\ x_{n-2}(k+1) \\ x_{n-1}(k+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & -a_3 & \cdots & -a_{n-1} \end{bmatrix} \begin{bmatrix} x_0(k) \\ x_1(k) \\ \vdots \\ x_{n-2}(k) \\ x_{n-1}(k) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ b_0 \end{bmatrix} u(k)$$

or using matrix notation

$$\mathbf{x}(k+1) = A\mathbf{x}(k) + Bu(k).$$

The solution to the n -th order difference equation is then

$$y_k = x_0(k) = [1, 0, \cdots, 0]\mathbf{x}(k) = \mathbf{c}^T \mathbf{x}(k).$$

Example 49.15. Given the following difference equation:

$$y_{k+2} + 0.2y_{k+1} + 0.3y_k = u_k.$$

Using the notation $x_0(k) = y_k$, $x_1(k) = y_{k+1}$ and $u(k) = u_k$. The state-space representation is

$$\begin{bmatrix} x_0(k+1) \\ x_1(k+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -0.3 & -0.2 \end{bmatrix} \begin{bmatrix} x_0(k) \\ x_1(k) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(k)$$

If the difference equation that we are considering is homogeneous, $u_k = 0$, then the matrix equation is reduce to

$$\mathbf{x}(k+1) = A\mathbf{x}(k).$$

Taking $k = 0$, we have

$$\begin{cases} \mathbf{x}(1) = A\mathbf{x}(0) = A^1\mathbf{x}(0) \\ \mathbf{x}(2) = A\mathbf{x}(1) = A^2\mathbf{x}(0) \\ \cdots \\ \mathbf{x}(k) = A\mathbf{x}(k-1) = A^k\mathbf{x}(0) \end{cases}$$

where $\mathbf{x}(0)$ is the initial condition. The matrix $\Phi(k) := A^k$ is called the *transition matrix*.

Since A is a constant matrix, we could calculate the transition matrix by

$$A^k = \alpha_0(k)\mathbf{I} + \alpha_1(k)A + \alpha_2(k)A^2 + \cdots + \alpha_{n-1}(k)A^{n-1}$$

where the values of the multipliers α_i are obtained by solving corresponding simultaneous equations of eigenvalues:

$$\lambda_j^k = \alpha_0(k) + \alpha_1(k)\lambda_j + \alpha_2(k)\lambda_j^2 + \cdots + \alpha_{n-1}(k)\lambda_j^{n-1}$$

After the transition matrix is obtained, then we have the solution to the homogeneous problem

$$y_k = \mathbf{c}^T \mathbf{x}(k)$$

where $\mathbf{c}^T = [1, 0, \cdots, 0]$ and $\mathbf{x}(k) = A^k\mathbf{x}(0)$.

Example 49.16. Solve the state-space model given by $\mathbf{x}(k+1) = A\mathbf{x}(k)$ where $A = \begin{bmatrix} \frac{1}{2} & 0 \\ -1 & \frac{1}{3} \end{bmatrix}$ and subject to initial condition $\mathbf{x}(0) = [1, 1]^T$.

First calculate the eigenvalues of the matrix A :

$$\begin{vmatrix} \frac{1}{2} - \lambda & 0 \\ -1 & \frac{1}{3} - \lambda \end{vmatrix} = 0 \Rightarrow \lambda_1 = \frac{1}{3}, \lambda_2 = \frac{1}{2}$$

Using the following equation we can determine the general form of A^k :

$$\begin{bmatrix} \lambda_1^k \\ \lambda_2^k \end{bmatrix} = \begin{bmatrix} 1 & \lambda_1 \\ 1 & \lambda_2 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} \Rightarrow \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} = \frac{1}{\lambda_2 - \lambda_1} \begin{bmatrix} \lambda_2 & -\lambda_1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} \lambda_1^k \\ \lambda_2^k \end{bmatrix} = \begin{bmatrix} \frac{\lambda_1^k \lambda_2 - \lambda_1 \lambda_2^k}{\lambda_2 - \lambda_1} \\ \frac{\lambda_2^k - \lambda_1^k}{\lambda_2 - \lambda_1} \end{bmatrix}$$

Hence,

$$\mathbf{x}(k) = A^k \mathbf{x}(0) = 6 \left\{ \left(\frac{1}{2 \cdot 3^k} - \frac{1}{3 \cdot 2^k} \right) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \left(\frac{1}{2^k} - \frac{1}{3^k} \right) \begin{bmatrix} \frac{1}{2} & 0 \\ -1 & \frac{1}{3} \end{bmatrix} \right\} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2^{-k} \\ 7 \cdot 3^{-k} - 6 \cdot 2^{-k} \end{bmatrix}$$

Chapter 50

Linear Difference System & State Space Method

$$\begin{aligned}\mathbf{x}(k+1) &= F\mathbf{x}(k) + G\mathbf{u}(k) \\ \mathbf{y}(k) &= C\mathbf{x}(k) + D\mathbf{u}(k)\end{aligned}\tag{50.1}$$

Observability Controllability

Part VIII

Probability and Statistics

Chapter 51

Probability Theory and Statistical Modelling

ch:prob+stat

I never had the chance to teach probability and statistics, so I will need to read the following references to prepare this topic.

Lebesgue Measure and Probability <https://www.math.purdue.edu/~banuelos/probability.pdf>

D. C. Montgomery & G. C. Runger. (2003), Applied Statistics & Probability for Engineers, 3rd Ed., John Wiley.

Bain, L.J. and Engelhardt, M., 2000. Introduction to Probability and Mathematical Statistics. 2nd ed. Boston, Massachusetts: Duxbury.

Hogg, R.V. and Tanis, E.A., 2010. Probability and Statistical Inference. 8th ed. Prentice Hall.

Wackerly, D.D., Mendenhall III, W. and Scheaffer, R.L., 2008. Mathematical Statistics with Applications. 7th ed. California: Duxbury/Thomson Learning.

Hogg, R.V., Craig, A. and McKean J.W., 2012. Introduction to Mathematical Statistics. 7th ed. Prentice Hall.

Milton, J.S., Liu, K.I. and Arnold, J.C., 2013. Probability and statistics. Boston: McGraw-Hill.

Kirkpatrick, L.A. and Feeney, B.C., 2013. A simple guide to IBM SPSS statistics for version 20.0. Belmont, CA: Thomson/Wadsworth.

Mann, P.S., 2012. Introductory statistics. 8th ed. New Jersey: John Wiley & Sons.

Data Wrangling and Analysis Examples??? https://github.com/zby0902/Monash_2018_S1

Identify problems and assumptions Formulate and problem solving Interpretation and verification of solution

Model fitting

Polynomial curve fitting Least square curve fitting Visualise model fitting and data Source of errors in model building Choosing best model

51.1 Population Space, Random Variables and Data

Population, sample, types of variables, sources of data.

- Sampling methods: simple random, systematic, stratified, cluster and quota sampling
- Presentation of data - Bar graphs, pie charts, dot plot, stem-and-leaf plot, box plot display.
- Frequency distribution, histogram, frequency polygon, ogives.
- Measures of location – mean, median and mode.
- Measures of dispersion – range, variance and standard deviation, interquartile range, percentile range, mean absolute deviation.
- Measure of skewness, coefficient of variation.
- Introduction to statistical software – plot graphs, perform exploratory data analysis (EDA).

51.2 General Probability Theory

- Properties of probability.
 - Methods of enumeration.
 - Sample spaces and events: Mutually Exclusive Events, Independence Events.
 - Rules of probability: The General Addition Rule, Multiplicative Rule. Marginal and Conditional Probability.
 - Bayes' Theorem.
 - Sigma algebra, Borel sets and Measure space
 - Finite and Infinite probability spaces
 - Lebesgue integral, comparison of Riemann and Lebesgue integrals
 - Random variables and distributions
 - Integrability of random variables
 - Fatou Lemma, Monotonic and Dominated Convergence Theorems
 - Computation of expectations
 - Law of Large Numbers
 - Central Limit Theorem
 - Equivalence of measures, Radon-Nikodym Theorem

51.3 Discrete Distributions

- Random variables, discrete probability densities, cumulative distribution.
 - Expected value of a function of a random variable.
 - Moments and moment generation functions.
 - Binomial, Poisson, Geometric, Negative Binomial, and Hypergeometric distribution.
 - Poisson approximation to Binomial distribution.
 - Probability generating functions.
 - Chebyshev's Theorem.

51.4 Continuous Distributions

- Continuous densities.
 - Cumulative distribution.
 - Expected value of a function of a random variable.
 - Mixed probability distributions.
 - Moment generating function.
 - Uniform, Exponential, Gamma, Chi-Square, Beta, Pareto, Normal and Lognormal distributions.
 - Normal probability rule, Chebyshev's inequality.
 - Normal approximation to Binomial and Poisson distribution.
 - Use statistical software to evaluate probabilities for discrete and continuous distributions.

51.5 Joint Distributions

- Joint distributions for discrete and continuous variables.
 - Marginal distributions.
 - The Expected Value and Variance of Linear Functions of Random Variables.
 - Covariance and correlation.
 - Conditional densities. Conditional expectation.
 - Joint Moment Generating Functions.
 - Multinomial probability distribution.
 - Bivariate normal distribution.

51.6 Distributions Functions of Random Variables

- Finding probability distribution of function of random variables by methods of distribution functions, transformations, moment-generating functions and bivariate transformation using Jacobians.
- Order statistics.
- Normal approximation for sum of random variables.

51.7 Inferences on the Mean of a Distribution

- Central limit theorem.
- Confidence interval for the mean (known and unknown variance).
- Element of a hypothesis test, Type I and Type II errors, power of the test, p-value.
- Hypothesis tests on the mean.
- Draw inferences by using statistical software.

51.8 Inference on Proportions

- Sample size.
- Confidence interval for proportion.
- Hypothesis tests on proportion.

51.9 Comparing Two Means and Two Proportions

- Inferences about the difference between two population means for large and independence samples.
- Inferences about the difference between two population means for small and independence samples (equal variance, unequal variance, and paired data): confidence interval and hypothesis tests.
- Inferences about the difference between two population proportions for large and independence samples.
- Making comparisons via statistical software.

51.10 The Analysis of Variance

- The analysis of variance procedure.
- Comparison of more than two means: Analysis of variance for a one-way layout.
- Analysis of variance table for a one-way layout.
- Statistical model for the one-way layout.
- Use statistical software to perform ANOVA.

Hypothesis testing

<http://bayesianbiologist.wordpress.com/2011/08/21/p-value-fallacy-on-more-or-less/>
P-value fallacy on More or Less

What we get from a p-value is the probability of observing something as extreme or more than our data, under the null hypothesis (written as $P(x \geq D|H_0)$). Isn't that awkward? No wonder it is so commonly misrepresented.

So, my word of caution is this: We have to remember that the p-value is only a statement of the likelihood of making an observation as extreme or more than your observation, if there is, in fact, no real effect present. We must be careful not to perform the tempting, but erroneous, logical inversion of using it to represent the probability that a hypothesis is, or is not true.

51.11 Chi-Square Tests

- Chi-squared goodness of fit test.
 - Tests for independence.
 - Tests for homogeneity.
- Use statistical software to perform chi-square tests.

51.12 Probability Models

- Probability and distributions
 - Moments
 - Moment Inequalities

51.13 Convergence

- Convergence in Distribution
 - Convergence in Probability
 - Central Limit Theorem

51.14 Statistical Models and Inferences

- Statistical models
 - Information and Sufficiency
 - Point estimation
 - Methods of Point Estimation

51.15 Criteria for Evaluating Point Estimators

- Mean Squared Error
 - Unbiasedness and Rao-Blackwellization Process
 - Cramer-Rao Lower Bound
 - Minimum Variance Unbiased Estimator (UMVUE)

51.16 Tests of Hypotheses

- Simple and Composite hypotheses
 - Type I and II error, power
 - Most Powerful Tests: Neymann-Pearson Lemma
 - Uniformly Most Powerful Tests
 - Generalized Likelihood Ratio Tests

51.17 Interval Estimation

- Sample statistic and its distribution
 - Standard error
 - Confidence Interval
 - Pivotal Quantity Method
 - Approximate Confidence Interval
 - General Method
 - Bayesian Interval Estimate

51.18 Information and Conditioning

- Information and sigma algebra generated by random variables
 - Filtration and adapted processes
 - Independence of sigma algebras and random variables
 - Joint distribution of several random variables
 - Normal and multivariate normal variables
 - Expectation, variance and covariance of random variables

51.19 General Conditional Expectation

- Conditional expectation of random variable
 - Properties of conditional expectation: measurability, partial averaging, linearity, towering, independence
 - Martingale and Markov processes

51.20 Single period multi-state model

- Pricing in one period binomial model and multistate model
 - Existence of state price vector
 - Characterization of no arbitrage for multi-state model
 - The risk-neutral measures

Chapter 52

Statistical Decisions

TO WORK ON. References

DeGroot, M.H., 2004. Optimal statistical decisions. New York: John Wiley & Sons.

Lindgren, B.W., 1971. Elements of decision theory. Macmillan Company.

Lindgren, B.W., 1976. Statistical theory. 4th ed. Boca Raton: Chapman & Hall/CRC.

DeGroot, M.H. and Schervish, M.J., 2014. Probability and statistics. 4th ed. Harlow: Pearson Education Limited.

Recommended by Dr Wong WY: David Blackwell-Girshick MA-Theory Of Games And Statistical Decisions-Wiley Publication In Statistics-Wiley 1954

Thomas S. Ferguson, 1967. Mathematical statistics: a decision theoretic approach. New York: Academic Press.

James O Berger-Statistical Decision Theory And Bayesian Analysis 2e-Springer 1985

Question: Relation to Online learning and reinforcement learning???

52.1 Chapter 1 Probability — Preparations

The aim of probability is to provide tools to measure uncertainty and to discover the laws that govern such a measurement. Statistician uses probability to model this variability and to quantify the resulting uncertainty. The discipline of probability also provides the bridge between descriptive and inferential statistics. So Dr Wong Wing Yue and Dr Leong Yoon Kwai introduce some basic concepts of probability as preparation for the study of inferential statistics for statistical decisions.

52.1.1 1.1 Betting Problems

Example 52.1 (Three Cards Problem). There are three cards.

- One red on both sides
- One black on both sides
- One red on one side and black on the other side

Picks one at random and places it flat on the table with only one side showing.

Example 52.2 (Monty Hall Problem). Imagine that you are so lucky to be chosen to participating in the following TV game:

The host shows you three closed doors, one of which hides the car of your dreams, shiny and desirable. Behind each of the other two doors, however, is standing a smelly goat. You will choose a door and win whatever is behind it. You decide on door and announce your choice. Whereupon the host opens one of the other two doors and reveals a goat. He then asks you if you would like to switch your choice to the unopened door that you did not at first choose. Is it to your advantage to switch?

Example 52.3 (Flipping a fair coin). Suppose a person suggests that you flip a fair coin 1000 times. He says that if the coin comes up heads at least 600 times, then he will pay you \$100, otherwise you have to pay him \$1.

- You might think that, while 500 heads is most likely, there is still a reasonable chance that 600 heads or more will appear, at least good enough to justify accepting that person's \$100 to \$1 bet. Agree?

All these example show:

- A good understanding of probability theory will allow you to correctly assess probabilities in everyday situations.
- It helps you in making good decision. It might even help you in making big money!

52.1.2 1.2 Probability models

A formal definition of probability begins with a random experiment.

Definition 52.1. Experiment: An experiment is an action or process that generates an outcome or an observation that is not predetermined.

Example 52.4. Example 1.2.1 (1) Toss a coin once. (Heads or Tails ?) (2) Roll a dice. (How many dots are there on the upper face ?) (3) Measure the lifetime of a car. (How long can it last ?)

Mathematical speaking, we always convert an experiment into a set S — a set in which all possible outcomes of the experiment are represented.

Definition 52.2. Sample space: The set which contains all possible outcomes of an experiment. Each outcome of an experiment is known as a sample point.

Remark 52.5. Remark The mathematical representation of an experiment is the sample S in which there corresponds one element to each of outcome of an experiment. It is intentional that we do not try to be more specific about S , and we thus permit several ways of choosing it. One reason for this ambiguity is mathematical convenience, but another, more important reason, relates to practical considerations.

Example 52.6. Example 1.2.2 Give a sample space that represents the experiment of recording the age of a randomly chosen Malaysian. Should we have tried to define S accurately, we would face several difficulties in solving this simple problem.

- If we count age in years, S should contain nonnegative integers.
- We ought to know the oldest age.
- We ought to make sure that we properly recorded all gaps in ages.
- S would change on the birthday as well as the death of some persons.

This means that a statistician planning to perform this experiment should be

- familiar with the entire population, including some personal data

in which case the experiment itself is of little interest. This shows the practical value of our requirement in S : Its elements should represent all possible outcomes of the experiment, but no further restriction is made on its members. Hence we can choose S as the infinite set $S = \{0, 1, 2, 3, \dots\}$

Example 52.7. Example 1.2.3 (1)

The sample space of the experiment of tossing a coin once is $S = H, T$ where $H = \text{Heads}$, $T = \text{Tails}$.

(2)

Roll a dice one, the sample space can be expressed as $S = 1, 2, 3, 4, 5, 6$.

(3)

Measure the lifetime of a car. The sample space is $S = [0, \infty)$

In general, a question posed in connection with an experiment is called an event. An event is a subset of the sample space S .

But that does not mean that all subsets of a sample space are events!

- An event is always related to an experiment has the mathematical consequence.
- An event E is said to be occurred if the outcome of the experiment is in E .
- We say that both events E and F occur if the outcome of the experiment is in $E \cap F$.
- We say either event E or F occurs if the outcome of the experiment is in $E \cup F$.
- We say that event E does not occur if E^c occurs.
- The set consists of all events of an experiment is called an event space.

Remark 52.8. Remark The choice of event space depends on the purpose performing the experiment.

Example 52.9. Example 1.2.4 Suppose we roll a die and want to know whether the outcome is odd or even. Then the events that we are interested in are $E = \{1, 3, 5\}$ and $F = \{2, 4, 6\}$

Remark 52.10. Remark

Intuitively we shall define probability in such a way that event that occurred more often will carry a higher probability.

Question: How to define probability of an event?

There are four very different ways of defining what is to be meant by the probability of an event.

Classical Probability (Equally Likely)

Example 52.11. Example 1.2.5 Toss a fair coin three times. What is the probability of getting at least 2 heads ?

- Sample Space : $S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$
- All outcomes are equally likely occurred. The event E of getting at least two heads is $1 / 2$

Example 52.12. Example 1.2.6 Suppose that a fair coin is tossed repeatedly until a Heads is observed. What is the chance that Heads is obtained at the 4th trial ?

$$P(\text{Obtaining the first Heads at the fourth trial}) = \frac{1}{2^4}$$

Example 52.13. Example 1.2.7 Suppose there are n students taking the course UECM2233 in Jan 2012 semester. What is the probability that there will be at least two students having the same birthday? We shall make the assumption that there are 365 days in a year. Let E_n denote the event that no two students has the same birthday. Then

$$P(E_n) = \frac{365 \times 364 \times \cdots \times (365 - (n - 1))}{365^n}, \quad n = 1, 2, \cdots, 365.$$

The following table gives the value of $P(E_n)$ and $P(E_n^c)$ for various values of n .

n	$P(E_n)$	$P(E_n^c)$
10	0.871	0.129
20	0.589	0.411
21	0.556	0.444
22	0.524	0.476
<u>23</u>	<u>0.493</u>	<u>0.507</u>
24	0.462	0.538
25	0.431	0.569
30	0.294	0.706
50	0.030	0.970
60	0.006	0.994
<u>68</u>	<u>0.001</u>	<u>0.999</u>

Remark 52.14. Remark

- There are many situations where the classical definition of probability is entirely appropriated.
- However, it has obvious limitation.
- What if the outcomes of the experiment are not equally likely? Or the number of outcomes is not finite?

Empirical Probability

The questions imposed above provide an impetus for the formation of a more general and more experimentally oriented definition of probability.

- Consider a sample space, S , and any event E , defined on S .
- If our experiment were performed one time, either E or E^c would be the outcome.
- If it were performed n times resulting E to occur m times and the ratio $\frac{m}{n}$ converges as n get large.
- The probability of E will be defined as

$$P(E) = \lim_{n \rightarrow \infty} \frac{m}{n}$$

This is known as the **empirical**, or a **posteriori** definition of probability. It is often credited to the German mathematician Richard von Mises.

Remark 52.15. Remark

- This definition is not 100
- There is a bit of inconsistency in extolling $\lim_{n \rightarrow \infty} \frac{m}{n}$ as a way of defining probability experimentally when the very act of repeating an experiment under identical conditions an infinite number of times is physically impossible.
- An unanswered question : How large n must be to give a good approximation for $\lim_{n \rightarrow \infty} \frac{m}{n}$?

Example 52.16. Example 1.2.8 Toss a coin in the air.

- Will it lands heads or tails?
- Sometimes it lands heads and sometimes tails.
- We cannot say what the next outcome will be.

- But if we toss a coin many times, a pattern emerges.
- Some diligent people have in fact made thousands of tosses:

(a)

The French naturalist Buffon (1707-1788) tossed a coin

4040 times. Result : 2048 heads 2048 Relative frequency : = 0.5069 4040 (b)

Around 1900, the English statistician Karl Pearson tossed a coin 24,000 times. Result : 12,012 heads 12012 Relative frequency : = 0.5005 24000

(c)

While imprisoned by the Germans during World War II, the English mathematician John Kerrich tossed a coin 10,000 times. Result : 5067 heads 5067 Relative frequency : = 0.5067 10000

The long-run relative frequency of heads tends to a number close to 0.50.

- If I were to toss such a coin once, I would say that the chance of getting a head is 0.50. Intuitively, we consider probability as a long-term relative frequency.

Subjective Probability

Subjective probability describes an individual's personal judgment about how likely a particular event is to occur.

Suppose we ask "What is the probability that nuclear war will break out in Asia sometime in the next 10 years?"

What is the chance that earthquake will hit Tokyo this year?

What is the chance that Malaysia Badminton team will bring back the Sudiman Cup from China? Like other probabilities, a subjective probability is conventionally expressed on a scale from 0 to 1; a rare event has a subjective probability close to 0, a very common event has a subjective probability close to 1. Example 1.2.8

In horse-racing, which horse will be the champ?

In Magnum Empty-Ekor betting, what 4-digit will most likely be drawn in the next draw?

What type of question will most likely be asked in the coming UECM2273 final examination?

Axiomatic Probability

In 1933 a Russian mathematician Andrei Kolmogorov reduced the behavior of probability function to a set of just three simple postulates. He showed that the following axioms are necessary and sufficient for characterizing probability function.

Axiom 1 $P(E) \geq 0$ for every event defined over sample space S .

Axiom 2 $P(S) = 1$

Axiom 3 If $E_1, E_2, \dots$ is a sequence of events such that $E_i \cap E_j = \emptyset$ for $i \neq j$, then

$$P(E_1 \cup E_2 \cup \dots) = \sum_{n=1}^{\infty} P(E_n).$$

52.1.3 1.3 Conditional Probability and Independent Events

Often events are related. If we know the relationship between the two events, then we can model the behavior of one event in terms of the other. Suppose that, for a given experiment, we are interested in calculating the probability of an event E . If we are told that an event F occurred, how does this affect $P(E)$?

Example 52.17. Example 1.3.1 Roll a fair die twice. The sample space of this experiment can be expressed as a set consisting the following 36 points :

(1, 1) (1, 2) (1, 3) (1, 4) (1, 5) (1, 6)
 (2, 1) (2, 2) (2, 3) (2, 4) (2, 5) (2, 6)
 (3, 1) (3, 2) (3, 3) (3, 4) (3, 5) (3, 6)
 (4, 1) (4, 2) (4, 3) (4, 4) (4, 5) (4, 6)
 (5, 1) (5, 2) (5, 3) (5, 4) (5, 5) (5, 6)
 (6, 1) (6, 2) (6, 3) (6, 4) (6, 5) (6, 6)

where the point (i, j) represents the outcome that “ i ” appeared in the first trial and we obtain “ j ” in the second trial.

Ask yourself the following question:

What is the probability that the sum of the two numbers observed is at least 11? Let A denote the event that the sum of the two numbers obtained is at least 11. Event $A = (5, 6), (6, 5), (6, 6)$ Answer : $P(A) =$

$$\frac{3}{36} = \frac{1}{12}$$

However, if we are told that the event B that the first trial produces a “6”, What will be the chance that occurs ? : Event $B = (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)$ $P(B) =$

$$\frac{6}{36} = \frac{1}{6}$$

Answer : Probability that the sum is at least 11 given that the first trial is a “6” is $\frac{1}{3}$ Now the concept of conditional probability is given as follows:

Definition 52.3. Definition 1.3.1 The conditional probability of E related to F is defined to be

$$P(E|F) = \begin{cases} \frac{P(E \cap F)}{P(F)}, & \text{if } P(F) > 0 \\ 0, & \text{if } P(F) = 0 \end{cases} \quad (52.1)$$

e:1.3.1

The notation “ $P(E|F)$ ” reads “ the conditional probability of E given F .”

Notice that

- the set function $P(\cdot | F)$ is a set function defined on $\mathcal{J}$.
- One can verify that $P(\cdot | F)$ is a probability measure defined on $\mathcal{J}$ with the property that
- $E \supseteq F \Rightarrow P(E|F) = 1; E \cap F = \emptyset \Rightarrow P(E|F) = 0$.

Equation (1.3.1) implies that

$$P(E \cap F) = P(E|F)P(F) \quad (1.3.2)$$

Remark 52.18. Remark If we define $0 \times \infty = 0$, equation (1.3.2) holds for the case even when $P(F) = 0$. Sometimes it happens that $P(E|F) = P(E)$. In this case, formula (1.3.2) becomes

$$P(E \cap F) = P(E)P(F) \quad (52.2)$$

e:sd:1.3.3

Definition 52.4. Definition 1.3.2 Two events E and F are said to be **(stochastically) independent** if

$$P(E \cap F) = P(E)P(F).$$

The concept of independence of two events can be generalized to a class of events.

Definition 52.5. Definition 1.3.3 A family of events $\{E_i : i \in I\}$ is said to be **(mutually) independent** if for any finite subset J of I ,

$$P(\bigcap_{j \in J} E_j) = \prod_{j \in J} P(E_j) \quad (52.3)$$

e:sd:1.3.4

The family $\{E_i : i \in I\}$ is said to be pair-wise independent whenever J consists of two elements.

Remark 52.19. Remark Mutually independence $\Rightarrow$ Pairwise independence
However the converse is not true.

Example 52.20. Example 1.3.2 Suppose that the sample space consists of four sample points, say $S = \{1, 2, 3, 4\}$ Define probability as

$$P(E) = \frac{\# \text{ of elements in } E}{4}$$

The events $E = \{1, 2\}$, $F = \{1, 3\}$, and $G = \{1, 4\}$ are pair-wise independent but not mutually independent.

52.1.4 1.4 Partition Theorem and Bayes's Formula

Let S be a sample space and P a probability defined on the event space of S .

A partition of S is a collection of nonempty disjoint events $\{E_i\}$ with union $\cup_i E_i = S$. The following partition theorem is extremely useful.

Theorem 52.21 (Law of Total Probability). *Theorem 1.4.1 If $\{E_1, E_2, \dots\}$ is a partition of Ω such that $P(E_i) > 0$ for each i , then*

$$P(F) = \sum_i P(F|E_i)P(E_i) \quad \text{for all event } F.$$

Partition theorem is closely related to the so-called "Bayes's formula":

If $\{E_1, E_2, \dots\}$ is a partition the sample space S such that $P(E_i) > 0$ for each i and F is an event with $P(F) > 0$, then

$$P(E_k|F) = \frac{P(F|E_k)P(E_k)}{\sum_i P(F|E_i)P(E_i)} \quad (\text{Bayes' Formula})$$

52.1.5 1.5 Random Variables

Definition 52.6. Definition 1.5.1 Let S be a sample space. A **random variable** X is a real function defined on S such that for any real number x , the set

$$\{s \in S : X(s) \leq x\}$$

is an event.

Remark 52.22. Remark (a) The concept of a random variable permits us to associate with any sample space S , a sample space that is a subset of real numbers, and in which the events of interest are subsets of real numbers. If such a subset of real numbers is denoted by A , and $E = X^{-1}(A)$, then we write

$$P_X(A) = P(X \in A) = P(E)$$

(b)

Student should not be too concerned about the concept of random variable. In this course, we just assumed that all functions defined on the sample space are random variables.

1.6 Discrete Random Variables

Definition 52.7. Definition 1.6.1 Random variable X defined over sample space S is said to be discrete if $X(S)$ is a countable set. The function

$$f(x) = P(X = x), \quad x \in X(S)$$

is called the **probability mass function (pmf)** of X .

The general properties of pmf are as follows:

- (a) $f(x) \geq 0$ for all $x \in X(S)$
- (b) $\sum_{x \in X(S)} f(x) = 1$

Example 52.23. Example 1.6.1 Flip a fair coin twice. Let X denote the number of Heads observed. Then $P(X = 0) = \frac{1}{4}$, $P(X = 1) = \frac{1}{2}$, and $P(X = 2) = \frac{1}{4}$

Definition 52.8. Definition 1.6.2 The cumulative distribution function (cdf) of a random variable X is defined to be

$$F(x) = P(X \leq x) = P(X^{-1}(-\infty, x])$$

The general properties of cdf are listed as follows:

- (a) $F(x) \geq 0$ for all $x \in X(S)$.
- (b) $x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$.
- (c) $\lim_{x \rightarrow -\infty} F(x) = 0$, $\lim_{x \rightarrow \infty} F(x) = 1$.
- (d) F is right continuous. That is, $\lim_{t \downarrow x} F(t) = F(x)$.

For real numbers $x < y$,

$$(-\infty, y] = (-\infty, x] \cup (x, y]$$

It follows that

$$P(x < X \leq y) = F(y) - F(x).$$

Let $x_1 < x_2 < \dots$ denote the image set of the discrete random variable X with pmf f and cdf F . The relationships between f and F are as follows:

- (a) $f(x_i) = F(x_i) - F(x_{i-1})$.
- (b) $F(x) = \sum_{x_i \leq x} f(x_i)$

Remark 52.24. Remark It is easier to understand the nature of a random variable and its probability distribution by considering the pmf, rather than the cdf.

Example 52.25. Example 1.6.2 Flip a fair coin twice. Let X denote the number of Heads observed. Then the pmf of X is given by

The cdf of X is

$$F(x) = \begin{cases} 0, & x < 0 \\ 1/4, & 0 \leq x < 1 \\ 3/4, & 1 \leq x < 2 \\ 1, & x \geq 2 \end{cases}$$

Definition 52.9. Definition 1.6.2 Let X be a discrete random variable with image set $\{x_1, x_2, \dots\}$. The expected value of X is defined by

$$E(X) = \sum_i x_i P(X = x_i)$$

provided that the sum converges absolutely.

Note: The terms **mean** and **expectation** also are often used.

The extension of the definition of expected value is given in the following theorem

Theorem 52.26. Theorem 1.6.1 If X is a discrete random variable and if $g(x)$ is any real-valued function, then

$$E(g(X)) = \sum_{x \in X(S)} g(x) P(X = x)$$

provided that the sum is absolutely convergent.

- The expected value of $(X - E(X))^2$, if exists, is called the variance of X and is denoted by

$$\sigma^2 = \text{Var}(X) = E[(X - E(X))^2]$$

- The positive square root of the variance of X is called the standard deviation of X .
- The variance of X can be calculated by means of the following formula:

$$\text{Var}(X) = E(X^2) - (E(X))^2.$$

eg:sd:1.6.3

Example 52.27. Example 1.6.3 Flip a fair coin twice and X denotes the number of Heads observed. Then

$$E(X) = 1 \text{ and } \text{Var}(X) = \frac{1}{2}$$

The following example shows that even the sum $\sum_{x \in X(S)} x P(X = x)$ is finite, but the expected value of X does not exist.

eg:sd:1.6.4

Example 52.28. Example 1.6.4 Let X denote a random variable with pdf given by

$$P(X = (-1)^n n) = \frac{A}{n^2}, \quad n = 1, 2, \dots; \quad A^{-1} = \sum_{n=1}^{\infty} \frac{1}{n^2}.$$

Then

$$\sum_{x=1}^{\infty} x P(X = x) < \infty$$

but

$$\sum_{x=1}^{\infty} x P(X = x) = \infty$$

The work in manipulating expected value can be greatly facilitated by making use of the following theorem.

Theorem 52.29. Theorem 1.6.2 For any random variable X and constants a and b ,

$$E(aX + b) = aE(X) + b, \quad \text{Var}(aX + b) = a^2 \text{Var}(X)$$

An important special case involves establishing a “standardised” variable. If $E(X) = \mu$, $\text{Var}(X) = \sigma^2$, and $Y = \frac{X - \mu}{\sigma}$, then $E(Y) = 0$ and $\text{Var}(Y) = 1$.

52.1.6 1.7 Continuous Random Variables

Definition 52.10. Definition 1.7.1 A random variable X is said to be (absolutely) continuous if there exists a function $f(x)$ such that the cdf of X is given by

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt \quad (52.4)$$

e:1.7.1

Note:

(a) (a) If F is (absolutely)continuous and f is continuous at x , then

$$f(x) = \frac{d}{dx}F(x)$$

(b) (b) For any real number a and any positive integer n ,

$$\{a\} \cup (-\infty, \frac{a-1}{n}) \subset (-\infty, a] P(X = a) + F(\frac{a-1}{n}) \leq F(a)$$

So if F is continuous at $x = a$, then $P(X = a) = 0$.

(c) (c) Any function f may be considered as a pdf if it produces a legitimate cdf when integrated as in formula (1.7.1).

The following conditions are necessary and sufficient for a function f to be a pdf of a random variable:

(a) $f(x) \geq 0$ for all x ;

(b) $\int_{-\infty}^{\infty} f(x)dx = 1$

52.1.7 1.8 Expected Values of Continuous Random Variables

Definition 52.11. Definition 1.8.1 The expected value of a continuous random variable X with pdf f is given by

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

provided that the integral is absolutely convergent.

For function of random variables we have the following theorem

Theorem 52.30. Theorem 1.8.1 If X is a continuous random variable with pdf $f(x)$ and $g(x)$ is a realvalued function, then

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

provided that the integral converges absolutely.

The definitions of variance and standard deviation and properties of expectation for discrete random variable hold for the continuous case. That is,

$$\text{Var}(X) = E(X^2) - [E(X)]^2, \quad E(a + bX) = a + bE(X) \quad \text{Var}(a + bX) = b^2 \text{Var}(X)$$

52.1.8 1.9 Some Special Distribution

In many applications by recognizing certain characteristics, it is possible to determine that the distribution has a known special form. In this chapter we shall develop some special discrete distributions

Special Discrete Distributions

(A) Bernoulli Distribution Let s denote the sample point of an experiment.

- (a) There are only two events of interest, say E or E^c
- (b) Event E occurs with probability p .

$$(c) X = \begin{cases} 1, & s \in E, \\ 0, & s \in E^c. \end{cases}$$

Then random variable is said to have a Bernoulli distribution with parameter p . The pdf of X is given by

$$P(X = 1) = p = 1 - P(X = 0) \quad (1.9.1) \quad \text{e:sd:1.9.1}$$

Equation (3.2.1) can also be expressed as

$$P(X = x) = px(1 - p)^{1 - x}, x = 0, 1; 0 \leq p \leq 1 \quad (1.9.2) \quad \text{e:sd:1.9.2}$$

In this case we denote $X \sim \text{BIN}(1, p)$. The mean, the variance, the moment generating function, and the probability generating function of X are respectively,

$$E(X) = p, \quad \text{Var}(X) = p(1 - p)$$

$\text{BIN}(1, p)$ is named Bernoulli distribution in honour of the German mathematician Jacob Bernoulli (1654–1705).

Remark 52.31. Remark Bernoulli random variable is a building block of binomial distribution.

(B) Binomial Distribution Suppose $X_1, X_2, \dots, X_n$ are n independent Bernoulli random variables with common parameter p . Let

$$X = X_1 + X_2 + \dots + X_n.$$

Random variable X is said to have a binomial distribution with parameters n and p . In many cases, n is assumed to be a known positive integer. The pdf of X is given by

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, \dots, n; \quad 0 \leq p \leq 1, \quad q = 1 - p \quad (52.7) \quad \text{eq:sd:pmf:binomial}$$

We write $X \sim \text{BIN}(n, p)$.

$$E(X) = np \quad \text{Var}(X) = np(1 - p)$$

The random generation, distribution function and quantile function are given by `rbinom`, `pbinom`, `qbinom` respectively. The pdf (52.7) is available in R through `dbinom(x, size=n, prob=p, log=FALSE)`.
 eq:sd:pmf:binomial

Remark 52.32. Remark: I should have apologized here that the concept “independence of random variables” has yet to be defined. However, binomial random variable can also be described as follows: Random variable X possesses a binomial distribution if the following conditions are satisfied: (1) The experiment consists of a fixed number n of identical trials. (2) Each trial can result in one of only two possible outcomes, called success or failure. (3) The probability of success p is constant from trial to trial. (4) The trials are independent. (5) X is defined to be the number of successes among the n trials.

(C) Geometric Distribution Suppose that a series of coin flipping can be represented by a sequence of independent Bernoulli random variables $X_1, X_2, \dots$ with common parameter p . Let X denote the number of trials on which the first Head occurs, then

$$P(X = x) = p(1 - p)^{x-1}, \quad x = 1, 2, \dots; \quad 0 < p \leq 1.$$

This distribution is referred to as the geometric distribution with parameter p . We write $X \sim \text{GEO}(p)$.

$$E(X) = \frac{1}{p}, \quad \text{Var}(X) = \frac{1-p}{p^2}$$

(D) Negative Binomial Distribution In repeated independent Bernoulli trials, let X denote the number of trials required to obtain r successes. Then the pdf of X is given by

$$P(X = x) = \binom{x-1}{r-1} p^r q^{x-r}, \quad x = r, r+1, \dots; \quad 0 < p \leq 1$$

We write $X \sim \text{NB}(r, p)$.

$$E(X) = \frac{r}{p}, \quad \text{Var}(X) = \frac{rq}{p^2}$$

Remark 52.33. Remark

- Negative binomial random variable with parameters r and p can be expressed as the sum of r copies of independent geometric random variables with common parameter p .
- Negative binomial problem sometimes is referred to as inverse binomial sampling. Suppose $X \sim \text{BIN}(n, p)$ and $Y \sim \text{NB}(r, p)$ where $0 < r \leq n$. Then

$$P(Y \leq n) = P(X \geq r)$$

In words, if r Heads have been observed in no more than n trials, then at least r Heads occur when a coin is flipped n times.

Example 52.34. Example 1.9.1 Suppose $X \sim \text{BIN}(3, p)$ and $Y \sim \text{NB}(2, p)$.

$$\begin{aligned} P(Y \leq 3) &= P(Y = 2) + P(Y = 3) = p^2 + 2p^2q = p^2(1 + 2q) = p^2(p + 3q) \\ P(X \geq 2) &= P(X = 2) + P(X = 3) = 3p^2q + p^3 = p^2(p + 3q) \end{aligned}$$

(D) Poisson distribution A number of probability distributions come about through limiting arguments applied to other distributions.

Suppose that $X \sim \text{BIN}(n, p)$ with $np = \lambda$, Then

$$\lim_{n \rightarrow \infty} P(X = x) = \lim_{n \rightarrow \infty} \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} = \frac{\lambda^x}{x!} \lim_{n \rightarrow \infty} \frac{n!}{n^x(n-x)!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} = e^{-\lambda} \frac{\lambda^x}{x!}.$$

To see how good is the approximation, let $X \sim \text{B}(1000, 0.001)$ with mean $\mu = 1$. Let $Y \sim \text{Poisson}(1)$. The probability functions of X and Y are computed as follows:

x	$P_{\text{Poisson}}(Y = x)$	$P_{\text{Binomial}}(X = x)$
0	0.367879	0.367695
1	0.367879	0.368063
2	0.183940	0.184032
3	0.061313	0.061283
4	0.015328	0.015290
5	0.003066	0.003049
6	0.000511	0.000506

Siméon-Denis Poisson (21 June 1781 – 25 April 1840), was a French mathematician, geometer, and physicist.

If $X \sim \text{Poisson}(\lambda)$, then $E(X) = \lambda = \text{Var}(X)$

- Poisson distribution can be used to model counts in areas or volumes, as well as in time.
- For example, we may use Poisson distribution to model
 - (a) (a) the number of flaws in a square yard of textile;
 - (b) (b) the number of bacteria colonies in a cubic centimeter of water;
 - (c) (c) the number of times a machine fails in the course of a workday;
 - (d) (d) the number of vehicles passing through a checkpoint in unit of time

1.10 Special Continuous Distributions

(A) Uniform Distribution Random variable X is said to be uniformly distributed over interval (a, b) if its pdf is given by

$$f(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{elsewhere} \end{cases}$$

We write $X \sim U(a, b)$.

$$E(X) = \frac{a+b}{2}, \quad \text{Var}(X) = \frac{(b-a)^2}{12}$$

(B) Exponential Distribution Random variable X is said to have an exponential distribution with mean β if its pdf is given by

$$f(x; \beta) = \frac{1}{\beta} e^{-x/\beta}, \quad x > 0, \beta > 0 \quad (3.3)$$

We write $X \sim \text{EXP}(\beta)$.

$$E(X) = \beta, \quad \text{Var}(X) = \beta^2$$

The exponential distribution has some “nice” properties.

- $P(X > x) = e^{-x/\beta}, x > 0$
- $P(X > a + b | X > a) = P(X > b)$ for positive values a and b .

Remark 52.35. Remark The second property listed above is known as “lack of memory”. It characterizes the exponential distribution.

(C) Gamma Distribution Let $\alpha > 0$ be a positive number. The symbol $\Gamma(\alpha)$ is known as gamma function at α and is defined by

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy$$

Note that the improper integral is convergent.

Some properties of gamma function $\Gamma(\alpha)$ are

- $\Gamma(\alpha + 1) = \alpha \Gamma(\alpha)$
- When n is a positive integer, $\Gamma(n + 1) = n!$

Note that the above improper integral is convergent. Formula (3.3.3) can be rewritten as

$$\int_0^{\infty} \frac{1}{\Gamma(\alpha)} y^{\alpha-1} e^{-y} dy = 1$$

In other words, when $\alpha > 0$,

$$g(y) = \begin{cases} \frac{1}{\Gamma(\alpha)} y^{\alpha-1} e^{-y}, & y > 0, \\ 0, & \text{elsewhere} \end{cases}$$

is a pdf. Let $x = \beta y$ with $\beta > 0$. Then

$$\int_0^{\infty} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x} dx = 1$$

Random variable X is said to have a gamma distribution with parameters $\alpha > 0$ and $\beta > 0$ if its pdf is given by

$$f(x; \alpha, \beta) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x}, & x > 0, \\ 0, & \text{elsewhere} \end{cases}$$

We write $X \sim \Gamma(\alpha, \beta)$

$$E(X) = \alpha\beta, \quad \text{Var}(X) = \alpha\beta^2$$

If r is a positive integer and $X \sim \Gamma(r/2, 2)$, then we say X has a χ^2 distribution with r degrees of freedom and we write $X \sim \chi^2(r)$.

(D) Beta Distribution Random variable X is said to have a beta distribution, denoted as $X \sim \text{Beta}(\alpha, \beta)$ if its pdf is given by

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad 0 < x < 1, \quad \alpha > 0, \quad \beta > 0.$$

The mean and variable of X are

$$E(X) = \frac{\alpha}{\alpha + \beta}, \quad \text{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

(E) Normal Distribution Random variable X is said to follow a normal distribution with mean μ and variance σ^2 if its pdf is given by

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{(x - \mu)^2}{2\sigma^2}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$$

We write $X \sim N(\mu, \sigma^2)$.

$$E(X) = \mu, \quad \text{Var}(X) = \sigma^2$$

- If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$
- The pdf and cdf of $N(0, 1)$ will be denoted as ϕ and Φ , respectively.

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp -\frac{x^2}{2}, \quad \Phi(z) = P(Z \leq z) = \int_{-\infty}^z \phi(x) dx$$

Note that

- ϕ is an even function, i.e., $\phi(-x) = \phi(x)$
- $\Phi(-z) = 1 - \Phi(z)$

52.1.9 1.11 Joint and Marginal Distributions

The preceding sections dealt with univariate distribution, i.e., with distribution of a single random variable. The joint distribution of two random variables is called a bivariate distribution. The bivariate distribution of random variables X and Y is specified at each point (x, y) by

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$$

If X and Y are discrete random variables, the joint probability function of X and Y is defined at each point (x, y) by

$$f_{X,Y}(x, y) = P(X = x, Y = y)$$

- A function $F(x, y)$ is a bivariate cdf if and only if
 - (a) $\lim_{x \rightarrow -\infty} F(x, y) = 0$ for all y
 - (b) $\lim_{y \rightarrow -\infty} F(x, y) = 0$ for all x
 - (c) $\lim_{\substack{x \rightarrow \infty \\ y \rightarrow \infty}} F(x, y) = 1$
 - (d) $F(b, d) - F(b, c) - F(a, d) + F(a, c) \geq 0$ for all $a < b$ and $c < d$.

1.12 Marginal and Conditional Distributions

Definition 52.12. Definition 1.12.1

Let $f(x_1, x_2)$ be the joint probability (density) function of discrete (continuous random) variables X_1 and X_2 . The marginal probability (density) functions of X_1 and X_2 are defined respectively as

$$f_1(x_1) = \begin{cases} \sum_{x_2} f(x_1, x_2), & X_i : \text{discrete}, i = 1, 2 \\ \int_{-\infty}^{\infty} f(x_1, x_2) dx_2, & X_i : \text{continuous}, i = 1, 2 \end{cases} \quad f_2(x_2) = \begin{cases} \sum_{x_1} f(x_1, x_2), & X_i : \text{discrete}, i = 1, 2 \\ \int_{-\infty}^{\infty} f(x_1, x_2) dx_1, & X_i : \text{continuous}, i = 1, 2 \end{cases}$$

Remark 52.36. Remark When integrating the joint pdf of X_1 and X_2 with respect to x_2 for finding the marginal pdf of X_1 , the limits of integration might not be from $-\infty$ to ∞ . More precisely, the limits of integration may depend on x_1 .

Example 52.37. Example 1.12.1 (a) Verify that $f(x_1, x_2) = 8x_1x_2$, $0 < x_1 < x_2 < 1$, is a joint pdf. (b) Find the marginal pdf's of X_1 and X_2 .

Solution. Solution (a) (i) Clearly $f(x_1, x_2) \geq 0$ for all x_1 and x_2 .

$$(ii) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 dx_2 = \int_0^1 \left(\int_0^{x_2} 8x_1x_2 dx_1 \right) dx_2 = \int_0^1 [4x_2x_1^2]_{x_1=0}^{x_1=x_2} dx_2 = \int_0^1 4x_2^3 dx_2 = 1$$

$$(b) f_1(x_1) = \int_{x_1}^1 8x_1x_2 dx_2 = 4x_1(1 - x_1^2), 0 < x_1 < 1.$$

$$f_2(x_2) = \int_0^{x_2} 8x_1x_2 dx_1 = 4x_2^3, 0 < x_2 < 1.$$

□

52.1.10 1.13 Independence of Random Variables

Definition 52.13. Definition 1.13.1 Random variables X_1 and X_2 are said to be independent if for any real numbers x and y ,

$$P(X_1 \leq x, X_2 \leq y) = P(X_1 \leq x)P(X_2 \leq y)$$

Note Let $f_i(x_i)$ denote the pdf (discrete or continuous) of X_i , $i = 1, 2$ and $f(x_1, x_2)$ the joint pdf of the X_i 's. Then X_1 and X_2 are independent if and only if

- (a) the support of $f = \{(x_1, x_2) : f(x_1, x_2) > 0\}$ is a Cartesian product, $S = S_1 \times S_2$, and

(b) $f(x_1, x_2) = f_1(x_1)f_2(x_2)$ for all $(x_1, x_2) \in S$.

eg:sd:1.13.1

Example 52.38. Example 1.13.1 It can be verified that

$$f(x_1, x_2) = 8x_1x_2, \quad 0 < x_1 < x_2 < 1$$

represents a joint pdf of two continuous random variables. However, the support of f is a triangular region and cannot be represented as a Cartesian product of two sets. Thus, the two random variables with f as their joint pdf are dependent.

52.1.11 1.14 Conditional distributions

Definition 52.14. Definition 1.14.1 Let X_1 and X_2 have joint pdf $f(x_1, x_2)$ and marginal pdf $f_1(x_1)$ and $f_2(x_2)$, respectively. The conditional pdf of X_2 given $X_1 = x_1$ is defined to be

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)} \text{ provided that } f_1(x_1) > 0$$

Likewise, the conditional pdf of X_1 given $X_2 = x_2$ is defined as

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)} \text{ provided that } f_2(x_2) > 0.$$

Note It follows from the definition that the joint pdf of X_1 and X_2 can be expressed as the product of conditional and marginal pdf as

$$f(x_1, x_2) = f(x_1|x_2)f_2(x_2) = f(x_2|x_1)f_1(x_1)$$

eg:sd:1.14.1

Example 52.39. Example 1.14.1 (Example 1.13.1 Revisited) The joint pdf of X_1 and X_2 , and the marginal pdf's are

$$f(x_1, x_2) = 8x_1x_2, \quad f_1(x_1) = 4x_1(1 - x_1^2),$$

$$0 \leq x_1 \leq x_2 \leq 1, \quad 0 \leq x_1 \leq 1$$

$$f_2(x_2) = 4x_2^3,$$

$$0 \leq x_2 \leq 1$$

Then

$$f(x_2|x_1) =$$

$$2x_2(1 - x_1^2)$$

$$f(x_1|x_2) =$$

$$2x_1x_2^2$$

,

$$x_1 \leq x_2 \leq 1$$

,

$$0 \leq x_1 \leq x_2$$

52.1.12 1.15 Properties of Expected Values

Let random variables $X_1, X_2, \dots, X_n$ have joint pdf $f(x_1, x_2, \dots, x_n)$ and let g be a real valued function of $(x_1, x_2, \dots, x_n)$. The expected value of $g(X_1, X_2, \dots, X_n)$, if exists is given by

$$E(g(X_1, X_2, \dots, X_n)) = \begin{cases} \sum_{x_1} \cdots \sum_{x_n} \sum g(x_1, x_2, \dots, x_n) f(x_1, x_2, \dots, x_n), & \text{discrete case} \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum g(x_1, x_2, \dots, x_n) f(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_n, & \text{continuous case} \end{cases}$$

In particular,

- $E(\sum_{i=1}^n a_i X_i) = \sum_{i=1}^n a_i E(X_i)$, where $a_1, a_2, \dots, a_n$ are constants.

- If $X_1, X_2, \dots, X_n$ are independent and g_i 's are real-valued functions, then $E(\prod_{i=1}^n g_i(X_i)) = \prod_{i=1}^n E(g_i(X_i))$ provided that expectations exist.

Definition 52.15. Definition 1.15.1 The covariance of random variables X_1 and X_2 is defined to be

$$\text{Cov}(X_1, X_2) = E[(X_1 - E(X_1))(X_2 - E(X_2))]$$

Some basic properties of covariance are as follows:

- $\text{Cov}(X_1, X_2) = E(X_1 X_2) - E(X_1)E(X_2)$
- $\text{Cov}(aX_1 + b, cX_2 + d) = ac \text{Cov}(X_1, X_2)$
- $\text{Cov}(X, X) = \text{Var}(X)$
- $\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) + 2 \text{Cov}(X_1, X_2)$
- $\text{Cov}(\sum_{i=1}^r X_i, \sum_{j=1}^s Y_j) = \sum_{i=1}^r \sum_{j=1}^s \text{Cov}(X_i, Y_j)$
- If X_1 and X_2 are independent, then $\text{Cov}(X_1, X_2) = 0$. However, the converse is not true.

eg:sd:1.15.1

Example 52.40. Example 1.15.1 Suppose $X \sim N(0, 1)$ and $Y = X^2$, then X and Y are dependent. But $\text{Cov}(X, Y) = 0$.

Definition 52.16. Definition 1.15.2 The correlation coefficient of random variables X and Y is defined to be

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}$$

Some important properties correlation coefficient are as follows:

- $-1 \leq \rho_{XY} \leq 1$
- $\rho_{XY} = 1$ if and only if there exists constant $a (\neq 0)$ and b such that $P(Y = aX + b) = 1$

Remark 52.41. Remark ρ_{XY} measure the strength of linear relationship between two random variables.

- If X and Y are independent, then $\rho_{XY} = 0$
- However, $\rho_{XY} = 0$ does not imply that there is no relationship between X and Y .

Example 52.42. Example 4.7.2 (Example 4.7.1 revisited) Suppose $X \sim N(0, 1)$ and $Y = X^2$. There is a perfect curvilinear (parabolic) relationship between X and Y . However, $\rho_{XY} = 0$.

1.16 Conditional Expectation

Definition 52.17. Definition 1.16.1 Let X and Y be two random variables. The conditional expectation of Y given $X = x$ is defined as

$$E^{Y|x}(Y) = E(Y|x) = \begin{cases} \sum_y y f(y|x), & X, Y : \text{discrete} \\ \int_{-\infty}^{\infty} y f(y|x) dy, & X, Y : \text{continuous} \end{cases}$$

provided that the sum or the integral converges absolutely.

Example 1.16.1 (Example 1.14.1 revisited) Let X and Y be continuous random variables with joint pdf

$$f(x, y) = 8xy, \quad 0 < x < y < 1$$

The conditional densities are

$$f(x|y) = \frac{4x^3}{y^4}, \quad 0 < x < y$$

$$f(y|x) = \frac{2y}{1-x^2}, \quad x < y < 1$$

$$E^{X|Y}(X) = E(X|y) = \int_0^y x f(x|y) dx = \frac{1}{y^4} \int_0^y 4x^4 dx = \frac{4}{5}y$$

$$E^{Y|X}(Y) = E(Y|x) = \frac{1}{1-x^2} \int_x^1 2y^2 dy = \frac{2(1-x^3)}{3(1-x^2)} = \frac{2}{3} \frac{1+x+x^2}{1+x}$$

The following result is useful in calculating expected value.

Theorem 52.43. *Theorem 1.16.1 Let X and Y be two random variable and $g(x, y)$ a real-valued function. Then*

$$E[g(X, Y)] = E[E^{Y|X}(g(X, Y))] = E[E^{X|Y}(g(X, Y))]$$

provided that the expectations are finite.

An interesting consequence of the above theorem is the following:

Corollary 52.44. *Corollary 4.8.1 $E(X) = E[E^{X|Y}(X)] = E[E(X|Y)]$*

Example 4.8.2 Roll a fair coin once. If the outcome is x , then a coin with the chance p of getting Head in a single trial is tossed x times. What is the expected number of Heads observed? Solution Let X denote the number obtained in rolling a fair die and Y the number of Heads observed. Then $E(Y) = E[E(Y|X)] = E[Xp] = pE(X) =$

$$7p \cdot 2$$

Definition 52.18. Definition 1.16.2 The conditional variance of Y given $X = x$ is defined to be

$$\text{Var}(Y|x) = E[(Y - E(Y|x))^2|x]$$

The relation between the unconditional variance and conditional variance of Y is stated below without proof.

Theorem 52.45. *Theorem 1.16.2 (a) $\text{Var}(Y) = \text{Var}(YX) + E[\text{Var}(Y|X)]$ (b) $\text{Var}(Y|X) \leq \text{Var}(Y)$*

52.2 Chapter 2 Decision Problems

52.2.1 2.1 Introduction

The science of statistics deals with the development of theories and techniques which are appropriate for making inferences under the certain conditions of uncertainty. Current statistical practice depends mainly on the formulation of probability models of various physical systems, on methods of collecting and analyzing data, and on the design of efficient and informative experiments. Some simple typical problems that the methods of statistics are designed to solve are:

- Decide on the basis of testing a few samples from a shipment of a certain drug whether the quality of the shipment is satisfactory.
- Predict on the basis of a small poll that the voter's preferences are on a vital issue.
- Decide on the basis of analyzing sonar signals whether a submarine is approaching.

Problems of this type can be formulated mathematically by considering the data that are to be used for analysing the problem as the observed values of a random variable X . The distribution of X is assumed to belong to a certain family of distributions, a particular member of which is specified when the value of a parameter θ is specified. The problem is to decide on the basis of the data which member of the family could represent the distribution of X . This is called the problem of parametric statistical inference.

The bulk of statistics that is taught concerns classical inference procedures, and so it is only natural that many people will try to use them to do everything, even to solve clear decision problems. One of the most commonly misused inference procedures is hypothesis testing as illustrated as follows:

Example 52.46. Example 2.1.1 Take a random sample of size n , say $X_1, \dots, X_n$, from the normal distribution $N(\mu, 1)$. It is desired to conduct a test of size $\alpha = 0.05$ for testing

$$H_0 : \mu = 0 \text{ against } H_1 : \mu \neq 0.$$

The likelihood ratio test of this hypotheses is to reject H_0 if the observed sample mean is “far away” from zero. More precisely, we shall reject H_0 if $|x| > \frac{1.96}{\sqrt{n}}$.

Now it is unlikely that the null hypothesis H_0 is true. Suppose that the true value of $\mu = 10^{-10}$ which while nonzero is probably a meaningless difference from zero in most practical contexts. However, if the sample size of large, say $n = 1024$ is to be taken, then

$$P(|\bar{X} - 10^{-10}| < 10^{-11}) = P(10^{12}|\bar{X} - 10^{-10}| > 10) = P(|Z| > 10) = 1$$

But then for

$$|\bar{x} - 10^{-10}| < 10^{-11} \Rightarrow \bar{x} > 10^{-10} - 10^{-11} = 9 \times 10^{-11} > 1.96 \times 10^{-12}$$

Hence the classical test will be virtually certain to reject H_0 .

This example shows that a “statistically significant” difference between the true parameter and the null hypothesis can be an unimportant difference practically. Likewise a difference that is not significant statistically can nevertheless be very important practically as demonstrated in the following example.

Example 52.47. Example 2.1.2 The effectiveness of a drug is measured by $X \sim N(\mu, 9)$. The following hypotheses are to be tested:

$$H_0 : \mu \leq 0 \text{ against } H_1 : \mu > 0$$

Suppose a random sample of size 9 gives sample mean $\bar{x} = 1$. Then

$$P(\bar{X} \geq 1 | \mu = 9) = P\left(\frac{\sqrt{9}(\bar{X} - 0)}{3} > 1\right) = P(Z > 1) = 0.1587$$

Hence there is no significant evident to reject hypothesis H_0 , say at level $\alpha = 0.05$. However, if 1 were practically important difference from 0, we would certainly be very interested in the drug.

Decision theory is an extension of testing, estimation students learnt in Probability and Statistics in year 1 courses (Chapter 51). Statistical decision theory enables us to model situations where classical theory is inadequate. The following examples illustrate towards what kind of problems this extension is directed.

eg:sd:2.1.3

Example 52.48. Example 2.1.3 A manufacturer has to decide on the basis of sales figures whether to expand production, reduce production, or to hold production steady. For each of these three possible actions and possible values of the true potential market for his product, the manufacturer can supply a rough measure of his potential loss or gain. In view of the number of actions the testing framework is inadequate. While estimation of the true market may be of interest, the mean-squared error is clearly not the right measure of performance.

eg:sd:2.1.4

Example 52.49. Example 2.1.4 A consumer organization preparing a report on air conditioners tests samples of several brands. On the basis of these samples, they want to give a ranking of the brands. Thus, if there are k different brand, there are $k!$ possible rankings. Again neither the estimation nor the testing framework seem appropriate.

The elements of decision theory are similar to those of the theory of games. In particular, decision theory may be considered as the theory of a two-person game, in which the nature takes the role of one of the players. The other player is usually referred as statistician.

We shall first consider some simple games involving two competing players.

52.2.2 2.2 Some Examples

Pure Strategies

eg:sd:2.2.1

Example 52.50. Example 2.2.1 (Matching Fingers)

In this very simple game each player chooses to present either one or two fingers. If the numbers presented match, Player I wins; if they differ, Player II wins. Since the outcomes are “win” or “lose”, we take the payoff (amount won by Player I) to be 1 if he wins and -1if he loses. The payoff matrix of this game is given as follows:

Player I (nature)	Player II (statistician)	
	a = 1	a = 2
$\theta = 1$	1	-1
$\theta = 2$	-1	1

The following example is another matching fingers game.

eg:sd:2.2.2

Example 52.51. Example 2.2.1 (Odd or Even)

Two contestants simultaneously put up either one or two fingers. One of the players, called him player I, wins if the sum of the digits showing is odd, and the other player, player II, wins if the sum of the digits showing is even. The winners in all cases receives in dollars the sum of the digits showing, this being paid to him by the loser. To create a triplet (Θ, A, L) , we call player I the nature and player II the statistician. Hence the set of nature’s strategies is $\Theta = \{1, 2\}$. The statistician also has set of strategies $A = \{1, 2\}$. The loss function of this game is tabulated as follows:

Strategies (θ) of nature	Strategies (a) of the statistician	
	1	2
1	-2	3
2	3	-4

eg:sd:2.2.3

Example 52.52. Example 2.2.3 (Tictactoe) Player I first in this familiar game and uses “X”, while plater II uses “O”. For convenience in describing strategies we number the nine spaces to be filled as shown below:

1	2	3
4	5	6
7	8	9

The strategy for player I could be as follows:

- 1st move : put X in 1

- 2nd move : place an X in 9 unless it is filled. If 9 is filled, place X in
- 3rd move : if 1 and 9 have X and 5 empty, place X in 5; If 1 and 7 have X and 4 is empty, place X in 4. Otherwise place X in the highest-numbered empty space.

Complete the game by placing an X in the highest-numbered empty space on each move. Conclusion : It is possible to give strategies for tictatoe which will always win or draw against any strategy of your opponent.

Mixed Strategies

Example 52.53. Example 2.2.4 The payoff matrix for the game of Matching Fingers (Example 52.50) is repeated here for references:

Strategies (θ) of nature	Strategies (a) of the statistician	
	1	2
1	-2	3
2	3	-4

Suppose Player I decides to choose the number of fingers he will present by tossing a fair coin (which he hides from Player II) and presenting one finger if heads, two if tails. His payoff against either of Player II's strategies is now a random variable. The expected value of that random variable tells him (by law of large numbers) the longterm average payoff he will receive. If Player II decides to put up one finger, the expected payoff is

$$1 \times P(\text{Player I puts up one finger}) + (-1) P(\text{Player I puts up two fingers}) = 0$$

If Player II decides to put up two fingers, then the expected payoff is $(-1) \times P(\text{Player I puts up one finger}) + 1 \times P(\text{Player I puts up two fingers}) = 0$ This means that by randomly putting up one or two fingers Player I can guarantee an expected payoff of 0, whichever strategy Player II chooses.

Example 52.54. Example 2.2.5 (Game with Bluffing) Two players each put an "ante" of a units into a pot ($a \geq 0$). Player I (nature) then draws a card from a deck, which gives him a winning or a losing card. Both players are aware of the probability p ($0 \leq p \leq 1$) that the card drawn is a winning card. Payer I then may bet b unit by putting b units into the pot or he may check. If player I checks, he wins the pot if he has a winning card and loses the pot if he has a losing card. If player I bets, player II may call and put b units in the pot or he may fold. If player II folds, player I wins the pot whatever card he has drawn. If player calls, player I wins if he has a winning card and loses it otherwise. If player I receives a winning card, it is clear that he should bet: if he checks, he automatically receives total winnings of a units, whereas if he bets, he will receive at least a units and possibly more. For the purpose of our discussion we assume that the rules of the game enforce this condition: that if I receives a winning card, he must bet. This will eliminate some obviously poor strategy from player I's strategy set. With this restriction, player I has two strategies

θ_1 : Bluff strategy : bet with a winning card, check with a losing card. θ_2 : Honest strategy : bet with a winning card, check with a losing card. The two strategies of player II are a_1 : Call strategy : if player I bets, call. a_2 : Fold strategy : if player I bets, fold.

The actual payoff in such a game is a random quantity determined by the chance moves. The following table shows player I's expected winnings and player II's expected losses. Player II (Statistician) Call Fold Player I (Nature) Bluff Honest

$$(2p - 1)(a + b) \quad (2p - 1)a + pb \\ a \quad (2p - 1)a$$

For example, If player I uses the honest strategy and player II uses the call strategy, player II's loss will be $a + b$ with probability p if player I receives a winning card - a with probability $1 - p$ if player I receives a losing card Thus the expected loss of player II is $(a + b)p - a(1 - p) = (2p - 1)a + pb$

The verification of the other expected payoff will be left as exercise.

52.2.3 2.3 Decision

In the previous section, we consider briefly some two-person games played against a rational opponent. On occasion, however, we formulated games in which the opponent was Nature. In such games the payoff depends on the player's action and also on a "state of nature" not chosen by a rational opponent.

eg:sd:2.3.1

Example 52.55. Example 2.3.1 A manufacturer of special-purpose electronic tubes has agreed to replace any which are defective. His gain on a tube sold is RM50. If a defective tube is sold, it must be replaced at a cost of RM100, so the manufacturer loses RM50 if he sells a defective tube. It is possible to detect defective tubes by testing, which cost RM10 per tube. A defective tube has a scrap value of RM20. The profit of the manufacturer depends on his own product inspection strategy and also on whether the particular component is or is not defective. There is no "rational" opponent plays against the manufacturer.

eg:sd:2.3.2

Example 52.56. Example 2.3.2

A person named Ajip has a congenital hearing defect in one ear, caused by malformation of the bones in the inner ear. A surgeon states that an operation is available to correct his defect, but that the operation does not always succeed. In fact,

the operation may correct Mr. Ajip hearing (θ_1),

have no effect (θ_2), or

destroy the partial hearing he now has (θ_3). The surgeon cannot predict in advance which of these states of nature holds in Mr. Ajip's case.

Mr. Ajip can of course agree to the operation (a_1),

Mr. Ajip can refuse to the operation (a_2).

Mr. Najip cannot easily assign dollar values to the possible outcomes, for he must consider the inconvenience of his present hearing loss and the greater inconvenience of a total hearing loss in one ear, as well as the cost of the operation. After careful consideration, however, he draws up a payoff matrix which reflects his personal feelings. It is given in the following table

States of Nature	Mr Ajip	
	Agree to the operation (a_1)	Refuse to the operation (a_2)
Operation has positive effect (θ_1)	25	-10
Operation has no effect (θ_2)	-15	-10
Operation has negative effect (θ_3)	-100	-10

This example illustrates three features which occur frequently in decision theory.

- (1) (1) The payoffs are personal assessment.
- (2) (2) Nature is not a conscious opponent.
- (3) (3) If Mr. Ajip wishes, he can often obtain additional information about the Nature's strategy (the "true" state of nature) by conducting experiments or otherwise taking observations.

52.2.4 2.4 Formulating Decision Problems

The term statistical decision refers to the class of statistical problems in which the statistician must gain information about certain critical parameters or states of nature in order to be able to make effective decisions in situations where the consequences of his decisions will depend on the true state of nature. The statistical decisions that will be discussed in this course can be applied in a wide variety of problems.

eg:sd:2.4.1

Example 52.57. Example 2.4.1

Car manufacturing company Proton knows that the national economic growth will be either recession (θ_1), mild (θ_2), or dynamic (θ_3). What is more is that the IMF forecasts that the probability of economic recession is 50% and that mild and dynamic growth each occurs 25%. Proton car company wants to determine the best production schedule to meet the demand of each type of economic conditions.

Statistical decision theory enables us to model situations where classical theory is inadequate. The following examples illustrate towards what kind of problems this extension is directed.

eg:sd:2.4.2

Example 52.58. Example 2.4.2 A manufacturer has to decide on the basis of sales figures whether to expand production, reduce production, or to hold production steady. For each of these three possible actions and possible values of the true potential market for his product, the manufacturer can supply a rough measure of his potential loss or gain. In view of the number of actions the testing framework is inadequate. While estimation of the true market may be of interest, the mean-squared error is clearly not the right measure of performance.

eg:sd:2.4.3

Example 52.59. Example 2.4.3 A consumer organisation preparing a report on air conditioners tests samples of several brands. On the basis of these samples, they want to give a ranking of the brands. Thus, if there are k different brands, there are $k!$ possible rankings. Again neither the estimation nor the testing framework seem appropriate.

Frequently problems which at first sight appear to be two-sided tests of a hypothesis really lead to one of three possible actions.

eg:sd:2.4.4

Example 52.60. Example 2.4.4 Suppose that we want to study if and how a new drug affects blood pressure. We may begin by testing the two-sided hypothesis that the drug has no effect against the alternative that it has some effect. But, of course, if the null hypothesis is rejected, we want to know whether the pressure has been raised or lowered. In effect we contemplate (at least) three actions, announcing “No effect (a_1),” “Increase blood pressure (a_2),” “Lowers blood pressure (a_3).”

The distinguish feature of decision problems is that the decision maker can obtain information about Nature’s pure strategy (the state of nature) from the outcome of a random experiment. More exactly, the decision maker has available a random experiment whose probability distribution depends on which state of nature happens to be the case.

eg:sd:2.4.5

Example 52.61. Example 2.4.5 A pharmaceutical manufacturer, ABC Lab, has developed a new drug to treat high blood pressure. In order to market the drug, ABC must be able to claim that it works for 80% of all patients with high blood pressure. If the drug is marketed and does work at least 80% of the time, ABC estimates a gain of RM50 million. If it is marketed and is less effective than claimed, an estimated loss of RM20 million in goodwill and penalties will result. If the drug is not marketed, RM2 million in development costs will be lost. ABC therefore decides to test the drug on 20 patients with high blood pressure before reaching a decision on marketing.

Decision theory consists of three basic elements:

- (1) (1) A nonempty set, Θ , of possible states of nature, sometimes referred as the parameter space.
- (2) (2) A nonempty set, A , of actions available to the statistician.
- (3) (3) A loss function $L(\theta, a)$, a real-valued function defined on $\Theta \times A$.

We call the triplet (Θ, A, L) a no-data decision problem. Such triplet is interpreted as follows: Nature use strategy θ and statistician employ strategy a , $L(\theta, a)$ represents the “loss” incurred by statistician.

When the decision maker has available a random experiment whose probability distribution depends on which state of nature happens to be the case, he can make use of the information obtained from the observation to select an appropriate action. So that in the selection of an action is not completely in the dark concerning the state of nature. The intelligent use of data will define a decision rule with lower expected loss than what would be incurred if data were not available. In this case we have a statistical decision problem.

Now we state explicitly what we must do to formulate a problem in statistical decision theory.

- (1) (1) S : The sample space of the observed random (vector) variable X .
- (2) (2) Θ : The parameter space Θ or the states of nature. Each state of nature $\theta \in \Theta$ corresponds to a different distribution for the random variable X . Set Θ is the set of pure strategies of the Nature.
- (3) (3) A : The set of all actions available to the decision maker.
- (4) (3) D : The set of all decision rules available to the decision maker. D is called the decision space.
- (5) (4) $L(\theta, a)$: The loss function, which measures the decision maker's loss when he takes action to deal with the problem he faced.

Note : In practice the decision maker often begins knowing θ , D , and A , and must choose a random experiment to provide more information.

Example 52.62. Example 2.4.6 A wildcat oil driller has studied geological condition surrounding production well and dry holes long and hard. He has observed two specific conditions which seem to be associated with oil. Call then conditions A and B, since he refuses to say what they are. More precisely, the wildcatter has observed that in the past dry holes the following relative frequencies of various conditions occurred.

Condition	A only	B only	Both A and B	Neither
Probability	0.1	0.2	0.1	0.6

In the past producing wells the corresponding probability distribution was as follows:

Condition	A only	B only	Both A and B	Neither
Probability	0.4	0.1	0.4	0.1

A producing well will bring the wildcatter a profit of RM1 million and a dry hole will cost him RM100,000. What should he to do when he observes A ? B ? Both? Neither?

52.3 Chapter 3 Utility

The study of decision problem requires a model of ordering among the various possible consequences of taking a decision. With such an ordering, it is the aim of the decision maker to achieve an optimal consequence. We use the term "prospect" to describe what one will be faced with as the result of making a decision. A prospect is quite involved and just about impossible to specify. Thus a mathematical analysis is simpler if it is possible to put prospects, not just in order, but on numerical scale – to assign a numerical measure to each prospect. Such an assignment will be called utility.

- Utility is a function, defined on various prospect, that measure the desirability of the prospect on a numerical scale.
- It should be pointed out that it is indeed not that easy to determine with precision just what that function is in a given case. For example, suppose one is sitting in a restaurant, choosing an entree from the choices steak, chicken, and fish. How to rank one's preference?

52.3.1 3.2 Preference

Every mathematical structure is based on axioms – propositions so basic and elementary that they must be accepted without proof. The axioms to be presented here, dealing with preferences, will imply the existence of a utility function, which provides a numerical scale in terms of which the consequences of actions will be assumed to be measured.

A notion of preference is given as follows:

“ $P_1 \succeq P_2$ ” means: Prospect P_1 is at least as desirable as prospect of P_2 .

“ $P_1 \sim P_2$ ” means: Prospects P_1 and P_2 are equally desirable.

Note: In terms of the notations used, “ $P_1 \sim P_2$ ” means $P_1 \succeq P_2$ and $P_2 \succeq P_1$. However, it doesn’t mean that they are the same.

“ $P_1 \succ P_2$ ” means Prospect P_1 is preferred over prospect P_2 .

Note: “ $P_1 \succ P_2$ ” means “ $P_1 \succeq P_2$ ” but not “ $P_1 \sim P_2$ ”.

The relation “ $\sim$ ” is

- reflexive : $P \sim P$
- symmetry : if $P \sim Q$, then $Q \sim P$
- transitive : if $P \sim Q$ and $Q \sim R$, then $P \sim R$.

Thus the collection of prospects can be partitioned into equivalence classes. It follows that

(a) (a) $P \sim Q$ and $Q \succeq R \Rightarrow P \succeq R$

(b) (b) $P \succeq Q$ and $Q \sim R \Rightarrow P \succeq R$

(c) (c) $P \succ Q$ and $Q \sim R \Rightarrow P \succ R$

(d) (d) $P \sim Q$ and $Q \succ R \Rightarrow P \succ R$

(e) (e) $P \succ Q$ and $Q \succ R \Rightarrow P \succ R$

Mixture of Prospect Let $p_i \geq 0$, $i = 1, 2, \dots$, and $p_1 + p_2 + \dots + p_k + \dots = 1$. If one is faced with probability p_i of prospect Q_i , $i = 1, 2, \dots$, it will be called a random prospect, and in particular it is said to be the mixture of the prospects $Q_1, Q_2, \dots$ and denoted

$$[Q_1, Q_2, \dots, Q_k, \dots]_{(p_1, p_2, \dots, p_k, \dots)}.$$

This mixture can also be denoted by $\sum_i p_i Q_i$. Notice that this notation represents a prospect, not a real number.

Example 52.63. Example 3.2.1 Suppose that a gambling house will pay off a number of dollars equal to the number of tosses of a fair coin necessary to get a heads. Then the prospect facing the bettor is

$$[\$1, \$2, \dots]_{(1/2, 1/2^2, \dots)}$$

52.3.2 3.3 Preference Axioms

Axiom 1: Given any prospects P and Q , then either

$$P \succ Q, \text{ or } P \prec Q, \text{ or } P \sim Q.$$

Axiom 2: If $P \succeq Q$ and $Q \succeq R$, then $P \succeq R$.

Axiom 3: If $P_1 \succ P_2$, then for any $0 < p \leq 1$ and any prospect P ,

$$[P_1, P]_{(p, 1-p)} \succ [P_2, P]_{(p, 1-p)}$$

Axiom 3’: If $P_i \succ Q_i$, $i=1, 2, \dots$, then for any set of probabilities $p_1, p_2, \dots$

$$[P_1, P_2, L]_{(p_1, p_2, L)} \succ [Q_1, Q_2, L]_{(p_1, p_2, L)}$$

eg:sd:3.2.1

Axiom 4: Given three prospects in order $P_1 \succ P_2 \succ P_3$, there exists $0 \leq r, s \leq 1$ such that

$$P_1 \sim [P_1, P_3](r, 1-r) \sim P_2 \sim [P_1, P_3](s, 1-s) \sim P_3$$

Remark:

- Axiom 1 states that, given any two prospects, either one is preferred over the other, or they are equally desirable.
- Axiom 2 is the axiom of transitivity of the preference relation. It precludes a type of inconsistency in the mathematical structure. However, people do sometimes act inconsistently : Choose coffee over tea, tea over milk, and yet choose milk over coffee.
- Axiom 3 and Axiom 3' say that improving some of the prospect in the mixture improves the mixture.

eg:sd:3.3.1

Example 52.64. Example 3.3.1 A young student just finishing college and about to embark on a career considers the following prospects: P_1 : A cash balance of \$100,000, P_2 : A cash balance of \$10,000 P_3 : A cash balance of \$1,000 Then the student would order the prospects as $P_1 \succ P_2 \succ P_3$.

Is there any mixture of P_1 and P_3 that is worse than P_2 ? How about comparing P_2 with the mixture $[P_1, P_3]_{(0.001, 0.999)}$?

52.3.3 3.4 Coding Intermediate Prospects

Question: If we are given a utility function u such that $u(P_1) = 1$ and $u(P_0) = 0$, then how to define $u(P)$ if $P_1 \succ P \succ P_0$? The following result provides an answer to this question.

Theorem 52.65. *Theorem 3.4.1 Given any two prospects P_0 and P_1 such that $P_0 \prec P_1$, and given any prospect such that $P_0 \prec P \prec P_1$, there exists a unique $0 \leq p \leq 1$ such that $[P_1, P_0](p, 1-p) \sim P$. (3.4.1)*

Proof By Axiom 4, there exists r such that

$[P_1, P_0](r, 1-r) \sim P$ Let $S = \{r : 0 \leq r \leq 1, [P_1, P_0](r, 1-r) \sim P\}$. Let p be the least upper bound of S . We claim that (3.4.1) holds. Suppose that $P \prec [P_1, P_0](p, 1-p)$. Then by Axiom 4, there exist $0 \leq s \leq 1$ such that

$$[P_1, P_0](p, 1-p) \sim [[P_1, P_0](p, 1-p), P_1](s, 1-s)$$

However

with

pP

$$[[P_1, P_0](p, 1-p), P_1](s, 1-s) = [P_1, P_0](sp + (1-s)p, s(1-p))$$

$sp + (1-s)p \leq p$. This contradicts to the assumption that p is an upper bound of S .

Similarly, preference $P \prec [P_1, P_0](p, 1-p)$ will to a conclusion that contradicts to the assumption that p is the least upper bound of set S . Remark This result means that the prospect intermediate to two prospect can be thought as a lined up on the scale of numbers from 0 to 1, each being identified with the number on that scale.

Recall that a utility function is a coding utility that gives a numerical representation of preferences. Thus one can construct a utility function by taking the utility of a prospect to be just the number on the scale from 0 to 1 where it happens to fall. In other words, if $P \sim [P_1, P_0]_{(p, 1-p)}$ we define

$$u(Q) = \begin{cases} 0, & Q \sim P_0 \\ p, & Q \sim P \\ 1, & Q \sim P_1 \end{cases} \quad (52.8) \quad \text{e:sd:3.4.2}$$

Example 52.66. Example 3.4.1 A dinner is to choose between : beef , chicken, and fish. His preferences are as follows: Beef f chicken f fish Upon introspection, he decides that a random prospect offering him beef with chance 0.25 and fish 0.75 is precisely as desirable as the prospect of chicken. That is, [Beef, Fish](0.25, 0.75) ~ Chicken If he assigns utility $u(\text{Beef}) = 1$ and $u(\text{Fish}) = 0$, then $u(\text{Chicken}) = 0.25$.

52.3.4 3.5 Properties of Utility Function

Theorem 52.67. *Theorem 3.5.1 Let utility function be defined by (3.4.2). Then for any prospects $P_1 \succ P$, $Q \succ P_0$,*

$$(A) \quad P \succ Q \Rightarrow u(P) \geq u(Q)$$

$$(B) \quad u([P, Q]_{(r, 1-r)}) = ru(P) + (1-r)u(Q)$$

Proof (A) Suppose P and Q are two prospect such that

$$P1fPfQfP0$$

It follows from Theorem 3.4.1 that there exists $0 \leq p, q \leq 1$ such that

$$[P1, P0](p, 1-p) = P \text{ and } [P1, P0](q, 1-q) = Q.$$

Then

$$PfQ \Rightarrow p = u(P) \geq u(Q) = q$$

(B) Suppose that $P1 \succ P \succ Q \succ P0$. Let $1 > p > q > 0$ be two numbers such that

$$P = [P1, P0](p, 1-p)f[P1, P0](q, 1-q) = Q$$

Then

$$[P, Q](r, 1-r) = [[P1, P0](p, 1-p), [P1, P0](q, 1-q)](r, 1-r) = [P1, P0](rp + (1-r)q, r(1-p) + (1-r)(1-q))(**)$$

which implies that

$$u([P, Q](r, 1-r)) = rp + (1-r)q = ru(P) + (1-r)u(Q)$$

Note(**) Write $P = pP1 + (1-p)P0$ and $Q = qP1 + (1-q)P0$, Then

$$[P, Q](r, 1-r) = rP + (1-r)Q = rpP1 + (1-p)P0 + (1-r)qP1 + (1-q)P0 = (rp + (1-r)q)P1 + (r(1-p) + (1-r)(1-q))P0 = [P1, P0](rp + (1-r)q, r(1-p) + (1-r)(1-q))$$

Remark: Property (B) can be expressed in terms of expectation:

The utility of a random prospect is the expected value of the utility As constructed above, the utility of prospect intermediate to two given prospects P1 and P0 is uniquely defined. But any linear function of this utility would also satisfy Properties (A) and (B). Let $v(P) = au(P) + b$ be a linear function of the utility function u with $a \neq 0$. The utility of P1 and P0 in this new scheme would be

$$v(P1) = a+b \quad v(P0) = b \quad \text{Clearly property (A) holds for utility function } v. \text{ Now}$$

$$\begin{aligned} & (\\ &) \\ & (\\ &) \end{aligned}$$

$$v[P, Q](r, 1-r) = av[P, Q](r, 1-r) + b = a(ru(P) + (1-r)u(Q)) + b = r(av(P) + b) + (1-r)(av(Q) + b) = rv(P) + (1-r)v(Q)$$

Hence utility function v also has property (B). Example 3.5.1 With reference to Example 3.4.1, suppose that the utilities for the prospect of the meats are $v(\text{Beef}) = 5$, $v(\text{Chicken}) = 2$, $v(\text{Fish}) = 1$. He would then be indifferent to chicken and a mixture of beef and fish, with probability p of beef, if $v([\text{Beef, Fish}](p, 1-p)) = pv(\text{Beef}) + (1-p)v(\text{Fish}) = v(\text{Chicken})$ $p(5) + (1-p)(1) = 2$

$$\text{or } 4p = 1 \text{ or } p = 0.25.$$

Notice that these utilities are essentially no different than those in Example 2.4.1, being obtainable by the linear transformation $v(P) = 4u(P) + 1$.

52.3.5 3.6 Utility for Arbitrary Prospects

Question: How to define utility for prospect Q that is not between P0 and P1?

We shall consider two cases: (A)

$Q \succ P_1 \succ P_0$: Prospect Utility

(B)

$P_1 \prec P_0 \prec Q$:

$P_0 \sim 0$

$p P_1 \sim p Q \sim 1$

?

$Q \sim ?$

Prospect Utility

8

$p P_0 \sim p P_1 \sim 0$

1

Case A: P_1 can be regarded as a mixture of the prospect P_0 and Q . Since there exists $0 \leq p \leq 1$ such that $P_1 = [Q, P_0]_{(p, 1-p)}$ By property (B) of Theorem 3.5.1, $u(P_1) = pu(Q) + (1-p)u(P_0)$ Because $u(P_0) = 0$ and $u(P_1) = 1$, we have $1 = px$ or $u(Q) =$

$1/p$

Case B: Let $u(Q) = x$. Let $0 \leq p \leq 1$ be chosen such that $P_0 = [P_1, Q]_{(p, 1-p)}$ Then we have $u(P_0) = pu(P_1) + (1-p)u(Q) = 0 = p + (1-p)x$ or

$u(Q) = -$

$p/(1-p)$

Conclusion: The extended utility function u with $u(P_0) = 0$ and $u(P_1) = 1$ is given by

$$u(Q) = \begin{cases} -\frac{p}{1-p}, & Q \prec P_0; P_0 = [P_1, Q]_{(p, 1-p)} \\ 0, & Q \sim P_0 \\ p, & Q = [P_1, P_0]_{(p, 1-p)} \\ 1, & Q \sim P_1 \\ \frac{1}{p}, & Q \prec P_1; P_1 = [Q, P_0]_{(p, 1-p)} \end{cases} \quad (52.9)$$

e:sd:3.6.1

Theorem 52.68. *Theorem 3.6.1 The extended utility function (3.6.1) satisfies the properties (A) and (B) listed in Theorem 3.5.1.*

Proof Let P and Q be two prospects. Let P_0^* and P_1^* be the least desirable and the most desirable prospects, respectively, among the prospect P , Q , P_0 , and P_1 . We seek for a utility function which could have defined by starting with P_0^* and P_1^* resulting in a function $u^*(\cdot)$ which is defined for all prospects between P_0^* and P_1^* , and obey the Property (A) and (B) stated in Theorem 3.5.1.

Case I : Suppose that $R \succ P_1$. Then $P_1 = [R, P_0]_{(1/u(R), 1-(1/u(R)))}$ so that

$$u^*(P_1) = \frac{1}{u(R)}u^*(R) + (1 - \frac{1}{u(R)})u^*(P_0)$$

or

$$u^*(R) = [u^*(P_1) - u^*(P_0)]u(R) + u^*(P_0) \quad (52.10)$$

e:sd:3.6.2

Case II : Suppose that $R \prec P_0$ (and $P_0 \sim pRpP_1^*$).

$$P_0 = [P_1, R]_{(-u(R)/(1-u(R)), 1-u(R))} u^*(P_0) = -u(R)u^*(P_1) + (1-u(R))u^*(R)$$

whence

$$u^*(R) = [u^*(P_1) - u^*(P_0)]u(R) + u^*(P_0) \quad (3.6.2')$$

which is the same as (3.6.2).

We conclude that in either case, $RfP1$ or $RpP0$, $u^*(R)$ and $u(R)$ are linearly related as:

$$u(R) = \alpha u^*(R) + \beta \quad (3.6.3)$$

with $\alpha - 1 = u^*(P1) - u^*(P0)$, $\beta = -\alpha u^*(P0)$ and u^* satisfies Properties (A) and (B) stated in Theorem 2.5.1.

Case (III) Suppose $P0 \succ P1$.

$$R = [P1, P0](r, 1-r)u^*(R) = ru^*(P1) + (1-r)u^*(P0) = r[u^*(P1) - u^*(P0)] + u^*(P0) = [u^*(P1) - u^*(P0)]u(R) + u^*(P0)$$

which is again same as (2.6.2). It follows that for any prospects

$$P0^*pP, QpP1^*u([P, Q](x, 1-x)) = \alpha u^*([P, Q](x, 1-x)) + \beta = \alpha xu^*(P) + (1-x)u^*(Q) + \beta x + \beta(1-x) = x\alpha u^*(P) + (1-x)\alpha u^*(Q) + \beta$$

which is Property (B) for the extended $u(\cdot)$. Moreover,

$$PpQ \Rightarrow u(P) = \alpha u^*(P) + \beta < \alpha u^*(Q) + \beta = u(Q)$$

Remark The utility function appears to be dependent on the prospects $P0$ and $P1$ chosen arbitrarily to be assigned utilities 0 and 1, respectively. The following theorem shows that utility is unique up to a linear transformation. **Theorem 3.6.2** Utility function is unique up to a linear transformation. **Proof** Suppose we start with prospects $Q0 \succ Q1$. Let us consider a special case where $P0$ and $P1$ are two prospects such that $P0 \succ P1 \succ Q0 \succ Q1$

$$, u(P1) = 1, u(P0) = 0$$

Let y and z be defined by the following equivalences $P1 \sim [Q0, P0](y, 1-y) \sim [Q1, P0](z, 1-z)$ so that $u(Q0) =$

$$1 - y \text{ and } u(Q1) = y - z$$

Consider any prospect P (for definiteness, let $P0 \succ P \succ P1$): $P0 \succ P \succ P1 \succ Q0 \succ Q1$ Let p and w be defined such that $P \sim [P1, P0](p, 1-p)$ and $Q0 \sim [Q1, P](w, 1-w)$ so that $u(P) = p$ and $v(P) =$

$$w - 1 - w$$

Then $1 = u(Q0) = wu(Q1) + (1-w)u(P) = w(y - z) + (1-w)p$ whence (solving for w and substituting in $v(P)$) $1 - w - yu(P) - u(Q0) = v(P) = w - 1 - w - 1 - u(Q1) - u(Q0) = y - p$

This shows that $v(P)$ is a linear function of $u(P)$, so constructed that $v(Q0) = 0$ and $v(Q1) = 1$. (Other cases are handled similarly) In summary, there has thus been established the following theorem.

11

Theorem 3.6.3 If Axiom 1 to 4 are accepted, there follows the existence of a finite number $u(P)$ associated with each prospect satisfying the properties (A) $PfQ \Rightarrow u(P) > u(Q)$,

$$(B) u[P, Q](\alpha, 1-\alpha) = \alpha u(P) + (1-\alpha)u(Q)$$

· Convex Combinations · Utility · Survey of probability theory · Personal probability

52.4 No-Data Problem

· Loss and regret · Randomized and mixed Actions · The Minimax principle · Bayes actions · Admissibility

52.5 Topic 3: Using Data in Decisions

· Risk function · Decision problems with finite action space · Randomized decision rules · Construction of Bayes decision procedures · Computation of the posterior distribution · Properties of decision functions

Chapter 53

Applied Statistical Models

TO WORK ON. References:

Montgomery, D.C., Peck, E.A., and Vining, G.G., 2012. Introduction to linear regression analysis. 5th ed. Hoboken, N. J.: Wiley-Interscience.

Makridakis, S., Wheelwright, S.C., and Hyndman, R.J., 1998. Forecasting: methods and applications. 3rd ed. Hoboken, N. J.: John Wiley & Sons.

Neter, J., Kutner, M.H., and Nachtsheim, C.J., 2013. Applied linear regression models. 5th ed. New York: McGraw-Hill.

Diebold, F.X., 2006. Element of forecasting. 4th ed. Mason, Ohio: Thomson/South-Western.

Chatfield, C., 2003. The analysis of time series: an introduction. 6th ed. London: Chapman & Hall.

Krause, A. and Olson, M., 2005. The basics of S-Plus. 4th ed. New York: Springer.

53.1 Multiple Linear Regression

sec:lin-reg

Merging the simple linear regression model to linear model!!!

Simple linear regression model.

- Estimation of parameters using method of least squares.
- Forecasting using the simple regression model.
- Interval estimation of the regression coefficients, variance and mean response.
- Prediction interval for future observation.
- Hypothesis test for regression coefficients.
- Linear correlation coefficient.
- Coefficient of determination.

Linear Regression

- Multiple regression model.
- Estimation of regression coefficients by method of least squares.
- Estimation of variance.
- Confidence intervals on the regression coefficients, mean response.
- Prediction interval for future observation.
- Hypothesis test for regression coefficient.
- Linear correlation coefficient.

53.2 Model Adequacy Checking

· Residual analysis: Methods for scaling residuals, residual plots, partial regression and partial residual plots.

- Direction and treatment of outliers.
- Test for lack of fit.
- Estimation of pure error from Near-Neighbours.

53.3 Transformations and Weighting to Correct Model Inadequacies

- Variance-stabilizing transformations.
 - Transformations to linearize model.
 - Transformations on y using Box-Cox model.
 - Transformations on the regressor variables.
 - Generalized and weighted least squares.

53.4 Exponential Smoothing Methods

- The forecasting scenario.
 - Averaging methods.
 - Single exponential smoothing and adaptive approach.
 - Holt's linear method.
 - Holt-Winters' trend and seasonality method.

53.5 Nonlinear Regression

53.6 MA, AR, ARMA, the Box-Jenkins Methodology for ARIMA Models and the Autocorrelation Function

- A white noise model. · The sampling distribution of autocorrelations. · Portmanteau test.
- The partial autocorrelation coefficient. · Examining stationarity of time series data. · MA, AR, ARMA and ARIMA models for times series data. · Identification. · Diagnostic checking. · Forecasting with MA, AR, ARMA and ARIMA.

Probably Relevant here or statistics?

<https://github.com/dicook/StatisticalThinking>

This repository is housing material for the course ETC2420/ETC5242 Statistical Thinking using Randomisation and Simulation taught in Spring 2017 in the Department of Econometrics and Business Statistics at Monash University.

* Topic 1: Simulation of games for decision strategies (2 weeks):

- Week 1, Class 1. Case studies in randomization using Australian election. What is randomness? (Include the draw vs flip coin tosses) - Week 1, Class 2. Case studies in randomization (Ch 2, Diez, Barr, Cetinkaya-Rundel). Hypothesis testing I.

- Lab 1: Introduction to R, functions, permutation, random number generation

- Week 2, Class 1. Case studies in randomization (Ch 2, Diez, Barr, Cetinkaya-Rundel). Hypothesis testing II. - Week 2, Class 2. Decision theory. Computing probabilities of outcomes. Zero-sum two-person: adding reward and loss, saddle points, domination. Criteria for making decisions: minimax, Bayes.

- Lab 2: Simulate Monty Hall in R

Vocabulary: permutation, association, hypothesis, p-value, pseudo-random number generator, simulation, event, probability, zero-sum two-person game, saddle point, domination, minimax, Bayes criterion

* Topic 2: Statistical distributions for decision theory (1.5 weeks)

- Week 3, Class 1: Random numbers Mapping random numbers to events for simulation Statistical distributions READING: CT6, Section 1.3-1.9

- Lab 3: Hypothesis testing using permutation

- Week 3, Class 2: Random variables Central limit theorem Density functions Quantiles Estimation Goodness of fit

- Week 4, Class 1:

- Lab 4: Finding the best distribution to model olympic medals, estimate number of medals Australia will earn?
- * Topic 3: Linear models for credibility theory (1.5 weeks) (Linear models)
 - review of regression - weighted regression - resampling methods for assessing parameter estimates: bootstrap - repeated measures, mixed effects models
- * Topic 4: Compiling data to problem solve (2 weeks)
 - types of data: sampling, survey, observational, experimental - working with temporal data, dates, times, seasonality, covariates - longitudinal data - working with maps and spatial data: chloropleth, point processes

Vocabulary: Data, information; population, sample; case, subject, sample, variable, feature; quantitative, integer, real-valued, categorical, ordinal, temporal, spatial,
- * Topic 5: Bayesian statistical thinking (1.5 weeks) - Charpentier Ch 3
 - (i) Introduction to Bayesian methods (ii) Conjugate priors, small sample examples (iii) MCMC (iv) Bayesian regression, and credibility
- * Topic 6: Temporal data and time series models (1.5 weeks)
 - Modeling time, autocorrelation, cross-correlation - Prospective life tables (Charpentier Ch 8)
- * Topic 7: Modeling risk and loss, with data and using randomization to assess uncertainty (2 weeks)

Chapter 54

Design and Analysis of Experiment

Montgomery, D.C., 2012. Design and analysis of experiments. 8th ed. New York: John Wiley.

Kuehl, R.O., 2000. Design of experiments: statistical principles of research design and analysis. 2nd ed. Pacific Grove, CA: Duxbury/Thomson Learning.

Berger, P.D. and Maurer, R.E., 2002. Experimental design. Belmont, CA: Duxbury/Thomson Learning.

54.1 Introduction

- Guidelines for designing experiments
 - Basic statistical concepts
 - Sampling and sampling distribution
 - Inferences about the differences in means, (randomized designs)
 - Inferences about the differences in means, (paired comparison designs)
 - Inferences about the variances of normal distributions

54.2 Experiments with single factor

- The analysis of variance
 - Analysis of the fixed effects model
 - Model adequacy checking
 - Practical interpretation of results
 - The random effects model
 - Choice of sample size
 - Discovering dispersion effects
 - Fitting response curves in the single-factor model
 - The regression approach to the analysis of variance

54.3 Randomized blocks, Latin squares and related designs

- Randomized complete block design
 - The Latin square design
 - The Graeco-Latin square design
 - Balanced incomplete block designs

54.4 Introduction to factorial designs

- Basic definitions and principles
 - The 2-factor factorial designs

- The general factorial designs
- Fitting response curves and surfaces
- Blocking in a factorial design

54.5 The 2^k factorial designs

- Sample statistic and its distribution
 - Standard error
 - Central Limit Theorem

Chapter 55

Statistical Simulation

1. Dagpunar, J.S., 2007. Simulation and Monte Carlo: With applications in finance and MCMC. New Jersey: Wiley-Interscience.
2. Ross, S.M., 2012. Simulation. 5th ed. San Diego, CA: Academic Press. ', 'Additional References:
3. Glasserman, P., 2010. Monte Carlo Methods in Financial Engineering. New York: Springer Verlag.
4. Chan, N.H. and Wong, H. Y., 2015. Simulation Techniques in Financial Risk Management. 2nd ed. New Jersey: Wiley-Interscience.
5. Kennedy, W. J. and Gentle, J. E., 1980. Statistical Computing. New York: Marcel Dekker.

55.1 Monte Carlo Approach

- Principles of Monte Carlo
 - Example Problems Solved Using Monte Carlo Simulation
 - Efficiency of Simulation Estimators

55.2 Random Number Generation

- Linear Congruential Generators (Section [3.9](#))
 - Theoretical tests for random numbers
 - Empirical tests.
 - Combinations of generators.
 - General methods for generating random variates
 - Inversion methods
 - Acceptance-Rejection Method
 - Generation of variates from standard distribution
 - Standard normal distribution
 - Lognormal distribution
 - Bivariate normal distribution
 - Generalized inverse Gaussian distribution

55.3 Variance Reduction Techniques

- Antithetic variates
 - Stratified sampling
 - Control variates
 - Latin Hypercube

55.4 Simulation of Stochastic Processes

- Brownian Motion
 - Asset price movements
 - Pricing simple derivatives and options
 - Option
 - Asian options
 - Basket options
 - Stochastic volatility

55.5 Discrete Event Simulation

- Poisson process
 - Time-dependent Poisson process
 - Poisson processes in the plane
 - Markov chains
 - Bayesian statistics
 - Markov chains and Metropolis-Hasting (MH) algorithms
 - Gibbs sampling

55.6 Applications in Finance and Insurance

Refer to Computational Finance???

- Interest rate modeling
 - Introduction
 - Computer implementation of interest rate models
- Options pricing
 - Path Dependent options
 - American options
- Risk Management
 - Value at risk (VaR)
 - Modeling heavy tails
 - Credit risk
- Insurance company solvency
 - Risk based capital
 - Solvency margin
 - Reinsurance strategies
 - Investment strategies
- Other factor

Chapter 56

Applications of Statistics to Econometrics

56.1 Non-negative Matrices

Non-negative matrices occur naturally in Markov Chains. Some models of Genetics are based on non-negative matrices. Leontief models in Economics derive sustenance from non-negative matrices. One of the most prominent results in the area of non-negative matrices is the Perron-Frobenius Theorem.

56.1.1 Theory

d:15.1.1

Definition 56.1. Let $A = (a_{ij}) \in M_n$ be Hermitian.

1. The matrix A is called *non-negative* if $a_{ij} \geq 0$ for all i, j . It is denoted as $A \geq_e 0$, or simply $A \geq 0$.
2. The matrix A is called *positive* if $a_{ij} > 0, \forall i, j$. It is denoted $A >_e 0$, or simply $A > 0$.

Remark 56.1. In most linear algebra books, $A \geq 0$ and $A > 0$ are used to indicate that A is *non-negative definite* and *positive definite* respectively.

The concepts of nonnegativity and positivity perfectly make sense for matrices not necessarily square. If $A = (a_{ij})$ and $B = (b_{ij})$ are two matrices of the same order, we say that $A \geq_e B$ or $A - B \geq_e 0$ if $a_{ij} \geq b_{ij}$ for all i, j .

Proposition 56.1. A is non-negative iff there exists a matrix U s.t. $A = U^*U$, where U is s.t. $\text{rank}A = \text{rank}U$. A is positive if $\text{rank}U = n$.

Proof. This proposition follows from the spectral decomposition of A :

$$A = \lambda_1^2 P_1 P_1^* + \cdots + \lambda_k^2 P_k P_k^*,$$

where $k = \text{rank}A$. In this form, $U = \lambda_1 P_1 P_1^* + \cdots + \lambda_k P_k P_k^*$. □

Proposition 56.2. Let $Y \in M_{n,r}$ and let $Y^\perp$ be the matrix whose columns are orthogonal to those of Y and s.t. $\text{rank}Y + \text{rank}Y^\perp = n$. Then

$$P_Y = Y(Y^*Y)^{-1}Y^* \tag{56.1}$$

e:14.1.2

and $I - P_Y$ are both non-negative.

Proposition 56.3. Let A be non-negative and consider a partition

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \quad (56.2) \quad \text{e:14.1.3}$$

where A_{11} and A_{22} are square matrices. Then $A_{11} - A_{12}A_{22}^{-1}A_{21}$ is non-negative and unique for any g -inverse A_{22}^{-} .

Proposition 56.4. Let A be positive and partition A^{-1} as

$$A^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{bmatrix} \quad (56.3)$$

where A_{11} is a square matrix. Then

1. $A^{11} = (A_{11} - A_{12}A_{22}^{-1}A_{21})$,
2. $A^{11} - A_{11}^{-1}$ is non-negative.

Definition 56.2. A non-negative matrix $A \in M_n (n \geq 2)$ is called *reducible* if there exists a permutation matrix $P \in M_n$ s.t.

$$PAP' = \begin{bmatrix} B & 0 \\ C & D \end{bmatrix}$$

where $B \in M_r$, $D \in M_{n-r}$ and $0 \in M_{r,n-r}$ is the null matrix and $r > 1$.

In the following propositions, let $n \geq 2$.

Proposition 56.5. Let $A \in M_n$ be non-negative. A is reducible iff there exists a nonempty proper subset I of $\{1, \dots, n\}$ s.t. $a_{ij} = 0$, $\forall i \notin I$ and $j \in I$.

One of the characteristic features of reducible matrices is the following. Suppose A is in the reduced form. Then A^k , for any $k \in \mathbb{N}^*$, is also reducible.

Definition 56.3. A non-negative matrix $A \in M_n$ is called *irreducible* if it is not reducible.

Proposition 56.6. Let $A = (a_{ij}) \in M_n$ be non-negative. The following statements are equivalent.

1. A is irreducible.
2. $(I + A)^{n-1}$ is positive.
3. For any i and j with $1 \leq i \leq n$ and $i \neq j \leq n$, there exists a positive integer k s.t. $k \leq n$ and $a_{ij}^{(k)} > 0$, where $A^k = (a_{ij}^{(k)})$.

□

There is another concept closely related to irreducibility. For primitive matrices, one could obtain a stronger version of Perron-Frobenius Theorem.

Definition 56.4. A non-negative matrix A is called *primitive* if A^k is positive for some positive integer k .

Let $A = (a_{ij})$. Define $m(A) = (|a_{ij}|)$.

Remark 56.2. Every primitive matrix is irreducible but the reverse is not true.

Proposition 56.7. 1. If A and B are two matrices s.t. AB is defined, then $m(AB) \leq_e [m(A)][m(B)]$.

2. If A is a square matrix, then $m(A^k) \leq_e [m(A)]^k$ for all positive integers k .

3. If A, B are square matrices of the same order s.t. $m(A) \leq_e m(B)$, then $\|m(A)\|_F \leq \|m(B)\|_F$.

p:15.1.7

Proposition 56.8. If $A, B \in M_n$ and $m(A) \leq B$, then $\rho_s(A) \leq \rho_s(m(A)) \leq \rho_s(B)$. In other words, the spectral radius is monotonically increasing on the set of all non-negative matrices in M_n .

p:15.1.8

Proposition 56.9. Let $A = (a_{ij}) \in M_n$ be non-negative. and B a principal submatrix of A . Then $\rho_s(B) \leq \rho_s(A)$. In particular, $\max_{1 \leq i \leq n} a_{ii} \leq \rho_s(A)$.

p:15.1.9

Proposition 56.10. Let $A = (a_{ij})$ be a non-negative matrix s.t. all the row sums of A are equal to the same number α . Then $\rho_s(A) = \alpha$. If all the column sums of A are equal to the same number β , then $\rho_s(A) = \beta$.

p:15.1.10

Proposition 56.11. Let $A = (a_{ij})$ be a non-negative matrix s.t. all the row sums are $r_1, \dots, r_n$ and column sums are $c_1, \dots, c_n$. Then

p:15.1.11

$$\min_{1 \leq i \leq n} r_i \leq \rho_s(A) \leq \max_{1 \leq i \leq n} r_i, \quad \min_{1 \leq i \leq n} c_i \leq \rho_s(A) \leq \max_{1 \leq i \leq n} c_i.$$

Proposition 56.12. Let $A = (a_{ij})$ be a non-negative matrix. Let $x' = (x_1, \dots, x_n)$ be positive. Then

p:15.1.12

$$1. \min_{1 \leq i \leq n} \frac{1}{x_i} \sum_{j=1}^n a_{ij} x_j \leq \rho_s(A) \leq \max_{1 \leq i \leq n} \frac{1}{x_i} \sum_{j=1}^n a_{ij} x_j,$$

$$2. \min_{1 \leq i \leq n} x_j \sum_{i=1}^n \frac{a_{ij}}{x_i} \leq \rho_s(A) \leq \max_{1 \leq j \leq n} x_j \sum_{i=1}^n \frac{a_{ij}}{x_i}.$$

Further, if $\alpha, \beta \geq 0$ and $\alpha x \leq_e Ax \leq_e \beta x$, then $\alpha \leq \rho_e(A) \leq \beta$. If $\alpha x < Ax$, then $\alpha < \rho_s(A)$; if $Ax <_e \beta x$, then $\rho_s(A) < \beta$. $\square$

Theorem 56.3 (Perron-Frobenius Theorem for Positive Matrices). Let $A = (a_{ij})$ be a positive matrix. Then the following statements are valid.

p:15.1.16

1. $\rho_s(A) > 0$. [For general matrix, it might be zero]
2. $\rho_s(A)$ is an eigenvalue of A . [For general matrix, all its eigenvalues may be < 0 .]
3. There exists a positive eigenvalue of A corresponding to the eigenvalues $\rho_s(A)$.
4. If μ is any other eigenvalue of A , then $|\mu| < \rho_s(A)$.
5. The multiplicity of the eigenvalue $\rho_s(A)$ is one.

Proof. 1. Since A is positive, by Proposition [56.11](#), $\rho_s(A) > 0$.

2. Let μ be an eigenvalue of A s.t. $|\mu| = \rho_s(A)$. Let x be a corresponding eigenvector and $|x| = m(x)$. Then

$$\rho_s(A)|x| = |\mu||x| = |\mu x| = |Ax| \leq_e |A||x| = A|x|.$$

Let $y = A|x| - \rho_s(A)|x|$. It is clear that $y \geq_e 0$.

If $y = 0$, then $\rho_s(A)$ is an eigenvalue of A with a corresponding eigenvector $|x|$. Since

$|x| \geq_e 0$, $|x| \neq 0$, and A is positive, $A|x| >_e 0$. By choosing $|x| = [\rho_s(A)]^{-1}A|x| >_e 0$, this establishes 2.

If $y \neq 0$. Set $z = A|x|$. It is positive. Since $A >_e 0$,

$$0 <_e Ay = A(A|x| - \rho_s(A)|x|) = Az - \rho_s(A)z,$$

which means that $Az >_e \rho_s(A)z$. By Proposition [56.12](#), $\rho_s(A) > \rho_s(A)$, which is impossible. Hence y has to be 0.

3. The proof is the same as 2.

4. Let μ be an eigenvalue of A s.t. $\mu \neq \rho_s(A)$. By the very definition of spectral radius, $|\mu| \leq \rho_s(A)$. We claim that $|\mu| < \rho_s(A)$.

Suppose $|\mu| = \rho_s(A)$. Let x be an eigenvector of A corresponding to the eigenvalue μ . Following the argument presented in the proof in 2., it follows that $|x| >_e 0$ and $|x|$ is an eigenvector of A corresponding to the eigenvalue $\rho_s(A)$. Let x_i be the i -th component of x . The equation $Ax = \mu x$ implies that

$$\rho_s(A)|x_i| = |\mu||x_i| = |\mu x_i| = \left| \sum_{j=1}^n a_{ij}x_j \right| \leq \sum_{j=1}^n a_{ij}|x_j| = \rho_s(A)|x_i|,$$

for each i . Thus equality must prevail throughout in the above. This means that the complex numbers $a_{ij}x_j$, $j = 1, \dots, n$ must lie on the same ray in the complex plane. Let θ be their common argument. Then $e^{-i\theta}a_{ij}x_j > 0$ for all j . Since $a_{ij} > 0$, we have $\omega = e^{-i\theta}x >_e 0$. The vector ω is also an eigenvector of A corresponding to the eigenvalue μ of A , i.e., $A\omega = \mu\omega$. Since $\omega >_e 0$, $\mu \geq 0$. Trivially, $\mu\omega \leq_e A\omega \leq_e \mu\omega$. By Proposition [56.12](#), $\mu \leq \rho_s(A) \leq \mu$. This contradiction establishes the claim.

5. ① Claim: If $0 <_e C \leq_e D$ and $\rho_s(C) = \rho_s(D)$, then $C = D$.

There exist positive eigenvectors x and y such that $Cx = \rho_s(C)x$ and $Dy = \rho_s(D)y$. We would like to show that x is also an eigenvector of D corresponding to the eigenvalue $\rho_s(C)$ of D . Note that $0 <_e \rho_s(C)x = \rho_s(D)x = Cx \leq_e Dx$. Let $z = Dx - \rho_s(D)x$. It is clear that $x \geq_e 0$. We claim that $z = 0$. Suppose $z \neq 0$. Let $y = Dx$. Note that $0 <_e Dz = D(y - \rho_s(D)x) = Dy - \rho_s(D)y$ leading to the inequality $Dy >_e \rho_s(D)y$. By Proposition [56.12](#), $\rho_s(D) < \rho_s(D)$. This contradiction establishes that $z = 0$. Thus the vector x is a common eigenvector of C and D corresponding to the same eigenvalue $\rho_s(C)$, i.e. $Cx = Dx$. Since $x >_e 0$ and $C \leq_e D$, we have $C = D$.

② Let $p(x) = |xI - A|$. Let A_{ii} be the principal submatrix of A obtained from A by deleting the i -th row and i -th column of A . Write

$$|xI - A_{ii}| = \prod_{j=1}^{n-1} (x - \mu_j),$$

where $\mu_1, \dots, \mu_{n-1}$ are the eigenvalues of A_{ii} . Since $\rho_s(A_{ii}) \leq \rho_s(A)$, by Proposition [56.9](#), $|\mu_i| \leq \rho_s(A)$ for every i . Since A is positive, $\rho_s(A_{ii}) < \rho_s(A)$. Equivalently, $|\mu_i| < \rho_s(A)$ for every i . Consequently, $|xI - A_{ii}| > 0$ if $x \geq \rho_s(A)$. In particular, $|\rho_s(A)I - A_{ii}| > 0$ for all i . Observe that

$$\frac{d}{dx}p(x)\Big|_{x=\rho_s(A)} = \sum_{i=1}^n |\rho_s(A)I - A_{ii}| > 0.$$

This shows that the eigenvalue $\rho_s(A)$ is simple. □

Theorem 56.4 (Perron-Frobenius Theorem). *Let $A = (a_{ij})$ be a non-negative irreducible matrix. Then:*

1. $\rho_s(A) > 0$.
2. $\rho_s(A)$ is an eigenvalue of A .
3. There exists a positive eigenvector x of A corresponding to the eigenvalue $\rho_s(A)$ of A .
4. The eigenvalue $\rho_s(A)$ is simple.

56.1.2 Leontief Models in Economics

The following examples can be found in Applications from C. Rorres, H. Anton (1984), *Applications of Linear Algebra*, 3e, John Wiley.

Closed Economy:

Suppose an economy has n industries and each industry produces only one commodity. Each industry requires commodities from all the industries, including itself. No input from outside the economy under focus is needed. The problem is to determine suitable “prices” to be charged for these commodities so that to each industry total expenditure equals total income. Such a price structure represents equilibrium for the economy. Let a_{ij} be the fraction of the total output of the j -th industry purchased by the i -th industry, $i, j = 1, \dots, n$. Clearly, $a_{ij} \geq 0$. Further, $a_{1j} + \dots + a_{nj} = 1$, $j = 1, \dots, n$. Let $A = (a_{ij})$, then A is a non-negative with each column sums up to unity. Consequently, $\rho_s(A) = 1$. Let p_i be the price for the i -th industry for its total output, $i = 1, \dots, n$. The equilibrium condition can be stated as follows:

$$a_{i1}p_1 + \dots + a_{in}p_n = p_i, \quad i = 1, \dots, n.$$

In matrix form, we have $Ap = p = \rho_s(A)p$, $p = (p_1, \dots, p_n)'$. A is called the *input-output matrix of the economy*. The objective is to determine the price vector p .

Hence, the “closed economy” problem is transformed to finding the eigenvector corresponding to eigenvalue $\rho_s(A)$ of positive matrix problem.

Example 56.5. Suppose an economy has 4 industries: a steel plant, an electricity generating plant, a coal mine, and an iron ore mining facility. 20% of the output of the steel plant is used by itself, 30% of the output of the steel plant is used by the electricity generating plant, 15% by the coal mine, and 35% by the iron ore facility. 20% of the electricity generated is used by the steel plant, 25% by itself, 25% by the coal mine, and 30% by the iron ore facility. 30% of the coal produced by the coal mine is used by the steel plant, 30% by the electricity generating plant, 20% by itself, and 20% by the iron ore facility. Finally 80% percent of iron ore produced by the iron ore mining facility is used by the steel plant, 10% by the electricity generating plant, 10% by the coal mine, and nothing for itself. Determine p .

Solution. The input-output matrix is

$$A = \begin{bmatrix} 0.20 & 0.20 & 0.30 & 0.80 \\ 0.30 & 0.25 & 0.30 & 0.10 \\ 0.15 & 0.25 & 0.20 & 0.10 \\ 0.35 & 0.30 & 0.20 & 0.00 \end{bmatrix}$$

Solving

$$(A - I)p = \begin{bmatrix} -0.80 & 0.20 & 0.30 & 0.80 \\ 0.30 & -0.75 & 0.30 & 0.10 \\ 0.15 & 0.25 & -0.80 & 0.10 \\ 0.35 & 0.30 & 0.20 & -1.00 \end{bmatrix} p = 0$$

using g-inverse (Proposition 4.15), we have

$$G = \begin{bmatrix} -0.51001246 & 0.34433457 & 0.21085197 & -0.04517408 \\ 0.29457841 & -0.77784061 & 0.10609758 & 0.37716463 \\ 0.3373457 & 0.15500836 & -0.84139111 & 0.34903704 \\ 0.2273383 & 0.16816659 & 0.18734924 & -0.58285413 \end{bmatrix}$$

$$p = (I_n - GA)z = \begin{bmatrix} 0.4728728 & 0.32109265 & 0.22741976 & 0.30731723 \\ 0.32109265 & 0.21803007 & 0.1544238 & 0.20867621 \\ 0.22741976 & 0.1544238 & 0.10937348 & 0.14779875 \\ 0.30731723 & 0.20867621 & 0.14779875 & 0.19972364 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} \approx \begin{bmatrix} 0.4728728 \\ 0.32109265 \\ 0.22741976 \\ 0.30731723 \end{bmatrix} t$$

where z is an arbitrary vector and $a, b, c, d, \in \mathbb{R}$. □

Open Economy:

Suppose an economy has n industries and each industry produces only one commodity. Portions of these outputs are to be used in the industries within the economy *but there is also some demand for these commodities outside the economy*. The prices of units of these commodities are fixed and known.

Let d_i be the monetary value of the output of the i -th industry demanded by sources outside the economy, $i = 1, \dots, n$. Let $d' = (d_1, \dots, d_n)$. Since the prices are known, the sources outside the economy can compute how much is the monetary value of the commodities they are seeking from the economy. Denote by c_{ij} , the monetary value of the output of the i -th industry needed by the j -th industry in order to produce one unit of monetary value of its output, $i, j = 1, \dots, n$. Let $C = (c_{ij})$. C is non-negative. The matrix C is called the *consumption matrix of the economy*. Normally, $\sum_{i=1}^n c_{ij} \leq 1$, for each j . If the sum is 1, the industry is not profitable. Finally, let x_i be the monetary value of the total output the i -th industry, $i = 1, \dots, n$.

The **objective** is to determine the values of x_i 's so that the needs of the industries within the economy and the demands from sources outside the economy are exactly met.

Let $x' = (x_1, \dots, x_n)$ and *non-negative*. Note that $\sum_{j=1}^n c_{ij}x_j$ is the monetary value of the output of the i -th industry needed by all the industries inside the economy. Consequently, $x - Cx$ represents the monetary values of excess output with the demand. The objective is to determine $x \geq 0$ so that

$$(I - C)x = d \tag{56.4}$$

e:15.2.2

is satisfied.

Theoretical result on the analysis of equation e:15.2.2 56.4: Let $C \in M_n$ be a non-negative matrix.

- $(I - C)^{-1}$ exists and is non-negative iff there exists a non-negative vector x s.t. $x > Cx$.
[Interpretation: there is some production schedule x s.t. each industry produces more than it consumes.]
- If each of the row (or column) of sums of C is less than 1, then $(I - C)^{-1}$ exists and is non-negative.

The essential differences between the closed model and open model are

1. In the closed model, the outputs of the industries are distributed among themselves. In the open model, an attempt is made to satisfy an outside demand for the outputs.
2. In the closed model, the outputs are fixed and the problem is to determine a price structure for the outputs so that the total expenditure and total income for each industries are equal. In the open model, the prices are fixed and the problem is to determine a production schedule meeting the internal and external demands.

Example 56.6. A town has three main industries: a coal mine, an electricity generating plant, and a local railroad. To mine \$1 worth of coal, the mining operation needs 25¢ worth of electricity and makes use of 25¢ worth of transportation facilities. To produce \$1 worth electricity, the generating plant requires 65¢ worth of coal, 5¢ worth of its own electricity, and 5¢ worth of transportation needs. To provide \$1 worth of transportation, the railroad requires 55¢ of coal for fuel and 10¢ worth of electricity. In a particular period of operation, the coal mine receives orders for \$50,000 of coal from outside and the generating plant receives orders for \$25,000 of electricity from outside. Determine how much each of these industries should produce in the period under focus so that internal and external demands are exactly satisfied.

Solution.

□

56.1.3 Markov Chains

Let $X_0, X_1, \dots$ be a stochastic process, i.e., a sequence of random variables. Assume that each random variable takes values in a finite set $\{1, \dots, k\}$. The set $\{1, \dots, k\}$ is called the *state space* of the process and members of the set are called the *states of the process*.

For every $n \geq 0$ and $i = 1, \dots, k$, let $p_i = P(X_0 = i)$. $p' = (p_1, \dots, p_k)$ is called the *initial distribution* of the process, i.e. p is the distribution of X_0 .

A physical evolution of a stochastic process can be described as follows. Suppose a particle moves over the states at times 0, 1, 2, $\dots$ in a random manner. Let X_n be the state in which the particle is at time n , $n \geq 0$. The joint distribution of the process X_n , $n \geq 0$ describes the random movement of the particle over time.

Definition 56.5. The process X_n , $n \geq 0$ is called a *Markov chain* if the conditional probability

$$\begin{aligned} & P(X_{n+1} = j | X_0 = i_0, X_1 = i_1, \dots, X_{n-1} = i_{n-1}, X_n = i) \\ &= \frac{P(X_0 = i_0, X_1 = i_1, \dots, X_{n-1} = i_{n-1}, X_n = i, X_{n+1} = j)}{P(X_0 = i_0, X_1 = i_1, \dots, X_{n-1} = i_{n-1}, X_n = i)} \\ &= P(X_{n+1} = j | X_n = i) \\ &= \frac{P(X_n = i, X_{n+1} = j)}{P(X_n = i)} =: p_{ij} \end{aligned}$$

d:15.3.1

for all $i_0, i_1, \dots, i_{n-1}, i, j \in \{1, \dots, k\}$ and $n \geq 0$. The numbers p_{ij} 's are called *one-step transition probabilities*. The number p_{ij} is the conditional probability of moving from the state i to state j in one step from time n to time $n + 1$.

Let $P = (p_{ij})$. The matrix P is called the *transition probability matrix*. It has a special property $\sum_{j=1}^k p_{ij} = 1$ for every i .

The knowledge of initial distribution p and the transition matrix P is enough to determine the joint distribution of any finite subset of the random variables X_n , $n \geq 0$:

$$X_n \sim (p')P^n.$$

Proposition 56.13. Let $q' = (q_1, \dots, q_k)$ be the left Perron vector of P , i.e. $q > 0$ and $q'P = q'$ and $\sum_{s=1}^k q_s = 1$. The right Perron vector of A is $x' = (1, 1, \dots, 1)$, i.e., $Px = x$.

If the transition matrix P is primitive, then

$$\lim_{n \rightarrow \infty} P^n = Q,$$

where all the rows of Q are identical and equal to q' . Consequently, the limiting distribution of X_n is q' , i.e. $\lim_{n \rightarrow \infty} p'P^n = q'$.

Remark 56.7. This is the fundamental theorem of Markov chains. It asserts that what ever the initial distribution p may be, the limiting distribution of X_n is q' , the left Perron vector of P . For primitive stochastic matrices, obtaining the left Perron vector of P is tantamount to solving the equations $q'P = q'$ in unknown q .

Example 56.8. Show that the transition matrix

$$\begin{bmatrix} 0.0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.0 \\ 0.5 & 0.0 & 0.5 \end{bmatrix}$$

is primitive. Obtain the limiting distribution of the Markov chain driven by the above transition matrix.

Solution. By definition, a non-negative matrix A is called *primitive* if A^k is positive for some positive integer k . In this case, let the matrix be P we calculate

$$P^2 = \begin{bmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}$$

which is a positive matrix with $k = 2$. Hence P is primitive.

Solving $q'P = q'$ or $P'q = q$, we obtain

$$q' = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}].$$

□

56.1.4 Genetic Models

Basic Problems in Genetics: Examine the propagation of traits over several generations of a population.

Approximate Fact: Inherited trait is governed by genes in chromosomes.

Consider the list of possible genotype of offspring along with their probabilities.

Genotypes of parents	Genotype of offspring		
	AA	Aa	aa
AA & AA	1	0	0
AA & Aa	1/2	1/2	0
AA & aa	0	1	0
Aa & Aa	1/4	1/2	1/4
Aa & aa	0	1/2	1/2
aa & aa	0	0	1

Note that no distinction is made of (father,mother) pairings of genotypes (AA, Aa) and (Aa,AA).

Let us look at a simple inbreeding model. Suppose in the 0-th generation, the population consists of a proportion of p_0 individuals of genotype AA, q_0 of genotype Aa, and r_0 of genotype aa. Clearly, $p_0 + q_0 + r_0 = 1$. Suppose mating takes place between individuals of the same genotype only. Let in the population,

- p_n = proportion of genotype AA in the n -th generation,
- q_n = proportion of genotype Aa in the n -th generation,
- r_n = proportion of genotype aa in the n -th generation.

We would like to determine the limiting behaviour of these proportions. In the first generation, we will have $p_1 = p_0 + \frac{1}{4}q_0$, $q_1 = \frac{1}{2}q_0$, $r_1 = r_0 + \frac{1}{4}q_0$. These equations can be rewritten as

$$\begin{bmatrix} p_1 \\ q_1 \\ r_1 \end{bmatrix} = \begin{bmatrix} 1 & 1/4 & 0 \\ 0 & 1/2 & 0 \\ 0 & 1/4 & 1 \end{bmatrix} \begin{bmatrix} p_0 \\ q_0 \\ r_0 \end{bmatrix} =: Ax_0.$$

Let $x_n = (p_n, q_n, r_n)'$, $n \geq 0$. It is clear that the vectors x_n 's are governed by the equation

$$x_n = A^n x_0.$$

Note that A is a **non-negative matrix**.

Determining the limiting behaviour of the sequence A^n , $n \geq 0$.

The eigenvalues of A are $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 1/2$ with the corresponding eigenvectors

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}.$$

Let $P = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & -2 \\ 0 & 1 & 1 \end{bmatrix}$, $\Lambda = \text{diag}(1, 1, 1/2)$. Then $P^{-1} = \frac{1}{2} \begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & 2 \\ 0 & -1 & 0 \end{bmatrix}$ and $A = P\Lambda P^{-1}$ implies

$$A^n = P\Lambda^n P^{-1} = \begin{bmatrix} 1 & 1/2 - (1/2)^{n+1} & 0 \\ 0 & (1/2)^n & 0 \\ 0 & 1/2 - (1/2)^{n+1} & 1 \end{bmatrix}$$

Hence,

$$p_n = p_0 + \left(\frac{1}{2} - \left(\frac{1}{2}\right)^{n+1}\right)q_0, \quad q_n = \left(\frac{1}{2}\right)^n q_0, \quad r_n = r_0 + \left(\frac{1}{2} - \left(\frac{1}{2}\right)^{n+1}\right)q_0.$$

Consequently

$$\lim_{n \rightarrow \infty} p_n = p_0 + \frac{q_0}{2}, \quad \lim_{n \rightarrow \infty} q_n = 0, \quad \lim_{n \rightarrow \infty} r_n = p_0 + \frac{q_0}{2}.$$

In the long run, individuals of genotype Aa will disappear! Only pure genotypes AA and aa will remain in the population.

Let us look at another model called selective breeding model. In the 0-th generation, a population has a proportion of p_0 individuals of genotype AA, q_0 of genotype Aa, and r_0 of genotype aa. In a special breeding program, all individuals are mated with individuals of genotype AA only. Let in the population,

- p_n = proportion of genotype AA in the n -th generation,
- q_n = proportion of genotype Aa in the n -th generation,
- r_n = proportion of genotype aa in the n -th generation.

We would like to examine the limiting behaviour of these proportions. In the first generation, we have

$$p_1 = p_0 + \frac{1}{2}q_0, \quad q_1 = r_0 + \frac{1}{2}q_0, \quad r_1 = 0.$$

These equations can be rewritten as

$$\begin{bmatrix} p_1 \\ q_1 \\ r_1 \end{bmatrix} = \begin{bmatrix} 1 & 1/2 & 0 \\ 0 & 1/2 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} p_0 \\ q_0 \\ r_0 \end{bmatrix} =: Ax_0.$$

Let $x_n = (p_n, q_n, r_n)'$, $n \geq 0$. It is clear that the vectors x_n 's are governed by the equation

$$x_n = A^n x_0.$$

Here A is a **non-negative matrix**.

Determining the limiting behaviour of the sequence A^n , $n \geq 0$.

The eigenvalues of A are $\lambda_1 = 1$, $\lambda_2 = 1/2$, $\lambda_3 = 0$ with the corresponding eigenvectors

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}.$$

Let $P = \begin{bmatrix} 1 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & 1 \end{bmatrix}$, $\Lambda = \text{diag}(1, 1/2, 0)$. Then $P^{-1} = \frac{1}{2} \begin{bmatrix} 1 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & 1 \end{bmatrix}$ and $A = P\Lambda P^{-1}$ implies

$$A^n = P\Lambda^n P^{-1} = \begin{bmatrix} 1 & 1 - (1/2)^n & 1 - 2^{1-n} \\ 0 & (1/2)^n & (1/2)^{n-1} \\ 0 & 0 & 0 \end{bmatrix}$$

Hence,

$$p_n = 1 - \left(\frac{1}{2}\right)^n q_0 - \left(\frac{1}{2}\right)^{n-1} r_0, \quad q_n = \left(\frac{1}{2}\right)^n q_0 + \left(\frac{1}{2}\right)^{n-1} r_0, \quad r_n = 0.$$

Consequently

$$\lim_{n \rightarrow \infty} p_n = 1, \quad \lim_{n \rightarrow \infty} q_n = 0, \quad \lim_{n \rightarrow \infty} r_n = 0.$$

In the long run, individuals of genotype Aa and aa will disappear! Only pure genotype AA will remain in the population.

56.1.5 Population Growth Models (Demography)

Important problem: Critical examination of how population grows over a period of interest.

The “Leslie Model” (http://en.wikipedia.org/wiki/Leslie_matrix) describes the growth of the female portion of a human or animal population. This section studies this model.

Suppose the maximum age attained by any female in the population is M years. We classify the female population into some k age classes of equal length. Say, the age classes are: $[0, M/k)$, $[M/k, 2M/k)$, $\dots$, $[(k-1)M/k, M]$. When the study is initiated, let $p_i^{(0)}$ = number of females in the population falling into the age group $[(i-1)M/k, iM/k)$, $i = 1, 2, \dots, k$.

The vector $p^{(0)} = (p_1^{(0)}, \dots, p_k^{(0)})$ is called the initial *age distribution* vector. The main objective is to examine how the age distribution changes over time. We would like to examine the age distribution of the female population at times $t_0 = 0$, $t_1 = M/k$, $t_2 = 2M/k$, etc. As time progresses, the composition of the classes varies because of three biological processes: birth, death and ageing. These biological processes may be described by the following demographic parameters:

- a_i = the average number of daughters born to a single female during the time she is in the i -th age class, $i = 1, 2, \dots, k$,
- b_i = the proportion of females in the i -th age class expected to survive and pass into the next class, $i = 1, 2, \dots, k$,

It is clear that $a_i \geq 0$ for every i . Assume that $0 < b_i \leq 1$ for every $i = 1, \dots, k-1$. Assume that at least one a_i is positive. Let $p^{(m)} = (p_1^{(m)}, \dots, p_k^{(m)})$ be the age distribution of the females at time t_m , where $p_i^{(m)}$ is the number of females in the i -th age class, $i = 1, \dots, k$. The vectors $p^{(m)}$'s satisfy the following recurrent relation:

$$\begin{aligned} p_1^{(m)} &= a_1 p_1^{(m-1)} + \dots + a_k p_k^{(m-1)}, \\ p_i^{(m)} &= b_{i-1} p_{i-1}^{(m-1)}, \quad i = 2, 3, \dots, k. \end{aligned}$$

These equations can be put in a succinct form:

$$p^{(m)} = Lp^{(m-1)} := \begin{bmatrix} a_1 & a_2 & a_3 & \cdots & a_{k-1} & a_k \\ b_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & b_2 & 0 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & b_{k-1} & 0 \end{bmatrix} p^{(m-1)}, \quad m = 1, 2, \dots$$

L is the so-called Leslie matrix. Clearly, L is **non-negative** matrix. It now follows that $p^{(m)} = L^m p^{(0)}$. In order to compute the powers of L , finding eigenvalues and eigenvectors of L is helpful, which can be done by solving

$$|L - \lambda I| = \lambda^k - a_1 \lambda^{k-1} - a_2 b_1 \lambda^{k-2} - a_3 b_1 b_2 \lambda^{k-3} - \cdots - a_k b_1 \cdots b_{k-1}.$$

Since at least one a_i is positive, there is at least one non-zero root of the polynomial equation. Consequently, the spectral radius $\rho_s(L) > 0$. The following are some facts about Leslie matrix:

1. The eigenvalue $\rho_s(L)$ is simple.
2. The vector $x := (1, b_1/\rho_s(L), b_1 b_2/(\rho_s(L))^2, \dots, b_1 b_2 \cdots b_{k-1}/(\rho_s(L))^{k-1})$ is positive and is an eigenvector corresponding to the eigenvalue $\rho_s(L)$.
3. If two successive entries a_i and a_{i+1} are positive, then $|\lambda| < \rho_s(L)$ for any eigenvalue λ of L different from $\rho_s(L)$.

If a_i and a_{i+1} are positive for some i , it now follows that $(\rho_s(L))^{-m} L^m p^{(0)}$ converges to a constant vector q which depends only on $p^{(0)}$. If m is large $p^{(m)} \approx (\rho_s(L))^m q \approx (\rho_s(L)) p^{(m-1)}$. This means that the age distribution is a scalar multiple of the preceding age distribution.

56.2 Linear Models

This section consider the statistical analysis of linear models and some general optimisation problems associated with it.

A *linear model* has the form

$$Y = X\beta + \epsilon \tag{56.5}$$

e:12.1.1

where Y is an n -vector random variable, X in an $n \times m$ matrix with $\text{rank} X = r \leq m$, β is an m -vector of unknown parameters and ϵ is an n -vector of error variables with $E(\epsilon) = 0$ and the covariance matrix $D(\epsilon) = \sigma^2 V$ is an $n \times n$ non-negative definite matrix.

The problem of interest are the estimation of the fixed parameters β and σ^2 and the estimation or the prediction of the random component ϵ .

A *mixed linear model* is of the form

$$Y = X\beta + U_1 \xi_1 + \cdots + U_k \xi_k + \epsilon \tag{56.6}$$

e:12.1.2

where Y, X, β and ϵ as in (56.5), $U_i \in M_{n, p_i}$, ξ_i is a p_i -vector random variable s.t. $E(\xi_i) = 0$ and

$$\text{Cov}(\xi_i) = \sigma_i^2 I_{p_i}, \quad \text{Cov}(\xi_i, \xi_j) = 0, \quad i \neq j, \quad \text{Cov}(\xi_i, \epsilon) = 0 \tag{56.7}$$

e:12.1.3

for $i = 1, \dots, k$.

The problem of interest are the estimation of the fixed parameters $\beta, \sigma^2, \sigma_1^2, \dots, \sigma_k^2$ and the prediction of the random components $\xi_1, \dots, \xi_k$ and ϵ .

56.2.1 Some Useful Lemmas

Let $Sp(A)$ and $\text{rank}A$ be the space generated by the columns of A and the rank of A . $A^\perp$ is a matrix of maximum rank s.t. $A'A^\perp = 0$. Let $V \in M_n$ be non-negative definite, $X \in M_{n,m}$ and $W = (V|X)$. Note that $\text{rank}W = \text{rank}(V + XX')$ and $Sp(V|X) = Sp(V + XX')$.

p:12.2.1

Lemma 56.1. *The linear equations in the matrices $L \in M_{n,k}$ and $\Lambda \in M_{m,k}$:*

$$\begin{aligned} VL + X\Lambda &= C_1 \\ X'L &= C_2 \end{aligned} \quad (56.8) \quad \text{e:12.2.1}$$

admit solutions for any $C_1 \in M_{n,k}$ and $C_2 \in M_{m,k}$ s.t. $Sp(C_1) \subset Sp(W) = Sp(V|X)$ and $Sp(C_2) \subset Sp(X')$, respectively.

Definition 56.6. Let $A, B \in M_n$. $A \leq_L B$ if $B - A = CC'$ for some matrix C , i.e. $B - A$ is a non-negative matrix.

The following lemma is concerned with the minimisation of a matrix function in the sense of Löwner.

p:12.2.2

Lemma 56.2. *Let $L_0 \in M_{n,k}$ and $\Lambda_0 \in M_{m,k}$ be a solution of the matrix equations in $L \in M_{n,k}$ and $\Lambda \in M_{m,k}$ given by:*

$$\begin{aligned} VL + X\Lambda &= F \\ X'L &= P \end{aligned} \quad (56.9) \quad \text{e:12.2.2}$$

where $Sp(F) \subset Sp(V|X)$ and $Sp(P) \subset Sp(X')$. Then for all L s.t. $X'L = P$, we have

$$L'VL - L'F - F'L \geq_L -F'L_0 - P'\Lambda_0 = -L'_0F - \Lambda'_0P. \quad (56.10) \quad \text{e:12.2.5}$$

56.2.2 Estimation in a Linear Model

We apply the result of Lemma [56.1](#) to estimate $X\beta$, ϵ and σ^2 , the triplet characterising the linear model

$$Y = X\beta + \epsilon, \quad E(\epsilon) = 0, \quad D(\epsilon) = \sigma^2V \quad (56.11) \quad \text{e:12.3.1}$$

where D denotes the dispersion (variance covariance) matrix.

d:12.3.1

Definition 56.7. A set of k parametric functions $P'\beta$ where $P \in M_{m,k}$ is said to be *unbiasedly estimable by linear functions of Y* if there exists $L \in M_{k,n}$ s.t. $E(L'Y) = P'\beta$, a sufficient condition for which is $Sp(P) \subset Sp(X')$.

Note that the entire m -vector parameter β is estimable only if $Sp(I) = Sp(X')$, i.e., when $\text{rank}X' = m$. The function $X\beta$ is always estimable.

Definition 56.8. A linear function L'_0Y is said to be the *minimum dispersion unbiased estimator* (MDUE) of an estimable function $P'\beta$ if $L'_0X = P'$ and

$$D(L'_0Y) = \sigma^2 L'_0V L_0 \leq_L \sigma^2 L'V L = D(L'Y)$$

d:12.3.2

for all L s.t. $L'X = P'$.

The following theorem provides the main result for computing the MDUE of estimable parametric functions of β .

p:12.3.1

Theorem 56.9. Let $L_0 \in M_{n,k}$ and $\Lambda_0 \in M_{m,k}$ be a solution of the equations:

$$\begin{aligned} VL + X\Lambda &= 0 \\ X'L &= X'. \end{aligned} \quad (56.12)$$

e:12.3.2

Then L'_0Y is the MDUE of $X\beta$ with the dispersion matrix, $D(L'_0Y) = -\sigma^2 X\Lambda_0$.

Proof. The covariance matrix of $L'Y$ is $\sigma^2 L'VL$. Using Lemma 56.1, the minimum of $L'VL$ with L subject to $X'L = X'$ is attained as stated in the theorem. $\square$

56.10. The following results are consequences of Theorem 56.9.

1. If $C_1, C_2, C_3, -C_4$ are the partitions of any g-inverse of the matrix of the equations in (56.12), then $L_0 = C_2X'$ and $\Lambda_0 = -C_4X'$ and the estimate of $X\beta$ is

$$X\hat{\beta} = L'_0Y = XC'_2Y, \quad (56.13)$$

e:12.3.3

and $D(X\hat{\beta}) = \sigma^2(-X\Lambda_0) = \sigma^2XC_4X'$.

2. Let $P \in M_{m,k}$ s.t. $Sp(P) \subset Sp(X')$ in which case the function $P'\beta$ is estimable and the MDUE of $P'\beta$ is $P'\hat{\beta}$ where $\hat{\beta} = C'_2Y$ with the covariance matrix $\sigma^2P'C_4P$.

p:12.3.2

Proposition 56.14. The linear predictor $L'Y$ of ϵ has a bounded mean dispersion error (MDE) for all β only if L is s.t. $X'L = 0$. Let (L_e, V_e) be any solution of the equation

$$\begin{aligned} VL + X\Lambda &= V \\ X'L &= 0. \end{aligned} \quad (56.14)$$

e:12.3.6

Then L'_0Y is the unbiased predictor of ϵ with the minimum MDE of $\sigma^2 X\Lambda_e$.

56.11. From Statements 56.9 and 56.14, the estimates of $X\beta$ and ϵ are L'_0Y and L'_eY , respectively. Do they add up to Y ? From $D(L'_0Y + L'_eY - Y) = \sigma^2(L'_0 + L'_e - I)'V(L'_0 + L'_e - I) = \sigma^2(L'_0 + L'_e - I)'(VL'_0 + VL'_e - V) = \sigma^2(L'_0 + L'_e - I)'X(-\Lambda_e - \Lambda_0) = \sigma^2(L'_0X + L'_eX - X)(-\Lambda_e - \Lambda_0) = 0$, we know that $L'_0Y + L'_eY = Y$ with probability one.

56.2.3 Estimation of Variance

Consider the linear model (56.11) with $V = I$. Let $Y'CY$ be a quadratic form in the data Y , for some symmetric matrix C , which we would like to propose as an estimator of σ^2 . The basic requirement is that it be unbiased. The following is a general result on the expected value of a quadratic form.

Proposition 56.15. Let Y be a random vector of order $n \times 1$ with mean vector μ and dispersion matrix σ^2V . Let C be any symmetric matrix of order $n \times n$. Then

$$E(Y'CY) = \sigma^2 \text{tr}(CV) + \mu'C\mu. \quad (56.15)$$

e:12.5.1

□

56.2.4 Method of MINQUE

The acronym of the title stands for *minimum norm quadratic unbiased estimation*. Consider the linear model (56.11) with $V = I$. In the previous subsection, under the pretext of multivariate normality, we obtained the minimum variance quadratic unbiased estimator of σ^2 . In this subsection, we will attempt to obtain an optimal quadratic estimator of σ^2 without invoking normality assumption.

Let $C = C^T$ be of order $n \times n$. The quadratic form $Y'CY = (X\beta + \epsilon)'C(X\beta + \epsilon) = \beta'X'CX\beta + \epsilon'CE + \beta'X'CE + \epsilon'CX\beta = \epsilon'CE$. Hence, $Y'CY - \epsilon'(n^{-1}I)\epsilon = \epsilon'(C - n^{-1}I)\epsilon$. The critical idea behind the method of MINQUE is to choose C s.t. it is close to $n^{-1}I$. Choosing the Frobenius norm, $\|B\|_F = \text{tr}(BB')$, the objective is to minimise $\|C - n^{-1}I\|_F$ subject to unbiasedness conditions that $CX = 0$ and $\text{tr}C = 1$. Consequently, this problem reduces to

$$\text{minimise } \text{tr}(C^2) \quad \text{subject to } CX = 0, \text{tr}C = 1.$$

In many variance components models, the error random vector has a structural representation, $\epsilon = U_1\xi_1 + \cdots + U_k\xi_k$, where U_i is a known matrix of order $n_i \times n_i$, ξ_i is a random vector of order $n_i \times 1$ with mean vector 0 and dispersion matrix $\sigma_i^2 I_{n_i}$, $i = 1, \dots, k$, and I_{n_i} is the identity matrix of order $n_i \times n_i$. Further $\xi_1, \dots, \xi_k$ are assumed to be independent. Under this assumptions, we have

$$D(\epsilon) = \sigma_1^2 U_1 U_1' + \cdots + \sigma_k^2 U_k U_k'.$$

If ξ_i is observable, then a natural estimator of σ_i^2 is $n_i^{-1} \xi_i' \xi_i$. Consequently, a natural estimator of $p_1 \sigma_1^2 + \cdots + p_k \sigma_k^2$ is

$$\sum_{i=1}^k \frac{p_i}{n_i} \xi_i' \xi_i = (\xi_1', \dots, \xi_k') D (\xi_1', \dots, \xi_k')', \quad (56.16) \quad \text{e:12.9.1}$$

where $D = \text{diag}\{\frac{p_1}{n_1} I_{n_1}, \dots, \frac{p_k}{n_k} I_{n_k}\}$.

$$\begin{aligned} Y'CY &= \beta'X'CX\beta + 2(U_1\xi_1 + \cdots + U_k\xi_k)'CX\beta + (U_1\xi_1 + \cdots + U_k\xi_k)'C(U_1\xi_1 + \cdots + U_k\xi_k) \\ &= (\xi_1', \dots, \xi_k') D_* (\xi_1', \dots, \xi_k')' \end{aligned} \quad (56.17) \quad \text{e:12.9.2}$$

where D_* is the block matrix whose (i, j) -th block is given by $U_i' C U_j$. The critical idea behind the method of MINQUE is to get the proposed invariant unbiased estimator of $p'\theta$ involving the data close to the natural estimator. This tantamount to getting D_* close to D in some norm. In the Frobenius norm, we have

$$\begin{aligned} \|D_* - D\|_F &= \text{tr}((D_* - D)(D_* - D)') = \text{tr}((D_* - D)^2) \\ &= \sum_{i=1}^k \text{tr}((U_i' C U_i - \frac{p_i}{n_i} I_{n_i})^2) + \sum_{i=1, i \neq j}^k \sum_{j=1, j \neq i}^k \text{tr}(U_i' C U_j U_j' C U_i) \\ &= \cdots \\ &= \text{tr}(C \sum_{i=1}^k V_i)^2 - \sum_{i=1}^k \frac{p_i^2}{n_i} = \text{tr}(C V_0)^2 - \sum_{i=1}^k \frac{p_i^2}{n_i} \end{aligned}$$

where $V_0 = V_1 + \cdots + V_k$. The objective now is to determine a symmetric matrix C which satisfies $CX = 0$ and $\text{tr}(C V_i) = p_i$, $i = 1, \dots, k$ and minimised $\text{tr}((C V_0)^2)$.

Since V_0 is positive, we can find a positive matrix B s.t. $B^2 = V_0$. Let $X_* = B^{-1}X$ and $V_{i*} = B^{-1}V_i B^{-1}$, $i = 1, \dots, k$. The optimisation problem translates into the following problem. Determine a symmetric matrix A s.t.

$$AX_* = 0, \quad \text{tr}(A V_{i*}) = p_i, \quad i = 1, \dots, k \quad (56.18) \quad \text{e:12.9.5}$$

and it minimises $\text{tr}(A^2)$. An optimal solution A_* is given by

$$A_* = \sum_{i=1}^k \lambda_i Q_* V_{i*} Q_*, \quad (56.19) \quad \text{e:12.9.6}$$

where $\lambda_1, \dots, \lambda_k$ satisfy the linear equations (trace minimisation problem):

$$\sum_{j=1}^k [\text{tr}(Q_* V_{i*} Q_* V_{j*})] \lambda_j = p_i, \quad i = 1, \dots, k$$

where $Q_* = I - X_*(X'_*X_*)^{-1}X'_*$. Once we compute A_* , the optimal invariant unbiased estimator $Y'C_*Y$ can be computed with $C_* = B^{-1}A_*B^{-1}$. We label $Y'C_*Y$ as an optimal estimator according to the criterion of MINQUE.

Consider a linear model with error random vector containing several variance components:

$$Y = X\beta + \epsilon, \quad E\epsilon = 0, \quad D(\epsilon) = \sigma_1^2 V_1 + \cdots + \sigma_k^2 V_k,$$

where $V_1, \dots, V_k$ are known non-negative matrices and $\sigma_1^2, \dots, \sigma_k^2$ are known. Let $p_1\sigma_1^2 + \cdots + p_k\sigma_k^2 = p'\theta$ be a linear combination of the variance components $\sigma_1^2, \dots, \sigma_k^2$, where $p' = (p_1, \dots, p_k)$ is known and $\theta = (\sigma_1^2, \dots, \sigma_k^2)$. We are seeking an optimal unbiased estimator $Y'CY$ of $p'\theta$. Unbiasedness means that we are seeking a symmetric matrix C satisfying

$$X'CX = 0 \quad \text{tr}(CV_i) = p_i, \quad \forall i. \quad (56.20) \quad \text{e:12.10.1}$$

Optimality means that we are seeking a symmetric matrix C s.t.

$$\text{tr}((CV_0)^2), \quad V_0 = V_1 + \cdots + V_k \quad (56.21) \quad \text{e:12.10.2}$$

is minimum. This is an optimisation problem with the following solution.

Theorem 56.12. Let V_0 be non-singular, B be positive definite square root of V_0 , i.e. $B^2 = V_0$. In this case $\text{tr}((CV_0)^2) = \text{tr}((BCB)^2)$. Let $A = BCB$, or $C = B^{-1}AB^{-1}$, $X_* = X'B^{-1}$, $V_{i*} = B^{-1}V_iB^{-1}$, $i = 1, \dots, k$.

The minimum of $\text{tr}(A^2)$ over all symmetric matrices A subject to the conditions $X'_*AX_* = 0$ and $\text{tr}(AV_{i*}) = p_i$ for all i , is attained at

$$A_* = \sum_{i=1}^k \lambda_i (V_{i*} - P_* V_{i*} P_*), \quad (56.22) \quad \text{e:12.10.3}$$

where $P_* = X_*(X'_*X_*)^{-1}X'_*$ is the projection operator associated with the design matrix X_* , and $\lambda_1, \dots, \lambda_k$ satisfy the linear equations:

$$\sum_{i=1}^k [\text{tr}(V_{i*}V_{j*} - P_* V_{i*} P_* V_{j*})] \lambda_i = p_j, \quad j = 1, \dots, k. \quad (56.23) \quad \text{e:12.10.4}$$

Once we obtain A_* , we can compute the required optimal matrix $C_* = B^{-1}A_*B^{-1}$. Hence, we can lay down an optimal unbiased estimator of $p'\theta$ as $Y'C_*Y$.

56.3 Econometrics

References:

- William Greene (2002), *Econometric Analysis* 5e (Very comprehensive).
- Marno Verbeek (2004), *A Guide to Modern Econometrics*, 2e, John Wiley & Sons.
- Baltagi, B. H. (2008), *Econometrics* 4e.
- *Handbook of Econometrics*, Volumes 1 to 4, Elsevier 1983
- Rachev, S. T. (2007) *Financial Econometrics from Basics to Advanced Modeling Techniques*, Wiley.
- *International Macroeconomics, Finance Theory and Econometric Methods*, 2000

56.3.1 Classical Approach: Linear Regression with OLS & GLS

The response of an economic data is related to the predictor as follows:

$$y_t = a + b_1x_{t1} + \cdots + b_qx_{tq} + \epsilon_t, \quad 1 \leq t \leq n \quad (*)$$

where

- y_t is the output/response
- x_{ti} is the input/predictor
- ϵ_t is an unobservable random errors with zero mean.

If $x_{t0} = 1$ then $(*)$ can be written as

$$Y = X\beta + \epsilon.$$

(a) If the variance is the same, i.e., $\text{Cov}(Y) = \sigma^2 I$, then

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

is the best estimate for $RSS = (Y - X\beta)^T(Y - X\beta)$.

(b) If the variance can be different, i.e., $\text{Cov}(Y) = V$, where V is symmetric and positive definite, then

$$\hat{\beta}_{GLS} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y$$

is the best estimate for the residual sum of squares $RSS = (Y - X\beta)^T(Y - X\beta)$.

Example 56.13 (Grant V. Farnsworth, *Econometrics in R*, 2008.10.26. Cross Sectional Regression). Suppose we have a data frame called `byu` containing columns for `age`, `salary`, and `exper`. We want to regress various forms of `age` and `exper` on `salary`. A simple linear regression might be

```
> lm(byu$salary ~ byu$age + byu$exper)
```

or alternately:

```
> attach(byu)
> lm(salary ~ age + exper)
```

The above command would correspond to the linear model

$$\text{salary} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{exper} + \epsilon \quad (56.24)$$

Using `lm()` results in an abbreviated summary being sent to the screen, giving only the coefficient estimates. For more exhaustive analysis, we can save the results in as a data member or “fitted model”:

```
> result <- lm(salary ~ age + exper + age*exper, data=byu)
> summary(result)
> myresid <- result$resid
> vcov(result)
```

The `summary()` command, run on raw data, such as `byu$age`, gives statistics, such as the mean and median (these are also available through their own functions, `mean` and `median`). When run on an `ols` object, `summary` gives important statistics about the regression, such as p -values and the R^2 .

The residuals and several other pieces of data can also be extracted from `result`, for use in other computations. The variance-covariance matrix (of the beta coefficients) is accessible through the `vcov()` command.

Notice that more complex formulae are allowed, including interaction terms (specified by multiplying two data members) and functions such as `log()` and `sqrt()`. Unfortunately, in order to include a power term, such as `age squared`, we must either first compute the values, then run the regression, or use the `I()` operator, which forces computation of its argument before evaluation of the formula

```
> salary$agesq <- (salary$age)^2
> result <- lm(salary ~ age + agesq + log(exper) + age*log(exper), data=byu)
```

or

```
> result <- lm(salary ~ age + I(age^2) + log(exper) + age*log(exper), data=byu)
```

Notice that if we omit the `I()` operator and don't explicitly generate a power term variable (like `agesq`) then `lm()` will not behave as expected (it just omits power terms whose first power is already included in the regression), but it will not give us an error or warning, so we must be careful.

To run a regression without an intercept, we simply specify the intercept explicitly, traditionally with a zero.

```
> result <- lm(smokes ~ 0 + male + female, data=smokerdata)
```

Extracting Statistics from the Regression

The most important statistics and parameters of a regression are stored in the `lm` object or the summary object.

```
> output <- summary(result)
> RSS <- deviance(result)
> LL <- logLik(result)
> DegreesOfFreedom <- result$df
> Yhat <- result$fitted.values
> Coef <- result$coefficients
> Resid <- result$residuals
> s <- output$sigma
> RSquared <- output$r.squared
> CovMatrix <- s^2*output$cov
> aic <- AIC(result)
```

Here `LL` is the log likelihood statistic, `Yhat` is the vector of fitted values, `Resid` is the vector of residuals, `s` is the estimated standard deviation of the errors (assuming homoskedasticity), `CovMatrix` is the variance-covariance matrix of the coefficients (also available via `vcov()`), `aic` is the Akaike information criterion and other statistics are as named.

Note that the AIC criterion is defined by R as

$$AIC = -2 \log L(p) + 2p$$

where p is the number of estimated parameters and $L(p)$ is the likelihood. Some econometricians prefer to call AIC/N the *information criterion*. To obtain the Bayesian Information Criterion (or Schwartz Bayesian Criterion) we use AIC but specify a different “penalty” parameter as follows

```
> sbc <- AIC(result, k=log(NROW(smokerdata)))
```

which means

$$SBC = -2 \log L(p) + p \log(N).$$

Example 56.14 (Handbook of Econometrics, Volume 1, Chapter 1). Consider a system of L simultaneous endogenous and K exogenous variables. We write $y_{\alpha l}$ and $x_{\alpha k}$ for the α th observation on the l th endogenous variable. Then the j th equation for observation α takes the ‘structural form’

$$\sum_{l=1}^L \gamma_{lj} y_{\alpha l} + \sum_{k=1}^K \beta_{kj} x_{\alpha k} = \epsilon_{\alpha j} \quad (56.25)$$

where $\epsilon_{\alpha j}$ is a random disturbance and the γ 's and β 's are coefficients. The above equation can be written as

$$\mathbf{y}'_{\alpha} \Gamma + \mathbf{x}'_{\alpha} \mathbf{B} = \epsilon'_{\alpha}$$

or in matrix form

$$\mathbf{Y} \Gamma + \mathbf{X} \mathbf{B} = \mathbf{E}$$

where $\mathbf{Y}$ and $\mathbf{X}$ are observation matrices of two sets of variables of order $n \times L$ and $n \times K$, respectively and $\mathbf{E}$ is an $n \times L$ disturbance matrix.

56.3.2 Time-Series Analysis

One objective of analysing economic data is to predict or forecast the future values of economic variables. One approach to do this is to build a more or less structural econometric model, describing the relationship between the variable of interest with other economic quantities, to estimate this model using a sample of data, and to use it as the basis for forecasting and inference. Although this approach has the advantage of giving economic content to one's predictions, it is not always very useful. For example, it may be possible to adequately model the contemporaneous relationship between unemployment and the inflation rate, but as long as we cannot predict future inflation rates we are also unable to forecast future unemployment.

In time-series approach, the current values of an economic variable are related to past values (either directly or indirectly). The emphasis is purely on making use of the information in past values of a variable for forecasting its future. In addition to producing forecasts, time series models also produce the distribution of future values, conditional upon the past, and can thus be used to evaluate the likelihood of certain events.

Topics:

- ARIMA models
- SEMTSA & Random Walk Hypothesis ???
- Co-integration test (comparing two time-series)

56.3.3 Qualitative Response (QR) Models

There are many settings in which the economic outcome we seek to model is a discrete choice among a set of alternatives, rather than a continuous measure of some activity. Consider, for example, modelling labour force participation, the decision of whether or not to make a major purchase, or the decision of which candidate to vote for in an election. For the first of these examples, intuition would suggest that factors such as age, education, marital status, number of children, and some economic data would be relevant in explaining whether an individual chooses to seek work or not in a given period. But something is obviously lacking if this example is treated as the same sort of regression model we used to analyse consumption or the costs of production or the movements of exchange rates.

This section examines a variety of what have come to be known as *qualitative response* (QR) models. There are numerous different types that apply in different situations. What they have in common is that they are models in which the dependent variable is an indicator of a discrete choice, such as a "yes or no" decision. In general, conventional regression methods are inappropriate in these cases.

Topics:

- Discrete Choice Models
- Limited dependent variables
- Probit, Logit Analysis
- Discriminant analysis
- Tobit models

56.3.4 Software: gretl (GUI) and R (CLI)

- <http://www.learneconometrics.com/>
- Youtube learning: Applied Econometrics using R
- <http://cran.r-project.org/web/views/Econometrics.html>

56.4 Real-Life Statistics

- Actuarial science: a discipline that applies mathematical and statistical methods to assess risk in the insurance and finance industries:
 - Life contingency
 - Loss model
- Biostatistics: a branch of biology that studies biological phenomena and observations by means of statistical analysis, and includes medical statistics.
- Chemometrics: a science of relating measurements made on a chemical system or process to the state of the system via application of statistical methods. It's heavily used in analytical chemistry and metabolomics.
- Environmental statistics: statistical analysis of weather, climate, air and water quality, plant and animal populations, etc.
- Epidemiology: the study of factors affecting the health and illness of populations, and serves as the foundation and logic of interventions made in the interest of public health and preventive medicine.
- Geostatistics: a branch of geography that deals with the analysis of data from disciplines such as petroleum geology, hydrogeology, hydrology, meteorology, oceanography, geochemistry, geography.
- Operational Research: an interdisciplinary branch of applied mathematics and formal science that uses methods such as mathematical modeling, statistics, and algorithms to arrive at optimal or near optimal solutions to complex problems.
- Psychometrics: the theory and technique of educational and psychological measurement of knowledge, abilities, attitudes, and personality traits.
- Quality control: reviews the factors involved in manufacturing and production; it can make use of statistical sampling of product items to aid decisions in process control or in accepting deliveries.
- Statistical finance/econophysics: an empirical attempt to shift finance from its normative roots to a positivist framework using exemplars from statistical physics with an emphasis on emergent or collective properties of financial markets.
- Business analytics: applies statistical methods to data sets (often very large) to develop new insights and understanding of business performance and opportunities.
- Web analytics (software engineering): Use in Internet marketing to gain insight into visitor behaviour and to increase conversion rate:
 - <https://bitbucket.org/akoha/django-lean/wiki/Home>: A framework for performing and analysing split-test experiments
 - Google Website Optimizer: Google's free website testing and optimisation tool, allows you to increase the value of your existing websites and traffic.
 - <http://exp-platform.com/default.aspx>: Microsoft's Experimentation Platform
 - Effective A/B Testing: <http://elem.com/~btilly/effective-ab-testing/>

Chapter 57

Methods for Quality Improvements

1. Montgomery, D.C., 2012. Introduction to statistical quality control. 7th ed. New York: John Wiley. 2. Mitra, A., 2005. Fundamentals of quality control and improvement. 2nd ed. Upper Saddle River, N.J.: Prentice Hall. ','Additional References: 3. Juran, J.M.E., 1999. Juran's quality handbook. 5th ed. New York: McGraw-Hill.

Introduction · Revision on some statistical concepts and techniques. · Causes of variation. · Methods for quality improvement: The magnificent seven. · Selection of rational samples. · Statistical basis of the control charts. · Analysis of patterns in control charts. · Maintenance of control charts.

Topic 2: Control Charts for Variables · Control charts for $\bar{x}$ and R. · The operating-characteristic function. · The average run length. · $\bar{x}$ -chart with probability limits. · Control charts for $\bar{x}$ and s. · The $\bar{x}$ and s charts with variable sample size. · The S2 control chart. · Control charts for individual measurements. · Moving range chart.

Topic 3: Control Charts for Attributes · Advantages and disadvantages of attributes charts. · Control charts for nonconforming (p-chart). · Control charts for number of nonconforming items (np-chart). · Control charts for nonconformities (c-chart). · Control charts for nonconformities per unit (u-chart). · The control charts with variable sample size (p-chart, u-chart) · Operating characteristic (OC) curves.

Topic 4: Process Capability Analysis · Natural tolerance limit. · Specification limits and control limits. · Process capability indices. · Process capability analysis procedures.

Topic 5: Special Process Control Procedures · The CUSUM control charts. · The EWMA control charts. · The moving average control chart. · The modified control chart. · The acceptance control chart.

Topic 6: Acceptance Sampling Plans for Attributes and Variables · Advantages and disadvantages of sampling. · Producer's risk and consumer's risk. · Type - A and Type - B OC curves. · Types of sampling plans. · Evaluating sampling plans. · Symbols and terms: acceptable quality level, average outgoing quality, LTPD, AOQL, ATI, ASN.

Contents for older syllabus UCCM3213???:

Chapter 1. Methods for Quality Improvement

Chapter 2. Control Charts for Variables

Chapter 3. Control Charts for Attributes

Chapter 4. Specifications and Tolerances

Chapter 5. Special Process Control Procedures

Chapter 6. Fundamental Concepts in Acceptance Sampling

Chapter 7. Acceptance Sampling by Attributes

Chapter 8. Acceptance Sampling by Variables

Reference Books:

1. Montgomery, D.C. (2005) Introduction to Statistical Quality Control, John Wiley.

2. Juran, J.M. Ed. Quality Control Handbook, McGraw-Hill.

3. Jerry Banks. Principles of Quality Control, John Wiley.

4. Feigenbaum, A.V. Total Quality Control, McGraw-Hill.

5. Grant, E.I., Leavenworth, R.S. (1999) Statistical Quality Control, McGraw-Hill.

57.1 Introduction

The notion of quality has been defined in different ways by various authors. Some frequently used definitions are

Quality is conformance to requirements or specifications. Crosby (1979)

Quality is fitness for use. Juran (1974)

The quality of a product or service is the fitness of that product or service for meeting or exceeding its intended use as required by the customers.

57.1.1 Dimensions of quality

The quality of a product can be evaluated in several ways. Garvin (1987) provides an excellent discussion of eight components or dimensions of quality.

1. Performance
 - * Primary product characteristics
 - * Will the product do the intended job?
2. Reliability
 - * Consistency of performance over time.
 - * How often does the product fail?
3. Durability
 - * The effective service life of the product, includes repair.
 - * How long does the product last?
4. Serviceability
 - * The speed, courtesy, competence and cost of repair.
 - * How easy is it to repair the product?
5. Aesthetics
 - * Sensory characteristics such as style, color, shape, exterior finish.
 - * What does the product look like?
6. Features
 - * Secondary characteristics, added features.
 - * What does the product do?
7. Perceived Quality
 - * Past reputation of the company concerning quality of its products.
 - * What is the reputation of the company or its products?
8. Conformance to Standards
 - * Is the product made exactly as the designer intended?

57.1.2 Three aspects of quality

Quality of Design

All goods and services are produced in various grades and level of quality. These variations in grades or level of quality are intentional and the appropriate technical term is quality of design.

Higher design quality usually means higher cost such as adding an extra feature of air-bag in cars.

It deals with the stringent conditions that the product or service must minimally possess to satisfy the requirements of the customer. The design should be the simplest and least expensive while still meeting the customers' expectations.

Quality of Conformance

It is how well the manufactured product or the service rendered conforms to the standards selected in the design phase.

The quality of design has an impact on the quality of conformance since one must be able to produce what was designed. If the product or service does not conform to the standards selected, the design phase is affected. This feedback suggests that the product be redesigned because the current design cannot be produced using existing capability.

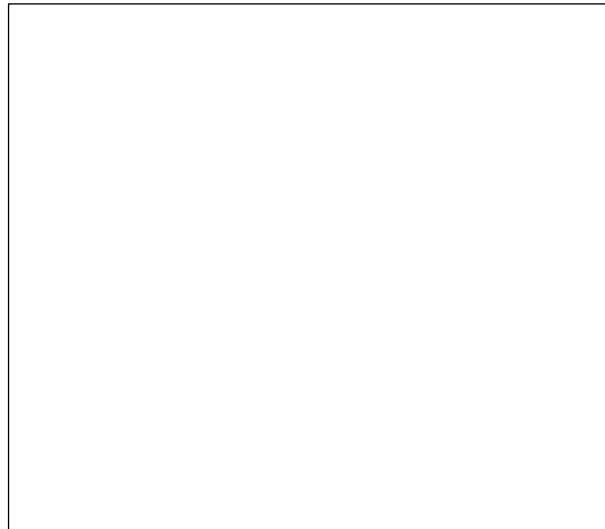
There should be a constant exchange of information between the design and manufacturing phases so that a feasible design can be achieved.

Quality of Performance

It is concerned with how well the product functions or service performs when put to use. It measures the degree to which the product or service satisfies the customer.

The final test of product or service acceptance always lies with the customers. If a product does not function well enough to meet these expectations, or if a service does not live up to customer standards, then adjustments need to be made in the design or conformance phase.

This feedback from the performance to the design phase may prompt a change in the design because the current design does not produce a product that performs adequately.



57.1.3 Quality characteristics

Every product possesses a number of elements that jointly describe what user or consumer thinks of as quality. These parameters are often called quality characteristics. Quality characteristics may be several types:

Structural characteristics : length, weight, strength, voltage

Sensory characteristics : taste, appearance, color

Time-oriented characteristics : warranty, reliability, maintainability

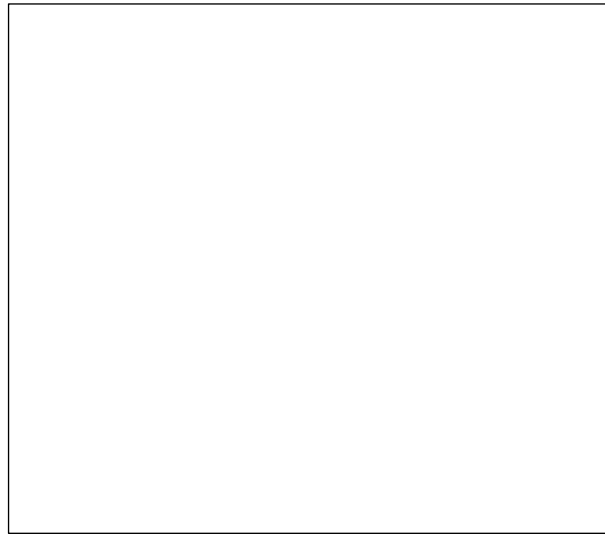
Ethical characteristics : honesty, courtesy, friendliness

For a manufactured product, the specifications are the desired measurements for the quality characteristics on the components and subassemblies that make up the product, as well as the desired values for the quality characteristics in the final product.

The largest allowable value for a quality characteristic is called the *upper specification limit* (USL) and the smallest allowable value for a quality characteristic is called the *lower specification limit* (LSL).

57.1.4 Types of variation in processes

A process is a sequence of operations by which inputs such as labour, materials and methods are transformed into outputs, in the form of products or services. A change in any one of these constitutes results in a new process. In any process, some variation in the quality measure from product to product is unavoidable.



The causes of the variation are usually:

Manpower

The personnel ability, physical and emotional stress can all contribute to variation in the product.

Machine

This source of variation includes tool wear, poor maintenance, machine vibration and inherent capability.

Material

Wrong material, lot-to-lot variation, multiple suppliers.

Method

Wrong planning or sequence, material handling.

Measurement

Faulty gauges, incorrect specifications. Lack of accuracy, repeatability and reproducibility in the measurement instrument.

Environment

Temperature, humidity, insufficient lighting.

The complex interactions of these 5M's and an E causes of variation are not easily understood. However, variation in the outcome of a process can be classified into two categories.

1. Common causes of variation due to factors that are inherent in the design of the system, and reflect the usual amount of variation to be expected.

2. Assignable causes or Special causes of variation due to unusual factors that are not part of the process design and nor ordinarily part of the process.

A stable process is one in which only common causes of variation affect the output quality. It is described as being in a state of statistical control. (or in control)

An unstable process is one in which both assignable causes and common causes affect the output quality. It is described as being out of control.

Quality improvement is the reduction of variability in processes and products.

57.2 Statistical Process Control

Statistical process control (SPC) refers to the application of the methods of statistical quality control (SQC) to the monitoring of processes. The purpose is to control the quality of product or service outputs from a process by maintaining control of the process.

SPC can be applied to any process. Its seven major tools are:

1. Histogram or stem-and-leaf display
2. Check sheet
3. Pareto chart
4. Cause-and-effect diagram / Ishikawa diagram
5. Defect concentration diagram
6. Scatter diagram
7. Control chart

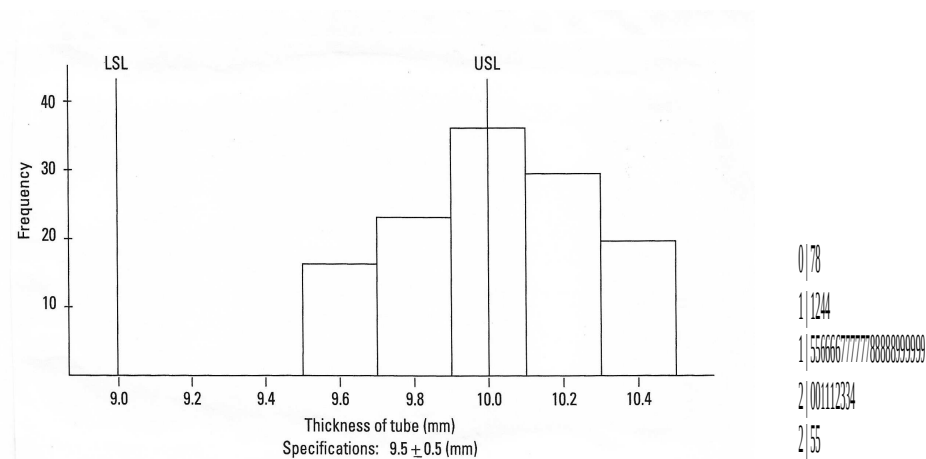
We often call these “the magnificent seven”.

57.2.1 The Magnificent Seven

Histogram or Stem-and-leaf display

Histograms or stem-and-leaf displays display large amounts of data that are difficult to interpret in their raw form.

A visual summary of the data in a histogram reveals whether the process is centered around a target value, the degree of variation in the data and whether the data meet the specifications.



Check sheets

Check sheets are used for collecting or gathering data in a logical format. It is the simplest yet most effective graphical methods. Observations are recorded as they happen, which reveals patterns or trends.

Part No: TAX-41
 Location: Bellevue
 Study Date: 6/5/1989
 Analyst: TCB

CHECK SHEET DEFECT DATA FOR 1988-1989 YTD

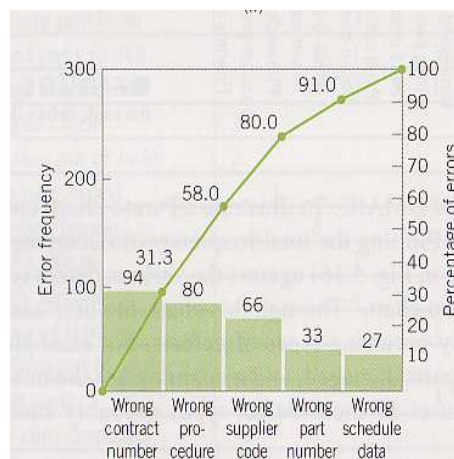
Defect	1988												1989					Total
	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	
Parts damaged		1		3	1	2		1		10	3		2	2	7	2		34
Machining problems			3	3				1	8		3		8	3				29
Supplied parts rusted			1	1		2	9											13
Masking insufficient		3	6	4	3	1												17
Misaligned weld	2																	2
Processing out of order	2														2			4
Wrong part issued		1						2										3
Unfinished fairing			3															3
Adhesive failure				1							1		2			1	1	6
Powdery alodine					1													1
Paint out of limits						1								1				2
Paint damaged by etching			1															1
Film on parts						3		1	1									5
Primer cans damaged								1										1
Voids in casting									1	1								2
Delaminated composite										2								2
Incorrect dimensions											13	7	13	1		1	1	36
Improper test procedure										1								1
Salt-spray failure													4					4
TOTAL	4	5	14	12	5	9	9	6	10	14	20	7	29	7	7	6	2	166

A check sheet to record defects on a tank used in an aerospace application.

When designing a check sheet, it is important to clearly specify the type of data to be collected, the part or operation number, the date, the analyst, and any other information useful on diagnosing the cause of poor performance

Pareto chart

The Pareto chart is simply a frequency distribution (or histogram) of problems, arranged in descending order of importance

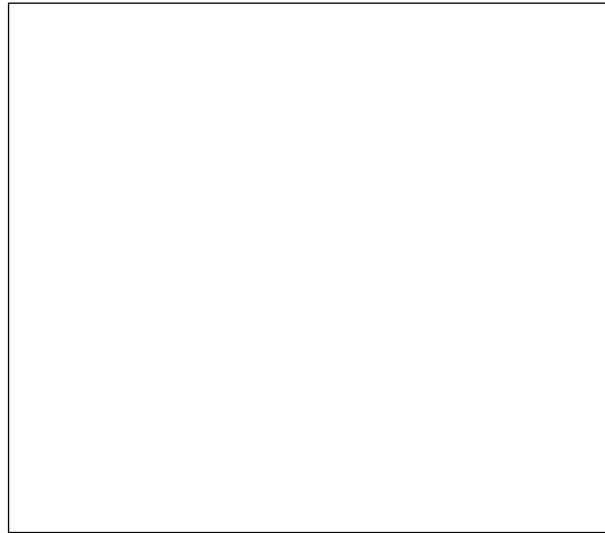


Cause-and-effect diagram / Ishikawa diagram

Cause-and-effect diagrams were developed by Kaoru Ishikawa in 1943 and thus are often called Ishikawa diagrams. They are also known as fishbone diagrams because of their appearance. Basically, cause-and-effect diagrams are used to identify and systematically list the different causes that can be attributed to a problem or an effect.

How to construct a cause-and-effect diagram

1. Define the problem or effect to be analyzed.
2. Draw the effect box and the center line.
3. Identify major causes. Specify the major potential cause and join them as boxes connected to the center line.
4. Brainstorm subcauses and identify them into the categories in Step 3.
5. Rank order the causes to identify those that seem most likely to impact the problem.
6. Take corrective action.



Defect concentration diagram

A defect concentration diagram is a picture of the unit, showing all relevant views. Then the various types of defects are drawn on the picture, and the diagram is analyzed to determine whether the location of the defects on the unit conveys any useful information about the potential causes of the defects.

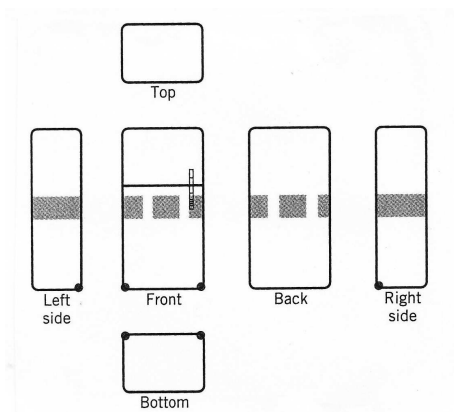


Figure 4-20 Surface-finish defects on a refrigerator.

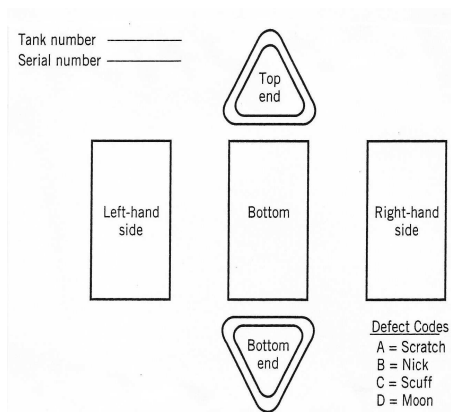


Figure 4-21 Defect concentration diagram for the tank.

Scatter diagram

Scatter diagram is the simplest form to depict the relationship between two variables. It is often used as follow-ups to a cause-and-effect analysis to determine whether a stated cause truly does impact the quality characteristic.

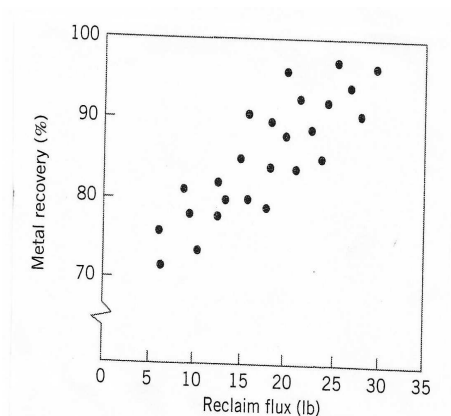


Figure 4-22 A scatter diagram.

Control chart

A control chart consist of: Upper Control Limit (UCL),
Lower Control Limit (LCL),
Center Line (CL),
Plot of quality characteristic measured form a sample versus sample number or time,
Sample points are connected by straight lines.

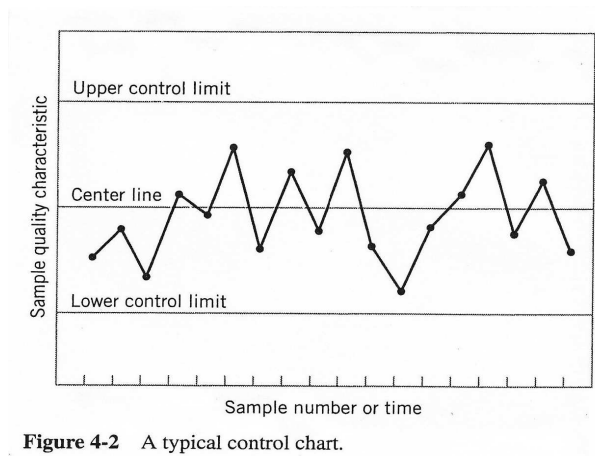


Figure 4-2 A typical control chart.

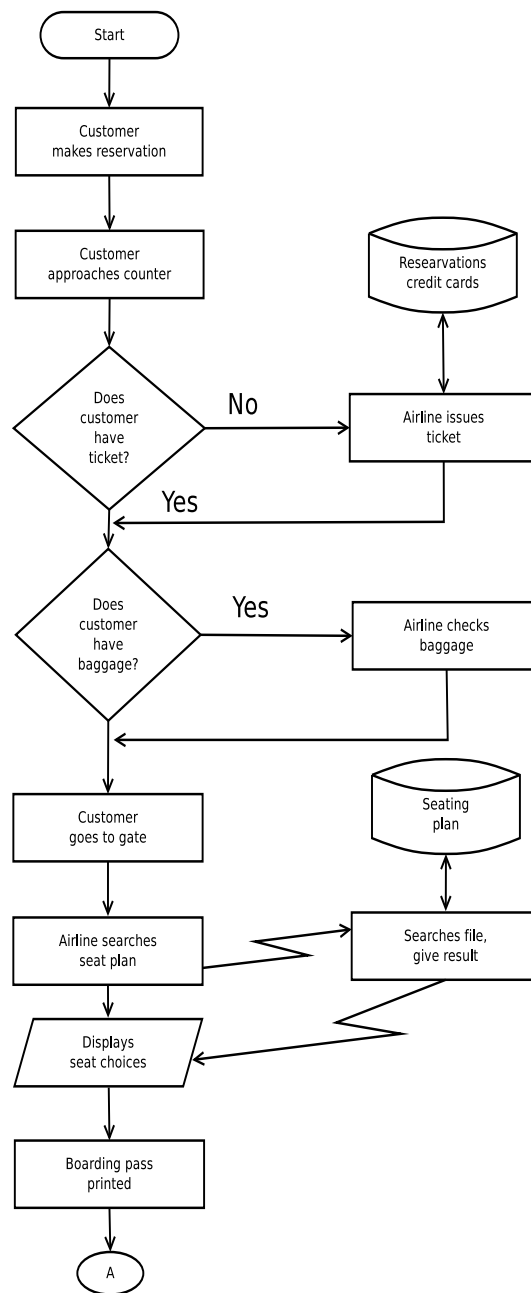
57.2.2 Some Other Methods

Boxplots

A boxplot shows the central tendency and dispersion of a data set and indicates the skewness (deviation from symmetry) and kurtosis (measure of tail length).

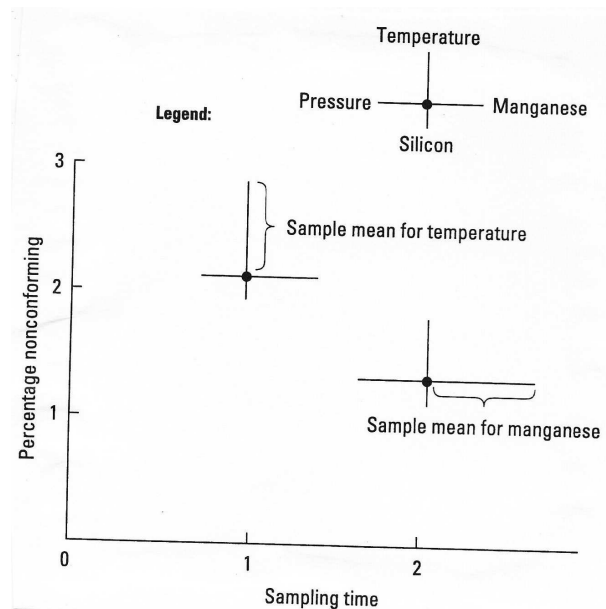
Flowcharts

Flowchart is the graphical method for displaying a system's operation and sequence. This is a very familiar type of flowchart as the program flowchart, sometimes called the logic diagram.



Star plots

Starplot is also known as radial plot, for which the different variables of interest correspond to different rays emanating from a star. The length of each ray represents the magnitude of the variable.



Quantile-quantile plots

A comparison of the distribution of two data sets can be made through the use of empirical quantile-quantile plots.

The quantiles of a data set are simply the ordered data values themselves. The p th quantile, $Q(p)$, is the data value such that the proportion of data values that are less than or equal to it is p . Thus, the median is $Q(0.5)$.

Example 57.1. Sample of a size 30 are collected from each of two mixers A and B in a production facility. The polypropylene content is measured for each sample. Determine if the distribution of polypropylene content is the same for the two mixers.

Mixer A					Mixer B				
48.6	49.6	48.3	48.9	48.9	49.8	47.9	47.6	48.5	48.9
49.5	49.3	49.4	49.4	49.2	47.7	49.6	49.5	48.3	48.2
54.5	54.6	54.9	52.9	53.2	52.4	52.0	50.2	50.5	53.2
55.5	53.5	53.5	54.1	53.7	52.7	50.9	51.9	51.6	51.2
49.8	51.5	52.8	52.1	49.9	47.4	47.0	47.6	45.8	45.4
50.9	51.4	50.2	50.4	49.8	46.2	45.9	46.4	46.8	47.2

Solution.



If the plotted points lie on, or close to, a straight line, the samples are assumed to be from the same distribution.

In case the samples from the two data sets are unequal in size, a minor modification is necessary.

Brainstorms

Brainstorming is a way to elicit a large number of ideas from a group of people in a short period of time.

Members of the group use their collective thinking power to generate ideas and unrestrained thoughts.

It is used to determine problems to work on, to find possible causes of a problem, to find solutions to a problem and to find ways to implement solutions

Quality circle

Quality circle is typically an informal group of people that consists of operators, supervisors, managers and so on, who get together to improve ways to make the product or deliver the service.

The concept behind is that, in most cases, the persons who are closest to an operation are in a better position to contribute ideas that will lead to an improvement in it.

It tries to overcome barriers that may exist within the prevailing organizational structure so as to foster an open exchange of ideas.

Key to its success is its participative style of management. The group members are actively involved in the decision-making process and therefore develop a positive attitude toward creating a better product or service.

Whatever suggestions that a quality circle comes up with will be examined by management for feasibility. Thus, members of management team must understand the workings and advantages of the proposed action clearly.

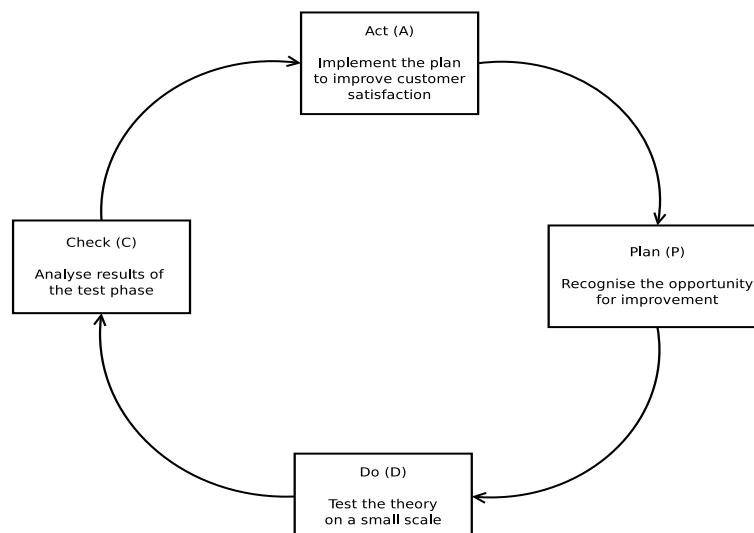
Deming's PDCA cycle

Deming cycle is a method that can aid management in stabilizing a process and pursuing continuous, never-ending process improvement.

It composed of 4 basic stages : Plan, Do, Check/Study and Act (the PDCA or PDSA cycle)

A plan is developed (Plan); the plan is tested on a trial basis using planned experiments (Do); the effects of the experiments are monitored (Check/Study); and appropriate actions are taken on the process (Act). These actions can lead to a new or modified plan, so the PDCA cycle continues forever in an uphill cycle of never-ending improvement.

The Deming cycle operates by recognizing that problems (opportunities for improvement) in a process are determined by difference between customer needs and process performance. If the difference is large, customer dissatisfaction may be high, but there's great opportunity for improvement. If the difference is small, the consequent opportunity for improvement is diminished.



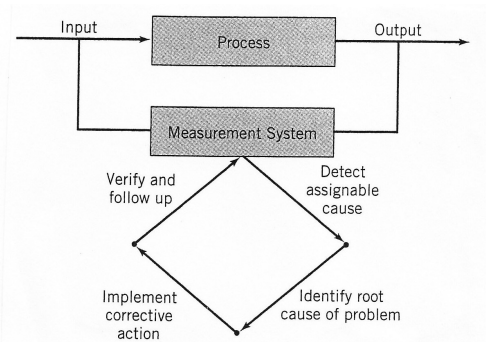


Figure 4-5 Process improvement using the control chart.

57.3 Introduction to ISO

What are standards?

Standards are documented agreements containing technical specifications or other precise criteria to be used consistently as rules, guidelines or definitions of characteristics, to ensure that material, products, processes and services are fit for their purpose.

Why is international standardization needed?

The existence of non-harmonized standards for similar technologies in different countries or regions can contribute to so-called “technical barriers to trade”. Export-minded industries have long sensed the need to agree on world standards to help rationalize the international trading process. This was the origin of the establishment of ISO. International standardization is well-established for many technologies in such diverse fields as information processing and communications, textiles, packaging, distribution of goods, energy production and utilization, shipbuilding, banking and financial services. It will continue to grow in importance for all sectors of industrial activity for the foreseeable future.

What is Quality?

Quality is the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs.

What is Quality System?

Product quality depends on many variables, such as the calibre of the components or materials used; type of equipment used in design, production, handling, installation, testing and shipping; the equipment calibration and maintenance procedures employed; the training and experience of production and supervisory personnel; the level of workmanship; and sometimes the environmental conditions (temperature, humidity, level of dust particles) in the area where the product is produced. The process, organizational structure, procedures and resources that manufacturers and suppliers use to control these variables to produce a product of consistent quality which meets defined specifications is called a quality system. The standards that are being adopted globally for quality systems are the ISO 9000 standards.

What is ISO?

Quality programs have been so effective in driving down costs and increasing competitiveness. Total quality systems are no longer an option - they are required. Companies without quality programs are at risk.

The International Standards Organization (ISO) is a worldwide federation of national standards bodies from some 130 countries, one from each country. ISO is a non-governmental organization established in 1947. The mission of ISO is to promote the development of standardization and related activities in the world with a view of facilitating the international exchange of goods and services and to developing cooperation in the spheres of intellectual, scientific, technological and economic activity.

Who makes up ISO?

A member body of ISO is the national body “most representative of standardization in its country”. Thus, only one body in each country may be admitted to membership of ISO. ISO is made up of its members which are divided into three categories:

A member body takes the responsibility for:

- * informing potentially interested parties in their country of relevant international standardization opportunities and initiatives;
- * ensuring that a concerted view of the country's interests is presented during international negotiations leading to standards agreements;
- * providing their country's share of financial support for the central operations of ISO, through payment of membership dues.

Who does the work?

The technical work of ISO is highly decentralized, carried out in hierarchy of some 2850 technical committees, subcommittees and working groups. In these committees, qualified representative of industries, research institutes, government authorities, consumer bodies and international organizations from all over the world come together as equal partners in the resolution of global standardization problems. Some 30,000 experts participate in meetings each year.

The major responsibility for administering a standards committee is accepted by one of the national standards bodies that make up the ISO membership - AFNOR, ANSI, BSI, etc. The member body holding the secretariat of a standards committee normally appoints one or two persons to do the technical and administrative work. A committee chairman assists committee members in reaching consensus. Generally, a consensus will mean that a particular solution to the problem at hand is the best possible one for international application at that time

The Central Secretariat in Geneva acts to ensure the flow of documentation in all directions, to clarify technical points with secretariats and chairmen, and to ensure that the agreements approved by the technical committees are edited, printed, submitted as draft International Standards to ISO member bodies for voting and published. Meetings of technical committees and subcommittees are convened by the Central Secretariat, which coordinates all such meetings with the committee secretariats before setting the date and place.

How are ISO standards developed?

ISO standards are developed according to the following principles:

- * Consensus
 - * The views of all interests are taken into account: manufacturers vendors and users, consumer groups, testing laboratories, governments, engineering professions and research organizations.
- * Industry-wide
 - * Global solutions to satisfy industries and customers worldwide.
- * Voluntary
 - * International standardization is market-driven and therefore based on voluntary involvement of all interests in the market place.

Generic management system standards

The vast majority of ISO standards are highly specific to a particular product, material or process. However, both ISO 9000 and ISO 14000 are known as generic management system standards.

Generic means that the same standards can be applied to any organization, large or small, whatever its product - including whether its "product" is actually a service - in any sector of activity and whether it is a business enterprise, a public administration or a government.

Management system refers to what the organization does to manage its processes or activities. In a very small organization, there is probably no "system", as such, just "our way of doing things" and "our way" is probably not written down but all in the manager's or owner's head. The larger the organization and the more people involved, the more the likelihood that there are some written procedures, instructions, forms or records. These help ensure that everyone is not just "doing his or her thing" and that there is a minimum of order in the way the organization goes about its business, so that, time, money and other resources are utilized efficiently.

To be really efficient and effective, the organization can manage its way of doing things by systemizing it. This ensures that nothing important is left out and that everyone is clear about who is responsible for doing what, when, how, why and where.

Management system standards provide the organization with a model to follow in setting up and operating the management system. This model incorporates the features which experts in the field have agreed upon as representing the state of the art. A management system which follows the model or “conforms to the standard” - is built on a firm foundation of state-of-the-art practices.

Large organizations, or ones with complicated processes, could not function well without management systems. Companies in such fields as aerospace, automobiles, defence or health products have been operating management systems for years.

ISO’s management system standard now make these successful practices available for all organizations.

ISO 9000 and ISO 14000

Both “ISO 9000” and “ISO 14000” are actually families of standards which are referred to under these generic titles for convenience. Both families consist of standards and guidelines relating to management systems and related supporting standards on terminology and specific tools such as auditing (the process of checking that the management system conforms to the standard).

ISO 9000 is primarily concerned with “quality management”. Like “beauty”, everyone may have his or her idea of what “quality” is. In plain language, the standardized definition of “quality” in ISO 9000 refers to all those features of a product (or service) which are required by the customer. “Quality management” means what the organization does to ensure that its products conform to the customer’s requirements.

ISO 14000 is primarily concerned with “environmental management”. In plain language, this means what the organization does to minimize harmful effects on the environment caused by its activities.

Both ISO 9000 and ISO 14000 concern the way an organization goes about its work and not directly the result of this work. In other words, they both concern processes and not product - at least, not directly. Nevertheless, the way in which the organization manages its processes is obviously going to affect its final product. In the case of ISO 9000, it is going to affect whether or not everything has been done to ensure that the product meets the customer’s requirements. On the case of ISO 14000, it is going to affect whether or not everything has been done to ensure a product will have the least harmful impact on the environment, either during production or disposal, either by pollution or by depleting natural resources.

However, neither ISO 9000 nor ISO 14000 are product standards. The management system standards in these families state requirements for what the organization must do to manage processes influencing quality (ISO 9000) or the processes influencing the impact of the organization’s activities on the environment (ISO 14000).

In both cases, the philosophy is that these requirements are generic. No matter what the organization is or does, if it wants to establish a quality management system or an environmental management system, then such a system has a number of essential features which are spelled out in ISO 9000 or ISO 14000.

In summary, the ISO 9000 family of standards represents an international consensus on good management practices with the aim of ensuring that the organization can time and time again deliver the product or services that meet the client’s quality requirements. These good practices have been distilled into a set of standardized requirements for a quality management system, regardless of what your organization does, its size, or whether it’s in the private, or public sector.

What Are The Benefits to Your Company?

The benefits of becoming certified vary from one organization to another.

To summarize, certification has provided our clients with:

- * market advantage over their competitors
- * higher and more visible quality
- * product/service consistency
- * reduced external quality audits
- * increased client satisfaction
- * improved documentation

- * continuous improvement of their products and services
- * better company image
- * reduced source inspections
- * increased efficiency in all activities and operations
- * positive cultural change
- * reduced scrap, waste, rework and repair
- * increased productivity and profitability
- * comply with customer requirements
- * increase capacity to do business
- * lower operating costs
- * improved customer service

Why should my organization implement ISO 9000?

The existence of an organization without customers, or with dissatisfied customers, is in peril! To keep customers - and to keep them satisfied - your product (which may, in fact, be a service) needs to meet their requirements. ISO 9000 provides a tried and tested framework for taking a systematic approach to managing your business processes (your organization's activities) so that they consistently turn out product conforming to the customer's expectations. And that means consistently happy customers!

How does the ISO 9000 family of standards work?

The requirements for a quality system have been standardized - but most of us like to think our business is unique. So how does ISO 9000 allow for the diversity of say, on the one hand, a "Mr. And Mrs." Enterprise, and on the other, a multinational manufacturing company with service components, or a public utility, or a government administration?

The answer is that ISO 9000 lays down what requirements your quality system must meet, but does not dictate how they should be met in your organization - which leaves great scope and flexibility for implementation in different business sectors and business cultures as well as different national cultures.

So, the ISO 9000 family includes standards that give organizations guidance and requirements on what constitutes an effective quality management system.

What if my organization implements ISO 9000?

The organization should carry out auditing of its ISO 9000-based quality system itself to verify that it is managing its processes effectively - or, to put it another way, to check that it is fully in control of its activities.

In addition, the organization may invite its clients to audit the quality system in order to give them confidence that the organization is capable of delivering products or services that will meet their requirements.

Lastly, the organization may engage the services of an independent quality system certification body to obtain an ISO 9000 certificate of conformity. The last option has proved extremely popular in the market-place because of the perceived credibility of an independent assessment. It may thus avoid multiple audits by the organization's clients, or reduce the frequency or duration of client audits.

The certificate can also serve as a business reference between the organization and potential clients, especially when supplier and client are new to each other, or far removed geographically, as in an export context.

Who evaluates quality system?

A manufacturer may choose to evaluate this own quality system. Such self-audits are usually major components of the quality system itself. Such self-audits can increase the confidence of management in its production system and demonstrate to its personnel that the firm is committed to quality management.

"Second party" evaluations are also common. In these cases, it is usually the buyer who requires and conducts quality system evaluations of his suppliers. These evaluations are mandatory only for companies wishing to become suppliers to that buyer.

"Third party" quality system evaluations and registrations may be voluntary or mandatory and are conducted by persons or organizations independent of both the supplier and the buyer.

‘ISO certificates’ don’t exist!

The standards which make up the ISO 9000 and ISO 14000 families are among the thousands of voluntary International Standards developed and published by ISO. ISO’s primary role is developing International Standards. However, it also develops Guides, when appropriate, that help to ensure good practice in standards related activities, and other documents that promote the development of standardization, or facilitate the implementation of standard.

ISO itself does not carry out assessments to check that its standards are being implemented by users in conformity with the requirements of the standards. Conformity assessment - as this process is known - is a matter for suppliers and their clients in the private sector, and of regulatory bodies when ISO standards have been incorporated into public legislation.

In addition, there exist many testing laboratories and certification bodies which offer independent (also known as “third party”) conformity assessment services to provide confirmation that products (including hardware, software and processed materials), services or systems measure up to ISO standards. Such organizations may perform these services under a mandate to a regulatory authority, or as a commercial activity, the aim of which is to create confidence between suppliers and their clients. In some countries, ISO members carry out conformity assessment, either on behalf of their respective governments, or as a business operation. ISO itself has no authority to control these activities.

However, in partnership with the IEC (International Electrotechnical Commission), ISO has developed ISO/IEC Guides covering various aspects of conformity assessment activities and the organizations that perform them. The voluntary criteria contained in these Guide represent an international consensus on what constitutes acceptable practice. Their use contributes to the consistency and coherence of conformity assessment worldwide and so facilitates trade across borders.

A point which logically follows on from the above remarks is that it is false to describe a company as “ISO-registered”, or to use the phrases such as “ISO certification”, “ISO certificates” and “ISO registration”. ISO operates no system or assessing the conformance of organizations’ management systems with standards in the ISO 9000 family, or the ISO 14000 family. ISO itself carries out neither ISO 9000 nor ISO 14000 audits and awards no certificates attesting to neither ISO 9000 nor ISO 14000 audits and awards no certificates attesting to conformity with the standards. There is no such thing as “ISO certification”, or “ISO registration”, whether in relation to ISO 9000, ISO 14000, or any other ISO standard.

Both ISO 9000 and ISO 14000 auditing and certification are carried out independently of ISO by certification bodies under their own responsibility.

Chapter 58

Applied Nonparametric Statistics

Experiments with response that can be ordered (ranked) but impossible to make statements such as “outcome A is twice as good as outcome B” (e.g. food A is twice as good as food B) occur in many social science and consumer preference study. Nonparametric statistical methods are useful for analysing this type of data.

Nonparametric statistics should be important in machine learning because a lot of the data has unknown distributions and nonparametric statistics do not assume known distributions.

TO WORK ON. References:

- Daniel, W.W., 1990. Applied nonparametric statistics. 2nd ed. Pacific Grove, CA:Duxbury.
- Conover, W.J., 1999. Practical nonparametric statistics. New York: Wiley.
- Wackerly, D.D., Mendenhall III, W., and Scheaffer, R.L., 2002. Mathematical statistics with applications. 6th ed. California: Duxbury/Thomson Learning. (Chapter 15???)

- Hypothesis testing, Estimation, Measurement scales for Nonparametric statistics???

58.1 Procedures that Utilize Data from a Single Sample

- Making inferences about a location parameter.
 - Making inferences about a population proportion.
 - One-sample runs test for randomness.
 - Cox-Stuart test for trend.

58.2 Procedures that Utilize Data from Two Independent Samples

- Making inferences about the difference between two location parameters.
 - Making inferences about the equality of two dispersion parameters.
 - Some miscellaneous two-sample tests.

58.3 Procedures that Utilize Data from Two Related Samples

- Procedures for testing hypotheses about location parameters.
 - Confidence interval procedures for the median difference.
 - Test for two related samples when data consist of frequencies.

58.4 Procedures that Utilize Data from Three or More Independent Samples

- Extension of the median test.
 - Kruskal-Wallis one-way analysis of variance by ranks.

- Jonckheere-Terpstra test for ordered alternatives.
- Multiple comparisons.

58.5 Procedures that Utilize Data from Three or More Related Samples

- Friedman two-way analysis of variance by ranks.
 - Multiple-comparison procedure for use with Friedman test.
 - Page's test for ordered alternatives.
 - Durbin test for incomplete block designs.
 - Cochran's test for related observations.

58.6 Goodness-of-Fit Tests

- Chi-square goodness-of-fit test.
 - Kolmogorov-Smirnov one-sample and two-sample tests.
 - Confidence band for the population distribution function.
 - Chi-square and Kolmogorov-Smirnov goodness-of-fit tests: a comparison.
 - Other goodness-of-fit tests.

58.7 Rank Correlation And Other Measures of Association

- Spearman rank correlation coefficient.
 - Kendall's tau.
 - Confidence interval for tau.
 - Kendall's coefficient of concordance.
 - Partial rank correlation.
 - Measures of association for contingency tables.
 - Other measures of association

Chapter 59

Stochastic Processes

TO WORK ON. References:

Ross, S.M., 2014. Introduction to probability models. 11th ed. San Diego, CA: Academic Press.

Daniel, J.W., 2008. Poisson processes and mixture distributions. SOA Study Note.

Papoulis, A. and Pillai, S.U., 2002. Probability, random variables and stochastic processes. New York: McGraw Hill.

59.1 Introduction

- Review of probability theory, jointly distributed random variables
 - Generating function
 - Limit theorems
 - Computing probabilities by conditioning
 - Computing expectations by conditioning
 - Examples of stochastic processes

59.2 Introduction to Markov Chain

- Chapman-Kolmogorov equations
 - Classification of states
 - Limiting probabilities
 - Mean time spent in transient states
 - Absorption probabilities
 - Time reversible Markov chains

59.3 Poisson Processes

- Exponential distribution
 - Properties of the exponential distribution
 - Convolutions of exponential random variables
 - Poisson processes
 - Counting processes
 - Definition of the Poisson process
 - Interarrival and waiting time distributions
 - Further properties of Poisson processes
 - Conditional distribution of the arrival times
 - Nonhomogeneous Poisson processes
 - Compound Poisson processes

59.4 Continuous-Time Markov Chain

- Continuous-time Markov Chain
 - Birth and death processes
 - Balance equations and stationary probabilities
 - The transition probability function
 - Kolmogorov's forward and backward equations
 - Renewal processes
 - Binomial / Bernoulli processes
 - Regenerative processes

59.5 Queuing Theory

- Cost equations
 - Steady-state probabilities
 - Network of queues (open systems, closed systems)
 - The model G/M/1
 - The G/M/k queue
 - The M/M/k queue

Part IX

Operations Research & Optimisation

Chapter 60

Optimisation

In engineering, one deals with many nonlinear optimisation problems such as finding the most cost effective or energy effective method to achieve an engineering goal. In this chapter, we learn simple numerical techniques to solve optimisation problems.

60.1 Numerical Optimisation Problems and Direct Search Methods

Direct search methods apply primarily to strictly unimodal 1-variable functions. The idea of these methods is to identify the *interval of uncertainty* that contains the optimal solution point. The procedure locates the optimum by iteratively narrowing the interval of uncertainty to any desired level of accuracy. We will discuss only the golden section method. This method is used to find the maximum value of a unimodal function $f(x)$ over a given interval $[a, b]$.

Definition 60.1. A function $f(x)$ is *unimodal* on $[a, b]$ if it has exactly one maximum (or minimum) on $[a, b]$.

60.2 Golden Section Search

The general steps for this method are as follows:

Let $f(x)$ be a unimodal function over a given interval $[a, b]$ with the optimal point x^* . Let $I_{k-1} = (x_L, x_R)$ be the current interval of uncertainty with $I_0 = [a, b]$. Define

$$x_1 = x_R - r(x_R - x_L), \quad x_2 = x_L + r(x_R - x_L)$$

where $r = \frac{\sqrt{5}-1}{2} = 0.618 \dots$ is the *golden ratio*. (Clearly, $x_L < x_1 < x_2 < x_R$.) The next interval of uncertainty I_k is determined in the following way:

1. If $f(x_1) > f(x_2)$, then $x_L < x^* < x_2$. Set $x_R = x_2$ and $I_k = [x_L, x_2]$.
2. If $f(x_1) < f(x_2)$, then $x_1 < x^* < x_R$. Set $x_L = x_1$ and $I_k = [x_1, x_R]$.
3. If $f(x_1) = f(x_2)$, then $x_1 < x^* < x_2$. Set $x_L = x_1$, $x_R = x_2$, and $I_k = [x_1, x_2]$.

Remark 60.1. 1. The choice of x_1 and x_2 ensures that $I_k \subset I_{k-1}$.

2. Let $L_k = \|I_k\|$, the length of I_k . Then the algorithm terminates at iteration k if $L_k < \epsilon$, the user specified level of accuracy.
3. It can be seen that $L_k = rL_{k-1} = r^k(b - a)$. Thus, the algorithm will terminate at k iterations where $L_k < \epsilon$.

Example 60.2. Find the maximum value of $f(x) = 3 + 6x - 4x^2$ on the interval $[0, 1]$ using the golden section search method with the final interval of uncertainty having a length less than 0.25.

Solution. Solving

$$r^k < 0.25$$

for the number k of iterations that must be performed, we obtain

$$k > 2.88.$$

Thus 3 iterations of golden section search must be performed.

- Iteration 1: $x_L = 0, x_R = 1$

$$x_1 = x_R - \frac{\sqrt{5}-1}{2}(x_R - x_L) = 0.381966,$$

$$x_2 = x_L + \frac{\sqrt{5}-1}{2}(x_R - x_L) = 0.618034$$

$$f(x_1) = 4.708204 < f(x_2) = 5.180340 \Rightarrow \text{take } x_L = x_1 = 0.381966 \text{ with the same } x_R.$$

- Iteration 2: $x_L = 0.381966, x_R = 1$.

$$x_1 = x_R - \frac{\sqrt{5}-1}{2}(x_R - x_L) = 0.618034,$$

$$x_2 = x_L + \frac{\sqrt{5}-1}{2}(x_R - x_L) = 0.763932$$

$$f(x_1) = 5.180340 < f(x_2) = 5.249224 \Rightarrow \text{take } x_L = x_1 = 0.618034 \text{ with the same } x_R \\ \text{i.e., } x_R = 1.$$

- Iteration 3: $x_L = 0.618034, x_R = 1$

$$x_1 = x_R - \frac{\sqrt{5}-1}{2}(x_R - x_L) = 0.763932,$$

$$x_2 = x_L + \frac{\sqrt{5}-1}{2}(x_R - x_L) = 0.854102,$$

$$f(x_1) = 5.249224 > f(x_2) = 5.206651 \Rightarrow \text{take } x_L \text{ remains the same with the same } x_R = x_2 = 0.854102 \text{ i.e., } x_L = 0.618034, x_R = 0.854102.$$

Thus, $I_3 = [0.618034, 0.854102]$ and $L_3 = 0.854102 - 0.618034 = 0.236068 < 0.25$ as wanted.

So the required maximum point lies within $I_3 = [0.618034, 0.854102]$. □

60.3 Simple Gradient Methods

Recall that the gradient vector of a function f at a point $\mathbf{x}$, $\nabla f(\mathbf{x})$, points in the direction of maximum increase (or the direction of steepest ascent). And, $-\nabla f(\mathbf{x})$, is the direction of maximum decrease (or the direction of steepest descent).

Theorem 60.3 (Method of Steepest Ascent/Gradient Method). *An algorithm for finding the nearest local maximum of a twice continuous differentiable function $f(\mathbf{x}_k)$, which presupposes that the gradient of the function can be computed. The method of steepest ascent, also called the gradient method, starts at a point $\mathbf{x}_0$ and, as many times as needed, moves from $\mathbf{x}_k$ to $\mathbf{x}_{k+1}$ by maximising along the line extending from $\mathbf{x}_k$ in the direction of $\nabla f(\mathbf{x}_k)$, the local uphill gradient. That is, we determine the value of t and the corresponding point*

$$\mathbf{x}_{k+1} = \mathbf{x}_k + t\nabla f(\mathbf{x}_k)$$

at which the function

$$g(t) = f(\mathbf{x}_{k+1}(t))$$

has a maximum. We take $\mathbf{x}_{k+1}$ as the next approximation after $\mathbf{x}_k$.

This method has the severe drawback of requiring a great many iterations (hence a slow convergence) for functions which have long, narrow valley structures.

Example 60.4. Use the method of steepest ascent to determine a maximum of $f(\mathbf{x}) = -x^2 - y^2$ starting from the point $\mathbf{x}_0 = (1, 1)$.

Solution. $\nabla f(\mathbf{x}) = (-2x, -2y)$

At $\mathbf{x}_0 = (1, 1)$, $\nabla f = (-2, -2)$.

The new approximation is

$$\mathbf{x}_1 = \mathbf{x}_0 + t\nabla f(\mathbf{x}_0)$$

where t is a value that will maximize the function

$$h(t) = f(\mathbf{x}_0 + t\nabla f(\mathbf{x}_0)) = f(1 - 2t, 1 - 2t) = -(1 - 2t)^2 - (1 - 2t)^2.$$

Solving $h'(t) = 0$, we obtain $4(1 - 2t) + 4(1 - 2t) = 0 \Rightarrow t = 1/2$.

Hence $\mathbf{x}_1 = (1, 1) + \frac{1}{2}(-2, -2) = (0, 0)$.

Now $\nabla f(\mathbf{x}_1) = \nabla f(0, 0) = (0, 0)$, and we terminate the algorithm. So the maximum point is $\mathbf{x}_1 = (0, 0)$ with maximum value $f(0, 0) = 0$. $\square$

Example 60.5. Starting at the point $(0, 0)$, use one iteration of the steepest-descent algorithm to approximate the minimum value of the function

$$f(x, y) = 2(x + y)^2 + (x - y)^2 + 3x + 2y.$$

Solution. $f_x = 4(x + y) + 2(x - y) + 3 = 6x + 2y + 3$, $f_y = 4(x + y) - 2(x - y) + 2 = 2x + 6y + 2$

$\nabla f(x, y) = (f_x, f_y) = (6x + 2y + 3, 2x + 6y + 2)$

Start with the initial approximation $\mathbf{x}_0 = (0, 0)$, we obtain the next approximation

$\mathbf{x}_1 = \mathbf{x}_0 + t\nabla f(\mathbf{x}_0) = (0, 0) + t(3, 2) = (3t, 2t)$ since $\nabla f(\mathbf{x}_0) = \nabla f(0, 0) = (3, 2)$.

Find t that minimises $h(t) = f(\mathbf{x}_1) = f(3t, 2t) = 2(5t)^2 + (t)^2 + 3(3t) + 2(2t) = 51t^2 + 13t$

Solve $h'(t) = 102t + 13 = 0 \Rightarrow t = -13/102 = -0.127$

Hence $\mathbf{x}_1 = (-0.382, -0.255)$ and $f(\mathbf{x}_1) = f(-0.382, -0.255) = -0.828 \approx \min f$. $\square$

60.4 Other topics

Parabolic Extrapolation

Conjugate Gradient Method

Minimisation with Constraints

Linear Programming, Simplex method and sensitivity analysis

Nonlinear programming

Integer Programming

Stochastic Problems

Chapter 61

Operations Research

TO WORK ON. References:

!!! Optimised Decision Makings under Uncertainty (Randomness)

<https://zayenz.se/blog/post/rotating-workforce-scheduling/>

<https://www.minizinc.org/>

<https://news.ycombinator.com/item?id=45772153>

* <https://buttondown.com/hillelwayne/archive/many-hard-leetcode-problems-are-easy-constraint/>

https://lobste.rs/s/8ubfdd/many_hard_leetcode_problems_are_easy

RamaMurthyP-OperationsResearch2e-2008

BruckerP-SchedulingAlgorithms5e-Springer2007

Winston, W.L., 2004. Operations Research: Applications and Algorithms. 4th ed. Belmont, CA: Brooks/Cole Cengage Learning.

Taha, H.A., 2003. Operations Research: An Introduction. 7th ed. Upper Saddle River, N.J.: Prentice Hall.

Hiller, F.S. and Lieberman, G.J., 2001. Introduction to operation research. 7th ed. New York: McGraw-Hill.

JaredL Cohon-MultiobjectiveProgrammingAndPlanning-AP1978 KaufmannA-FaureR-IntroductionToOp AP1968

JorgeNocedal-StephenJWright-NumericalOptimization-1999

AnthonyV.Fiacco-GarthP.McCormick-NonlinearProgrammingSequentialUnconstrainedMinimizationTech ClassicsInAppliedMathematics-SocietyForIndustrialMathematics1987 AnthonyV.Fiacco-MathematicalProgra

LectureNotesInPureAndAppliedMathematics195-MarcelDekker1998 DrewFudenberg-JeanTirole-

GameTheory-TheMITPress1991 OlviL.Mangasarian-NonlinearProgramming-ClassicsInAppliedMathematics

SocietyForIndustrialAndAppliedMathematics1987 StephenP.Bradley-ArnoldoC.Hax-ThomasL.Magnanti-

AppliedMathematicalProgramming-1977

61.1 Linear Programming

- The graphical solution of two-variable linear programming problems
 - Case studies of a diet problem; a working-scheduling problem; a capital budgeting problem
 - Using linear programming to solve multi-period decision problems with application in financial models and work scheduling

61.1.1 The Simplex Algorithm and Goal Programming

- Converting linear programming to standard form
 - Direction of unboundedness
 - Optimal feasible solutions to a linear programming problem
 - The Simplex Algorithm
 - Using the simplex algorithm to solve minimization problem
 - Alternative optimal solutions

- Unbounded linear programming problem
- Degeneracy and convergence of the simplex algorithm
- The Big M method
- Unrestricted-in-sign variables
- Multi-attribute decision making in the absence of uncertainty

61.1.2 Sensitivity Analysis and Duality

- A graphical approach introduction to sensitivity analysis
 - Finding the dual of a linear programming
 - Economic interpretation of the dual problem
 - The dual theorem and its consequences
 - Shadow prices
 - Complementary slackness
 - The dual simplex method

61.2 Transportation, Assignment and Transshipment Problems

- Formulating transportation problems
 - Finding basic feasible solutions for transportation problems
 - The transportation simplex method
 - Sensitivity analysis for transportation problems
 - Problems on assignment and transshipment

61.3 Integer Programming

- Introduction to integer programming
 - Formulating integer programming
 - The branch-and-bound method for solving integer programming problems
 - Solving combinatorial optimization problems by the branch-and-bound method
 - Implicit enumeration

61.4 Game Theory

- Two-person zero-sum and constant-sum games
 - Saddle points
 - Randomized strategies, domination, and graphical solution
 - Linear Programming and zero-sum games.
 - Two-person nonconstant-sum games

61.5 Inventory Models

- Deterministic EOQ inventory models
 - Continuous-Rate EOQ Model
 - Quantity Discount Model
 - Shortages Model

61.6 Nonlinear Programming

- Local extremums
 - Concave and convex functions
 - Solving NLP with one variable
 - Golden Section Search

- Unconstrained optimization of several variables
- The method of steepest ascent
- Lagrange multipliers
- The Kuhn-Tucker conditions

61.7 Queuing Theory

- Modeling arrival and service processes
 - Birth-death processes
 - The M/M/1/FCFS/ ∞/∞ model
 - The M/M/c/FCFS/ ∞/∞ model

61.8 Decision Analysis

- Decision criteria
 - Utility theory
 - Decision trees
 - Bayes rule and decision trees

61.9 Dynamic Programming

Additional Reference: DreyfusSE-LawAM-TheArtAndTheoryOfDynamicProgramming-AP1977

- Deterministic DP: Two puzzles
- Problems on network, inventory, resource allocation
- Problems on equipment replacement
- Formulating DP recursions
- Wagner-Whitin algorithm
- Silver-Meal heuristic

61.10 Network Models

- CPM, PERT, minimum cost network flow problems
 - Minimum spanning tree problems
 - Shortest-path problems
 - Maximum flow problems

References

- 2015-whitepaper M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL <http://tensorflow.org/>. Software available from tensorflow.org.
- bib:qm E. S. Abers. *Quantum Mechanics*. Prentice-Hall/Pearson Education, Inc., Upper Saddle River, NJ, 2004. (A mathematical physics book that uses bra-ket $|\psi\rangle$ notation).
- and_datab:da119 S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Database*. Addison-Wesley, 1995. <http://webdam.inria.fr/Alice/>.
- apl_variab:da42 M. J. Ablowitz and A. S. Fokas. *Complex Variables: Introduction and Applications*. Cambridge University Press, 2003.
- amowitz+stegun M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth dover printing, tenth gpo printing edition, 1964. (http://en.wikipedia.org/wiki/Abramowitz_and_Stegun).
- on_pricin:da93 Y. Achdou and O. Pironneau. *Computational Methods for Option Pricing*, volume 30 of *Frontiers in Applied Mathematics*. Philadelphia: Society for Industrial and Applied Mathematics, 2005.
- n1990numerical F. Acton. *Numerical Methods that Work*. MAA spectrum. Mathematical Association of America, 1990. ISBN 9780883854501. URL <http://books.google.com.my/books?id=YHXU4W3Ez2MC>.
- 9:_compl:math3 L. V. Ahlfors. *Complex Analysis: An Introduction to the Theory of Analytic Functions of One Complex Variable*. Auckland: McGraw-Hill Book Co., New York, 3rd ed. edition, 1979. ISBN 0-07-000657-1.
- CS:114768:da54 A. V. Aho and J. D. Ullman. *Foundations of Computer Science*. Computer Science Press, Inc., New York, NY, USA, 1992. ISBN 0-7167-8233-2.
- _elemen_method J. E. Akin. *Application and Implementation of Finite Element Methods*. Academic Press, 1982.
- ts_applic:da14 M. A. Al-Gwaiz. *Sturm-Liouville Theory and its Applications*. Springer-Verlag London Limited, 2008.
- erson1999lapack E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition, 1999. ISBN 0-89871-447-8 (paperback). URL <http://www.netlib.org/lapack/lug/>.
- t_quadr_method A. D. Andricopoulos, M. Widdicks, P. W. Duck, and D. P. Newton. Universal option valuation using quadrature methods. *Journal of Financial Economics*, 67:447–471, 2003.
- dynam_system_i D. V. Anosov and V. I. Arnold, editors. *Dynamical Systems I*. Number 1 in Encyclopaedia of Mathematical Sciences. Springer-Verlag Berlin Heidelberg, 1988.

- ppl_variab:da10 M. Y. Antimirov, A. A. Kolyshkin, and R. Vaillancourt.
- Complex Variables*
- . Academic Press, 1998.

am_logics:da118 A. W. Appel, R. Dockins, A. Hobor, L. Beringer, J. Dodds, G. Stewart, S. Blazy, and X. Leroy.
- Program Logics For Certified Compilers*
- . Cambridge University Press, 2014.

ldl992ordinary V. Arnold.
- Ordinary Differential Equations*
- . Springer Textbook Series. New York, 1992. ISBN 9783540548133. URL
- <http://books.google.com.my/books?id=JUoyqlW7PZgC>
- .

numerical:da27 U. Ascher, R. Mattheij, and R. Russell.
- Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*
- . Classics in applied mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 1994. ISBN 9781611971231. URL
- <http://books.google.com.my/books?id=0RvOrnO4WgwC>
- .

er_series:da28 N. H. Asmar.
- Partial Differential Equations with Fourier Series and Boundary Value Problems*
- . Pearson Education, Inc., Upper Saddle River, NJ, 2nd edition, 2005.

mer_analy:da67 K. Atkinson and W. Han.
- Theoretical Numerical Analysis: A Functional Analysis Framework*
- . Texts in Applied Mathematics 39. Springer Science+Business Media, LLC, 3rd ed. edition, 2009.

b_system:da117 K. J. Åström and R. M. Murray.
- Feedback Systems: An Introduction for Scientists and Engineers*
- . Princeton University Press, 2008.

ue_probl:da107 P. B. Bailey, L. F. Shampine, and P. E. Waltman.
- Nonlinear Two-Point Boundary Value Problems*
- . Academic Press, 1968.

mpl_analy:da88 J. Bak and D. Newman.
- Complex Analysis*
- . Springer Science+Business Media, LCC, 3rd ed. edition, 2010.

ly_power_func W. Balser.
- From Divergent Power Series to Analytic Power Functions*
- . Number vol. 1582 in Lecture Notes in Math. Springer Verlag, New York, 1994.

r_system:da108 W. Balser.
- Formal Power Series and Linear Systems Of Meromorphic Ordinary Differential Equations*
- . Springer-Verlag New York, Inc., 2000.

em_logic:da119 D. W. Barnes and J. M. Mack.
- An Algebraic Introduction to Mathematical Logic*
- , volume 22 of
- Graduate Texts in Mathematics*
- . Springer-Verlag New York Inc., 1975.

templates R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. V. der Vorst.
- Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods, 2nd Edition*
- . SIAM, Philadelphia, PA, 1994.

met_model:da22 R. H. Bartels, J. C. Beatty, and B. A. Barsky.
- An Introduction to Splines for use in Computer Graphics and Geometric Modeling*
- . Morgan Kaufmann Publishers, Inc., 1987.

roof_logic:da08 J. Barwise and J. Etchemendy.
- Language, Proof and Logic*
- . CSLI Publications, 1999. (a introduction to logic).

am_astrod:da111 R. R. Bate, D. D. Mueller, and J. E. White.
- Fundamentals of Astrodynamics*
- . Dover Publications, Inc. New York, 1971.

ic_engin:da108 K. J. Beers.
- Numerical Methods for Chemical Engineering: Applications in MATLAB*
- . Cambridge University Press, New York, NY, 2006. ISBN 9780521859714.

ien_engin-da11 W. W. Bell.
- Special Functions for Scientists and Engineers*
- . D. Van Nostrand Company Ltd, London, 1968a.

ien_engin:da11 W. W. Bell.
- Special Functions for Scientists and Engineers*
- . D. Van Nostrand Company Ltd, 1968b.

- pex_optim:da118 A. Ben-Tal and A. Nemirovski.
- Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*
- . MPS/SIAM Series on Optimization. Society for Industrial and Applied Mathematics, 2001.

apl_variab:da84 C. A. Berenstein and R. Gay.
- Complex Variables: An Introduction*
- . Graduate Texts in Mathematics. Springer-Verlag New York Inc., 1991.

c_lagran_inter J.-P. Berrut and L. N. Trefethen. Barycentric lagrange interpolation.
- SIAM Review*
- , 46(3): 501–517, 2004.

t_volum_i:da118 D. P. Bertsekas.
- Dynamic Programming and Optimal Control Volume I*
- . Athena Scientific, Belmont, Massachusetts, 1995a.

t_volum_ii:da118 D. P. Bertsekas.
- Dynamic Programming and Optimal Control Volume II*
- . Athena Scientific, Belmont, Massachusetts, 1995b.

p_method:da117 D. P. Bertsekas.
- Constrained Optimization and Lagrange Multiplier Methods*
- . Athena Scientific, Belmont, Massachusetts, 1996.

ffer_equat:da75 D. Betounes.
- Differential Equations: Theory and Applications*
- . Springer Science+Business Media, LLC, 2 edition, 2010.

ntr_estim:da92 J. T. Betts.
- Practical Methods for Optimal Control and Estimation Using Nonlinear Programming*
- . Society for Industrial and Applied Mathematics, 2nd edition, 2010.

or_mappin:da91 L. Bieberbach.
- Conformal Mapping*
- . Chelsea Publishing Company, New York, 1964.

it_theor_tech J. Bird.
- Electrical Circuit Theory and Technology*
- . Routledge (Taylor & Francis Group), 5th ed. edition, 2014.

chin_learn:da66 C. M. Bishop.
- Pattern Recognition and Machine Learning*
- . Information Science and Statistics. Springer Science+Business Media, LLC, 2006.

tw_engin:da12e D. Bjørner.
- Software Engineering 1: Abstraction and Modelling*
- . Springer-Verlag Berlin Heidelberg, 2006.

bjoerner09 N. Bjørner and L. de Moura. z3
- ¹⁰
- : Applications, enablers, challenges and directions. In
- Sixth International Workshop on Constraints in Formal Verification Grenoble, France*
- , 2009.

o_corpor_liabil F. Black and M. Scholes. The pricing of options and corporate liabilities.
- Journal of Political Economy*
- , 81(3):637–654, May–Jun. 1973. URL
- <http://EconPapers.repec.org/RePEc:ucp:jpolec:v:81:y:1973:i:3:p:637-54>
- .

n_comput:da117 F. J. Blanco-Silva.
- Learning SciPy for Numerical and Scientific Computing*
- . PACKT Publishing Ltd., 2013.

ve_phenom:da83 N. Bleistein.
- Mathematical Methods for Wave Phenomena*
- . Academic Press, Inc., 1984.

d_differ_equat G. W. Bluman and S. C. Anco.
- Symmetry and Integration Methods for Differential Equations*
- . Springer-Verlag New York, Inc., 2002.

t_differ_equat G. W. Bluman and S. Kumei.
- Symmetries and Differential Equations*
- . Springer-Verlag, 1989.

n_physic_sci M. L. Boas.
- Mathematical Methods in the Physical Sciences*
- . John Wiley & Sons, Inc., ???, 3rd edition, 2006.

fourier_analy A. Boggess and F. J. Narcowich.
- A First Course in Wavelets with Fourier Analysis*
- . Prentice-Hall, Inc., Upper Saddle River, NJ, 2001.

t_differ_equat G. Boole. A treatise on differential equations, 1859. URL
- <http://www.archive.org/details/atreatiseondiff06boolgoog>
- .

- comput_logic:1 G. S. Boolos, J. P. Burgess, and R. C. Jeffrey. *Computability and Logic*. Cambridge University Press, 5th edition, 2007.
- formal_logic:da13 R. Bornat. *Proof and Disproof in Formal Logic: An introduction for programmers*. Oxford University Press, 2005.
- func_relat_integ:1 D. Borwein and J. M. Borwein. Some remarkable properties of sinc and related integrals. *The Ramanujan Journal*, 5(1):73–89, March 2001.
- mechat_system:1 E.-K. Boukas and F. M. AL-Sunni. *Mechatronic Systems: Analysis, Design and Implementation*. Springer-Verlag Berlin Heidelberg, 2011a.
- and_implem:da106 E.-K. Boukas and F. M. AL-Sunni. *Mechatronic Systems Analysis Design And Implementation*. Springer-Verlag Berlin Heidelberg, 2011b.
- and_value_prob:1 W. E. Boyce and R. C. DiPrima. *Elementary Differential Equations and Boundary Value Problems*. John Wiley & Sons, Inc., Hoboken, N. J., 10th ed. edition, 2012. URL [\url{http://www.amazon.com/Elementary-Differential-Equations-Boundary-Problems/dp/0470458313/}](http://www.amazon.com/Elementary-Differential-Equations-Boundary-Problems/dp/0470458313/).
- convex_optim:da92 S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- and_its_applic:1 R. N. Bracewell. *The Fourier Transform and Its Applications*. McGraw-Hill, ???, 3rd edition, 2000.
- earn_opencv:da37 G. Bradski and A. Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly Media, Inc., 2008.
- an_econom:da12e P. Brandimarte. *Numerical Methods in Finance and Economics: A MATLAB-Base Introduction*. John Wiley & Sons, Inc., 2nd ed. edition, 2006.
- al_proces:da11 D. Brandwood. *Fourier Transforms in Radar and Signal Processing*. Artech House, Inc., 2003.
- ffer_equat:da09 F. Brauer and J. A. Nohel. *The Qualitative Theory of Ordinary Differential Equations: An Introduction*. W. A. Benjamin, Inc. New York, 1969. A nice book which is published by Dover Publications, Inc.
- at_their_applic:1 M. Braun. *Differential Equations and Their Applications: An Introuction to Applied Mathematics*. Number 11 in Text in Applied Mathematics. Springer Science+Business Media, Inc., 4th edition, 1993.
- al_value_prob:1 K. E. Brenan, S. L. Campbell, and L. R. Petzold. *Numerical Soltion of Initial-Value Problems in Differential-Algebraic Equations*. SIAM, 1996.
- en_method:da111 S. C. Brenner and L. Ridgway Scott. *The Mathematical Theory of Finite Element Methods*. Texts in Applied Mathematics 15. Springer Science+Business Media, LLC, 3rd ed. edition, 2008.
- elemen_method:1 S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer Science+Business Media, LLC, 3rd edition, 2008.
- ts_applic:da68 E. O. Brigham. *The Fast Fourier Transform and Its Applications*. Prentice-Hall, Englewood Cliffs, NJ, 1988.
- mathem:da12m2 I. N. Bronshtein, K. A. Semendyayev, G. Musiol, and H. Muehlig, editors. *Handbook of Mathematics*. Springer-Verlag Berlin Heidelberg, 5th ed. edition, 2007.
- and_value_prob:1 J. W. Brown and R. V. Churchill. *Fourier Series and Boundary Value Problems*. McGraw-Hill, New York, 7th edition, 2008.
- heff_ode_solver:1 P. N. Brown, G. D. Byrne, and A. C. Hindmarsh. **VODE**, a variable-coefficient ode solver. *SIAM J. Sci. Stat. Comput.*, 10:1038–1951, 1989.

- in_simul_equat C. G. Broyden. A class of methods for solving nonlinear simultaneous equations. *Mathematics of Computation*, 19(92):577–593, Oct. 1965.
- _diff:da12m2 M. Bücker, G. Corliss, P. Hovland, U. Naumann, and B. Norris, editors. *Automatic Differentiation: Applications, Theory and Implementations*. Springer-Verlag Berlin Heidelberg, 2006.
- 5:_numer_analy R. L. Burden and J. D. Faires. *Numerical Analysis*. Boston, MA: Brooks/Cole Cengage Learning, 10th edition, 2015. URL <https://sites.google.com/site/numericalanalysis1burden/>.
- _m_physic:da114 E. Butkov. *Mathematical Physics*. Addison-Wesley Publishing Company, Inc., 1968.
- _ies_spher:da13 W. E. Byerly. *An Elementary Treatise on Fourier’s Series and Spherical, Cylindrical, and Ellipsoidal Harmonics, with Applications to Problems in Mathematical Physics*. Ginn and Company, 1893.
- :_spect_method C. G. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral Methods: Fundamentals in Single Domains*. Springer-Verlag Berlin Heidelberg, 2006.
- _pl_variab:da28 C. Carathéodory. *Theory of Functions of a Complex Variable, Volume 1 & 2*. Chelsea Publishing Company, New York, 1954. (A very nice book on basic complex analysis).
- _m_illus:da118 B. Casselman. *Mathematical Illustrations: A Manual of Geometry and PostScript*. Cambridge University Press, 2005. <https://www.math.ubc.ca/~cass/graphics/manual/>.
- _m_simul:da111 F. E. Cellier and E. Kofman. *Continuous System Simulation*. Springer Science+Business Media, Inc., 2006.
- _eat_mass_trans Y. A. Cengel and A. J. Ghajar. *Heat and Mass Transfer: Fundamentals & Applications*. McGraw-Hill, 4th ed. edition, 2011.
- _n_of_theor_and P. D. Chateau and D. W. Zachmann. *Schaum’s Outline of Theory and Problems of Partial Differential Equations*. McGraw-Hill, New York, 1986.
- _chen15:_mxnet T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. In *Neural Information Processing Systems, Workshop on Machine Learning Systems*, 2015. URL <https://github.com/apache/incubator-mxnet>.
- _er_handb:da109 W.-K. Chen, editor. *The Circuits and Filters Handbook: Fundamentals of Circuits and Filters*. CRC Press, Taylor & Francis Group, LLC, 3rd ed. edition, 2009.
- _ree_body_probl A. Chenciner. Poincaré and the three-body problem. *Séminaire Poincaré XVI*, pages 44–133, 2012.
- _m_scien_comput E. W. Cheney and D. R. Kincaid. *Numerical Mathematics of Scientific Computing*. Boston, MA: Brooks/Cole, Cengage Learning., 7th edition, 2012. (The authors homepage are <http://www.cs.utexas.edu/~kincaid/> and <http://www.ma.utexas.edu/CNA/DMY/ewc.html>).
- _at_applic:da38 C. Chicone. *Ordinary Differential Equations with Applications*. Number 34 in Texts in Applied Mathematics. Springer-Verlag New York, Inc., 1999.
- _them_logic12e2 I. Chiswell and W. Hodges. *Mathematical Logic*. Oxford University Press, 2007. (a really nice and rigorous introduction to logic).
- _b_applic:da111 R. V. Churchill and J. W. Brown. *Complex Variables and Applications*. McGraw Hill Higher Education, Boston, 8th edition, 2008.
- _us_effec:da120 W. F. Clocksin. *Clause and Effect: Prolog Programming for the Working Programmer*. Springer-Verlag Berlin Heidelberg, 1997.

- er_equat:da117 E. A. Coddington and N. Levinson. *Theory of Ordinary Differential Equations*. McGraw-Hill Book Company, Inc., 1984.
- l_elemen_func W. J. Cody, Jr and W. Waite. *Software Manual for the Elementary Functions*. Prentice-Hall, Inc., 1980.
- hon_hdf5:da117 A. Collette. *Python and HDF5: Unlocking Scientific Data*. O' Reilly Media, Inc., 2014.
- f_guided_waves R. E. Collin. *Field Theory of Guided Waves*. IEEE Press, ???, 2nd edition, 1991. A good book that demonstrates how various techniques in PDE and complex analysis could be applied to the study of travelling waves and waveguides. The second chapter contains SL problem and the expansion of $\delta(x)$ in terms of eigenfunctions which is applied in Green's function.
- y_algor_approac S. D. Conte and C. de Boor. *Elementary Numerical Analysis, An Algorithm Approach*. Auckland: McGraw-Hill, 3rd edition, 1981.
- pl_variab:da08 J. B. Conway. *Functions of One Complex Variable I*, volume 11 of *Graduate Texts in Mathematics*. Springer Science & Business Media, Inc., 2nd ed. edition, 1978.
- variab_ii:da78 J. B. Conway. *Functions of One Complex Variable II*, volume 159 of *Graduate Texts in Mathematics*. Springer-Verlag New York Inc., 1995.
- hod_finan:da12 G. Cornuejols and R. Tütüncü. *Optimization Methods in Finance*. Cambridge University Press, 2007.
- em_physic:da08 R. Courant and D. Hilbert. *Methods of Mathematical Physics*, volume I. Interscience, New York, 1953.
- 6:_what_mathem R. Courant, H. Robbins, and I. Stewart. *What is Mathematics*. Oxford University Press, 2nd edition, 1996.
- it_volum_algor T. N. Croft. *Unstructured Mesh Finite Volume Algorithms For Swirling Turbulent Reacting Flows*. PhD thesis, University of Greenwich, 1998. URL <http://gala.gre.ac.uk/6371/>.
- od_scien2:da117 R. Dautray and J.-L. Lions. *Mathematical Analysis and Numerical Methods for Science and Technology Volume 2: Functional and Variational Methods*. Springer-Verlag Berlin Heidelberg, 1988.
- s_linear_system T. A. Davis. *Direct Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, 2006.
- 491491.2491498 T. A. Davis. Algorithm 930: Factorize: An object-oriented linear system solver for matlab. *ACM Trans. Math. Softw.*, 39(4):28:1–28:18, July 2013. ISSN 0098-3500. 10.1145/2491491.2491498. URL <http://doi.acm.org/10.1145/2491491.2491498>.
- ct_analy:da114 B. Daya Reddy. *Introductory Functional Analysis: With Applications to Boundary Value Problems and Finite Elements*. Springer Science+Business Media New York, 1998.
- _b_splin_basic C. de Boor. *B(asic)-Spline Basics*. MRC TSR. Mathematics Research Center, University of Wisconsin-Madison, 1986. URL <http://books.google.com.my/books?id=TkA-OAAACAAJ>.
- ide_splin:da47 C. de Boor. *A Practical Guide to Splines*, volume 27 of *Applied Mathematical Sciences*. Springer-Verlag New York, Inc., 2001.
- is_modul_theor L. de Moura and N. Bjørner. Satisfiability modulo theories: An appetizer. In M. V. M. Oliveira and J. Woodcock, editors, *Formal Methods: Foundations and Applications*, pages 23–36, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-10452-7.
- U:1096889:da27 J. E. Dennis, Jr. and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations (Classics in Applied Mathematics, 16)*. Society for Industrial and Applied Mathematics, 1996. ISBN 0898713641.

- er_equat:da105 P. Deuffhard and F. Bornemann. *Scientific Computing with Ordinary Differential Equations*, volume 42 of *Texts in Applied Mathematics*. Springer-Verlag New York, Inc., 2002.
- en_comput:da105 P. Deuffhard and A. Hohmann. *Numerical Analysis in Modern Scientific Computing: An Introduction*, volume 43 of *Text in Applied Mathematics*. Springer-Verlag New York, Inc., 2 edition, 2003.
- mer_solut_pdes P. Deuffhard and M. Weiser. *Adaptive Numerical Solution of PDEs*. Walter de Gruyter GmbH & Co. KG, Berlin/Boston, 2012.
- com_fluid_flow M. O. Deville, P. F. Fischer, and E. H. Mund. *High-Order Methods for Incompressible Fluid Flow*. Cambridge University Press, 2004.
- 0:_centur_loop F. Diacu. A century-long loop: The solution of the n -body problem. *The Mathematical Intelligencer*, 22(2):19–25, 2000.
- ic_fittin_splin P. Dierckx. *Curve and Surface Fitting with Splines*. Oxford University Press, New York, 1993.
- ied_optim:da47 U. Diwekar. *Introduction to Applied Optimization*, volume 22 of *Springer Optimization and Its Applications*. Springer Science+Business Media, LLC, 2nd edition, 2008.
- ac_trans:da109 G. Doetsch. *Introduction to the Theoy and Application of the Laplace Transformation*. Springer-Verlag Berlin Heidelberg New York, 1974.
- bab_count:da78 J. L. Doob. *Classical Potential Theory and Its Probabilistic Counterpart*. Springer-Verlag Berlin Heidelberg New York, 2001.
- r_system:da109 R. C. Dorf and R. H. Bishop. *Modern Control Systems*. Pearson Education, Inc., 12th ed. edition, 2011.
- s_mappin:da110 T. A. Driscoll and L. N. Trefethen. *Schwarz-Christoffel Mapping*. Cambride University Press, 2002.
- ter_class:da11 R. O. Duda, P. E. Har, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, 2001.
- ffer_equat:da32 D. G. Duffy. *Transform Methods for Solving Partial Differential Equations*. Chapman & Hall/CRC, 2004.
- nan_engin:da63 D. J. Duffy. *Finite Difference Methods in Financial Engineering: A Partial Differential Equation Approach*. John Wiley & Sons, Ltd, 2006.
- es_integ:da12ml H. Dym and H. P. McKean. *Fourier Series and Integrals*. Academic Press, 1972.
- r_mathem_applic S. S. Epp. *Discrete Mathematics with Applications*. Cengage Learning Inc., 5th edition, 2020.
- git_arith:da70 M. D. Ercegovac and T. Lang. *Digital Arithmetic*. Morgan Kaufmann Publishers, 2004.
- y_trans_volum_i A. Erdelyi. *Tables of Integral Transforms, Volume I*. McGraw-Hill Book Company, Inc., 1954.
- gar_poten:da104 G. C. Evans. *The Logarithmic Potential: Discontinuous Dirichlet and Neumann Problems*. American Mathematical Society, 1927.
- ffer_equat:da08 L. C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. AMS, Providence, RI, 1998.
- faires2003 J. D. Faires and R. Burden. *Numerical Methods, 3rd Ed*. Brooks/Cole (Thomson Learning, Inc.), 2003.
- scien_and_engin S. J. Farlow. *Partial Differential Equations for Scientists and Engineers*. Dover, New York, 1993.

- erm_struc_model D. Filipović. *Term-Structure Models: A Graduate Course*. Springer-Verlag Berlin Heidelberg, 2009.
- ts_finit_elemen J. Fish and T. Belytschko. *A First Course in Finite Elements*. John Wiley & Sons, Ltd, 2007.
- o:_compl_variab S. D. Fisher. *Complex Variables*. Dover, 1999.
- princ_of_elect R. Fitzpatrick. *Maxwell's Equations and the Principles of Electromagnetism*. Infinity Science Press, LLC, Hingham, Massachusetts, 2008.
- bal_optim:dall18 C. A. Floudas and P. M. Pardalos. *Handbook of Test Problems in Local and Global Optimization*, volume 33 of *Nonconvex Optimization and Its Applications*. Springer-Science+Business Media Dordrecht, 1999.
- numer_method_c B. H. Flowers. *An Introduction to Numerical Methods in C++*. Oxford University Press, 2000.
- fer_equat:dall14 G. B. Folland. *Introduction to Partial Differential Equations*. Princeton University Press, 2nd ed. edition, 1995. (Chapter 1 introduces the Lewy example which tells that a general PDE may have no solution even if the coefficients are smooth).
- atom_func:da38 L. R. Ford. *Automorphic Functions*. McGraw-Hill, 1929.
- nduc_sets:da08 T. Forster. *Logic, Induction and Sets*. London Mathematical Society Student Texts. Cambridge University Press, 2003. 10.1017/CBO9780511810282. URL <https://www.dpmms.cam.ac.uk/~tff/baby.html>.
- aly_theor_heat J. Fourier. *The Analytical Theory of Heat*. At The University Press, Cambridge, 1878.
- ycomjnl/4.3.265 J. G. F. Francis. The qr transformation a unitary analogue to the lr transformation—part 1. *The Computer Journal*, 4(3):265–271, 1961. 10.1093/comjnl/4.3.265. URL [+http://dx.doi.org/10.1093/comjnl/4.3.265](http://dx.doi.org/10.1093/comjnl/4.3.265).
- ycomjnl/4.4.332 J. G. F. Francis. The qr transformation—part 2. *The Computer Journal*, 4(4):332–345, 1962. 10.1093/comjnl/4.4.332. URL [+http://dx.doi.org/10.1093/comjnl/4.4.332](http://dx.doi.org/10.1093/comjnl/4.4.332).
- 05complex:da62 E. Freitag and R. Busam. *Complex Analysis*. Universitext (Berlin. Print). Springer, 2005. ISBN 9783540257240. URL <http://books.google.com.my/books?id=M-fI9tPrBRIC>.
- mpl_analy:da62 E. Freitag and R. Busam. *Complex Analysis*. Springer-Verlag Berlin Heidelberg, 2nd ed. edition, 2009.
- put_scien:da08 J. H. Gallier. *Logic for Computer Science: Foundations of Automatic Theorem Proving*. Dover Publications, 2015.
- g_maple_matlab W. Gander and J. Hřebíček. *Solving Problems in Scientific Computing Using Maple and MATLAB*. Springer-Verlag Berlin Heidelberg, 3rd edition, 1997.
- _comput_statist J. E. Gentle. *Computational Statistics*. Statistics and Computing. Springer Science+Business Media, LLC, 2009. DA68.
- e08:_mesh_gener P.-L. George and S. Frey. *Mesh Generation: Application to Finite Elements*. John Wiley & Sons, Ltd., 2nd edition, 2008.
- an_vibrat:dall18 M. Géradin and D. J. Rixen. *Mechanical Vibrations: Theory and Application to Structural Dynamics*. John Wiley & Sons, Ltd, 3rd edition, 2015. (amazingly good engineering text).
- ed_numer_analy C. F. Gerald and P. O. Wheatley. *Applied Numerical Analysis*. Boston: Pearson/Addison-Wesley, 7th edition, 2004.
- ect_engin:da86 W. J. Gibbs. *Conformal Transformations in Electrical Engineering*. The British Thomson-Houston Co., Ltd., 1958.

- 9:_microm:da70 V. Giurgiutiu and S. E. Lyshevski. *Micromechatronics: Modeling Analysis and Design with MATLAB*. CRC Press, Taylor & Francis Group, LLC, 2nd ed. edition, 2009.
- at_topic:da12m1 H.-J. Glaeske, A. P. Prudnikov, and K. A. Skòrník. *Operational Calculus and Related Topics*. Chapman & Hall/CRC, Taylor and Francis Group, LLC, 2006.
- Golub:1996:MC G. H. Golub and C. F. V. Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. The Johns Hopkins University Press, Baltimore, MD, USA, third edition, 1996. ISBN 0-8018-5413-X (hardcover), 0-8018-5414-8 (paperback).
- at_scilab:da118 C. Gomez, C. Bunks, J.-P. Chancelier, F. Delebecque, M. Goursat, R. Nikoukhah, and S. Steer. *Engineering and Scientific Computing with Scilab*. Springer Science+Business Media New York, 1999.
- ue_probl:da114 E. A. González-Velasco. *Fourier Analysis and Boundary Value Problems*. Academic Press, Inc., 1995.
- ep_learn:da117 I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. Massachusetts Institute of Technology, 2016. URL <http://www.deeplearningbook.org/>.
- em_desig:da110 G. C. Goodwin, S. F. Graebe, and M. E. Salgado. *Control System Design*. Prentice Hall, 2000.
- hem_analy:da52 E. Goursat. A course in mathematical analysis, 1904. URL <https://archive.org/details/courseinmathemat02gouruoft>.
- an_mathem:da26 T. Gowers, J. Barrow-Green, and I. Leader. *The Princeton Companion to Mathematics*. Princeton University Press, Princeton, 2008. ISBN 9781400830398 1400830397.
- ateg_trans:da79 U. Graf. *Introduction to Hyperfunctions and Their Integral Transforms: An Applied and Computational Approach*. Birkhäuser / Springer Basel AG, 2010.
- ier_analy:da41 L. Grafakos. *Classical and Modern Fourier Analysis*. Pearson Education, Inc., Upper Saddle River, NJ, 2004. (<http://www.math.missouri.edu/~eloukas/FourierAnalysis.html>).
- er_harmon:da78 H. Groemer. *Geometric Applications of Fourier Series and Spherical Harmonics*. Cambridge University Press, 1996.
- l_differ_equat C. Grossmann and H.-G. Roos. *Numerical Treatment of Partial Differential Equations*. Springer-Verlag Berlin Heidelberg, 2007.
- moon_earth_sun M. C. Gutzwiller. Moon-earth-sun: The oldest three-body problem. *Review of Modern Physics*, 70(2):589–, April 1998.
- ffer_equat:da11 R. Haberman. *Elementary Applied Partial Differential Equations with Fouries Series and Boundary Value Problems*. Prentice-Hall, Englewood Cliffs, New Jersey, 2nd ed. edition, 1987.
- d_iterat_method L. A. Hageman and D. M. Young. *Applied Iterative Methods*. Academic Press, 1981.
- iffer_equat_ii E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer-Verlag Berlin Heidelberg, second revised edition, 1996.
- differ_equat_i E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Number 8 in Springer Series in Comput. Mathematics. Springer-Verlag, second revised edition, 1993.
- et_numer_integ E. Hairer, C. Lubich, and G. Warner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer-Verlag Berlin Heidelberg, 2002.
- and_engin:da59 R. W. Hamming. *Numerical Methods for Scientists and Engineers*. Dover Publications, Inc., New York, 2nd edition, 1986.
- od_physic-da10 C. Harper. *Analytic Methods in Physics*. Wiley-VCH Verlag Berlin GmbH, 1999.

- om_reason:da66 J. Harrison. *Handbook of Practical Logic and Automated Reasoning*. Cambridge University Press, 2009.
- art1978computer J. F. Hart. *Computer approximations*. SIAM series in applied mathematics. R. E. Krieger Pub. Co., 1978. ISBN 9780882756424. URL <http://books.google.com.my/books?id=7vVQAAAAMAAJ>.
- ffer_equat:da09 P. Hartman. *Ordinary Differential Equations*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 2nd ed. edition, 1982. ISBN 9780898719222. URL <https://books.google.com.my/books?id=NEkkJ9309okC>.
- is_learn:da115 T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer Science+Business Media, LLC, 2nd edition, 2008.
- ut_graph_openg D. Hearn and M. P. Baker. *Computer Graphics with OpenGL*. Pearson Education, 3rd ed. edition, 2004.
- MR3043640 F. Hecht. New development in freefem++. *J. Numer. Math.*, 20(3-4):251–265, 2012. ISSN 1570-2820.
- ten_theor:da66 L. L. Helms. *Introduction to Potential Theory*. Wiley-Interscience, 1969.
- n_differ_equat P. Henrici. *Discrete Variable Methods in Ordinary Differential Equations*. John Wiley & Sons, 1962.
- mpl_analy:da08 P. Henrici. *Applied and Computational Complex Analysis Volume 1: Power Series, Integration, Conformal Mapping, Location of Zeros*. John Wiley & Sons, Inc., 1974. (generalised to Banach algebra).
- aly_volum-da08 P. Henrici. *Applied and Computational Complex Analysis Volume 2: Special Functions, Asymptotics, Continued Fractions*. John Wiley & Sons, Inc., 1977.
- en_numer_analy P. Henrici. *Essentials of Numerical Analysis*. Wiley, 1982.
- sen02:b_splin K. Henriksen and K. Erleben. B-splines. Technical report, Department of Computer Science, University of Copenhagen, August 2002. DIKU Teknisk rapport 02/17.
- gic_every:da08 R. A. Herrmann. Logic for everyone. arXiv:0601709, 29 Jan 2006.
- er_method:da108 J. S. Hesthaven and T. Warburton. *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications*, volume 54 of *Texts in Applied Mathematics*. Springer Science+Business Media, LLC, 2008.
- igham:1996:ASN N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1996. ISBN 0-89871-355-2 (paperback). Typeset with L^AT_EX2e.
- igham:2002:ASN N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, second edition, 2002a. ISBN 0-89871-521-0.
- er_algor:da12m1 N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, 2nd. edition, 2002b.
- rc_lagran_inter N. J. Higham. The numerical stability of barycentric lagrange interpolation. *IMA Journal of Numerical Analysis*, 24(4):547–556, 2004.
- nct_theor:da73 E. Hille. *Analytic Function Theory Volumes 1 & 2*. Ginn and Company, 1959.

- chem_logic:da98 P. G. Hinman. *Fundamentals of Mathematical Logic*. A. K. Peters, Ltd., 2005. (abstract mathematics).
- 0:_catbox:da83 W. Hochstättler and A. Schliep. *CATBox: An Interactive Course in Combinatorial Optimization*. Springer-Verlag Berlin Heidelberg, 2010.
- an:243156:da08 J. D. Hoffman. *Numerical Methods for Engineers and Scientists*. McGraw-Hill, New York, NY, 1992.
- ta_analy:da117 M. H. Holmes. *Introduction to Scientific Computing and Data Analysis*, volume 13 of *Texts in Computational Science and Engineering*. Springer International Publishing Switzerland, 2016.
- _matlab:da12m1 E. Holzbecher. *Environmental Modeling Using MATLAB®*. Springer-Verlag Berlin Heidelberg, 2007.
- ard_probl:da84 J. Hromkovič. *Algorithmics for Hard Problems: Introduction to Combinatorial Optimization, Randomization, Approximation, and Heuristics*. Texts in Theoretical Computer Science. Springer-Verlag Berlin Heidelberg New York, 2001.
- er_equat1:da109 J. H. Hubbard and B. H. West. *Differential Equations: A Dynamical Systems Approach — Ordinary Differential Equations*. Springer-Verlag New York, Inc., 1991.
- er_equat2:da109 J. H. Hubbard and B. H. West. *Differential Equations: A Dynamical Systems Approach — Higher Dimensional Systems*. Springer-Verlag New York, Inc., 1995.
- en_method:da111 T. J. R. Hughes. *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis*. Prentice-Hall, Inc., 1987.
- en_advec:da111 W. Hundsdorfer and J. Verwer. *Numerical Solution of Time-Dependent Advection Diffusion Reaction Equations*. Number 33 in Springer Series in Computational Mathematics. Springer-Verlag Berlin Heidelberg, 2003.
- ut_scienc:da100 M. Huth and M. Ryan. *Logic in Computer Science: Modelling and Reasoning about Systems*. Cambridge University Press, 2nd edition, 2004.
- ed_differ_equat P. E. Hydon. *Symmetry Methods for Differential Equations: A Beginner's Guide*. Cambridge: Cambridge University Press, 2000.
- er_method:da107 E. Isaacson and H. B. Keller. *Analysis of Numerical Methods*. John Wiley & Sons, New York, 1966.
- ss_elect:da104 J. D. Jackson. *Classical Electrodynamics*. John Wiley & Sons, Inc., 3rd ed. edition, 1999.
- er_engin_mathem G. James. *Modern Engineering Mathematics*. Pearson Education Limited, 5th ed. edition, 2015.
- ert2010gnuplot P. Janert. *Gnuplot in Action: Understanding Data with Graphs*. Manning Pubs Co Series. Manning Publications, 2010. ISBN 9781933988399. URL <http://books.google.com.my/books?id=wSbKQwAACAAJ>.
- vol_exper:da86 R. C. Jennison. *Fourier Transforms and Convolutions for the Experimentalist*. Pergamon Press, 1961.
- jia2014caffe Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. URL <https://github.com/BVLC/caffe>{and}<http://caffe.berkeleyvision.org>.
- l_differ_equat F. John. *Partial Differential Equations*. Number 1 in Applied mathematical Sciences. Springer-Verlag, New York, 4th edition, 1982.

- p_r_differ_equat R. M. Johnson.
- Linear Differential and Difference Equations: A Systems Approach for Mathematicians and Engineers*
- . Chichester: Albion Publishing Limited, 1997.

set_theor:da119 P. T. Johnstone.
- Notes on Logic and Set Theory*
- . Cambridge University Press, 1987.

fer_equat:da114 D. W. Jordan and P. Smith.
- Nonlinear Ordinary Differential Equations: An Introduction for Scientists and Engineers*
- . Oxford University Press, 4th ed. edition, 2007.

r_method_softw D. Kahaner, C. Moler, and S. Nash.
- Numerical Methods and Software*
- . Prentice-Hall, Inc., 1989.

en_guide_wavel G. Kaiser.
- A Friendly Guide to Wavelets*
- . Birkhäuser, 1994.

high_perfor_low L. Ke.
- Design and Analysis of High Performance Low Noise Oscillators and Phase Lock Loops*
- . PhD thesis, School of Electronics and Computer Science, University of Southampton, United Kingdom, 2010.

lin_equat:da08 C. T. Kelly.
- Iterative Methods for Linear and Nonlinear Equations*
- . Society for Industrial and Applied Mathematics, 1995.

method_mechan N. Kikuchi.
- Finite Element Methods in Mechanics*
- . Cambridge University Press, 1986. (Using Basic Programming Language).

in_series:da77 K. Knopp.
- Theory and Application of Infinite Series*
- . Blackie & Son Limited, London and Glasgow, 1944.

knuth86:_metaf D. E. Knuth.
- MetaFont: the Program*
- . Addison-Wesley, 1986.

KnuthTAOCP1 D. E. Knuth.
- The Art of Computer Programming, Volume 1: Fundamental Algorithms, 3rd ed.*
- Addison-Wesley, 1998a.

KnuthTAOCP2 D. E. Knuth.
- The Art of Computer Programming, Volume 2: Seminumerical Algorithms, 3rd ed.*
- Addison-Wesley, 1998b.

r_repres:da111 H. Kober.
- Dictionary of Conformal Representations*
- . Dover, 1957.

r_differ_equat M. Kohn.
- Global Analysis in Linear Differential Equations*
- . Kluwer Academic Publishers, 1999.

l_differ_equat D. A. Kopriva.
- Implementing Spectral Methods for Partial Differential Equations: Algorithms for Scientists and Engineers*
- . Springer Science + Business Media B.V., 2009.

ier_analy:da71 T. W. Körner.
- Fourier Analysis*
- . Cambridge University Press, 1988.

em_physic:da11 N. S. Koshlyakov, M. M. Smirnov, and E. B. Gliner.
- Differential Equations of Mathematical Physics*
- . North-Holland, 1964.

at_applic:da86 M. Kracht and E. Kreyszig.
- Methods of Complex Analysis in Partial Differential Equations with Applications*
- . John Wiley & Sons, Ltd., 1988. (A very theoretical book).

an_engin_mathem E. Kreyszig.
- Advanced Engineering Mathematics*
- . John Wiley & Sons, Hoboken, N. J., 10th ed. edition, 2011.

fer_equat-da104 G. Kristensson.
- Second Order Differential Equations: Special Functions and Their Classification*
- . Springer-Verlag, 2010.

rdw_verif:da112 T. Kropf.
- Introduction to Formal Hardware Verification*
- . Springer-Verlag Berlin Heidelberg, 1999.

OVSKAYA1962637 V. Kublanovskaya. On some algorithms for the solution of the complete eigenvalue problem.
- USSR Computational Mathematics and Mathematical Physics*
- , 1(3):637 – 657, 1962. ISSN 0041-5553.
- [http://dx.doi.org/10.1016/0041-5553\(63\)90168-X](http://dx.doi.org/10.1016/0041-5553(63)90168-X)
- . URL
- <http://www.sciencedirect.com/science/article/pii/004155536390168X>
- .

- with_valid:da35 U. Kulisch. *Computer Arithmetic and Validity: Theory, Implementation, and Applications*. Walter de Gruyter GmbH & Co. KG. 10785 Berlin, Germany, 2008.
- en_engin:da104 Y. K. Kwok. *Applied Complex Variables for Scientists and Engineers*. Cambridge University Press, 2010.
- lamb53:_hydrod H. Lamb. *Hydrodynamics*. Cambridge University Press, 1953. URL <https://archive.org/details/hydrodynamics02lambgoog>. pages 110–120.
- f_system:da109 L. Lamport. *Specifying Systems: The TLA+ Language and Tools for Hardware and Software Engineers*. Addison-Wesley, 2003.
- ng97:underanaly S. Lang. *Undergraduate Analysis*. Undergraduate Texts in Mathematics. Springer Science+Business Media, Inc., 2nd ed. edition, 1997. ISBN 9780387948416.
- mpl_analy:da16 S. Lang. *Complex Analysis*. Graduate Text in Mathematics. Springer-Verlag New York Inc., 4th edition, 1999.
- lax2007linear P. D. Lax. *Linear Algebra and Its Applications*. Number v. 10 in Linear algebra and its applications. Wiley, 2007. ISBN 9780471751564. URL <http://books.google.com.my/books?id=e7FJM6aqZD8C>.
- er2004numerical J. J. Leader. *Numerical Analysis and Scientific Computation*. Pearson Addison Wesley, 2004. ISBN 9780321223357. URL <http://books.google.com.my/books?id=y-XEGAAACAAJ>.
- l_differ_equat R. J. Leveque. *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady State and Time-Dependent Problems*. Society for Industrial and Applied Mathematics, 2007.
- eat_fluid_flow R. W. Lewis, P. Nithiarasu, and K. N. Seetharamu. *Fundamentals of the Finite Element Method for Heat and Fluid Flow*. John Wiley & Sons, Inc., 2004. (Using Fortran Programming Language).
- er_funct:da114 M. J. Lighthill. *An Introduction to Fourier Analysis and Generalised Functions*. Cambridge University Press, 1958.
- ffer_equat:da25 J. D. Logan. *A First Course in Differential Equations*. Undergraduate Texts in Mathematics. Springer Science+Business Media, Inc., 2006.
- er_equat:da114 J. D. Logan. *Applied Partial Differential Equations*. Undergraduate Texts in Mathematics. Springer International Publishing Switzerland, 3rd ed. edition, 2015.
- id_dynam_tech R. Löhner. *Applied Computational Fluid Dynamics Techniques: An Introduction Based on Finite Element Methods*. John Wiley & Sons, Ltd, 2008. (Using Fortran Programming Language).
- _physic_biolog A. J. Lotka. *Elements of Physical Biology*. Baltimore: Williams and Wilkins Company, 1925.
- am_system:da111 D. G. Luenberger. *Introduction to Dynamic Systems: Theory, Models, and Applications*. John Wiley & Sons, Inc., 1979.
- onom_analy:da84 M. Luptáčík. *Mathematical Optimization and Economic Analysis*, volume 36 of *Springer Optimization and Its Applications*. Springer Science+Business Media, LLC, 2010.
- in_comput:da94 Y.-D. Lyuu. *Financial Engineering and Computation: Principles, Mathematics, Algorithms*. Cambridge University Press, 2002.
- al_proces:da63 S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1988.
- chem_logic:da73 Y. I. Manin. *A Course in Mathematical Logic*, volume 53 of *Graduate Texts in Mathematics*. Springer-Verlag New York Inc., 1977.

- thesis/2006Mann S. Mann. *A Blossoming Development of Splines*. Synthesis Lectures on Computer Graphics and Animation. Morgan & Claypool Publishers, 2006. URL <http://dx.doi.org/10.2200/S00041ED1V01200607CGR001>.
- er_equat:da12g P. A. Markowich. *Applied Partial Differential Equations: A Visual Approach*. Springer-Verlag Berlin Heidelberg, 2007.
- joy_of_fourier R. J. Marks II. *Handbook of Fourier Analysis and Its Applications*. Oxford University Press, Inc., 2009.
- pl_variab:da92 A. I. Markushevich. *Theory of Functions of a Complex Variable Volume 1*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1965. (translated by Richard A. Silverman).
- pl_analy:da112 J. E. Marsden. *Basic Complex Analysis*. W. H. Freeman and Company., 3rd ed. edition, 1973.
- marsh2005applied D. Marsh. *Applied Geometry for Computer Graphics and CAD*. Springer Undergraduate Mathematics Series. Springer, 2005. ISBN 9781852338015. URL <http://books.google.com.my/books?id=Q\YN2L39\YC>.
- rs2004numerical J. H. Mathews and K. D. Fink. *Numerical Methods Using MATLAB*. Featured Titles for Numerical Analysis Series. Pearson Prentice Hall, 2004. ISBN 9780130652485. URL <http://books.google.com.my/books?id=E5lRAAAAMAAJ>.
- y_mathem_engin J. H. Mathews and R. W. Howell. *Complex Analysis for Mathematics and Engineering*. Johns and Bartlett Publishers, 3rd ed. edition, 1997.
- or_pract:da117 R. Mattheij and J. Molenaar. *Ordinary Differential Equations in Theory and Practice*. Society for Industrial and Applied Mathematics, 2002.
- m_finit_elemen C. Mattiussi. An analysis of finite volume, finite element, and finite difference methods using some concepts from algebraic topology. *Journal of Computational Physics*, 133:289–309, 1997.
- l_differ_equat R. C. McOwen. *Partial Differential Equations*. Prentice Hall, 2nd ed. edition, 2003.
- er_harmon:da78 T. M. McRobert. *Spherical Harmonics: An Elementary Treatise on Harmonic Functions with Applications*. Methuen & Co. Ltd., 2nd ed. edition, 1947.
- onlin_scienc_faq J. Meiss. Nonlinear science faq, 2003. URL <http://www.faqs.org/faqs/sci/nonlinear-faq/>.
- m_prolog:da119 D. Merritt. *Building Expert Systems in Prolog*. Amzi! inc., 2000.
- in_devel:da119 I. Millington. *Game Physics Engine Development: How to Build a Robust Commercial Grade Physics Engine for Your Game*. Morgan Kaufmann Publishers, 2nd edition, 2010.
- hod_resid:da11 D. S. Mitrinović and J. D. Kečkić. *The Cauchy Method of Residues: Theory and Applications*. D. Reidel Publishing Company, 1984.
- 4dynamik:da108 M. Mitschke and H. Wallentowitz. *Dynamik der Kraftfahrzeuge*. VDI-Buch. Springer, 2004. ISBN 9783540420118. URL <http://books.google.com.my/books?id=i1ld5BFEoe0C>.
- d_maxwel_equat P. Monk. *Finite Element Methods for Maxwell's Equations*. Oxford University Press, 2003.
- ftw_guide:da46 J. J. Moré and S. J. Wright. *Optimization Software Guide*. Society for Industrial and Applied Mathematics, 1993.
- ic_part_i:da10 P. M. Morse and H. Feshbach. *Methods of Theoretical Physics, Part I*. McGraw-Hill Book Company, Inc., 1953a.
- c_part_ii:da10 P. M. Morse and H. Feshbach. *Methods of Theoretical Physics, Part II: Chapters 9 to 13*. McGraw-Hill Book Company, Inc., 1953b.
- 07:digit_contr K. M. Moudgalya. *Digital Control*. John Wiley & Sons, Ltd., 2007.

- p_g_autom_comput D. E. Muller. A method for solving algebraic equations using an automatic computer.
- Mathematical Tables and Other Aids to Computation*
- , 10(208–215), 1956.

:_elemen_func J.-M. Muller.
- Elementary Functions: Algorithms and Implementation*
- . Birkhauser Boston, 1997.

dras_pearl:da16 D. Mumford, C. Series, and D. Wright.
- Indra’s Pearls: The Vision of Felix Klein*
- . Cambridge University Press, 2002.

symp_analy-da25 J. D. Murray.
- Asymptotic Analysis*
- . Number 48 in Applied Mathematical Sciences. Springer-Verlag, 2 edition, 1984.

s_physic:da114 J. Nearing.
- Mathematical Tools for Physics*
- . Dover Publications, 2010. (
- <http://www.physics.miami.edu/~nearing/mathmethods/>
-).

998visual:da110 T. Needham.
- Visual Complex Analysis*
- . Clarendon Press, 1998. ISBN 9780198534464. URL
- <http://books.google.com.my/books?id=ogz5FjmiqlQC>
- .

or_mappin:da60 Z. Nehari.
- Conformal Mapping*
- . Dover, 1952.

mpl_analy:da08 R. Nevanlinna and V. Paatero.
- Introduction to Complex Analysis*
- . Birkhäuser Verlag Basel, 1964. (A book of nice geometrical flavour).

it_elemen_java G. Nikishkov.
- Programming Finite Elements in Java*
- . Springer-Verlag London Limited, 2010.

t_system:da119 N. Nisan and S. Schocken.
- The Elements of Computing Systems: Building a Modern Computer from First Principles*
- . MIT Press, 2005.

m_nonlin_equat T. Noda. The steffensen iteration method for systems of nonlinear equations. II.
- Proc. Japan Acad.*
- , 63(Series A), 1987.

n_numer_recip_c B. A. Ødegaard. Financial numerical recipes in c++, June 2014. URL
- http://finance.bi.no/~bernt/gcc_prog/
- .
- http://finance.bi.no/~bernt/gcc_prog/index.html
- .

tr_engin:da109 K. Ogata.
- Modern Control Engineering*
- . Prentice-Hall, Inc., 4th ed. edition, 2002.

bib:olver2007 P. J. Olver.
- Applied Mathematics Lecture Notes*
- .
- <http://www.math.umn.edu/~olver/appl.html>
- , May 2007.

bib:oneil2003 P. V. O’Neil.
- Advanced Engineering Mathematics*
- . Brooks/Cole, Thomson Learning, ???, 5th edition, 2003.

l_differ_equat P. V. O’Neil.
- Beginning Partial Differential Equations*
- . John Wiley & Sons, Inc., 2nd ed. edition, 2008.

n_mathem:da110 P. V. O’Neil.
- Advanced Engineering Mathematics*
- . Cengage Learning, Stamford, 7th ed. edition, 2012. URL
- <http://www.amazon.com/Advanced-Engineering-Mathematics-Peter-ONeil/dp/1111427410/>
- .

open_logic_text Open Logic Project. The open logic text.
- <http://openlogicproject.org/>
- , 2019.

at_proof:da119 P. Ordning.
- 99 Variations on a Proof*
- . Princeton University Press, 2019.

unct_most_gener W. F. Osgood. On the existence of the green’s function for the most general simply connected plane region.
- Transactions of the American Mathematical Society*
- , 1(3):pp. 310–314, 1900.

t_physic:da111 T. Pang.
- An Introduction to Computational Physics*
- . Cambridge University Press, 2nd ed. edition, 2006.

bin_optim:da50 C. H. Papadimitriou and K. Steiglitz.
- Combinatorial Optimization: Algorithms and Complexity*
- . Prentice-Hall, Inc., New Jersey, 1982.

- for_mappin:da92 N. Papamichael and N. Stylianopoulos. *Numerical Conformal Mapping: Domain Decomposition and the Mapping of Quadrilaterals*. World Scientific, 2010.
- comput_arith:da24 B. Parhami. *Computer Arithmetic: Algorithms and Hardware Designs*. Oxford University Press, 2000.
- net_eigen_probl B. N. Parlett. *The Symmetric Eigenvalue Problem*, volume 20 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, 1998.
- re2017automatic A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in PyTorch. In *NIPS 2017 Autodiff Workshop*, 2017. URL <https://openreview.net/forum?id=BJJsrmfCZ>.
- ar1980numerical S. Patankar. *Numerical Heat Transfer and Fluid Flow*. Electro Skills Series. Hemisphere Publishing Corporation, 1980. ISBN 9780070487406. URL <http://books.google.com.my/books?id=N2MVAQAAIAAJ>.
- prox_polyn:da11 G. M. Phillips. *Interpolation and Approximation by Polynomials*. Springer-Verlag New York, 2003.
- ens1983quadpack R. Piessens, E. de Doncker-Kapenga, C. W. Überhuber, and D. K. Kahaner. *Quadpack: a subroutine package for automatic integration*. Springer series in computational mathematics. Springer-Verlag, 1983. ISBN 9783540125532. URL <http://books.google.com.my/books?id=m2AZAQAAIAAJ>. <http://www.netlib.org/quadpack>.
- nt_method:da119 P. M. Prenter. *Splines and Variational Methods*. John Wiley & Sons, Inc., 1975.
- Press:1992:NRC W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge, UK, second edition, 1992a. ISBN 0-521-43108-5, 0-521-43720-2.
- Press:1992:NRF W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in FORTRAN: The Art of Scientific Computing*. Cambridge University Press, Cambridge, UK, second edition, 1992b. ISBN 0-521-43064-X (book), 0-521-43721-0 (example book) 0-521-43717-2 (diskette), 0-521-43719-9 (diskette), 0-521-43716-4 (diskette).
- ress:1992:NRFa W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in FORTRAN: the art of scientific computing*. Cambridge University Press, Cambridge, UK, second edition, Oct. 1992c. ISBN 0-521-43064-X (book), 0-521-43721-0 (example book) 0-521-43717-2 (diskette), 0-521-43719-9 (diskette), 0-521-43716-4 (diskette). URL <http://www.cbooks.com/sqlnut/SP/search/gtsumt?source=&isbn=052143064X>; <http://www.cbooks.com/sqlnut/SP/search/gtsumt?source=&isbn=0521437164>; <http://www.cbooks.com/sqlnut/SP/search/gtsumt?source=&isbn=0521437172>; <http://www.cbooks.com/sqlnut/SP/search/gtsumt?source=&isbn=0521437199>. Includes MacIntosh disk.
- ress:1996:NRFa W. H. Press, S. A. Teukolsky, and M. Metcalf. *Numerical Recipes in Fortran 90: The Art of Parallel Scientific Computing*, volume 2 of *Fortran Numerical Recipes*. Cambridge University Press, Cambridge, UK, second edition, Sept. 1996a. ISBN 0-521-57439-0 (hardcover). URL <http://www.cbooks.com/sqlnut/SP/search/gtsumt?source=&isbn=0521574390>; http://www.nr.com/nronline_switcher.html.
- ress:1996:NRFb W. H. Press, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in Fortran 77 and Fortran 90: The Art of Scientific and Parallel Computing*. Cambridge University Press, Cambridge, UK, Sept. 1996b. ISBN 0-521-57440-4. URL <http://www.cbooks.com/sqlnut/SP/search/gtsumt?source=&isbn=0521574404>; <http://www.nr.com/>; http://www.nr.com/nronline_switcher.html.
- re:2007:1403886 W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 3 edition, 2007. ISBN 0521880688, 9780521880688.

- olac_trans:da36 A. P. Prudnikov, Y. A. Brychkov, and O. I. Marichev. *Direct Laplace Transforms*, volume 4 of *Integrals and Series*. Gordon and Breach Science Publishers, CRC1992.
- lab_octav:da38 A. Quarteroni and F. Saleri. *Scientific Computing with MATLAB and Octave*. Springer-Verlag Berlin Heidelberg, 2nd edition, 2006.
- l_differ_equat A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations*. Springer-Verlag Berlin Heidelberg, 2008.
- oint_view:da36 H. Rademacher. *Higher Mathematics from an Elementary Point of View*. Birkhäuser Boston, Inc., 1983.
- er_funct:da113 M. Rahman. *Applications of Fourier Transforms to Generalized Functions*. WIT Press, 2011.
- ces_simul:da20 W. F. Ramirez. *Computational Methods in Process Simulation*. Elsevier Ltd., 2nd edition, 1998. A very nice book with concrete application of numerical methods in process engineering system.
- mpl_plane:da79 T. Ransford. *Potential Theory in the Complex Plane*. Cambridge University Press, 1995.
- an_vibrat:da118 S. S. Rao. *Mechanical Vibrations (in SI Units)*. Pearson Education, Inc., 6th edition, 2018.
- aly_desig:da92 R. Raut and M. N. S. Swamy. *Modern Analog Filter Analysis and Design: A Practical Approach*. Wiley-VCH Verlag GmbH & Co. KGaA, 2010.
- hem_logic:da98 W. Rautenberg. *A Concise Introduction to Mathematical Logic*. Universitext. Springer Science+Business Media, LLC, 3rd edition, 2010. (abstract mathematics).
- ion_model:da94 R. Rebonato. *Interest Rate Option Models: Understanding, Analysing and Using Models for Exotic Interest-Rate Options*. John Wiley & Sons, 2nd ed. edition, 1998.
- en_method:da108 J. N. Reddy. *An Introduction to the Finite Element Method*. McGraw-Hill, 3rd edition, 2006.
- physic_ii:da04 M. Reed and B. Simon. *Methods of Modern Mathematical Physics II: Fourier Analysis, Self-Adjointness*. Academic Press, Inc., 1975.
- c_compl_spaces R. Remmert. From riemann surfaces to complex spaces. In *Matériaux pour l'histoire des mathématiques au XXe siècle - Actes du colloque à la mémoire de Jean Dieudonné*, volume 3 of *Séminaires et Congrès*, pages 203–241, 1998. http://www.emis.de/journals/SC/1998/3/html/smf_sem-cong_3_203-241.html.
- er_given_quant B. Riemann. On the number of prime numbers less than a given quantity. *Monatsberichte der Berliner Akademie*, 1859. Translated by Davide R. Wilkins, 1998.
- rsic_engin:da12 K. F. Riley, M. P. Hobson, and S. J. Bence. *Mathematical Methods for Physics and Engineering*. Cambridge University Press, 3rd ed. edition, 2006.
- fer_equat:da47 J. C. Robinson. *An Introduction to Ordinary Differential Equations*. Cambridge University Press, 2004.
- rsic_engin:da28 B. W. Roos. *Analytic Functions and Distributions in Physics and Engineering*. John Wiley & Sons, Inc., 1969.
- esson_i_wish_i G.-C. Rota. Ten lessons i wish i had learned before i started teaching differential equations, 1997. URL <http://www.ega-math.narod.ru/Tasks/GCRota.htm>.
- en_softw_desig D. Rouson, J. Xia, and X. Xu. *Scientific Software Design: The Object-Oriented Way*. Cambridge University Press, 2011.
- r_qualit_probl J. Roux. Mathematical models in air quality problems. In UNESCO-EOLSS, editor, *Encyclopedia of Life Support Systems*, volume II of *Mathematical Models*. 2015. URL <http://www.eolss.net/>.

- mpl_analy:da09 W. Rudin. *Real and Complex Analysis*. McGraw-Hill, New York, 2nd ed. edition, 1987.
- l_proces_wavel O. Ryan. Linear algebra, signal processing, and wavelets. a unified approach. Technical report, Centre of Mathematics for Applications, University of Oslo, 2014. Lecture Notes for MAT-INF2360.
- in_scienc:da112 E. B. Saff and A. D. Snider. *Fundamentals of Complex Analysis with Applications to Engineering and Science*. Pearson/Prentice Hall, Upper Saddle River, N. J., 3rd ed. edition, 2003.
- at_action:da109 S. Salsa. *Partial Differential Equations in Action: From Modelling to Theory*. Springer-Verlag Italia, Milano, 2008.
- ce_method:da114 J. Samuel S. Holland. *Applied Analysis by the Hilbert Space Method: An Introduction with Applications to the Wave, Heat, and Schrödinger Equations*. Marcel Dekker, Inc., 1990.
- erson16:_armad C. Sanderson and R. Curtin. Armadillo: a template-based c++ library for linear algebra. *Journal of Open Source Software*, 1:26–, 2016.
- fer_equat_model W. E. Schiesser and G. W. Griffiths. *A Compendium of Partial Differential Equation Models: Method of Lines Analysis with Matlab*. Cambridge University Press, 2009.
- lac_trans:da47 J. L. Schiff. *Laplace Transformation: Theory and Applications*. Undergraduate Texts in Mathematics. Springer-Verlag New York, Inc., 1999.
- 901zeitschrift O. Schlömilch, B. Witzschel, E. Kahl, R. Mehmke, and C. Runge. *Zeitschrift für Mathematik und Physik*. Number v. 46. B. G. Teubner., 1901. URL <http://books.google.com.my/books?id=QL0KAAAAIAAJ>.
- is_probl:da123 U. Schöning and J. Torán. *The Satisfiability Problem: Algorithms and Analysis*. Lehmanns Media: Berlin, 2013.
- n_differ_equat B. J. Schroers. *Ordinary Differential Equations: A Practical Guide*. Cambridge University Press, 2011.
- _valuat_warran E. S. Schwartz. The valuation of warrants: Implementing a new approach. *Journal of Financial Economics*, 4:77–93, 1977.
- pl_number:da42 H. Schwerdtfeger. *Geometry of Complex Numbers: Circle Geometry, Moebius Transformation, Non-Euclidean Geometry*. Dover Publications, Inc., New York, 1979.
- s_chemis_physic J. H. Seinfeld and S. N. Pandis. *Atmospheric Chemistry and Physics*. John Wiley and Sons, Inc., 1998. Reference book about the chemistry of the troposphere.
- at_finan:da12m2 R. U. Seydel. *Tools for Computational Finance*. Springer-Verlag Berlin Heidelberg, 3rd edition, 2006.
- _solvin_pdes_c Y. Shapira. *Solving PDEs in C++: Numerical Methods in a Unified Object-Oriented Approach*. Society for Industrial and Applied Mathematics, 2006.
- l_differ_equat R. E. Showalter. *Hilbert Space Methods for Partial Differential Equations*, volume 01 of *Electron. J. Diff. Eqns., Monograph*. European Mathematical Society, 1994. (<http://www.emis.de/journals/EJDE/Monographs/Monographs/01/abstr.html>).
- drog_atom:da60 S. F. Singer. *Linearity, Symmetry and Prediction in the Hydrogen Atom*. Springer Science+Business Media, Inc., 2005.
- _volum_ii:da08 V. I. Smirnov. *A Course of Higher Mathematics Volume II*. Pergamon Press, 1964a.
- chem_volum_iii2 V. I. Smirnov. *A Course of Higher Mathematics Volume III Part 2*. Pergamon Press, 1964b.
- l_differ_equat G. D. Smith. *Numerical Solution Of Partial Differential Equations: Finite Difference Methods*. Oxford University Press, 3rd edition, 1985.

- ic_mechan:da47 M. V. Soare, P. P. Teodorescu, and I. Toma. *Ordinary Differential Equations with Applications to Mechanics*. Springer, 2007.
- le_ecolog_model K. Soetaert and P. M. J. Herman. *A Practical Guide to Ecological Modelling: Using R as a Simulation Platform*. Springer Science+Business Media B. V., 2009.
- st_physic:da110 A. Sommerfeld. *Partial Differential Equations in Physics*. Academic Press Inc., 1949.
- er_analy:da113 E. M. Stein and R. Shakarchi. *Fourier Analysis: An Introduction*. Princeton Lectures in Analysis Volume I. Princeton University Press, 2003.
- art_prolog:da08 L. Sterling and E. Shapiro. *The Art of Prolog: Advanced Programming Techniques*. The MIT Press, 2nd edition, 1999.
- _volum_ii:da67 G. W. Stewart. *Matrix Algorithms Volume II: Eigensystems*. Society for Industrial and Applied Mathematics, 2001.
- er_analy:da119 J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*, volume 12 of *Text in Applied Mathematics*. Springer-Verlag New York, Inc., 3rd edition, 2002.
- geb_its_applic G. Strang. *Linear Algebra and Its Applications*. Brooks Cole, 1988. Strang’s book is something of a classic though some may find it to be a little too formal.
- geb_geodes_gps G. Strang and K. Borre. *Linear Algebra, Geodesy, and GPS*. Wellesley-Cambridge Press, 1997.
- ss_mechan:da72 D. Strauch. *Classical Mechanics: An Introduction*. Springer-Verlag Berlin Heidelberg, 2009.
- er_equat:da107 W. A. Strauss. *Partial Differential Equations: An Introduction*. John Wiley & Sons, Inc., 1992.
- l_multip_integ A. H. Stroud. *Approximate Calculation of Multiple Integrals*. Prentice-Hall, 1971.
- ss_quad_r Formul A. H. Stroud and D. Secrest. *Gaussian Quadrature Formulas*. Prentice-Hall, 1966.
- StuaVo08 A. M. Stuart and J. Voss. Matrix analysis and algorithms. lecture notes, 2009.
- es79:_n_dimen_n M. Stynes. An n-dimensional bisection method for solving systems of n equations in n unknowns. *Applicable Analysis*, 9(4):295–296, 1979. 10.1080/00036817908839278. URL <http://www.tandfonline.com/doi/abs/10.1080/00036817908839278>.
- gin_scien:da25 K.-T. Tang. *Mathematical Methods for Engineers and Scientists 3: Fourier Analysis, Partial Differential Equations and Variational Methods*. Springer-Verlag Berlin Heidelberg, 2007.
- cial_funct-da80 N. M. Temme. *Special Functions: An Introduction to Classical Functions of Mathematical Physics*. John Wiley & Sons, Inc., 1996a.
- cial_funct:da80 N. M. Temme. *Special Functions: An Introduction to Classical Functions of Mathematical Physics*. John Wiley & Sons, Inc., 1996b.
- er_equat:da108 M. Tenenbaum and H. Pollard. *Ordinary Differential Equations: An Elementary Textbook for Students of Mathematics, Engineering, and the Sciences*. Dover Publications, Inc. New York, 1985, 1963.
- cgal:eb-17b The CGAL Project. *CGAL User and Reference Manual*. CGAL Editorial Board, 4.11 edition, 2017. URL <http://doc.cgal.org/4.11/Manual/packages.html>.
- 160502688short Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016. URL <http://arxiv.org/abs/1605.02688>. Software: <http://deeplearning.net/software/theano>.
- 00:_inter_revis P. Thévenaz, T. Blu, and M. Unser. Interpolation revisited. *IEEE Transactions on Medical Imaging*, 19(7), July 2000.

- ut_physic:da76 J. M. Thijssen. *Computational Physics*. Cambridge University Press, 1999.
- on1985numerical J. Thompson, Z. Warsi, and C. Mastin. *Numerical Grid Generation: Foundations and Applications*. North-Holland, 1985. ISBN 9780444009852. URL <http://books.google.com.my/books?id=GjjvAAAAMAAJ>.
- matem_ui_fizik A. N. Tihonov and A. A. Samarskii. “*Uravneniya Matematicheskoi Fiziki*”. ???, 1977. There are very nice pictures describing the behaviour of wave solutions. English translation available.
- ic_model_model M. M. Tiller. *Introduction to Physical Modeling with Modelica*. Norwell, Massachusetts: Kluwer Academic Publishers, 2001. ISBN 0-7923-7367-7.
- eor_funct:da11 E. C. Titchmarsh. *The Theory of Functions*. Oxford University Press, 2nd ed. edition, 1939.
- gen_expan:da11 E. C. Titchmarsh. *Eigenfunction Expansions: Associated with Second-order Differential Equations Part I*. Oxford University Press, 1962.
- fourier_series G. P. Tolstov. *Fourier Series*. Dover, 1976. (translated by Richard A. Silverman. It has an amazing chapter on Bessel functions).
- it_elemen:da63 J. Topper. *Financial Engineering with Finite Elements*. John Wiley & Sons Ltd, 2005.
- od_matlab:da24 L. N. Trefethen. *Spectral Methods in MATLAB*. SIAM, Philadelphia, 2000.
- 8:_numer_analy L. N. Trefethen. *Numerical Analysis*. Princeton University Press, 2008.
- ect_pseud:da75 L. N. Trefethen and M. Embree. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, 2005.
- db_system:da119 J. D. Ullman and J. Widom. *A First Course in Database Systems*. Pearson Education, Inc., 3th edition, 2008.
- de_simpl:da112 D. C. Ullrich. *Complex Made Simple*. Number Volume 97 in Graduate Studies in Mathematics. Providence, Rhode Island: American Mathematical Society, 2008.
- log_progr:da87 T. van Le. *Techniques of PROLOG PROGRAMMING: With Implementation of Logical Negation and Quantified Goals*. John Wiley & Sons, Inc., 1993.
- 07introduction H. Versteeg and W. Malalasekera. *An Introduction to Computational Fluid Dynamics: The Finite Volume Method*. Pearson Education Limited, 2007. ISBN 9780131274983. URL <http://books.google.com.my/books?id=RvBZ-UMpGzIC>.
- matem_ui_fizik V. S. Vladimirov. *Obobshchennye Funkcii v Matematicheskoi Fizike*. Nauka, Moskva, 1979.
- hatis:1988:ACM M. N. Vrahatis. Algorithm 666: CHABIS: a mathematical software package for locating and evaluating roots of systems of nonlinear equations. *ACM Transactions on Mathematical Software*, 14(4):330–336, Dec. 1988. ISSN 0098-3500 (print), 1557-7295 (electronic). <http://doi.acm.org/10.1145/50063.51906>. URL <http://www.acm.org/pubs/citations/journals/toms/1988-14-4/p330-vrahatis/>.
- _elemen_method P. Šolín. *Partial Differential Equations and the Finite Element Method*. John-Wiley & Sons, Inc., Hoboken, New Jersey, 2006. A very good introductory text to FEM with the first chapter introducing examples of not well-defined PDEs.
- or_bessel_funct G. N. Watson. *Treatise On The Theory of Bessel Functions*. Cambridge University Press, 1922.
- ssel_funct:da10 G. N. Watson. *Treatise On The Theory of Bessel Functions*. Cambridge University Press, 2nd ed. edition, 1944.
- ffer_equat:da10 H. F. Weinberger. *A First Course in Partial Differential Equations with Complex Variables and Transform Methods*. Dover Publications, Inc., New York, 1995.

- anon_form:da38 S. H. Weintraub. *Jordan Canonical Form: Application to Differential Equations*. Morgan & Claypool, 2008.
- _comput_mathem R. E. White. *Computational Mathematics: Models, Methods, and Analysis with MATLAB and MPI*. Chapman & Hall/CRC, 2004.
- er_analy:da112 E. T. Whittaker and G. N. Watson. *A Course of Modern Analysis: An Introduction to the General Theory of Finite Processes and of Analytic Functions; with an Account of the Principal Transcendental Functions*. Cambridge University Press, 3rd ed. edition, 1920a.
- er_analy:da112 E. T. Whittaker and G. N. Watson. *A Course Of Modern Analysis*. Cambridge University Press, 3rd ed. edition, 1920b.
- 8:_gener_funct Wikipedia. Generalized function, 2008a. http://en.wikipedia.org/wiki/Generalized_function.
- 8:microl_analy Wikipedia. Microlocal analysis, 2008b. http://en.wikipedia.org/wiki/Microlocal_analysis.
- edia13:_butter Wikipedia. Butterfly curve (transcendental), 2013. URL http://en.wikipedia.org/wiki/Butterfly_curve_%28transcendental%29. [Online; accessed 27-Dec-2013].
- ck_nicol_method Wikipedia. Crank-nicolson method, 2015. URL http://en.wikipedia.org/wiki/Crank-Nicolson_method. [Online; accessed 2-Apr-2015].
- ikipedia:_hermit Wikipedia. Hermite polynomials, 2015.02.28. URL http://en.wikipedia.org/wiki/Hermite_polynomials.
- :_spher_harmon Wikipedia. Spherical harmonics, 2016. URL https://en.wikipedia.org/wiki/Spherical_harmonics. [Online; accessed 18-March-2016].
- 1988:AEP:59657 J. H. Wilkinson, editor. *The Algebraic Eigenvalue Problem*. Oxford University Press, Inc., New York, NY, USA, 1988. ISBN 0-198-53418-3.
- in_mathem:da94 P. Wilmott, J. Dewynne, and S. Howison. *Option Pricing: Mathematical Models and Computation*. ?, 1994.
- nan_deriv:da93 P. Wilmott, S. Howison, and J. Dewynne. *The Mathematics of Financial Derivatives: A Student Introduction*. Cambridge University Press, New York, 1995.
- od_numer_analy A. Wood. *Introduction to Numerical Analysis*. Harlow: Addison-Wesley, 2000.
- od_higher_dimen G. Wood. The bisection method in higher dimensions. *Mathematical Programming*, 55(1-3): 319–337, 1992. ISSN 0025-5610. 10.1007/BF01581205. URL <http://dx.doi.org/10.1007/BF01581205>.
- es_system:da47 R. Woods, J. McAllister, G. Lightbody, and Y. Yi. *FPGA-based Implementation of Signal Processing Systems*. John Wiley & Sons, Ltd, 2008.
- _r_rmetr:da113 D. Würtz, Y. Chalabi, W. Chen, and A. Ellis. *Portfolio Optimization with R/Rmetrics*. Rmetrics Association & Finance Online, Zurich, 2009.
- od_physic:da12 H. W. Wyld. *Mathematical Methods for Physics*. Perseus Books Publishing, L.L.C., 1999.
- er_equat_engin W.-C. Xie. *Differential Equations for Engineers*. Cambridge University Press, 2010.
- er_geomet:da42 I. M. Yaglom. *Complex Numbers in Geometry*. Academic Press, New York and London, 1968.
- Yang2014695 X. I. Yang and R. Mittal. Acceleration of the jacobi iterative method by factors exceeding 100 using scheduled relaxation. *Journal of Computational Physics*, 274(0):695 – 708, 2014. ISSN 0021-9991. <http://dx.doi.org/10.1016/j.jcp.2014.06.010>. URL <http://www.sciencedirect.com/science/article/pii/S0021999114004173>.

- punct_analy:da50 E. Zeidler.
- Applied Functional Analysis: Applications to Mathematical Physics*
- . Number 108 in Applied Mathematical Sciences. Springer-Verlag New York, Inc., 1995.

trial_func:da66 S. Zhang and J. Jin.
- Computation of Special Functions*
- . John Wiley & Sons, Inc., 1996.

t_differ_method Y.-L. Zhu, X. Wu, I.-L. Chern, and Z.-Z. Sun.
- Derivative Securities and Difference Methods*
- . Springer Science+Business Media New York, 2nd ed. edition, 2013.

t_elemen_method O. C. Zienkiewicz, R. L. Taylor, and J. Z. Zhu.
- The Finite Element Method: Its Basis and Fundamentals*
- . Elsevier Butterworth-Heinemann, 6th edition, 2005.

t_model_applic D. G. Zill.
- A First Course in Differential Equations with Modeling Applications*
- . Brooks/Cole Cengage Learning, Boston, MA, 10th ed. edition, 2013.
- http://www.cengage.com/math/book_content/0495108243_zill/projects_archive/debv4e/AModelingApplication2.pdf
- .

t_model_applic D. G. Zill.
- A First Course in Differential Equations with Modeling Applications*
- . Brooks/Cole, Cengage Learning, Boston, MA, 11th ed. edition, 2017. URL
- <http://www.amazon.com/Course-Differential-Equations-Modeling-Applications/dp/1305965728>
- .

zlatev81:_y12m Z. Zlatev, J. Wasniewski, and K. Schaumburg.
- Y12M: Solution of Large and Sparse Systems of Linear Algebraic Equations, Documentation of Subroutines*
- . Lecture Notes in Computer Science. Springer-Verlag Berlin Heidelberg, 1981.

tfer_equat:da10 D. Zwillinger.
- Handbook of Differential Equations*
- . Academic Press, 3rd edition, 1997.

er1998handbook D. Zwillinger.
- Handbook of Differential Equations*
- . Academic Press, 1 edition, 1998. ISBN 9780127843964. URL
- <http://books.google.com.my/books?id=YD4PL0GZgOcC>
- .