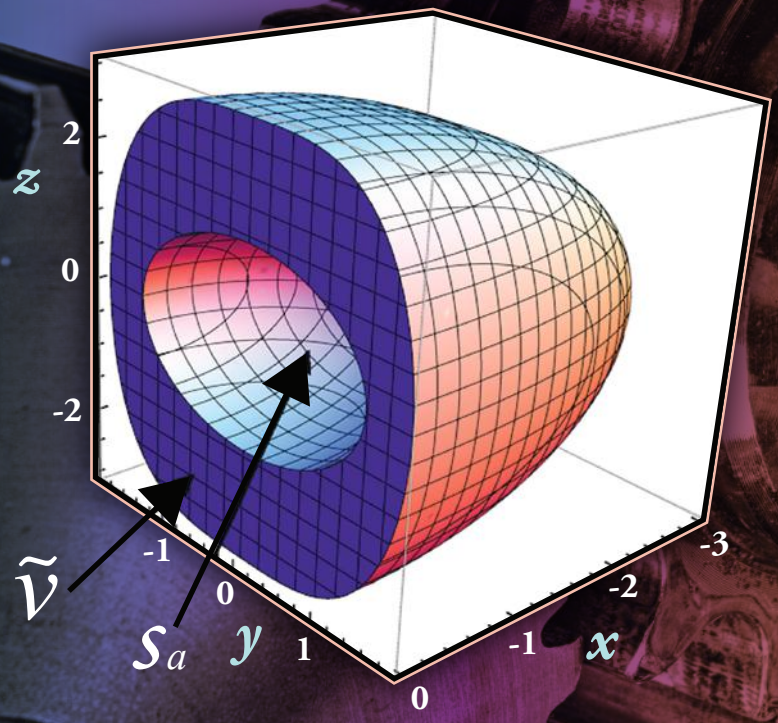
ADVANCED ENGINEERING MATHEMATICS



Larry Turyn

CRC Press
Taylor & Francis Group

www.TechnicalBooksPDF.com

ADVANCED ENGINEERING MATHEMATICS

Larry Turyn

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

www.TechnicalBooksPDF.com

MATLAB® is a trademark of The MathWorks, Inc. and is used with permission. The MathWorks does not warrant the accuracy of the text or exercises in this book. This book's use or discussion of MATLAB® software or related products does not constitute endorsement or sponsorship by The MathWorks of a particular pedagogical approach or particular use of the MATLAB® software.

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2014 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Version Date: 20130801

International Standard Book Number-13: 978-1-4822-1939-5 (eBook - PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

www.TechnicalBooksPDF.com

Contents

Preface xix
Acknowledgments xxv

1. Linear Algebraic Equations, Matrices, and Eigenvalues 1
1.1 Solving Systems and Row Echelon Forms 1
1.1.1 Matrices 2
1.1.2 Augmented Matrices 5
1.1.3 Row Reduced Echelon Form (RREF) 6
1.1.4 Problems 13
1.2 Matrix Addition, Multiplication, and Transpose 16
1.2.1 Special Kinds of Matrices 21
1.2.2 Powers of a Matrix 23
1.2.3 Transpose 24
1.2.4 Elementary Matrices 24
1.2.5 Problems 25
1.3 Homogeneous Systems, Spanning Set, and Basic Solutions 26
1.3.1 Problems 32
1.4 Solutions of Nonhomogeneous Systems 33
1.4.1 Problems 36
1.5 Inverse Matrix 37
1.5.1 Row Reduction Algorithm for Constructing the Inverse 40
1.5.2 Inverse of a Partitioned Matrix 43
1.5.3 Problems 44
1.6 Determinant, Adjugate Matrix, and Cramer's Rule 47
1.6.1 Adjugate Matrix 53
1.6.2 Cramer's Rule 54
1.6.3 Problems 57
1.7 Linear Independence, Basis and Dimension 61
1.7.1 Linear Independence 62
1.7.2 Vector Spaces and Subspaces 65
1.7.3 Problems 69
Key Terms 71
MATLAB® Commands 72
Mathematica Commands 72
References 73

2. Matrix Theory 75
2.1 Eigenvalues and Eigenvectors 75
2.1.1 The Adjugate Matrix Method for Finding an Eigenvector 79
2.1.2 Complex Numbers 82
2.1.3 Complex Eigenvalues and Eigenvectors 83
2.1.4 Eigenvalues and Eigenvectors of Triangular and Diagonal Matrices 85

www.TechnicalBooksPDF.com

2.1.5 MATLAB® and *Mathematica*™ 86
2.1.6 Problems 87
2.2 Basis of Eigenvectors and Diagonalization 91
2.2.1 Diagonalizing a Matrix 95
2.2.2 Deficient Eigenvalues 98
2.2.3 Problems 99
2.3 Inner Product and Orthogonal Sets of Vectors 101
2.3.1 Orthogonal Set of Vectors 105
2.3.2 The Gram–Schmidt Process 106
2.3.3 Orthogonal Projections 109
2.3.4 Problems 112
2.4 Orthonormal Bases and Orthogonal Matrices 114
2.4.1 Orthogonal Sets and Bases 114
2.4.2 Orthogonal Matrices 116
2.4.3 Appendix 118
2.4.4 Problems 119
2.5 Least Squares Solutions 121
2.5.1 The Normal Equations 123
2.5.1.1 Least Squares Solution and Orthogonal Matrices 127
2.5.2 Problems 129
2.6 Symmetric Matrices, Definite Matrices, and Applications 131
2.6.1 A Spectral Theorem 131
2.6.1.1 A Spectral Formula 135
2.6.1.2 Positive Definite and Positive Semi-Definite Matrices 136
2.6.1.3 Application to $A^2, A^{-1}, \sqrt{A}$ 137
2.6.1.4 Application to Least Squares Solutions 138
2.6.2 Further Study of Positive Definite Matrices 139
2.6.2.1 Vibrations and the Generalized Eigenvalue Problem 140
2.6.2.2 Positive Definiteness and Determinants 141
2.6.3 Problems 146
2.7 Factorizations: QR and SVD 148
2.7.1 QR Factorization 148
2.7.2 QR and Solving Systems 150
2.7.3 QR and Least Squares Solutions 151
2.7.4 SVD 151
2.7.5 SVD and L.S.S. 158
2.7.6 Moore–Penrose Generalized Inverse 159
2.7.7 Problems 163
2.8 Factorizations: LU and Cholesky 165
2.8.1 LU Factorizations 165
2.8.2 Cholesky Factorizations 168
2.8.3 Problems 169
2.9 Rayleigh Quotient 170
2.9.1 A Rayleigh Theorem 172
2.9.2 Problems 174
2.10 Short Take: Inner Product and Hilbert Spaces 175
2.10.1 Linear Functionals and Operators 177
2.10.2 Norm and Bounded Linear Operators 179
2.10.3 Convergence, Cauchy Completeness, and Hilbert Spaces 183

www.TechnicalBooksPDF.com

2.10.4 Bounded Linear Functionals and Operator Adjoint 186
2.10.5 Application to Signal Restoration 187
2.10.6 Projection and Minimization 188
2.10.7 Weak Convergence and Compactness 189
2.10.8 Problems 191
Key Terms 192
MATLAB® Commands 194
Mathematica Commands 194
References 194

3. Scalar ODEs I: Homogeneous Problems 195
3.1 Linear First-Order ODEs 195
3.1.1 Scalar ODEs 195
3.1.2 Linear First-Order ODEs 196
3.1.3 Steady-State and Transient Solutions 202
3.1.4 Problems 205
3.2 Separable and Exact ODEs 209
3.2.1 Separable ODEs 209
3.2.2 Exact ODEs 211
3.2.3 Existence of Solution(s) of an IVP 215
3.2.4 Problems 219
3.3 Second-Order Linear Homogeneous ODEs 222
3.3.1 Spring–Mass–Damper Systems 222
3.3.2 Series *RLC* Circuit 225
3.3.3 The Underdamped Case 230
3.3.4 The Amplitude and Phase Form 232
3.3.5 Figures of Merit in Graphs of Underdamped Solutions 235
3.3.6 The Critically Damped Case 237
3.3.7 The Wronskian Determinant 238
3.3.8 Problems 239
3.4 Higher-Order Linear ODEs 244
3.4.1 The Zoo of Solutions of LCCHODEs 250
3.4.2 Differential Operator Notation 251
3.4.3 Shift Theorem 253
3.4.4 Problems 254
3.5 Cauchy–Euler ODEs 255
3.5.1 Problems 260
Key Terms 261
Mathematica Command 262
Reference 262

4. Scalar ODEs II: Nonhomogeneous Problems 263
4.1 Nonhomogeneous ODEs 263
4.1.1 Special Case: RHS from the Zoo and Constant Coefficients on LHS 264
4.1.2 The Method of Coefficients to Be Determined 265
4.1.3 Justification for the Method 269
4.1.4 Using a Shift Theorem 270
4.1.5 Problems 271

www.TechnicalBooksPDF.com

4.2 Forced Oscillations 273
4.2.1 The Resonance Case 274
4.2.2 Steady-State Solution, Frequency Response, and Practical Resonance 276
4.2.3 Maximum Frequency Response 281
4.2.4 Beats Phenomenon, Fast and Slow Frequencies, and Frequency Response 283
4.2.5 Problems 287
4.3 Variation of Parameters 291
4.3.1 Method of Variation of Parameters 294
4.3.2 Problems 298
4.4 Laplace Transforms: Basic Techniques 299
4.4.1 Problems 305
4.5 Laplace Transforms: Unit Step and Other Techniques 307
4.5.1 Writing a Function in Terms of Step Function(s) 308
4.5.2 Graph of a Solution of an ODE Involving a Step Function 310
4.5.3 Convolution 312
4.5.4 Convolution and Particular Solutions 314
4.5.5 Delta "Functions" 318
4.5.6 Laplace Transform of a Periodic Function 321
4.5.7 Remarks 321
4.5.8 Problems 321
4.6 Scalar Difference Equations 323
4.6.1 General Solution and the Casorati Determinant 330
4.6.2 Nonhomogeneous Linear Difference Equation 334
4.6.3 The Method of Undetermined Coefficients 334
4.6.4 Problems 338
4.7 Short Take: z-Transforms 340
4.7.1 Sinusoidal Signals 344
4.7.2 Steady-State Solution 345
4.7.3 Convolution and z-Transforms 348
4.7.4 Transfer Function 348
4.7.5 Problems 349
Key Terms 349
References 351

5. Linear Systems of ODEs 353
5.1 Systems of ODEs 353
5.1.1 Systems of Second-Order Equations 357
5.1.2 Compartment Models 358
5.1.3 Problems 360
5.2 Solving Linear Homogenous Systems of ODEs 362
5.2.1 Fundamental Matrix and e^{tA} 368
5.2.2 Equivalence of Second-Order LCCHODE and LCCHS in $\mathbb{R}^2$ 372
5.2.3 Maclaurin Series for e^{tA} 375
5.2.4 Nonconstant Coefficients 376
5.2.5 Problems 377
5.3 Complex or Deficient Eigenvalues 381

www.TechnicalBooksPDF.com

5.3.1 Complex Eigenvalues 381
5.3.2 Solving Homogeneous Systems of Second-Order Equations 385
5.3.3 Deficient Eigenvalues 387
5.3.4 Laplace Transforms and e^{tA} 390
5.3.5 Stability 391
5.3.6 Problems 392
5.4 Nonhomogeneous Linear Systems 395
5.4.1 Problems 401
5.5 Nonresonant Nonhomogeneous Systems 403
5.5.1 Sinusoidal Forcing 408
5.5.2 Problems 411
5.6 Linear Control Theory: Complete Controllability 412
5.6.1 Some Other Control Problems 419
5.6.2 Problems 421
5.7 Linear Systems of Difference Equations 422
5.7.1 Color Blindness 423
5.7.2 General Solution and the Casorati Determinant 425
5.7.3 Complex Eigenvalues 427
5.7.4 Equivalence of Second-Order Scalar Difference Equation and a System in $\mathbb{R}^2$ 428
5.7.5 Ladder Network Electrical Circuits 429
5.7.6 Stability 434
5.7.7 Problems 436
5.8 Short Take: Periodic Linear Differential Equations 439
5.8.1 The Stroboscopic, or "Return," Map 441
5.8.2 Floquet Representation 442
5.8.3 Stability 445
5.8.4 Hill's Equation 447
5.8.5 Periodic Solution of a Nonhomogeneous ODE System 448
5.8.6 Problems 451
Key Terms 454
MATLAB® Commands 455
References 455

6. Geometry, Calculus, and Other Tools 457
6.1 Dot Product, Cross Product, Lines, and Planes 457
6.1.1 Dot Product and Cross Product 457
6.1.2 Lines 459
6.1.3 Planes 459
6.1.4 Problems 461
6.2 Trigonometry, Polar, Cylindrical, and Spherical Coordinates 463
6.2.1 Cylindrical Coordinates 465
6.2.2 Spherical Coordinates 465
6.2.3 Right-Handed Orthogonal Bases for $\mathbb{R}^3$ 467
6.2.4 Orthonormal Basis in Spherical Coordinates 468
6.2.5 Relationships to the Standard o.n. Basis 470
6.2.6 Problems 471

www.TechnicalBooksPDF.com

6.3 Curves and Surfaces 471
6.3.1 Curves and Calculus 474
6.3.2 Zhukovskii Airfoil 477
6.3.3 Surfaces 478
6.3.4 Problems 482
6.4 Partial Derivatives 485
6.4.1 Linear Approximation 486
6.4.2 Multivariable Chain Rules 489
6.4.3 Gradient Vector in $\mathbb{R}^3$ 492
6.4.4 Scalar Potential Functions 493
6.4.5 Problems 495
6.5 Tangent Plane and Normal Vector 498
6.5.1 Problems 504
6.6 Area, Volume, and Linear Transformations 505
6.6.1 Linear Transformations 511
6.6.2 Linear Transformations, Area, and Volume 514
6.6.3 Change of Variables, Area, and Volume 516
6.6.4 Element of Surface Area 519
6.6.5 Problems 520
6.7 Differential Operators and Curvilinear Coordinates 522
6.7.1 Properties of the Operators *grad*, div, and *curl* 524
6.7.2 Curvilinear Coordinates 525
6.7.3 Differential Operators in Curvilinear Coordinates 529
6.7.4 Summary of Operators in Cylindrical Coordinates 533
6.7.5 Summary of Operators in Spherical Coordinates 533
6.7.6 Problems 534
6.8 Rotating Coordinate Frames 537
6.8.1 ODEs Describing Rotation 537
6.8.2 Velocity and Acceleration 540
6.8.3 Velocity and Acceleration in a Rotating Frame Whose Origin Is Moving 543
6.8.4 Problems 544
Key Terms 546
Mathematica Command 548
Reference 548

7. Integral Theorems, Multiple Integrals, and Applications 549
7.1 Integrals for a Function of a Single Variable 549
7.1.1 Improper Integrals 553
7.1.2 Problems 554
7.2 Line Integrals 555
7.2.1 Line Integrals of Vector-Valued Functions 560
7.2.2 Fundamental Theorem of Line Integrals 563
7.2.3 Path Direction 565
7.2.4 Other Notations 565
7.2.5 Problems 567
7.3 Double Integrals, Green's Theorem, and Applications 570
7.3.1 Double Integral as Volume 574
7.3.2 Polar Coordinates 580

www.TechnicalBooksPDF.com

7.3.3 Green's Theorem 582
7.3.4 Comparison with Single Variable Calculus 584
7.3.5 Green's Theorem for an Annulus 585
7.3.6 Green's Theorem in Polar Coordinates 586
7.3.7 Problems 589
7.4 Triple Integrals and Applications 594
7.4.1 Cylindrical Coordinates 597
7.4.2 Spherical Coordinates 602
7.4.3 Problems 605
7.5 Surface Integrals and Applications 607
7.5.1 Surface Integral of a Scalar-Valued Function 611
7.5.2 Surface Integral of a Vector Field 613
7.5.3 Problems 617
7.6 Integral Theorems: Divergence, Stokes, and Applications 621
7.6.1 The Divergence Theorem in $\mathbb{R}^2$ 623
7.6.2 Euler's Conservation Equation 624
7.6.3 Stokes' Theorem 627
7.6.4 Problems 632
7.7 Probability Distributions 635
7.7.1 Joint Distribution 643
7.7.2 Problems 645
Key Terms 645
Mathematica Commands 647
Reference 647

8. Numerical Methods I 649
8.1 Solving a Scalar Equation 649
8.1.1 Newton–Raphson Method 649
8.1.2 Modified Newton's Method 653
8.1.3 Secant Method 654
8.1.4 Fixed Point Problem Iteration 655
8.1.5 Aitken's δ^2 Method 656
8.1.6 Newton's Method versus Other Methods 657
8.1.7 Trouble 658
8.1.8 Problems 658
8.2 Solving a System of Equations 661
8.2.1 Newton's Method in $\mathbb{R}^n$ 661
8.2.2 Newton–Kantorovich Theorem 664
8.2.3 Fixed Point Problem Iteration 666
8.2.4 Secant Method for Systems of Equations 666
8.2.5 Problems 668
8.3 Approximation of Integrals 669
8.3.1 Approximation Error for Quadrature Rules 672
8.3.2 Iteration of Quadrature 674
8.3.3 Theory versus Practice 675
8.3.4 Problems 676
8.4 Numerical Solution of $A\mathrm{x} = \mathrm{b}$ 677

www.TechnicalBooksPDF.com

8.4.1 Partial Pivoting in the Gauss–Jordan Algorithm 679
8.4.2 Iterative Methods for Solving $Ax = b$ 682
8.4.4 Problems 686
8.5 Linear Algebraic Eigenvalue Problems 687
8.5.1 Elementary Method 687
8.5.2 Power Methods 688
8.5.3 Deflation and Similarity 690
8.5.4 Using Similarity Transformations 690
8.5.5 Background 691
8.5.6 QR Algorithm 693
8.5.7 Problems 694
8.6 Approximations of Derivatives 695
8.6.1 Problems 699
8.7 Approximate Solutions of ODE-IVPs 699
8.7.1 Runge–Kutta Methods 704
8.7.2 Multistep Methods 705
8.7.3 Predictor–Corrector Methods 706
8.7.4 Systems of ODEs 706
8.7.5 Taylor's Formula Method 706
8.7.6 Numerical Instability and Stiffness 707
8.7.7 Problems 709
8.8 Approximate Solutions of Two Point BVPs 712
8.8.1 An ODE-BVP Eigenvalue Problem 715
8.8.2 Using IVP Numerical Methods to Solve BVPs 716
8.8.3 Finding Periodic Solutions of Linear Problems 717
8.8.4 Problems 718
8.9 Splines 719
8.9.1 Cubic B-Splines 722
8.9.2 Nonuniform Splines 727
8.9.3 Spline Approximation of a Curve in $\mathbb{R}^n$ 729
8.9.4 Surface Splines 731
8.9.5 Triangular Surface Patches 731
8.9.6 Problems 733
Key Terms 733
MATLAB® Commands 736
Mathematica Command 736
References 736

9. Fourier Series 737
9.1 Orthogonality and Fourier Coefficients 737
9.1.1 Introduction 737
9.1.2 Convergence of Fourier Series 739
9.1.3 Orthogonality and Calculating Fourier Coefficients 741
9.1.4 Even and Odd Functions and Their Fourier Series Coefficients 744
9.1.5 Finding the Fourier Series in a Special Case 748
9.1.6 Periodic Extension of a Function Given for $0 < x < 2L$ 748
9.1.7 Other Kinds of Fourier Series 751
9.1.8 Problems 752

www.TechnicalBooksPDF.com

9.2 Fourier Cosine and Sine Series 755
9.2.1 Fourier Cosine Series 756
9.2.2 Fourier Sine Series 758
9.2.3 Fourier Analysis and Oscillations 762
9.2.4 Problems 765
9.3 Generalized Fourier Series 767
9.3.1 Other Boundary Conditions 770
9.3.2 Periodic Boundary Conditions and the Full Fourier Series 775
9.3.3 Problems 776
9.4 Complex Fourier Series and Fourier Transform 777
9.4.1 The Fourier Transform 780
9.4.2 Convolution 785
9.4.3 Problems 788
9.5 Discrete Fourier and Fast Fourier Transforms 790
9.5.1 Convolution and Auto-Correlation 798
9.5.2 Fast Fourier Transform 799
9.5.3 Problems 803
9.6 Sturm–Liouville Problems 805
9.6.1 Other Sturm–Liouville Problems 808
9.6.2 A Composite Media Problem 810
9.6.3 Fourth-Order ODE-BVP 813
9.6.4 Problems 815
9.7 Rayleigh Quotient 818
9.7.1 Problems 822
9.8 Parseval's Theorems and Applications 824
9.8.1 Best Approximation by a Partial Sum of a Fourier Series 827
9.8.2 Complex Fourier Series 829
9.8.3 Fourier Transforms 830
9.8.4 Problems 831
Key Terms 831
Mathematica Command 833
References 833

10. Partial Differential Equations Models 835
10.1 Integral and Partial Differential Equations 835
10.1.1 Maxwell's Equations of Electromagnetism 838
10.1.2 Continuum Mechanics 840
10.1.3 Problems 845
10.2 Heat Equations 846
10.2.1 Steady-State Temperature 849
10.2.2 Lower Dimensional Problems 849
10.2.3 Composite Rod 854
10.2.4 Problems 855
10.3 Potential Equations 860
10.3.1 Magnetostatics 860
10.3.2 Boundary Conditions 861
10.3.3 Properties of Solutions 861
10.3.4 Problems 865

www.TechnicalBooksPDF.com

10.4 Wave Equations 868
10.4.1 Guitar String 868
10.4.2 Vibrating String 869
10.4.3 Speed of Sound 872
10.4.4 Linear Elasticity 876
10.4.5 Linear Elastostatics 879
10.4.6 Problems 880
10.5 D'Alembert Wave Solutions 882
10.5.1 Zero Initial Velocity 883
10.5.2 Writing the Solution Using Step Functions 888
10.5.3 Support 888
10.5.4 Zero Initial Displacement 889
10.5.5 Problems 895
10.6 Short Take: Conservation of Energy in a Finite String 897
10.6.1 Problems 900
Key Terms 900
Reference 902

11. Separation of Variables for PDEs 903
11.1 Heat Equation in One Space Dimension 903
11.1.1 Easy Initial Conditions 912
11.1.2 Composite Rod 913
11.1.3 Time-Dependent Boundary Conditions 914
11.1.4 Problems 918
11.2 Wave Equation in One Space Dimension 924
11.2.1 Problems 928
11.3 Laplace Equation in a Rectangle 932
11.3.1 Using Clairvoyance to Choose Alternatives to $\cosh\left(\frac{n\pi x}{L}\right)$ and $\sinh\left(\frac{n\pi x}{L}\right)$ 937
11.3.2 Contour Plot and 3D Plot Using *Mathematica* 940
11.3.3 Problems 946
11.4 Eigenvalues of the Laplacian and Applications 948
11.4.1 Application to Time-Dependent Heat Flow in a Slab 951
11.4.2 Special Case 952
11.4.3 Application to Transverse Vibrations of a Rectangular Membrane 952
11.4.4 Application to Steady-State Temperature in a Slab with a Steady Source or Sink 953
11.4.5 Application to Surface Waves 954
11.4.6 Problems 956
11.5 PDEs in Polar Coordinates 958
11.5.1 Laplace Equation in Polar Coordinates 958
11.5.2 Heat Equation in Polar Coordinates 967
11.5.3 Problems 969
11.6 PDEs in Cylindrical and Spherical Coordinates 973
11.6.1 Spherical Coordinates 981
11.6.2 Polar Coordinates Again 985
11.6.3 Problems 986

www.TechnicalBooksPDF.com

Key Terms 989
Mathematica Commands 989
References 989

12. Numerical Methods II 991
12.1 Finite Difference Methods for Heat Equations 991
12.1.1 Incompatibility of Initial Condition with Boundary Conditions 996
12.1.2 Time-Dependent Boundary Conditions 997
12.1.3 Other Boundary Conditions 997
12.1.4 Nonlinearity 998
12.1.5 Problems 1000
12.2 Numerical Stability 1001
12.2.1 Crank–Nicholson Method 1005
12.2.2 Problems 1007
12.3 Finite Difference Methods for Potential Equations 1007
12.3.1 Other Boundary Conditions 1010
12.3.2 Problems 1011
12.4 Finite Difference Methods for the Wave Equation 1014
12.4.1 Scalar Hyperbolic Problem 1014
12.4.2 Lax Scheme 1016
12.4.3 Problems 1017
12.5 Short Take: Galerkin Method 1018
12.5.1 A Generalization of the Galerkin Method 1021
12.5.2 The Galerkin Method for PDEs 1022
12.5.3 Nonlinear Problems 1023
12.5.4 Problems 1024
Key Terms 1025
Reference 1025

13. Optimization 1027
13.1 Functions of a Single Variable 1027
13.1.1 Global Optimization Result 1029
13.1.2 Convex Functions and Optimization 1029
13.1.3 Problems 1030
13.2 Functions of Several Variables 1032
13.2.1 Global Optimization and Lagrange Multipliers 1038
13.2.2 Numerical Minimization and Steepest Descent Methods 1045
13.2.3 Problems 1046
13.3 Linear Programming Problems 1048
13.3.1 Slack Variables and Standard Form 1050
13.3.2 Application: Structural Optimization 1054
13.3.3 Problems 1054
13.4 Simplex Procedure 1055
13.4.1 Unit Cost Reduction 1056
13.4.2 Problems 1060

www.TechnicalBooksPDF.com

13.5 Nonlinear Programming 1062
13.5.1 Dual LP Problem 1068
13.5.2 Application: Geometric Tolerancing 1071
13.5.3 Problems 1072
13.6 Rayleigh–Ritz Method 1073
13.6.1 Other Eigenvalues 1075
13.6.2 Hilbert Space Eigenvalue Problems 1077
13.6.3 Problems 1079
Key Terms 1079
Mathematica Commands 1080
References 1080

14. Calculus of Variations 1081
14.1 Minimization Problems 1081
14.1.1 The Rayleigh–Ritz Method 1085
14.1.2 Problems 1087
14.2 Necessary Conditions 1088
14.2.1 Euler–Lagrange Equations 1089
14.2.2 Natural Boundary Condition 1094
14.2.3 Hamilton's Principle 1095
14.2.4 Hamilton's Principle for Continuous Media 1097
14.2.5 Problems 1098
14.3 Problems with Constraints 1100
14.3.1 Differential Equation Constraints 1103
14.3.2 Problems 1105
14.4 Eigenvalue Problems 1106
14.4.1 An ODE-BVP 1106
14.4.2 An Eigenvalue Problem for the Laplacian 1108
14.4.3 Sturm–Liouville Problem 1109
14.4.4 Problems 1110
14.5 Short Take: Finite Element Methods 1111
14.5.1 *Mathematica* Commands 1113
14.5.2 Rayleigh–Ritz, Galerkin, and Least Squares 1116
14.5.3 Finite Elements for PDEs 1118
14.5.4 Problems 1118
Key Terms 1119
Mathematica Commands 1120
References 1121

15. Functions of a Complex Variable 1123
15.1 Complex Numbers, Roots, and Functions 1123
15.1.1 Polar Forms 1123
15.1.2 Roots 1127
15.1.3 Functions 1130
15.1.4 Problems 1133
15.2 Derivative and the Cauchy–Riemann Equations 1134
15.2.1 Derivatives 1138
15.2.2 Cauchy–Riemann Equations 1140

www.TechnicalBooksPDF.com

15.2.3 Orthogonal Families of Curves and an Application to Fluid Flow 1143
15.2.4 Appendix 1145
15.2.5 Problems 1145
15.3 Analyticity, Harmonic Function, and Harmonic Conjugate 1147
15.3.1 Harmonic Functions 1150
15.3.2 Harmonic Conjugate 1152
15.3.3 Problems 1155
15.4 Elementary Functions 1157
15.4.1 Arg(z) 1157
15.4.2 Exp(z) 1159
15.4.3 Log(z) 1160
15.4.4 Branches of Logarithms 1164
15.4.5 Power Functions 1164
15.4.6 Problems 1169
15.5 Trigonometric Functions 1171
15.5.1 Problems 1174
15.6 Taylor and Laurent Series 1175
15.6.1 Taylor Series 1178
15.6.2 Laurent Series 1179
15.6.3 Product of Taylor Series 1185
15.6.4 Problems 1185
15.7 Zeros and Poles 1186
15.7.1 Singularities 1188
15.7.2 Problems 1193
15.8 Complex Integration and Cauchy's Integral Theorem 1194
15.8.1 Integration on a Closed Contour 1198
15.8.2 Cauchy–Goursat Integral Theorem 1200
15.8.3 Problems 1204
15.9 Cauchy's Integral Formulas and Residues 1205
15.9.1 Use of a Dumb-Bell Contour 1207
15.9.2 Integration of a Laurent Series 1210
15.9.3 Cauchy's Residue Theorem 1212
15.9.4 Problems 1213
15.10 Real Integrals by Complex Integration Methods 1214
15.10.1 Integration of Periodic Functions 1214
15.10.2 Improper Integrals over $(-\infty, \infty)$, $[0, \infty)$, or $(-\infty, 0]$ 1215
15.10.3 Cauchy Principal Value 1218
15.10.4 Hilbert Transform 1219
15.10.5 Problems 1220
Key Terms 1221

16. Conformal Mapping 1223
16.1 Conformal Mappings and the Laplace Equation 1223
16.1.1 Linear Mappings 1223
16.1.2 Harmonic Functions 1226
16.1.3 Elementary Functions 1227
16.1.4 Möbius Transformations 1228
16.1.5 Problems 1229

www.TechnicalBooksPDF.com

16.2 Möbius Transformations 1230
16.2.1 Circles, Lines, and Möbius Transformations 1230
16.2.2 Mapping Two Given Circles to Two Concentric Circles 1233
16.2.3 Some Useful Facts about Möbius Transformations 1240
16.2.4 Möbius Transformation to or from a Line 1241
16.2.5 Problems 1243
16.3 Solving Laplace's Equation Using Conformal Maps 1245
16.3.1 Boundary Values on a Circle 1248
16.3.2 The Joukowsky Map 1251
16.3.3 Zhukovskii Airfoils 1253
16.3.4 Lift on Zhukovskii Airfoils 1255
16.3.5 Problems 1259
Key Terms 1260
References 1261

17. Integral Transform Methods 1263
17.1 Fourier Transform 1263
17.1.1 Convolution 1268
17.1.2 Problems 1268
17.2 Applications to Partial Differential Equations 1270
17.2.1 Fourier Cosine and Sine Transforms 1272
17.2.2 Problems 1279
17.3 Inverse Laplace Transform 1283
17.3.1 Solving a Wave Equation 1288
17.3.2 Problems 1292
17.4 Hankel Transforms 1294
17.4.1 Problems 1298
Key Terms 1300
References 1301

18. Nonlinear Ordinary Differential Equations 1303
18.1 Phase Line and Phase Plane 1303
18.1.1 Equilibria 1306
18.1.2 Qualitative Study: The Phase Line 1307
18.1.3 Qualitative Study: The Phase Plane for LCCHS 1309
18.1.4 Saddle Point Case 1309
18.1.5 Problems 1311
18.2 Stability of an Equilibrium Point 1312
18.2.1 Stability from Linearization 1315
18.2.2 Using $r(t)$ 1318
18.2.3 Problems 1319
18.3 Variation of Parameters Using Linearization 1321
18.3.1 Saddle Point Theorem 1321
18.3.2 Periodic Solutions 1326
18.3.3 Problems 1328
18.4 Liapunov Functions 1328
18.4.1 Definite Functions 1331
18.4.2 Liapunov Functions and Quadratic Forms 1334
18.4.3 Instability 1336

www.TechnicalBooksPDF.com

18.4.4 Stability of Another Equilibrium Point 1338
18.4.5 Problems 1339
18.5 Short Take: LaSalle Invariance Principle 1341
18.5.1 Stability of a Set 1343
18.5.2 Problems 1343
18.6 Limit Cycles 1344
18.6.1 Periodic Linearization 1346
18.6.2 Linearization about a Periodic Solution 1346
18.6.3 Levinson–Smith Theorem 1347
18.6.4 Hopf Bifurcation 1349
18.6.5 Problems 1353
18.7 Existence, Uniqueness, and Continuous Dependence 1354
18.7.1 Continuous Dependence 1359
18.7.2 Problems 1362
18.8 Short Take: Horseshoe Map and Chaos 1364
18.9 Short Take: Delay Equations 1367
18.9.1 Characteristic Equation 1369
18.9.2 Euler's Method 1370
18.9.3 Problems 1370
Key Terms 1371
Mathematica Commands 1372
Reference 1372

Appendix A: Partial Fractions 1373

Appendix B: Laplace Transforms Definitions and Derivations 1379

Appendix C: Series Solutions of ODEs 1387

Index 1407

www.TechnicalBooksPDF.com

www.TechnicalBooksPDF.com

Preface

Purpose

Advanced Engineering Mathematics (AEM) can be used in a course for engineering students who are at the beginning graduate or advanced undergraduate level. It could also be used in a course for junior undergraduate engineering students who have undergone an elementary course on ordinary differential equations (ODEs) and matrices. In addition, this book could be used by undergraduate engineering students in a variety of their post-calculus mathematics courses and by undergraduate students pursuing applied mathematics courses.

This book aims to (1) be comprehensive and self-contained, (2) be relatively "lean and lively," (3) have an appropriately varied pace, (4) be accessible and well written, and (5) have a large choice of appropriate and varied homework problems. It is designed for a heterogeneous group of engineering students in order to extend and enrich their knowledge and to introduce them to new topics. Students who use this book will become well prepared for their engineering courses.

Learning Features

This book successfully blends intuition and logical reasoning. It deals with more advanced material than most textbooks of its kind but does so in a way that is accessible to advanced undergraduate audiences. It helps students understand the basic "what" and "why" questions and learn material at several levels, thus extending their capabilities. Software packages evolve and are even replaced, but the "what" and "why" questions they address are more constant. The habits needed to discern these questions will serve engineers well as they progress through their careers.

For most engineering students a deductive, "theorem/proof/special case" style of exposition is alien to their ways of learning things. But most people appreciate the need to explain things that they are interested in. Sometimes, a plausibility argument is the most accessible explanation. Most engineering graduate students need more practice with logical arguments that explain why techniques are correct or at least plausible, and this will enhance their problem-solving and communication skills.

Often an example leads to its standardization in a definition or a theorem with general applicability. The style of exposition is usually inductive rather than deductive. Also, in order to not overwhelm them, I show students the difficulties gradually and often begin with analogies to familiar topics. It is precisely the students who are less well prepared who need this book the most.

As Epsteen (1913) wrote, "The professor of engineering is certainly on firm ground when he takes the stand that the mathematics taught to his students should not be too abstract on the one hand nor too concrete on the other. If the subject matter is too abstract it is unintelligible or uninteresting to the beginner; if it is too concrete the science degenerates to the mere performing of certain mechanical operations to a common tool instead of a valuable instrument." This is still a very good guide to follow.

www.TechnicalBooksPDF.com

As Henderson (1997) wrote when discussing a survey of what businesses looked for in hiring bachelor's degree holders: "Although engineers may not often need to develop novel mathematical techniques, an ability to read, interpret and implement such techniques is still a vital part of engineering research and development."

In this book, the role of theory is to organize results, provide solution techniques, illuminate what to calculate, and to assure when it is best to calculate. Theorems allow us to avoid reinventing the wheel" and thus are part of a style of establishing formulas and other results that is analogous to the engineering style of standardization. In general, theory in this book directly relates to methods. Usually, "theory" consists of derivations of useful identities.

My choice of what to explain is driven by what I can reasonably expect the readers to explain when they do the problems. What most readers learn from the text connects to what they really learn from working on the problems.

A lot of what engineers do these days is to use software packages; some engineers in research and development also help create software. As much as possible, software should not be a "black box" for the user. While this book is not a book about software packages, I want to give students the mathematical tools they need to understand what their software hopes to do and sometimes even how the software does it. In context, I show students useful *Mathematica*™ and MATLAB® commands in strategic places.

Examples

There are three kinds of examples in the book. Woven into the narrative are examples that give background knowledge or develop the fundamentals of a subject or method. Other examples serve as models for students' work on homework problems. Occasionally, examples point out the limitations of a method or indicate further directions in a topic beyond the scope of this book.

Often an example, or sequence of examples of increasing depth, leads to its standardization in definitions and theorems with general applicability.

Problems

One of the most important things about a textbook is the problems. Usually, students will have reason to feel that they truly understand the material while or after they do homework problems based on, or related to, the narrative of the book.

Most of the problems are based on what is discussed in the narrative; these may be called "exercises." A few problems provide a structured way of filling in some of the details in the narrative. Other problems explore topics related to, but not directly based on, the material in the narrative; these problems are referred to as "complements."

I believe students will find that the problems vary in difficulty, point of view, and style. This will help them learn thoroughly and assess their learning, and this will make the book good preparation for their use of mathematics in engineering courses.

Some of the problems are derivations, usually requiring manipulations of formulas as students do in science and engineering courses to get new, useful formulas. While this may strike some as being "theory," asking students to derive things is essential to measuring their understanding of what they have learned and determining if they are likely to be able to use that knowledge in future courses.

www.TechnicalBooksPDF.com

Appendices

There are three appendices. Appendix A develops the technique of partial fractions that are useful (a) in solving problems using Laplace transforms in Sections 4.4 and 4.5 and (b) in calculating contour integrals in the complex plane in Sections 15.8, 15.9, and 17.3. Students will also be familiar with using partial fractions to evaluate integrals in calculus courses.

Appendix B provides the definition of the Laplace transform and the derivations of its properties that are used in Sections 4.4 and 4.5.

Appendix C discusses series solutions of ordinary differential equations. Bessel functions and Legendre polynomials discussed in this appendix are used in solving partial differential equations in Sections 11.5 and 11.6.

Ancillaries and Supplements

Supplemental resources for the book, including an answer key for odd-numbered problems, as well as complete solutions of selected homework problems, can be found at URL http://www.crcpress.com/product/isbn/9781439834473.

Developmental Plan

Over the past 14 years, I have been either developing or teaching new courses of mathematics for graduate students of engineering, in consultation with professors of mechanical engineering. Also, I have taught much of that material in other courses, at both the undergraduate and graduate levels, for decades.

Many colleagues and independent reviewers have helped me in developing this book. I also appreciate help from my MTH 399/599, MTH 699, and MTH 304/504 and 305/605 students. Professors Antonio Mastroberardino and Peter Olszewski, both from Penn State Erie, The Behrend College, assisted with accuracy checking.

Most of the book's chapters underwent classroom testing by one or more professors, including Professors Lop-Fat Ho, David Miller, and me at Wright State University, as well as by Professor Paul Eloe at the University of Dayton.

Professors Vasilios Alexiades, University of Tennessee–Knoxville; Markus Bussmann, University of Toronto; Paul Eloe, University of Dayton; Harry Hardee, New Mexico State University; Allen Hunt, Wright State University; Thomas Pence, Michigan State University; Allen Plotkin, San Diego State University; Carl Prather, Virginia Polytechnic Institute; Scott Strong, Colorado School of Mines; Hooman Tafreshi, Virginia Commonwealth University; Thad Tarpey, Wright State University; James T. Vance, Jr., Wright State University; Aleksandra Vinogradov, Montana State University; and Dr. Glenn Stoops reviewed and commented on early drafts of the chapters of the book.

Professors Yuqing Chen, Weifu Fang, Ann Farrell, Qingbo Huang, Terry McKee, Munsup Seoh, and James T. Vance, Jr., all from Wright State University, helped me check the first page proofs.

Scott Isenberg served as developmental editor for the project. The feedback he and the reviewers gave me were essential in improving the book. It was helpful for me to get constructive criticism, and it was heartening for me to read some good reviews. I appreciate early encouragement and help from Bill Stenquist.

www.TechnicalBooksPDF.com

Guided Tour

In order to be comprehensive, this book provides a wide spectrum of the mathematical tools beginning graduate engineering students need to use. The core chapters cover ordinary differential equations, matrix/linear algebra, Fourier series and transforms, numerical methods, and partial differential equations.

This book is similar to many AEM books in that it covers many more topics than individual professors are likely to use. Professors will typically pick and choose from among noncore chapters according to the needs of their students. Also, students can use the brief, "Short takes" sections for projects or as introductions to more advanced topics.

In order to cover a comprehensive range of topics, these courses must have a rapid pace. While a lot of the material covered is a review of mathematics coursework taken by typical undergraduate engineering students, I have added many topics that are new to most students. I have found that engineering students progress rapidly through many of the basic topics but need more time and effort for the more highly enriched or new topics.

In general, I have tried to keep material in different chapters as independent as possible. Occasionally, it made sense to refer the reader to an example or problem in a previous chapter, for example, Example 15.67 in Section 15.10 refers back to Example 7.3 in Section 7.1. Similarly, the use of "clairvoyance" in Example 17.9 in Section 17.3 refers to the discussion in Section 11.3.

Better prepared students can leave out almost all of Chapters 1, 3, 4, 6, and 7, which review material most engineers have learned in the first two or three years of their undergraduate education. Those chapters do contain some sections you may consider adding to your syllabi because they are less likely to have been in students' background or because they may be particularly useful to review:

- Section 1.7: Linear Independence, Basis, and Dimension
- Section 3.2: Separable and Exact ODEs
- Section 3.5: Cauchy–Euler ODEs
- Section 4.2: Forced Oscillations
- Section 4.3: Variation of Parameters
- Section 4.6: Scalar Difference Equations
- Section 4.7: Short Take: z-Transforms
- Section 6.6: Area, Volume, and Linear Transformations
- Section 6.7: Differential Operators and Curvilinear Coordinates
- Section 6.8: Rotating Coordinate Frames
- Section 7.6: Integral Theorems: Divergence, Stokes, and Applications
- Section 7.7: Continuous Probability Distributions

In almost all cases, you may omit a "Short take" section and not worry that the omission will prevent you from covering subsequent material. The only exception is that Section 12.5, "Short Take: Galerkin Method," leads directly to Section 14.5, "Short Take: Finite Element Methods."

Notations

The symbol ◯ is written at the end of the work on an example problem, and the symbol □ is written at the end of the derivation or other explanation of a theorem, lemma, or

www.TechnicalBooksPDF.com

corollary. Some of the example problems are derivations or explanations, but I still use ○ to show their end.

The symbol $\triangleq$ means "is defined by." For example, Euler's formula is stated as

$$e^{i\theta} \triangleq \cos\theta + i\sin\theta,$$

that is, $e^{i\theta}$ is defined to be equal to $\cos\theta + i\sin\theta$. Usually, in a statement with $\triangleq$, the left-hand side is defined by the right-hand side.

When reference is made to an equation it is inside parentheses. For example, in Section 1.1 reference is made to (1.5).

Further, if reference is made to an equation, example, definition, theorem, lemma, or corollary in another section, then that section number is mentioned. For example, in Section 1.2.1 reference is made to "Example 1.5 in Section 1.1."

Some Suggested Courses

A one-semester course for junior or senior undergraduate students whose background knowledge is good could include all or parts of Chapters 2, 5, 9, 10, 11, and 15.

A one-semester course for undergraduates whose background knowledge is not as good could include all of Chapters 1, 3, 4, 6, and 7, and parts of Chapters 2 and 5, particularly as preparation for a second semester course.

A two-semester course sequence for beginning graduate students who need to fill a lot of gaps in their background could be as follows:

1. First semester: All of Chapters 3 and 7, and parts of Chapters 1, 2, 4, 5, and 6
2. Second semester: Chapters 9, 10, and 11, and other sections and chapters

A one-semester course for advanced undergraduates or beginning graduate students could consist of parts of Chapters 2 and 5, and all or parts of Chapters 9, 10, 11, and other sections and chapters.

The "other sections and chapters" could include numerical methods in Chapters 8 and/or 12; optimization and calculus of variations in Chapters 13 and/or 14; complex variables and applications in Chapters 15, 16, and/or 17; and nonlinear ordinary differential equations in Chapter 18.

In addition, you can use selective chapters of the book to cover the material in courses such as "differential equations and matrix algebra," "linear algebra and applications," "ordinary differential equations," "partial differential equations and Fourier series," "numerical methods," "applied mathematics," and "complex variables and applications."

MATLAB® is a registered trademark of The MathWorks, Inc. For product information, please contact:

The MathWorks, Inc.
3 Apple Hill Drive
Natick, MA, 01760-2098 USA
Tel: 508-647-7000
Fax: 508-647-7001
E-mail: info@mathworks.com
Web: www.mathworks.com

www.TechnicalBooksPDF.com

References

Epsteen, S. Minimum courses in engineering mathematics. *Am. Math. Monthly* 20 (1913) 47–52.
Henderson, K. Educating electrical and electronic engineers. *Eng. Sci. Educ. J.* 6 (1997) 95–98.

www.TechnicalBooksPDF.com

Acknowledgments

I appreciate the help I have had from all of the people mentioned earlier in the "Developmental Plan" section. They have played a crucial role in improving the book as well as influencing its conceptual basis. I also appreciate the help and support that Buzz Reed gave me as well as the hard work put in by my editor, Jonathan Plant, along with his excellent staff at CRC Press/Taylor & Francis Group.

I am profoundly grateful for the support and help of my family and friends, many of whom were directly involved in the development of this book. I am also very grateful to all my teachers and mentors for sharing their knowledge with me.

I have saved decades of files of problems I have used in courses. I have tried to give credit to the distinctive work of others in creating good problems. I apologize in advance if I have neglected to do so for some of the problems.

The publisher's website will have what I hope will be only a short compendium of corrections and other changes. Experience has taught me that it is very difficult to cleanse all errors from a book. Please let me know if you find any error!

You may also have specific suggestions about what material should be added to the book, or left out, or reorganized. You may find a particular explanation strange or poorly presented, or you may think of a better explanation. Please give me those specific suggestions for improvement!

Of course, I take full responsibility for the book as it is, and I look forward to your comments.

Larry Turyn
Dayton, Ohio

www.TechnicalBooksPDF.com

www.TechnicalBooksPDF.com

1

Linear Algebraic Equations, Matrices, and Eigenvalues

1.1 Solving Systems and Row Echelon Forms

We begin with solving systems of linear algebraic equations. You will use matrix methods often to solve problems in your engineering courses and in the real world, although a software package may hide this fact from you. Numerical approximations of differential equations, optimization, data analysis, and applications to vibrations and circuits have methods and algorithms that have at their core systems of linear algebraic equations and matrix methods.

For the system of linear algebraic equations

$$\left\{\begin{array}{rcrcrl} x_1 & - & x_2 & + & x_3 & = 0 \\ -2x_1 & + & 2x_2 & - & x_3 & = 0 \\ & & 3x_2 & + & 2x_3 & = 4 \end{array}\right\} \tag{1.1}$$

in unknowns x_1, x_2, x_3, adding two times the first equation to the second equation gives an "equivalent system,"

$$\left\{\begin{array}{rcrcrl} x_1 & - & x_2 & + & x_3 & = 0 \\ & & & & x_3 & = 0 \\ & & 3x_2 & + & 2x_3 & = 4 \end{array}\right\}.$$

The operation of adding two times the first equation to the second equation is an analogue of what we will call an "elementary row operation" later in this section. By a **solution** of (1.1), we mean a point (x_1, x_2, x_3) whose values for x_1, x_2, and x_3 make *all* of the equations in system (1.1) true simultaneously. By **equivalent system**, we mean a system of linear algebraic equations that has exactly the same solutions as the original system (1.1). From now on, to save writing, we may write "system" when we mean "system of linear algebraic equations."

In general, a system in unknowns $x_1, x_2, \ldots, x_n$ has the form

$$\left\{\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \qquad\qquad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{array}\right\}. \tag{1.2}$$

www.TechnicalBooksPDF.com

1.1.1 Matrices

A **matrix** is a rectangular array, such as

$$\begin{bmatrix} 1 & 2 & 0 \\ 6 & -5 & 7 \end{bmatrix}.$$

In this example, the **first row** has entries $1,\ 2,\ 0$, and the second row has entries $6, -5,\ 7$. We say that the first row is **higher up** than the second row. Also, in this example, the **first column** has entries $1, 6$; the second column has entries $2, -5$; and the third column has entries $0,\ 7$. We say the first column is to the **left** of the second column, etc. The (i, j) **entry** of a matrix is the entry in the ith row and the jth column. In the example, the $(2, 3)$ entry is 7.

A matrix having exactly m rows and exactly n columns is said to be of **size** $m \times n$. A matrix is **square** if $m = n$.

A **zero row** has all entries being 0, and a **nonzero row** has at least one nonzero entry. The leftmost nonzero entry of a nonzero row is called a **leading entry**. If a leading entry is the number 1, we call it a **leading 1**.

Definition 1.1

A matrix is in a **row echelon form** if it satisfies *all three* of the following properties:

The higher up a leading entry is, the further to the left it is. (1.3)

Any column with a leading entry has zeros below the leading entry. (1.4)

Any zero row(s) is at the bottom of the matrix. (1.5)

The **pivot positions** of a matrix are the locations of the leading entries in a row echelon form of that matrix.

Here are examples of matrices in row echelon form:

$$\begin{bmatrix} 1 & 5 & 3 \\ 0 & -2 & -6 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} \sqrt{2} & 3 & 0 \\ 0 & -1 & b \\ 0 & 0 & 0 \end{bmatrix},$$

where b is an unspecified constant. Here are examples of matrices that are not in row echelon form:

$$\begin{bmatrix} 0 & 1 & 1 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ \vdots \\ n \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} 1 & 0 & 2 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}.$$

Each of the latter three examples violates at least one of properties (1.3)–(1.5).

We listed the three properties (1.3)–(1.5) in that order because we will see that property (1.3) influences our work the most.

www.TechnicalBooksPDF.com

If a matrix is in a row echelon form, every column having a leading entry is called a **pivot column**. Often, we will circle the leading entries in, and thus identify, the pivot columns, for example,

$$\begin{bmatrix} \textcircled{1} & 5 & -3 \\ 0 & \textcircled{2} & -6 \end{bmatrix}.$$

Definition 1.2

If a matrix C is in a row echelon form, then the **rank** of C is the number of its pivot columns.

The definition of rank will be extended to all matrices later, in Definition 1.6.
Given an $m \times (n+1)$ matrix

$$C = \begin{bmatrix} c_{11} & . & . & . & c_{1,n+1} \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ c_{m1} & . & . & . & c_{m.n+1} \end{bmatrix},$$

we can write $C = \begin{bmatrix} C_{1*} \\ \vdots \\ C_{m*} \end{bmatrix}$, where the rows of C are denoted by

$$C_{1*} = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1,n+1} \end{bmatrix}, \; C_{2*} = \begin{bmatrix} c_{21} & c_{22} & \dots & c_{2,n+1} \end{bmatrix}, \dots .$$
$$C_{m*} = \begin{bmatrix} c_{m1} & c_{m2} & \dots & c_{m,n+1} \end{bmatrix}.$$

Definition 1.3

The **elementary row operations** are to

$$\text{Exchange the } i\text{th and } j\text{th rows, where } i \neq j, \text{notated } R_i \leftrightarrow R_j. \tag{1.6}$$
$$\text{Add a multiple of the } i\text{th row } \textit{into} \text{ the } j\text{th row, notated } kR_i + R_j \to R_j. \tag{1.7}$$
$$\text{Multiply the } i\text{th row by a } \textit{nonzero} \text{ constant, notated } kR_i \to R_i. \tag{1.8}$$

If any elementary row operation(s) is performed on matrix C to obtain matrix $\widetilde{C}$, we say C and $\widetilde{C}$ are row equivalent, and we write $C \sim \widetilde{C}$. If matrix C is not already in a row echelon form, we can perform a succession of elementary row operations on C to find matrix $\widetilde{C}$ that is a **row echelon form of** C, that is, row equivalent to C and in a row echelon form.

The following is a method for choosing what sequence of row operations to do on C to find a row echelon form of C. Essentially, the method discovers which of the columns in a row echelon form should contain leading entries, that is, finds the pivot columns.

www.TechnicalBooksPDF.com

Here is a method: If the leftmost *nonzero* column of C is the *jth* column, then it will contain a leading entry, *and* the first $j-1$ columns will *not* be pivot columns. If the $(1,j)$th entry is zero, perform an interchange of rows to produce a row equivalent matrix, $\widetilde{C}_1$, whose $(1,j)$th entry is not zero and is thus a leading entry.* After that, add multiples of the first row to the other rows to produce zeros below the leading entry.

Second, find the next pivot column, if possible: It should be the leftmost column that is both to the right of the jth column and has a nonzero entry in the second or lower row. This second pivot column will be, say, the kth column. If the $(2,k)$th entry is zero, perform an interchange of rows to produce a row equivalent matrix, $\widetilde{C}_2$, whose $(2,k)$th entry is not zero and is thus a leading entry. After that, add multiples of the second row to the other rows to produce zeros below the leading entry.

After that, continue finding the pivot columns, producing leading entries in the third row, fourth row, etc., if possible. Of course, eventually there will be no further pivot columns to be found, and the matrix produced is an echelon form of C. If the total number of pivot columns is ℓ, then the first ℓ rows in an echelon form will have leading entries. Any rows below the ℓth row will be zero rows.

In fact, if the original matrix C is a zero matrix, it is already in row echelon form, and there are no pivot columns. Similarly, if the matrix

$$\widetilde{C}_2 = \begin{bmatrix} 0 & 0 & 2 & -1 & -3 \\ 0 & 0 & 0 & 5 & 7 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

is produced by a succession of elementary row operations on C, then the algorithm stops because $\widetilde{C}_2$ is already in a row echelon form.

Example 1.1

Use elementary row operations on matrix C as follows to find a matrix that is row equivalent to C and is in a row echelon form. [Note: There can be more than one correct final conclusion for this problem.]

Method:

$$C = \begin{bmatrix} 1 & -1 & 2 & 0 \\ -2 & 2 & -1 & 0 \\ -3 & 3 & 2 & 4 \end{bmatrix} \underset{\substack{2R_1+R_2 \to R_2 \\ 3R_1+R_3 \to R_3}}{\sim} \begin{bmatrix} 1 & -1 & 1 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 8 & 4 \end{bmatrix}$$

$$\underset{-\frac{8}{3}R_2+R_3 \to R_3}{\sim} \begin{bmatrix} \textcircled{1} & -1 & 1 & 0 \\ 0 & 0 & \textcircled{3} & 0 \\ 0 & 0 & 0 & \textcircled{4} \end{bmatrix}. \bigcirc$$

* If the method is implemented on a computer, the basic advice is to get an entry in the $(1,1)$ position of greatest absolute value, although the "implicit partial pivoting" technique is a little more complicated than that. The reasons why will be discussed in Chapter 8.

www.TechnicalBooksPDF.com

1.1.2 Augmented Matrices

For system (1.1), that is,

$$\left\{\begin{array}{rcrcrl} x_1 & - & x_2 & + & x_3 & = 0 \\ -2x_1 & + & 2x_2 & - & x_3 & = 0 \\ & & 3x_2 & + & 2x_3 & = 4 \end{array}\right\},$$

we can assemble all of the coefficients that multiply the variables, as well as the right-hand sides, in the *augmented matrix*

$$C = \left[\begin{array}{rrr|c} 1 & -1 & 1 & 0 \\ -2 & 2 & -1 & 0 \\ 0 & 3 & 2 & 4 \end{array}\right]. \tag{1.9}$$

We perform the elementary row operations $2R_1 + R_2 \to R_2, R_2 \leftrightarrow R_3$, *in that order*, giving the augmented matrix

$$\left[\begin{array}{rrr|c} \textcircled{1} & -1 & 1 & 0 \\ 0 & \textcircled{3} & 2 & 4 \\ 0 & 0 & \textcircled{1} & 0 \end{array}\right] \tag{1.10}$$

that corresponds to the system

$$\left\{\begin{array}{rcrcrl} x_1 & - & x_2 & + & x_3 & = 0 \\ & & 3x_2 & + & 2x_3 & = 4 \\ & & & & x_3 & = 0 \end{array}\right\}. \tag{1.11}$$

So, simple operations on system (1.1) exactly correspond to elementary row operations on the corresponding augmented matrix in (1.9).

Theorem 1.1

A succession of elementary row operations does not change the set of solutions of a system, that is, two row equivalent matrices are the augmented matrices of two equivalent systems.

In general, given (1.2), a system of m equations in n unknowns, we can form the corresponding $m \times (n+1)$ augmented matrix

$$C = \left[\begin{array}{ccccc|c} a_{11} & . & . & . & a_{1n} & b_1 \\ . & . & & & . & . \\ . & & . & & . & . \\ . & & & . & . & . \\ a_{m1} & . & . & . & a_{mn} & b_m \end{array}\right]$$

and then perform a succession of elementary row operations to obtain $\widetilde{C}$, a matrix in a row echelon form.

www.TechnicalBooksPDF.com

For the moment, we will discuss a special case of this situation: suppose that $m = n$ and that the first n columns of B are pivot columns. In this *special* case, B has the form

$$\left[\begin{array}{ccccccc|c} \tilde{c}_{11} & \tilde{c}_{12} & \tilde{c}_{13} & . & . & . & \tilde{c}_{1n} & \tilde{c}_{1,n+1} \\ 0 & \tilde{c}_{22} & \tilde{c}_{23} & . & . & . & \tilde{c}_{2n} & \tilde{c}_{2,n+1} \\ 0 & 0 & \tilde{c}_{33} & . & . & . & \tilde{c}_{3n} & \tilde{c}_{3,n+1} \\ . & . & . & . & & & . & . \\ . & . & . & & . & & . & . \\ . & . & . & & & . & . & . \\ 0 & 0 & 0 & . & . & . & \tilde{c}_{nn} & \tilde{c}_{n,n+1} \end{array}\right],$$

where $\tilde{c}_{nn} \neq 0$. (Why?) We can use the last row of $\widetilde{C}$ to solve for x_n, the next to last row to solve for x_{n-1}, etc. The latter process is called "back substitution" because, after solving the last equation for x_n, we substitute that into the next to last equation to solve for x_{n-1}, etc. This process of reduction to a row echelon form, followed by back substitution, is called the **Gaussian elimination method** for solving a system of equations.

Example 1.2

Use the Gaussian elimination method to solve system (1.1).

Method: System (1.1) has augmented matrix (1.9), and elementary row operations reduce it to (1.10), which corresponds to system (1.11), that is,

$$\left\{\begin{array}{rcrcrl} x_1 & - & x_2 & + & x_3 & = 0 \\ & & 3x_2 & + & 2x_3 & = 4 \\ & & & & x_3 & = 0 \end{array}\right\}.$$

Solving the last equation gives $x_3 = 0$. Substituting $x_3 = 0$ into the next to last equation gives $x_2 = \frac{1}{3}(4 - 2 \cdot 0) = \frac{4}{3}$. Substituting $x_2 = \frac{4}{3}$, $x_3 = 0$ into the first equation gives $x_1 = x_2 - x_3 = \frac{4}{3} - 0 = \frac{4}{3}$. The system has exactly one solution: $(x_1, x_2, x_3) = \left(\frac{4}{3}, \frac{4}{3}, 0\right)$. ○

Now, instead of using back substitution, we could have continued row reducing the augmented matrix.

1.1.3 Row Reduced Echelon Form (RREF)

If a leading entry is the number 1, we call it a **leading 1**.

Definition 1.4

If a matrix is in a row echelon form, that is, satisfies properties (1.3) through (1.5), and *also* satisfies two other properties,

$$\text{Each leading entry is a leading 1, and} \tag{1.12}$$

$$\text{any column with a leading 1 has zeros above the leading 1,} \tag{1.13}$$

then we say that the matrix is in RREF.

www.TechnicalBooksPDF.com

The **Gauss–Jordan method** for solving a system of equations row reduces the augmented matrix to RREF.

In MATLAB® there is a command, RREF, that reduces a given matrix to RREF.

Example 1.3

Use the Gauss–Jordan method to solve system (1.1).

Method: Continuing from (1.10),

$$\left[\begin{array}{rrr|c} 1 & -1 & 1 & 0 \\ -2 & 2 & -1 & 0 \\ 0 & 3 & 2 & 4 \end{array}\right] \sim \left[\begin{array}{rrr|c} 1 & -1 & 1 & 0 \\ 0 & 3 & 2 & 4 \\ 0 & 0 & 1 & 0 \end{array}\right] \underset{\substack{-2R_3+R_2 \to R_2 \\ -R_3+R_1 \to R_1}}{\sim} \left[\begin{array}{rrr|c} 1 & -1 & 0 & 0 \\ 0 & 3 & 0 & 4 \\ 0 & 0 & 1 & 0 \end{array}\right]$$

$$\underset{\substack{\frac{1}{3}R_2 \to R_2 \\ R_2+R_1 \to R_1}}{\sim} \left[\begin{array}{ccc|c} \textcircled{1} & 0 & 0 & \frac{4}{3} \\ 0 & \textcircled{1} & 0 & \frac{4}{3} \\ 0 & 0 & \textcircled{1} & 0 \end{array}\right] = \text{RREF}(C), \tag{1.14}$$

which corresponds to the system

$$\left\{\begin{array}{rrrl} x_1 & & & = \frac{4}{3} \\ & x_2 & & = \frac{4}{3} \\ & & x_3 & = 0 \end{array}\right\}.$$

The system has exactly one solution, namely, $(x_1,\ x_2,\ x_3) = \left(\frac{4}{3},\ \frac{4}{3},\ 0\right)$. ○

We could have read the solution directly from (1.14) without bothering to write the corresponding system of equations.

Example 1.4

Solve the system of equations

$$\left\{\begin{array}{rcrcrl} x_1 & + & 2x_2 & - & 2x_3 & = -6 \\ -2x_1 & - & 4x_2 & + & 5x_3 & = 17 \\ x_1 & + & 2x_2 & - & x_3 & = -1 \end{array}\right\}. \tag{1.15}$$

Method:

$$\left[\begin{array}{rrr|c} 1 & 2 & -2 & -6 \\ -2 & -4 & 5 & 17 \\ 1 & 2 & -1 & -1 \end{array}\right] \underset{\substack{2R_1+R_2 \to R_2 \\ -R_1+R_3 \to R_3}}{\sim} \left[\begin{array}{rrr|c} 1 & 2 & -2 & -6 \\ 0 & 0 & 1 & 5 \\ 0 & 0 & 1 & 5 \end{array}\right]$$

$$\underset{\substack{-R_2+R_3 \to R_3 \\ 2R_2+R_1 \to R_1}}{\sim} \left[\begin{array}{ccc|c} \textcircled{1} & 2 & 0 & 4 \\ 0 & 0 & \textcircled{1} & 5 \\ 0 & 0 & 0 & 0 \end{array}\right],$$

www.TechnicalBooksPDF.com

which corresponds to the system

$$\left\{\begin{array}{rcrl} x_1 & + & 2x_2 & = 4 \\ & & x_3 & = 5 \\ & & 0 & = 0 \end{array}\right\}.$$

This system has infinitely many solutions: The second equation gives $x_3 = 5$, and the first equation gives $x_1 = 4 - 2x_2$. For *any* value of x_2, we have a distinct solution of the original system. The solutions are

$$(x_1,\ x_2,\ x_3) = (4 - 2c_1,\ c_1,\ 5),$$

where c_1 is an arbitrary constant. ◯

We recall from integral calculus and previous study of ordinary differential equations that arbitrary constants, such as c_1, often appear when we solve mathematical problems.

But why not make x_1 be the arbitrary constant and solve for x_2 in terms of x_1, specifically $x_1 + 2x_2 = 4$ gives $x_2 = \frac{1}{2}(4 - x_1)$? Other than wanting to avoid a fraction in the solution, it seems to be just as good. But we will see that to make the method more systematic, we introduce two more definitions that will "standardize" our work and thus make it both more reliable and more accessible to the reader. By the way, the engineering profession favors establishing standards for exactly the same reasons.

Definition 1.5

Suppose an augmented matrix C is $m \times (n+1)$ and has $\text{RREF}(C) = C^{RR}$. For $1 \le k \le n$,

(a) x_k is a **free variable** if the kth column of C^{RR} is not a pivot column.
(b) x_k is a **basic variable** if the kth column of C^{RR} is a pivot column.

Note that entries in the kth column of an augmented matrix C^{RR} multiply x_k in the corresponding system.

We have an immediately useful result:

Theorem 1.2

The $m \times (n+1)$ augmented matrix has the $(n+1)$th column of C^{RR} being a pivot column if, and only if, the corresponding system of equations has no solution.

The $(n+1)$th column of C^{RR} is a pivot column exactly when the bottom nonzero row of C^{RR} is $[0 \quad 0 \quad \cdots \quad 0|①]$.

Theorem 1.3

A system (1.2) has either no solution, exactly one solution, or infinitely many solutions.

www.TechnicalBooksPDF.com

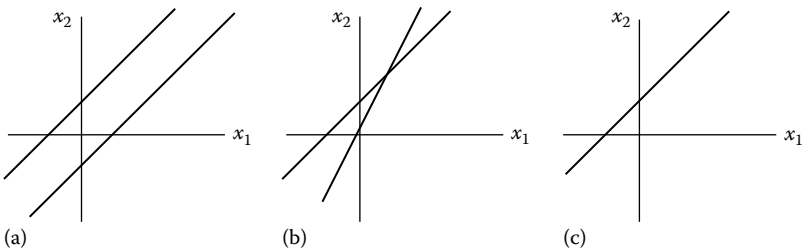


FIGURE 1.1
Theorem 1.3. (a) No solution, (b) exactly one solution, (c) infinitely many solutions.

Why? Theorem 1.2 tells us when a system has no solution. On the other hand, if a system has at least one solution, either there is no free variable, in which case there is exactly one solution of the system, or there is a free variable, in which case there are infinitely many solutions. □

The set of points (x_1, x_2) that satisfy a single algebraic equation is a line in the x_1x_2 plane. The points (x_1, x_2) that satisfy a system of two algebraic equations, that is, satisfy both of the equations "simultaneously," can be illustrated by drawing two lines in a plane. Figure 1.1 illustrates Theorem 1.3.

From now on, we will add to both the Gauss–Jordan and the Gaussian elimination methods the requirement that we will solve for all of the basic variables in terms of the free variables and that we will replace the latter by arbitrary constants, just as we did in Example 1.4.

Theorem 1.4

Any given matrix has exactly one RREF.

Why? See (Yuster 1984).

Because of this theorem, we refer to *the* RREF of a matrix, as opposed to *a* row echelon form.

Associated with the RREF are many useful results.

Definition 1.6

Given any matrix C, we define $rank(C)$ to be the rank of $\text{RREF}(C)$, that is, its number of pivot columns.

Theorem 1.5

If $\widetilde{C}$ is row equivalent to C, then (a) $\text{RREF}(\widetilde{C}) = \text{RREF}(C)$, and (b) $rank(\widetilde{C}) = rank(C)$.

www.TechnicalBooksPDF.com

Theorem 1.6

If A is $m \times n$, then $rank(A) \leq \min\{m, n\}$.

Why? The $rank(A)$ is the number of leading ones in RREF(A). Each leading 1 is in one column, so the number of leading ones is less than or equal to the number of columns of A, that is, less than or equal to n. On the other hand, each row can have at most one leading 1, so the number of leading ones is less than or equal to the number of rows, that is, less than or equal to m. Because $\min\{m, n\}$ is either m or n, both of which are greater than or equal to $rank(A)$, the result follows. □

Now, associated with a system (1.2), that is,

$$\left\{\begin{array}{c} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{array}\right\},$$

are five types of matrices. Denote

$$A = \begin{bmatrix} a_{11} & . & . & . & a_{1n} \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ a_{m1} & . & . & . & a_{mn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ . \\ . \\ . \\ b_m \end{bmatrix}.$$

We call A the **matrix of coefficients**, $\mathbf{x}$ the vector of **unknowns**, and $\mathbf{b}$ the vector of **right-hand sides**. The **augmented matrix** is

$$C = [A \mid \mathbf{b}\,] = \left[\begin{array}{ccccc|c} a_{11} & . & . & . & a_{1n} & b_1 \\ . & . & & & . & . \\ . & & . & & . & . \\ . & & & . & . & . \\ a_{m1} & . & . & . & a_{mn} & b_m \end{array}\right],$$

and the latter is RREF(C).

The next result is called a "lemma" because it helps to establish the theorem after that.

Lemma 1.1

$$\text{RREF}([A \mid \mathbf{b}\,]) = [\text{RREF}(A) \mid \mathbf{c}^{RR}\,], \text{ where } \mathbf{c}^{RR} = \begin{bmatrix} c^{RR}_{1,n+1} \\ c^{RR}_{2,n+1} \\ . \\ . \\ . \\ c^{RR}_{m,n+1} \end{bmatrix}.$$

www.TechnicalBooksPDF.com

Why? The elementary row operations that reduce A to its RREF also reduce the augmented matrix, $[A \mid \mathbf{b}\,]$, to a matrix that is in reduced row echelon form, except for possibly its $(n+1)$th column.

Theorem 1.7

There are three cases for solutions of a linear system:

(a) If $rank([A \mid \mathbf{b}\,]) > rank(A)$, then (1.2) has no solution.
(b) If $rank([A \mid \mathbf{b}\,]) = rank(A) = n$, then (1.2) has exactly one solution.
(c) If $rank([A \mid \mathbf{b}\,]) = rank(A) < n$, then (1.2) has infinitely many solutions, written in terms of $(n - rank(A))$ arbitrary constants corresponding to the free variables.

Example 1.5

Study system (1.15) using Theorem 1.7.

Method: Example 1.4 gave the elementary row operations to find that

$$[A \mid \mathbf{b}] = \left[\begin{array}{rrr|r} 1 & 2 & -2 & -6 \\ -2 & -4 & 5 & 17 \\ 1 & 2 & -1 & -1 \end{array}\right] \sim \cdots \sim \left[\begin{array}{ccc|c} \textcircled{1} & 2 & 0 & 4 \\ 0 & 0 & \textcircled{1} & 5 \\ 0 & 0 & 0 & 0 \end{array}\right] = [\text{RREF}(A) \mid \mathbf{c}^{RR}]$$

has $rank([A \mid \mathbf{x}]) = rank(A) = 2 < 3 = n$. So, system (1.15) has infinitely many solutions with one arbitrary constant, according to Theorem 1.7(c). ◯

Example 1.6

Study the system

$$\left\{\begin{array}{rcrcrcl} x_1 & + & 2x_2 & - & 2x_3 & = & -6 \\ -2x_1 & - & 4x_2 & + & 5x_3 & = & 17 \\ x_1 & & 2x_2 & - & x_3 & = & 0 \end{array}\right\} \tag{1.16}$$

using Theorem 1.7.

Method: The coefficient matrix A is the same for system (1.16) as for system (1.15), so we can use the same elementary row operations as in Example 1.4. We get

$$[A \mid \mathbf{x}] = \left[\begin{array}{rrr|r} 1 & 2 & -2 & -6 \\ -2 & -4 & 5 & 17 \\ 1 & 2 & -1 & 0 \end{array}\right] \sim \cdots \sim \left[\begin{array}{ccc|c} \textcircled{1} & 2 & 0 & 0 \\ 0 & 0 & \textcircled{1} & 0 \\ 0 & 0 & 0 & \textcircled{1} \end{array}\right] = [\text{RREF}(A) \mid \mathbf{c}^{RR}],$$

so $rank([A|\mathbf{x}]) = 3 > 2 = rank(A)$. According to Theorem 1.7(a), system (1.16) has no solution. ◯

Note that system (1.16) has no solution *even though* x_2 is a free variable.

In Example 1.6, we could have instead used Theorem 1.2; the latter is the reason why Theorem 1.7(a) is true.

Generally, computer methods use the Gaussian elimination method followed by back substitution, not the Gauss–Jordan method, because for "large" matrices, back substitution uses fewer operations than the alternative. Nevertheless, for small systems that we solve

www.TechnicalBooksPDF.com

by hand, the Gauss–Jordan method is perfectly fine. Also, we will see in Section 1.5 that the Gauss–Jordan method will be used to find the "inverse" of a square matrix.

Example 1.7

For the direct current (DC) circuit shown in Figure 1.2, set up a system of three equations for the loop currents I_1, I_2, I_3; write the corresponding augmented matrix; and then find the exact solution in terms of the unspecified voltages V_1, V_2.

Method: The Kirchhoff voltage law applied to the three loops, starting at the upper right and proceeding clockwise, gives the three equations

$$\left\{\begin{array}{c} 2I_1 + 3(I_1 - I_2) + 1 \cdot I_1 = V_1, \\ 3(I_2 - I_1) + 7I_2 + 9I_2 + 10(I_2 - I_3) = 0, \\ 10(I_3 - I_2) + 5I_3 + 6I_3 = V_2. \end{array}\right\}.$$

The corresponding augmented matrix is

$$\left[\begin{array}{ccc|c} 6 & -3 & 0 & V_1 \\ -3 & 29 & -10 & 0 \\ 0 & -10 & 21 & V_2 \end{array}\right].$$

If we do the row operation $\frac{1}{2}R_1 + R_2 \to R_2$ followed by $\frac{4}{11}R_2 + R_3 \to R_3$, then we have a row echelon form

$$\left[\begin{array}{ccc|c} 6 & -3 & 0 & V_1 \\ 0 & \frac{55}{2} & -10 & \frac{1}{2}V_1 \\ 0 & 0 & \frac{191}{11} & \frac{2}{11}V_1 + V_2 \end{array}\right].$$

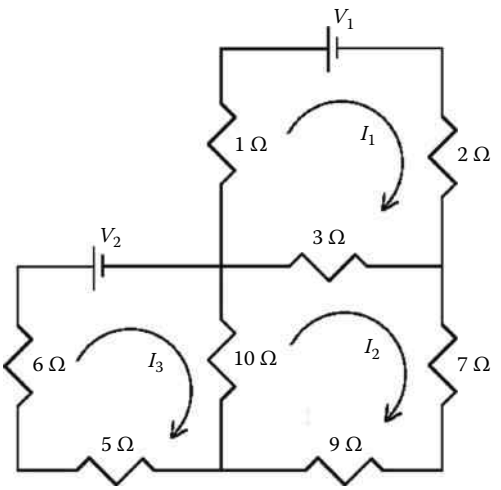


FIGURE 1.2
Example 1.7.

www.TechnicalBooksPDF.com

Using back substitution gives, successively, the exact loop currents in terms of V_1, V_2:

$$I_3 = \frac{2}{191}V_1 + \frac{11}{191}V_2, \quad I_2 = \frac{21}{955}V_1 + \frac{4}{191}V_2, \quad \text{and} \quad I_1 = \frac{509}{2865}V_1 + \frac{2}{191}V_2. \bigcirc$$

Computer methods may put in the pivot positions, entries whose relative magnitudes are as large as possible, in order to reduce "numerical error." Computer methods may do "partial pivoting," that is, row interchanges that move relatively larger numbers into pivot positions, or even "complete pivoting," which both exchanges rows and exchanges columns. (Exchanging columns corresponds to exchanging the index of two variables, so such a method must keep track of the variable indices and return them to their original indices in the final conclusion.)

1.1.4 Problems

In problems 1–3, for each system, use a row reduction algorithm to find all solutions.

1. $\left\{\begin{matrix} x & - & y & = -7 \\ 3x & - & 4y & = 11 \end{matrix}\right\}$.
2. $\left\{\begin{matrix} x_1 & - & x_2 & + & 2x_3 & = -1 \\ 2x_1 & + & x_2 & + & x_3 & = -1 \\ -x_1 & + & 3x_2 & - & 4x_3 & = \frac{5}{3} \end{matrix}\right\}$.
3. $\left\{\begin{matrix} & - & x_2 & + & 4x_3 & = 1 \\ -x_1 & + & 3x_2 & + & 2x_3 & = 0 \\ 2x_1 & & & - & x_3 & = -2 \end{matrix}\right\}$.
4. Find the RREF of the matrix

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 0 & 1 & 1 & 1 \\ 2 & 0 & 0 & 1 \end{bmatrix}.$$

5. Suppose that $[A \mid \mathbf{b}]$ can be row reduced to

$$\begin{bmatrix} 1 & 2 & 3 & -1 & \vert 4 \\ 0 & 0 & 1 & 0 & \vert 5 \\ 0 & 0 & 0 & 0 & \vert r \end{bmatrix}.$$

   (a) If $r = 6$, find all solutions of $A\mathbf{x} = \mathbf{b}$.
   (b) If $r = 0$, find all solutions of $A\mathbf{x} = \mathbf{b}$.

6. A system $A\mathbf{x} = \mathbf{b}$ had its augmented matrix row reduced to $\begin{bmatrix} 1 & 2 & 0 & \vert r_1 \\ 0 & 0 & 1 & \vert r_2 \\ 0 & 0 & 0 & \vert r_3 \end{bmatrix}$.

www.TechnicalBooksPDF.com

(a) Find all solutions of $A\mathbf{x} = \mathbf{b}$, if

$$\begin{bmatrix} r_1 \\ r_2 \\ r_3 \end{bmatrix} = \begin{bmatrix} -1 \\ 3 \\ 0 \end{bmatrix}.$$

(b) Find all solutions of $A\mathbf{x} = \mathbf{b}$, if

$$\begin{bmatrix} r_1 \\ r_2 \\ r_3 \end{bmatrix} = \begin{bmatrix} -1 \\ 3 \\ 2 \end{bmatrix}.$$

7. Write down three different 3×4 matrices that are in RREF and whose rank is exactly three.

8. Let $A = \begin{bmatrix} 1 & 0 & \blacksquare \\ \blacksquare & 1 & 0 \\ 0 & \blacksquare & 1 \end{bmatrix}$.

   (a) Replace the three ■s by three different positive integers, each ≥ 2. Write down your A.

   (b) For the matrix A you wrote in part (a), use the Gauss–Jordan or the Gaussian elimination method to solve the system

$$A\mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

9. The Chevford Company produces hybrid trucks and cars at three plants: In an hour, Plant I produces 4 cars and 4 trucks, Plant II produces 4 cars and 1 truck, and Plant III produces 2 cars and 3 trucks. Currently, Plant I runs 7 hours a day, Plant II runs 6 hours a day, and Plant III runs 9 hours a day.

   As part of a program of regular maintenance, Plant I must be shut down for a week. The Company will make up for the shutdown of Plant I by adjusting the number of hours per day that the two remaining plants will run. What should be the total number of hours per day that Plants II and III should run while Plant I is shut down, in order that *at least* the same total production should be achieved? Turn this problem into a system of linear equations and then solve it.

10. The foods puffed rice, rolled oats, bran flakes, and corn flakes contain the nutrients protein, carbohydrates, and calories. Specifically, one cup of puffed rice contains* 1 g of protein, 13 g of carbohydrates, and 60 cal, one cup of rolled oats contains 5 g of protein, 23 g of carbohydrates, and 130 cal, one cup of bran flakes contain 4 g of protein, 28 g of carbohydrates, and 105 cal, and one cup of corn flakes contain 2 g of protein, 20 g of carbohydrates, and 95 cal. Can the foods puffed rice, rolled oats, and bran flakes be combined to yield a new cereal with the same nutritional content as corn flakes? If so, give your answer in the form 106 cups of corn flakes =__ cups of puffed rice +__ cups of rolled oats+__ cups of bran flakes.

* Nutritional information from the 1985 World Almanac.

www.TechnicalBooksPDF.com

Solve Problem 1.1.4.10* by first turning it into a system of linear equations and then solving it.

11. A dietician is planning a meal to supply certain quantities of calcium, potassium, and magnesium. Three foods will be used. One hundred grams of food #1 contains 40 mg of calcium, 20 mg of potassium, and 40 mg of magnesium. One hundred grams of food #2 contains 70 mg of calcium, 10 mg of potassium, and 30 mg of magnesium. One hundred grams of food #3 contains 50 mg of calcium, 40 mg of potassium, and 60 mg of magnesium. If the meal is to contain exactly 120 mg of calcium, 30 mg of potassium, and 70 mg of magnesium, use a system of equations to find the amounts of foods #1, #2, and #3 that should be consumed in a meal to exactly meet the dietary requirements for calcium, potassium, and magnesium. Do state clearly what quantities your variables represent.
12. Suppose Midwestern University is considering how much of the area of its central quadrangle to turn into landscaped parking and how much to turn into a parking garage. Landscaped parking spaces require 50 ft^2 of surface area per car, and parking garage spaces require 10 ft^2 of surface area per car. The total quadrangle area is 8000 ft^2, and they want to have a total of 250 parking spaces. Turn this problem into a system of linear equations and then solve it.
13. Assume A is an $m \times n$ matrix, the system $A\mathbf{x} = \mathbf{b}$ has infinitely many solutions and $m > n$. For each of (a)–(e), decide whether it must be true, must be false, or may be true or may be false.
    (a) A has a row of zeros.
    (b) $[\,A \mid \mathbf{b}\,]$ has a row of zeros.
    (c) The RREF of A has a row of zeros.
    (d) The RREF of $[\,A \mid \mathbf{b}\,]$ has a row of zeros.
    (e) Replace the assumption $m > n$ by the assumption $m = n = 2$. Then the system $A\mathbf{x} = \mathbf{b}$ can be represented geometrically as two lines that are identical.
14. For the DC circuit shown in Figure 1.3, set up a system of three equations for the loop currents I_1, I_2, I_3; write the corresponding augmented matrix; and then find the exact solution in terms of the unspecified voltages V_1, V_2. On your picture, indicate which loop current is which.
15. Suppose that a system $A\mathbf{x} = \mathbf{b}$ has its augmented matrix row equivalent to

$$\left[\begin{array}{cccc|c} 1 & 0 & 0 & 2 & 5 \\ 0 & 0 & 1 & 0 & 6 \\ 0 & 2 & 3 & -4 & 7 \\ 0 & 4 & 6 & -8 & 14 \end{array}\right].$$

    (a) Use elementary row operations to further row reduce the aforementioned to find RREF($[\,A \mid \mathbf{b}\,]$) and RREF(A), and label which is which.
    (b) Find all solutions of $A\mathbf{x} = \mathbf{b}$.
    (c) Find the ranks of RREF($[\,A \mid \mathbf{b}\,]$) and RREF(A), and label which is which if they are not equal.

* This is a problem we wrote that appeared in Lederer (1989).

www.TechnicalBooksPDF.com

With this definition, we can write compactly the system of equations

$$\left\{\begin{array}{ll} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n & = b_1 \\ \vdots & \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n & = b_m \end{array}\right\}$$

as

$$A\mathbf{x} = \mathbf{b}.$$

In the definition of multiplication the n's "match," that is, A is $m \times \underline{n}$ and $\mathbf{x}$ is $\underline{n} \times 1$.

We can write the matrix A in terms of its columns, specifically

$$A = \left[\mathbf{A}_{*1} \,\vdots\, \mathbf{A}_{*2} \,\vdots\, \cdots \,\vdots\, \mathbf{A}_{*n}\right], \tag{1.18}$$

where its columns are denoted by

$$\mathbf{A}_{*j} = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{bmatrix}, \; j = 1, \ldots, n.$$

We write the columns in boldface type because they are vectors.

Equation (1.18) is an example of writing a matrix as a **partitioned matrix**.

The next result will be very useful in Section 1.7 for understanding what matrices do, but the result here comes from simple calculations.

Lemma 1.2

$$A\mathbf{x} = x_1\mathbf{A}_{*1} + x_2\mathbf{A}_{*2} + \cdots + x_n\mathbf{A}_{*n}. \tag{1.19}$$

Why? Because

$$A\mathbf{x} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{bmatrix} = \begin{bmatrix} x_1a_{11} \\ x_1a_{21} \\ \vdots \\ x_1a_{m1} \end{bmatrix} + \begin{bmatrix} x_2a_{12} \\ x_2a_{22} \\ \vdots \\ x_2a_{m2} \end{bmatrix} + \cdots + \begin{bmatrix} x_na_{1n} \\ x_na_{2n} \\ \vdots \\ x_na_{mn} \end{bmatrix}$$

$$= x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} + \cdots + x_n \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix} = x_1\mathbf{A}_{*1} + x_2\mathbf{A}_{*2} + \cdots + x_n\mathbf{A}_{*n}. \; \square$$

Now that we have a definition for multiplication of an $m \times n$ matrix times an n-vector, that is, multiplication of an $m \times n$ matrix times an $n \times 1$ matrix, we can generalize to define multiplication of an $m \times n$ matrix A by an $n \times p$ matrix B.

www.TechnicalBooksPDF.com

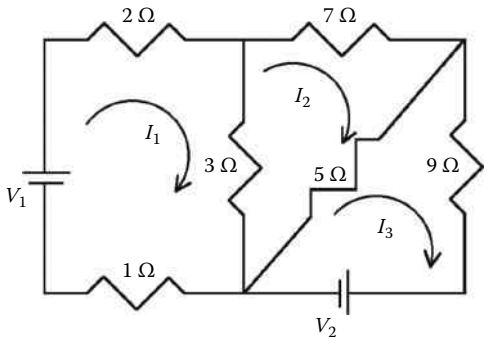


FIGURE 1.3
Problem 1.1.4.14.

1.2 Matrix Addition, Multiplication, and Transpose

If $A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$ and $C = \begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & & \vdots \\ c_{m1} & \cdots & c_{mn} \end{bmatrix}$ are two matrices of the same size, we can define their sum by $A + C \triangleq \begin{bmatrix} a_{11} + c_{11} & \cdots & a_{1n} + c_{1n} \\ \vdots & & \vdots \\ a_{m1} + c_{m1} & \cdots & a_{mn} + c_{mn} \end{bmatrix}$.

This defines **matrix addition**.

We've defined

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \text{ and } \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}.$$

The set of all such n-vectors $\mathbf{x}$, having real number entries, is called $\mathbb{R}^n$.

We define multiplication of a matrix times a vector, that is, $A\mathbf{x}$, by

Definition 1.7

$$A\mathbf{x} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \triangleq \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{bmatrix}. \tag{1.17}$$

www.TechnicalBooksPDF.com

Definition 1.8

Matrix multiplication is defined by

$$AB = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} b_{11} & \cdots & b_{1p} \\ \vdots & & \vdots \\ b_{n1} & \cdots & b_{np} \end{bmatrix}$$

$$\triangleq \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + \cdots + a_{1n}b_{n1} & \cdots & a_{11}b_{1p} + a_{12}b_{2p} + \cdots + a_{1n}b_{np} \\ \vdots & & \vdots \\ a_{m1}b_{11} + a_{m2}b_{21} + \cdots + a_{mn}b_{n1} & \cdots & a_{m1}b_{1p} + a_{m2}b_{2p} + \cdots + a_{mn}b_{np} \end{bmatrix}.$$

Example 1.8

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 7 & 10 \\ 8 & 11 \\ 9 & 12 \end{bmatrix} = \begin{bmatrix} 1\cdot 7 + 2\cdot 8 + 3\cdot 9 & 1\cdot 10 + 2\cdot 11 + 3\cdot 12 \\ 4\cdot 7 + 5\cdot 8 + 6\cdot 9 & 4\cdot 10 + 5\cdot 11 + 6\cdot 12 \end{bmatrix}$$

$$= \begin{bmatrix} 50 & 68 \\ 122 & 167 \end{bmatrix}. \bigcirc$$

Recall that we can write a matrix in terms of its rows, for example,

$$A = \begin{bmatrix} \mathbf{A}_{1*} \\ - - \\ \mathbf{A}_{2*} \\ - - \\ \vdots \\ - - \\ \mathbf{A}_{m*} \end{bmatrix}$$

has rows denoted by $\mathbf{A}_{i*} = [a_{i1} \quad a_{i2} \quad \cdots \quad a_{in}]$, $i = 1, \ldots, m$.

Similarly, we can write a matrix in terms of its columns, for example, $B = \left[\mathbf{B}_{*1} \;\vdots\; \mathbf{B}_{*2} \;\vdots\; \cdots \;\vdots\; \mathbf{B}_{*p} \right]$ has columns denoted by $\mathbf{B}_{*j} = \begin{bmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{nj} \end{bmatrix}$, $j = 1, \ldots, p$.

www.TechnicalBooksPDF.com

The product of a row vector with n entries, that is, a $1 \times n$ matrix, and a column vector with n entries, that is, a $n \times 1$ matrix, is a special case of matrix multiplication, for example,

$$\mathbf{A}_{i*} \bullet \mathbf{B}_{*j} = \begin{bmatrix} a_{i1} & a_{i2} & \cdots & a_{in} \end{bmatrix} \begin{bmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{nj} \end{bmatrix} = \begin{bmatrix} a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} \end{bmatrix}.$$

We use a $\bullet$ to remind ourselves of the dot product used for vectors in physics. We may consider a 1×1 matrix to be a scalar, that is, a number, so we may write

$$\mathbf{A}_{i*} \bullet \mathbf{B}_{*j} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj}.$$

Theorem 1.8

If A is $m \times n$ and B is $n \times p$,

$$AB = \begin{bmatrix} a_{11} & . & . & . & a_{1n} \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ a_{m1} & . & . & . & a_{mn} \end{bmatrix} \begin{bmatrix} b_{11} & . & . & . & b_{1p} \\ & & & & \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ & & & & \\ b_{n1} & . & . & . & b_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{1*} \bullet \mathbf{B}_{*1} & . & . & . & \mathbf{A}_{1*} \bullet \mathbf{B}_{*p} \\ & . & & & . \\ & & . & & . \\ & & & . & . \\ \mathbf{A}_{m*} \bullet \mathbf{B}_{*1} & . & . & . & \mathbf{A}_{m*} \bullet \mathbf{B}_{*p} \end{bmatrix}.$$

In general, $AB \neq BA$, that is, the left–right order of matrix multiplication matters. For one thing, the sizes of A and B may allow only one of AB and BA to exist. And, even if they both exist, AB and BA may be of different sizes, for example, reusing the result of Example 1.8, we note the following:

Example 1.9

For the matrices $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$ and $B = \begin{bmatrix} 7 & 10 \\ 8 & 11 \\ 9 & 12 \end{bmatrix}$, check to see if $AB = BA$.

Method: $AB = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 7 & 10 \\ 8 & 11 \\ 9 & 12 \end{bmatrix} = \begin{bmatrix} 50 & 68 \\ 122 & 167 \end{bmatrix}$ and

$$BA = \begin{bmatrix} 7 & 10 \\ 8 & 11 \\ 9 & 12 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} = \begin{bmatrix} 47 & 64 & 81 \\ 52 & 71 & 90 \\ 57 & 78 & 99 \end{bmatrix},$$

so $AB \neq BA$. ◯

In fact, $AB = BA$ is only possible if A and B are both square and of the same size. But having A and B both of size $n \times n$ doesn't even come close to guaranteeing $AB = BA$. This reminds me of a street sign I saw that read "Don't even think of parking here."

www.TechnicalBooksPDF.com

Example 1.10

For the matrices $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}$, check to see if $AB = BA$.

Method: $AB = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix} \neq \begin{bmatrix} 23 & 34 \\ 31 & 46 \end{bmatrix} = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, so $AB \neq BA$. ◯

If we choose at "random" two $n \times n$ matrices A and B, there is as much chance of AB equaling BA as there is of flipping an infinitesimally thin coin and having it land standing on its side.

The following results are as amazingly useful as they are simple!

Theorem 1.9

If A is $m \times n$ and B is $n \times p$,

$$AB = A\left[\mathbf{B}_{*1} \,\vdots\, \mathbf{B}_{*2} \,\vdots\, \cdots \,\vdots\, \mathbf{B}_{*p}\right] = \left[A\mathbf{B}_{*1} \,\vdots\, A\mathbf{B}_{*2} \,\vdots\, \cdots \,\vdots\, A\mathbf{B}_{*p}\right], \tag{1.20}$$

and

$$AB = \begin{bmatrix} \mathbf{A}_{1*} \\ - - \\ \mathbf{A}_{2*} \\ - - \\ \vdots \\ - - \\ \mathbf{A}_{m*} \end{bmatrix} B = \begin{bmatrix} \mathbf{A}_{1*}B \\ - - - \\ \mathbf{A}_{2*}B \\ - - - \\ \vdots \\ - - - \\ \mathbf{A}_{m*}B \end{bmatrix}. \tag{1.21}$$

Why? The (i, j)th entry of AB is $a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj}$. On the other side of (1.20), the jth column of $\left[A\mathbf{B}_{*1} \,\vdots\, A\mathbf{B}_{*2} \,\vdots\, \cdots \,\vdots\, A\mathbf{B}_{*p}\right]$ is

$$A\mathbf{B}_{*j} = \begin{bmatrix} a_{11}b_{1j} + a_{12}b_{2j} + \cdots + a_{1n}b_{nj} \\ a_{21}b_{1j} + a_{22}b_{2j} + \cdots + a_{2n}b_{nj} \\ \vdots \\ a_{m1}b_{1j} + a_{m2}b_{2j} + \cdots + a_{mn}b_{nj} \end{bmatrix},$$

whose ith entry down is also $a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj}$. The explanation for (1.21) is similar. □

www.TechnicalBooksPDF.com

1.2.1 Special Kinds of Matrices

Definition 1.9

The $n \times n$ **identity** matrix is $I_n \triangleq \begin{bmatrix} 1 & 0 & . & . & . & 0 & 0 \\ 0 & 1 & . & . & . & 0 & 0 \\ . & . & . & & & . & . \\ . & . & & . & & . & . \\ . & . & & & . & . & . \\ 0 & 0 & . & . & . & 1 & 0 \\ 0 & 0 & . & . & . & 0 & 1 \end{bmatrix}$, or I, for short.

Definition 1.10

The $m \times n$ **zero** matrix is $O_{mn} \triangleq \begin{bmatrix} 0 & . & . & . & 0 \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ 0 & . & . & . & 0 \end{bmatrix}$, or O, for short.

We have the following facts:

$$AI_n = A,\ I_nB = B,\ AO_{nq} = O_{mq},\ O_{qn}B = O_{qp}$$

for any $m \times n$ matrix A and $n \times p$ matrix B. Because of these facts about multiplication, we call I_n an "identity" matrix, and we call O_{mn} a zero matrix.

It is useful to have another notation for a matrix:

$$A = \begin{bmatrix} a_{11} & . & . & . & a_{1n} \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ a_{m1} & . & . & . & a_{mn} \end{bmatrix} = \left[a_{ij} \right]_{\substack{1 \le i \le m \\ 1 \le j \le n}}.$$

Definition 1.11

$D = \left[d_{ij} \right]_{\substack{1 \le i \le m \\ 1 \le j \le n}}$ is called a **diagonal** matrix if $d_{ij} = 0$ for all (i, j) with $i \neq j$. If D is a diagonal matrix, we may write $D = diag(d_{11}, d_{22}, \ldots, d_{nn})$.

So, for example, $I = diag(1, \ldots, 1)$ and $O = diag(0, \ldots, 0)$.

www.TechnicalBooksPDF.com

Example 1.11

$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -3 \end{bmatrix}$ is a diagonal matrix; $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -3 & 0 \end{bmatrix}$ is not a diagonal matrix. ◯

Theorem 1.10

If $A = \left[\mathbf{A}_{*1} \mid \mathbf{A}_{*2} \mid \cdots \mid \mathbf{A}_{*n} \right]$ and $D = diag(d_{11}, d_{22}, \ldots, d_{nn})$, then

$$AD = \left[d_{11}\mathbf{A}_{*1} \mid d_{22}\mathbf{A}_{*2} \mid \cdots \mid d_{nn}\mathbf{A}_{*n} \right].$$

Why? You will explain why Theorem 1.10 is true in Problem 1.2.5.13.

Definition 1.12

(a) $U = \left[u_{ij} \right]_{\substack{1 \le i \le m \\ 1 \le j \le n}}$ is called an **upper triangular** matrix if $u_{ij} = 0$ for all (i, j) with $i > j$. For example, for $m = n$, an upper triangular matrix has the form

$$\begin{bmatrix} u_{11} & u_{12} & u_{13} & . & . & . & u_{1n} \\ 0 & u_{22} & u_{23} & . & . & . & u_{2n} \\ 0 & 0 & u_{33} & . & . & . & u_{3n} \\ . & . & . & . & & & . \\ . & . & . & & . & & . \\ . & . & . & & & . & . \\ 0 & 0 & 0 & . & . & . & u_{nn} \end{bmatrix}.$$

For $m > n$, an upper triangular matrix has the form

$$\begin{bmatrix} u_{11} & u_{12} & u_{13} & . & . & . & u_{1n} \\ 0 & u_{22} & u_{23} & . & . & . & u_{2n} \\ 0 & 0 & u_{33} & . & . & . & u_{3n} \\ . & . & . & . & & & . \\ . & . & . & & . & & . \\ . & . & . & & & . & . \\ 0 & 0 & 0 & . & . & . & u_{nn} \\ 0 & 0 & 0 & . & . & . & 0 \\ . & . & . & . & & & . \\ . & . & . & & . & & . \\ . & . & . & & & . & . \\ 0 & 0 & 0 & . & . & . & 0 \end{bmatrix}.$$

(b) $L = \left[\ell_{ij} \right]_{\substack{1 \le i \le m \\ 1 \le j \le n}}$ is called a **lower triangular** matrix if $\ell_{ij} = 0$ for all (i, j) with $i < j$.

www.TechnicalBooksPDF.com

For example, for $m = n$, a lower triangular matrix has the form

$$\begin{bmatrix} \ell_{11} & 0 & 0 & . & . & . & 0 \\ \ell_{21} & \ell_{22} & 0 & . & . & . & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} & . & . & . & 0 \\ . & . & . & . & & & . \\ . & . & . & & . & & . \\ . & . & . & & & . & . \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & . & . & . & \ell_{nn} \end{bmatrix}.$$

Example 1.12

$\begin{bmatrix} 1 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$ is an upper triangular matrix. Recall that we saw this matrix, in the role of RREF(A), in Example 1.5 in Section 1.1. ○

We have a more general result:

Theorem 1.11

All matrices in row echelon form are upper triangular. In particular, all matrices in RREF are upper triangular.

1.2.2 Powers of a Matrix

Definition 1.13

Given any *square* matrix A, $A^0 \triangleq I$, $A^2 \triangleq AA$, $A^3 \triangleq AA^2, \ldots, A^{r+1} \triangleq AA^r, \ldots$ are called the **powers** of A.

Fact: $A^r A^s = A^{r+s}$ for any positive integers r and s.

Example 1.13

Does $(A + B)^2 = A^2 + 2AB + B^2$ for all matrices A and B?

Method: No, because $(A+B)^2 = (A+B)(A+B) = AA+AB+BA+BB = A^2+AB+BA+B^2$. In general, $(A + B)^2 \neq A^2 + 2AB + B^2$ because, in general, $BA \neq AB$. ○

There is one other property of matrix multiplication that we will use often:

Theorem 1.12

(Associativity): If the matrix multiplications CA and AB both make sense, that is, the dimensions of the matrices allow those multiplications, then $(CA)B = C(AB)$.

www.TechnicalBooksPDF.com

Because it doesn't matter where the parentheses are, as long as the order of the matrices stays the same, we often write CAB instead of $(CA)B$ or $C(AB)$.

1.2.3 Transpose

Definition 1.14

Given any $m \times n$ matrix $A = \left[a_{ij}\right]_{\substack{1 \le i \le m \\ 1 \le j \le n}}$, the **transpose** of A is the $n \times m$ matrix $A^T = \left[a_{ji}\right]_{\substack{1 \le j \le n \\ 1 \le i \le m}}$.

Example 1.14

$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$. Note that the rows of A become the columns of A^T and the columns of A become the rows of A^T. ○

Theorem 1.13

Assuming the sizes of A and B allow them to exist, then (a) $(A + B)^T = A^T + B^T$ and (b) $(AB)^T = B^T A^T$.

1.2.4 Elementary Matrices

It turns out that all three kinds of elementary row operations on a matrix can be implemented by multiplication on the left by "elementary matrices." For example,

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \underset{2R_1 + R_2 \to R_2}{\sim} \begin{bmatrix} a & b \\ 2a + c & 2b + d \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

so multiplication on the left by $\begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}$ implements the row operation $2R_1 + R_2 \to R_2$.

By the way, a matrix such as $\begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}$ and generalizations to 3×3 matrices are called "shear transformations" and come up in many subjects, including solid mechanics and computer graphics.

Here are some other elementary matrices and the corresponding elementary row they implement:

www.TechnicalBooksPDF.com

$$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \text{ implements } 2R_1 \to R_1.$$

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ implements } R_1 \leftrightarrow R_2.$$

1.2.5 Problems

1. Find another example of 2×2 matrices A and B for which $AB \neq BA$.
2. Suppose A is $n \times n$, $(A - A^T)(A - A^T) = 0$, and $A^2 = A$. Why must $(AA^T)^2 = AA^T$?
3. Find an example of 2×2 nonzero matrices A and B for which $AB = 0$.
4. Suppose A and B are 2×2 matrices, $rank(A) = 1$, and $rank(B) = 1$.
   (a) Find a specific example of A and B for which $rank(A + B) = 2$.
   (b) Find a specific example of A and B for which $rank(A + B) = 1$.
   (c) Find a specific example of A and B for which $rank(A + B) = 0$.
5. For $A = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & 2 \end{bmatrix}$, find A^2 and A^3.
6. If A is $m \times n$, B is $n \times p$, and the jth column of B is zero, why is the jth column of AB zero?
7. If A and B are both upper triangular matrices, why is AB also upper triangular?
8. If $D = diag(d_{11}, d_{22}, \ldots, d_{nn})$ is a diagonal matrix, why is D^k a diagonal matrix for every positive integer k? Find a formula for D^k in terms of the $d_{11}, d_{22}, \ldots, d_{nn}$.
9. True or false: If a 2×2 matrix A has $rank(A) = p$, then it must be true that $rank(A^2) = p$ also. If true, why? If false, give a specific counter example.
10. For the matrices of Example 1.8, verify that $(AB)^T = B^T A^T$.
11. For each of (a), (b), and (c), give the 3×3 elementary matrix that implements the elementary row operation: (a) $-2R_1 + R_3 \to R_3$, (b) $\frac{1}{2} R_2 \to R_2$, and (c) $R_2 \leftrightarrow R_3$.
12. Suppose L, D, U are lower triangular, diagonal, and upper triangular matrices, respectively. Explain why (a) LD and DL are lower triangular and (b) UD and DU are upper triangular, at least for those sizes of matrices for which the products that exist.

    While you may use specific examples to learn about this problem, your explanations should be in general, that is, not for specific matrices or even specific sizes of matrices. [Hint: Use Theorem 1.9, for example,

$$L\left[\mathbf{D}_{*1} \mid \mathbf{D}_{*2} \mid \cdots \mid \mathbf{D}_{*n}\right] = \left[L\mathbf{D}_{*1} \mid L\mathbf{D}_{*2} \mid \cdots \mid L\mathbf{D}_{*n}\right],$$

    and then use Lemma 1.2.]
13. Use Theorem 1.9 and Lemma 1.2 to explain why Theorem 1.10 is true.

www.TechnicalBooksPDF.com

1.3 Homogeneous Systems, Spanning Set, and Basic Solutions

Definition 1.15

A **homogeneous system** has all zeros on the right-hand sides, that is, has the form $A\mathbf{x} = \mathbf{0}$.

Example 1.15

$$\text{Solve} \begin{Bmatrix} 2x_1 & +4x_2 & +x_3 & = 0 \\ -x_1 & -3x_2 & +x_3 & = 0 \\ x_1 & +2x_2 & & = 0 \end{Bmatrix}.$$

Method:

$$\begin{bmatrix} 2 & 4 & 1 & \vert\, 0 \\ -1 & -3 & 1 & \vert\, 0 \\ 1 & 2 & 0 & \vert\, 0 \end{bmatrix} \underset{\substack{\frac{1}{2}R_1 + R_2 \to R_2 \\ -\frac{1}{2}R_1 + R_3 \to R_3}}{\sim} \begin{bmatrix} \textcircled{2} & 4 & 1 & \vert\, 0 \\ 0 & \textcircled{1} & 1.5 & \vert\, 0 \\ 0 & 0 & \textcircled{\frac{1}{2}} & \vert\, 0 \end{bmatrix}.$$

Back substitution gives $x_3 = 0$; substituting that gives $x_2 = 0$, and then substituting $x_3 = x_2 = 0$ gives $x_1 = 0$. There is exactly one solution: $(x_1, x_2, x_3) = (0, 0, 0)$. ◯

Example 1.16

Solve

$$\begin{Bmatrix} x_1 & +x_2 & +x_3 & & = 0 \\ x_1 & & -x_3 & & = 0 \\ 2x_1 & & -2x_3 & +x_4 & = 0 \\ -3x_1 & -3x_2 & -3x_3 & & = 0 \end{Bmatrix}. \tag{1.22}$$

Method:

$$\begin{bmatrix} 1 & 1 & 1 & 0 & \vert\, 0 \\ 1 & 0 & -1 & 0 & \vert\, 0 \\ 2 & 0 & -2 & 1 & \vert\, 0 \\ -3 & -3 & -3 & 0 & \vert\, 0 \end{bmatrix} \underset{\substack{-R_1 + R_2 \to R_2 \\ -2R_1 + R_3 \to R_3 \\ 3R_1 + R_4 \to R_4}}{\sim} \begin{bmatrix} 1 & 1 & 1 & 0 & \vert\, 0 \\ 0 & -1 & -2 & 0 & \vert\, 0 \\ 0 & -2 & -4 & 1 & \vert\, 0 \\ 0 & 0 & 0 & 0 & \vert\, 0 \end{bmatrix}$$

$$\underset{\substack{-2R_2 + R_3 \to R_3 \\ R_2 + R_1 \to R_1 \\ -R_2 \to R_2}}{\sim} \begin{bmatrix} \textcircled{1} & 0 & -1 & 0 & \vert\, 0 \\ 0 & \textcircled{1} & 2 & 0 & \vert\, 0 \\ 0 & 0 & 0 & \textcircled{1} & \vert\, 0 \\ 0 & 0 & 0 & 0 & \vert\, 0 \end{bmatrix}.$$

www.TechnicalBooksPDF.com

The latter is in RREF. We have circled the pivot positions, which are in the first, second, and fourth columns, so x_3 is the only free variable. Using the corresponding system of equations,

$$\left\{\begin{array}{rcrcl} x_1 & & -x_3 & = & 0 \\ & x_2 & +2x_3 & = & 0 \\ & & x_4 & = & 0 \\ & & 0 & = & 0 \end{array}\right\},$$

we solve for the basic variables, x_1, x_2, and x_4, in terms of the free variable: $x_4 = 0$, $x_2 = -2x_3$, and $x_1 = x_3$. The solutions are $(x_1,\ x_2,\ x_3,\ x_4) = (x_3, -2x_3,\ x_3,\ 0) = (c_1, -2c_1,\ c_1,\ 0)$, where c_1 is an arbitrary constant. ◯

We can also write the solutions in vector form:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} c_1 \\ -2c_1 \\ c_1 \\ 0 \end{bmatrix} = c_1 \begin{bmatrix} 1 \\ -2 \\ 1 \\ 0 \end{bmatrix}$$

that is,

$$\mathbf{x} = c_1\mathbf{x}^{(1)}, \quad \text{where} \quad \mathbf{x}^{(1)} \triangleq \begin{bmatrix} 1 \\ -2 \\ 1 \\ 0 \end{bmatrix} \quad \text{and} \quad c_1 \text{ is an arbitrary constant.} \tag{1.23}$$

Theorem 1.14

(Solution of homogeneous systems): A homogeneous system $A\mathbf{x} = \mathbf{0}$ always has at least $\mathbf{x} = \mathbf{0}$ as a solution. A homogeneous system has either infinitely many solutions or $\mathbf{0}$ as the only solution, the so-called **trivial solution**.

Theorem 1.15

(Linearity principle). If A is an $m \times n$ matrix, $\mathbf{x}$ and $\mathbf{y}$ are any n-vectors, and α is any scalar, then

(a) $A(\mathbf{x} + \mathbf{y}) = (A\mathbf{x}) + (A\mathbf{y})$, and
(b) $A(\alpha\mathbf{x}) = \alpha(A\mathbf{x})$.

Definition 1.16

The **general linear combination** of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_\ell$ is

$$\mathbf{v} = c_1\mathbf{v}_1 + \cdots + c_\ell\mathbf{v}_\ell,$$

where $c_1, \ldots, c_\ell$ are arbitrary constants.

www.TechnicalBooksPDF.com

Definition 1.17

The **general solution** of a homogeneous system of equations $A\mathbf{x} = \mathbf{0}$ has the form $\mathbf{x} = c_1\mathbf{x}^{(1)} + \cdots + c_\ell\mathbf{x}^{(\ell)}$ if for every solution $\mathbf{x}^*$, there are values of scalars $c_1, \ldots, c_\ell$ giving $\mathbf{x}^* = c_1\mathbf{x}^{(1)} + \cdots + c_\ell\mathbf{x}^{(\ell)}$.

So, for example, the general solution of system (1.22) is given by (1.23). We see that a general solution of a homogeneous system has the form of a general linear combination of vectors. In Chapters 3, 5, and 11, we will see how this generalizes to solving linear ordinary differential equations and even linear partial differential equations.

Example 1.17

Solve

$$\left\{\begin{array}{rrrrr} x_1 & +x_2 & & -x_4 & = 0 \\ & & x_3 & +2x_4 & = 0 \end{array}\right\}. \tag{1.24}$$

Method: The corresponding augmented matrix, $\left[\begin{array}{ccccc|c} \textcircled{1} & 1 & 0 & -1 & 0 \\ 0 & 0 & \textcircled{1} & 2 & 0 \end{array}\right]$, is already in RREF and has its pivot positions circled. The basic variables are x_1, x_3, and the free variables x_2, x_4 can be replaced by arbitrary constants c_1, c_2, respectively. The solutions are

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -x_2 + x_4 \\ x_2 \\ -2x_4 \\ x_4 \end{bmatrix} = \begin{bmatrix} -c_1 + c_2 \\ c_1 \\ -2c_2 \\ c_2 \end{bmatrix} = \begin{bmatrix} -c_1 \\ c_1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} c_2 \\ 0 \\ -2c_2 \\ c_2 \end{bmatrix}$$

$$= c_1 \begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + c_2 \begin{bmatrix} 1 \\ 0 \\ -2 \\ 1 \end{bmatrix},$$

that is,

$$\mathbf{x} = c_1 \begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + c_2 \begin{bmatrix} 1 \\ 0 \\ -2 \\ 1 \end{bmatrix} \triangleq c_1\mathbf{x}^{(1)} + c_2\mathbf{x}^{(2)}, \tag{1.25}$$

where c_1, c_2 are arbitrary constants, which gives the general solution of system (1.24). ◯

Definition 1.18

Suppose that a set of vectors, $\mathcal{W}$, can be written as a general linear combination of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_\ell$, that is, $\mathcal{W} = \{\mathbf{v} : \mathbf{v} = c_1\mathbf{v}_1 + \cdots + c_\ell\mathbf{v}_\ell,$ arbitrary scalars $c_1, \ldots, c_\ell\}$. Then we say $\{\mathbf{v}_1, \ldots, \mathbf{v}_\ell\}$ is a **spanning set** for $\mathcal{W}$, and we write

$$\mathcal{W} = Span\{\mathbf{v}_1, \ldots, \mathbf{v}_\ell\}.$$

www.TechnicalBooksPDF.com

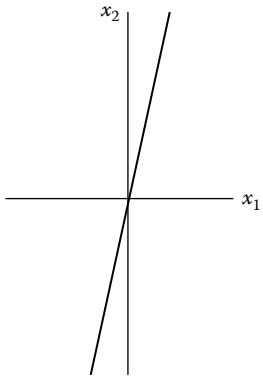


FIGURE 1.4
Line through the origin, spanned by one vector.

Shown in Figure 1.4 is a line spanned by a single vector in $\mathbb{R}^2$. Shown in Figure 1.5 is the plane

$$\mathcal{W} = Span\left\{\begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}\right\}$$

in $\mathbb{R}^3$ with many linear combinations, of various lengths, shown.

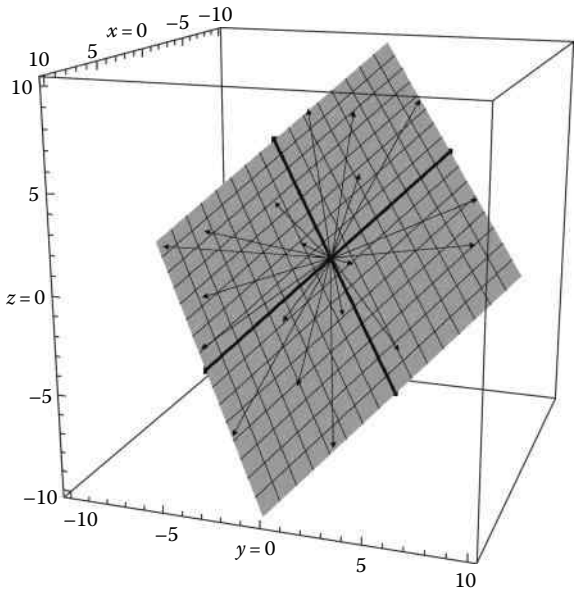


FIGURE 1.5
Plane spanned by two vectors.

www.TechnicalBooksPDF.com

Example 1.18

(Example 1.17 again) Find a spanning set for the solution set of system (1.24).

Method: In Example 1.17, we saw that all solutions of (1.24) can be written in the form (1.25). So,

$$\left\{\begin{bmatrix}-1\\1\\0\\0\end{bmatrix}, \begin{bmatrix}1\\0\\-2\\1\end{bmatrix}\right\}$$

is a spanning set for the solution set of system (1.24). ○

Example 1.19

(Example 1.16 again) Find a spanning set for the solution set of system (1.22).

Method: In Example 1.16, we saw that all solutions of (1.22) can be written in the form (1.23). So, a spanning set for the solution set of system (1.22) is given by

$$\left\{\begin{bmatrix}1\\-2\\1\\0\end{bmatrix}\right\}. \bigcirc$$

Definition 1.19

Suppose $\mathcal{W}$ is the set of all solutions of a homogeneous system $A\mathbf{x} = \mathbf{0}$. Row reduction of the augmented matrix $[A \mid \mathbf{0}]$ to its RREF decides which are the free variables. Suppose $\mathcal{W}$ has a spanning set $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\ell)}\}$ corresponding to those free variables. Then we call the set $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\ell)}\}$ a **complete set of basic solutions**, and we call each of $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\ell)}$ a **basic solution**.

So, the number of basic solutions we need to give a general solution of $A\mathbf{x} = \mathbf{0}$ is the same as the number of free variables for that system of equations.

Notice that in Example 1.17, the solutions are

$$\mathbf{x} = c_1 \begin{bmatrix}-1\\1\\0\\0\end{bmatrix} + c_2 \begin{bmatrix}1\\0\\-2\\1\end{bmatrix}.$$

If we let $c_1 = 1$ and $c_2 = 0$, we have the first basic solution, $\mathbf{x}^{(1)} = \begin{bmatrix}-1\\1\\0\\0\end{bmatrix}$.

If we let $c_1 = 0$ and $c_2 = 1$, we have the second basic solution, $\mathbf{x}^{(2)} = \begin{bmatrix}1\\0\\-2\\1\end{bmatrix}$.

www.TechnicalBooksPDF.com

We have a general principle that we will see over and over again in this book, in a variety of situations. This general principle will tie together many different mathematical techniques!

Theorem 1.16

(Linear superposition principle): If a homogeneous system $A\mathbf{x} = \mathbf{0}$ has solutions $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\ell)}$, then

$$\mathbf{x} = c_1\mathbf{x}^{(1)} + \cdots + c_\ell\mathbf{x}^{(\ell)} \tag{1.26}$$

also solves $A\mathbf{x} = \mathbf{0}$, for any values of scalars $c_1, \ldots, c_\ell$. Further, if those vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\ell)}$ are all of the basic solutions, then (a) $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\ell)}\}$ gives a spanning set for the set of solutions, and (b) for the set of solutions, there is no spanning set with fewer than ℓ vectors.

The explanation for Theorem 1.16 would follow from Theorem 1.41 in Section 1.7.

Definition 1.20

The **nullity** of a matrix A is the number of vectors needed to span the solution set of $A\mathbf{x} = \mathbf{0}$. We denote the nullity of A by $\nu(A)$.

We note that $\nu(A)$ is the number of free variables for the system $A\mathbf{x} = \mathbf{0}$.

Theorem 1.17

(Rank and nullity): If A is $m \times n$, then $n = rank(A) + \nu(A)$.

All this theorem says is that for the homogeneous system $A\mathbf{x} = \mathbf{0}$, the number of variables, n, equals the sum of the number of basic variables, $rank(A)$, plus the number of free variables, $\nu(A)$. Nevertheless, Theorem 1.17 is fundamental to the study of solving systems of linear algebraic equations.

Corollary 1.1

If A is $m \times n$, the homogeneous system $A\mathbf{x} = \mathbf{0}$ has a nontrivial solution for $\mathbf{x}$ if, and only if, $rank(A) < n$.

Example 1.20

Find the nullity of the matrix $A = \begin{bmatrix} 1 & 2 & -2 \\ -2 & -4 & 5 \\ 1 & 2 & -1 \end{bmatrix}$ and verify that the result of Theorem 1.17 holds true.

www.TechnicalBooksPDF.com

Method: Similar to our work on Example 1.4 in Section 1.1, we have

$$\begin{bmatrix} 1 & 2 & -2 & | & 0 \\ -2 & -4 & 5 & | & 0 \\ 1 & 2 & -1 & | & 0 \end{bmatrix} \underset{\substack{2R_1+R_2\to R_2 \\ -R_1+R_3\to R_3}}{\sim} \begin{bmatrix} 1 & 2 & -2 & | & 0 \\ 0 & 0 & 1 & | & 0 \\ 0 & 0 & 1 & | & 0 \end{bmatrix}$$

$$\underset{\substack{-R_2+R_3\to R_3 \\ 2R_2+R_1\to R_1}}{\sim} \begin{bmatrix} \textcircled{1} & 2 & 0 & | & 0 \\ 0 & 0 & \textcircled{1} & | & 0 \\ 0 & 0 & 0 & | & 0 \end{bmatrix}.$$

Because there are two basic variables, $rank(A) = 2$; because there is exactly one free variable, the nullity of A is $\nu(A) = 1$. The matrix A has three columns, so $n = 3$. Indeed, $n = 3 = 2 + 1 = rank(A) + \nu(A)$, in agreement with the conclusion of Theorem 1.17. ◯

In any specific example, we can discuss basic solutions of $A\mathbf{x} = \mathbf{0}$ and the nullity of A only after performing row reduction on $[\,A \mid \mathbf{0}\,]$ (or row reduction on A, because the last column of zeros in $[\,A \mid \mathbf{0}\,]$ never changes during the row reduction).

1.3.1 Problems

In problems 1–3, for each given matrix A, use the row reduction algorithm to find the general solution of the system $A\mathbf{x} = \mathbf{0}$.

1. $A = \begin{bmatrix} 1 & 3 & -1 & 1 \\ 2 & 4 & 0 & 3 \end{bmatrix}$.
2. $A = \begin{bmatrix} 1 & 2 & 0 & 3 \\ 0 & 1 & 1 & -1 \\ 1 & 0 & -2 & 5 \\ 0 & 1 & 1 & -1 \end{bmatrix}$.
3. $A = \begin{bmatrix} 0 & 0 & 1 & -1 \\ 1 & 2 & 3 & 1 \\ -1 & 3 & 2 & 4 \end{bmatrix}$.
4. Suppose A is an $m \times n$ matrix and $A\mathbf{x} = \mathbf{0}$ has a nontrivial solution. Which of the following *must* be true? For each one, explain why or why not.
   (a) $rank(A) < n$, (b) $rank(A) < m$, (c) $A\mathbf{x} = \mathbf{0}$ has infinitely many solutions, (d) $m \geq n$, and (e) $n \geq m$.
5. Suppose that (i) $\mathbf{w}$ is a linear combination of $\mathbf{v}_1, \mathbf{v}_2$; (ii) $\mathbf{v}_1$ is a linear combination of $\mathbf{u}_1, \mathbf{u}_2$; and (iii) $\mathbf{v}_2$ is a linear combination of $\mathbf{u}_1, \mathbf{u}_2$. Why must $\mathbf{w}$ be a linear combination of $\mathbf{u}_1, \mathbf{u}_2$?
6. Assume A is an $m \times n$ matrix and $\mathbf{x}$ is a vector that satisfies $A\mathbf{x} = \mathbf{0}$ and $\mathbf{x} \neq \mathbf{0}$.
   For each of (a)–(e), decide whether it must be true, must be false, or may be true and may be false.
   (a) $rank(A) < n$.
   (b) $rank(A) < m$.
   (c) $A\mathbf{x} = \mathbf{0}$ has infinitely many solutions.

www.TechnicalBooksPDF.com

(d) $m > n$.

(e) $n > m$.

7. (a) Write down an example of a 3×3 matrix A whose rank is exactly two, is in RREF, and is *not* a diagonal matrix.

   (b) For the matrix A you wrote in part (a), let $B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - A$. Find all solutions of the system $B\mathbf{x} = \mathbf{0}$ *and* find all basic solutions of that system.

8. (a) Write down a 4×5 matrix A that satisfies all of the following properties: (i) A is in RREF, (ii) A has exactly three pivot positions, and (iii) at least two entries of A are 2 and at least two entries of A are -1.

   (b) For the matrix A you wrote down in part (a), find the rank and nullity of A.

1.4 Solutions of Nonhomogeneous Systems

A non homogeneous system has the form

$$A\mathbf{x} = \mathbf{b}, \tag{1.27}$$

where $\mathbf{b} \neq \mathbf{0}$.

Definition 1.21

$\mathbf{x}_p$ is a **particular solution** of (1.27) if $A\mathbf{x}_p = \mathbf{b}$.

Example 1.21

(Example 1.3 in Section 1.1 again) $\mathbf{x}_p = \begin{bmatrix} 4/3 \\ 4/3 \\ 0 \end{bmatrix}$ is a particular solution of

$$\begin{bmatrix} 1 & -1 & 1 \\ -2 & 2 & -1 \\ 0 & 3 & 2 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 0 \\ 0 \\ 4 \end{bmatrix}. \bigcirc$$

Example 1.22

(Example 1.4 in Section 1.1 again) $\mathbf{x}_p = \begin{bmatrix} 4 \\ 0 \\ 5 \end{bmatrix}$ is a particular solution of

$$\begin{bmatrix} 1 & 2 & -2 \\ -2 & -4 & 5 \\ 1 & 2 & -1 \end{bmatrix} \mathbf{x} = \begin{bmatrix} -6 \\ 17 \\ -1 \end{bmatrix}. \bigcirc \tag{1.28}$$

www.TechnicalBooksPDF.com

Recall from Theorem 1.15(a) in Section 1.3 the linearity principle that $A(\mathbf{x}+\mathbf{y}) = (A\mathbf{x}) + (A\mathbf{y})$. Suppose $\mathbf{x}_p$ is a solution of a nonhomogeneous system (1.27) and $\mathbf{x}_h$ satisfies the corresponding homogeneous system $A\mathbf{x}=\mathbf{0}$, that is, $A\mathbf{x}_h = \mathbf{0}$. By "corresponding," we mean that the matrix A is the same. We get

$$A(\mathbf{x}_p + \mathbf{x}_h) = (A\mathbf{x}_p) + (A\mathbf{x}_h) = \mathbf{b} + \mathbf{0} = \mathbf{b},$$

so $\mathbf{x} \triangleq \mathbf{x}_p + \mathbf{x}_h$ is a solution of the nonhomogeneous system $A\mathbf{x} = \mathbf{b}$.

Recall from Definition 1.17 in Section 1.3 that $\mathbf{x}$ is a "general solution" for the system $A\mathbf{x}=\mathbf{0}$ if $\mathbf{x} = c_1\mathbf{x}^{(1)} + \cdots + c_\ell\mathbf{x}^{(\ell)}$ has the property that for every $\mathbf{x}^*$ that solves $A\mathbf{x}=\mathbf{0}$, there are values for $c_1, \ldots, c_\ell$ for which $\mathbf{x}^* = c_1\mathbf{x}^{(1)} + \cdots + c_\ell\mathbf{x}^{(\ell)}$. For nonhomogeneous systems, similar to this, we have

Definition 1.22

$\mathbf{x} = \mathbf{x}_p + c_1\mathbf{x}^{(1)} + \cdots + c_\ell\mathbf{x}^{(\ell)}$ is a **general solution** for $A\mathbf{x} = \mathbf{b}$ if for every $\mathbf{x}^*$ that solves (1.27), there are values for $c_1, \ldots, c_\ell$ for which $\mathbf{x}^* = \mathbf{x}_p + c_1\mathbf{x}^{(1)} + \cdots + c_\ell\mathbf{x}^{(\ell)}$.

Theorem 1.18

If $\mathbf{x}_p$ is any particular solution of (1.27) and $\mathbf{x}_h = c_1\mathbf{x}^{(1)} + \cdots + c_\ell\mathbf{x}^{(\ell)}$ is a general solution of the corresponding homogeneous system, then

$$\mathbf{x} = \mathbf{x}_p + \mathbf{x}_h = \mathbf{x}_p + c_1\mathbf{x}^{(1)} + \cdots + c_\ell\mathbf{x}^{(\ell)} \tag{1.29}$$

is a general solution of (1.27).

Example 1.23

(Example 1.22 again) $\mathbf{x}_p = \begin{bmatrix} 4 \\ 0 \\ 5 \end{bmatrix}$ is a particular solution of system (1.28), and the corresponding homogeneous system, $\begin{bmatrix} 1 & 2 & -2 \\ -2 & -4 & 5 \\ 1 & 2 & -1 \end{bmatrix}\mathbf{x} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$, has general solution $\mathbf{x}_h = c_1\begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}$, as we can see from the row reduction done in Example 1.20 in Section 1.3. The general solution of system (1.28) is

$$\mathbf{x} = \begin{bmatrix} 4 \\ 0 \\ 5 \end{bmatrix} + c_1\begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}, \tag{1.30}$$

where c_1 is an arbitrary constant. ◯

Figure 1.6 shows the line of solutions $\mathbf{x}^* = \mathbf{x}_p + c_1\mathbf{x}^{(1)} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$ for Example 1.23.

www.TechnicalBooksPDF.com

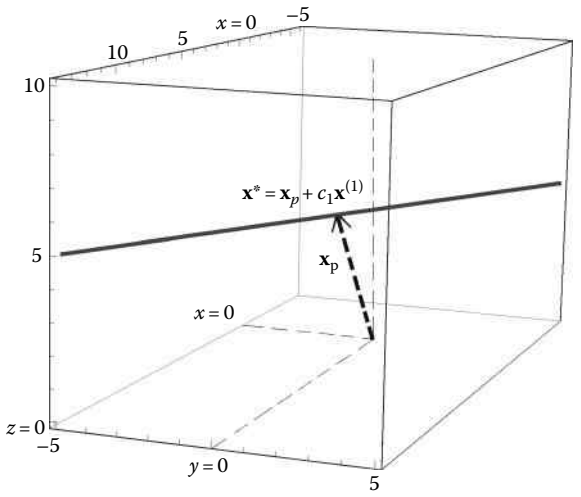


FIGURE 1.6
Line solving nonhomogeneous system.

Note that (1.30) agrees with our conclusion for Example 1.22, as well as for Example 1.4 in Section 1.1. In fact, while Theorem 1.18 announces a great principle, the *easiest* way to solve system (1.28) is, instead, to do the row reduction we did for Example 1.4 in Section 1.1:

Example 1.24

(Example 1.4 in Section 1.1 again) Solve system (1.28) *and* write the solution in the form (1.29).

Method: To summarize the work done for Example 1.4 in Section 1.1,

$$\left[\begin{array}{rrr|r} 1 & 2 & -2 & -6 \\ -2 & -4 & 5 & 17 \\ 1 & 2 & -1 & -1 \end{array}\right] \underset{\substack{2R_1+R_2\to R_2 \\ -R_1+R_3\to R_3 \\ -R_2+R_3\to R_3 \\ 2R_2+R_1\to R_1}}{\sim} \left[\begin{array}{rrr|r} \textcircled{1} & 2 & 0 & 4 \\ 0 & 0 & \textcircled{1} & 5 \\ 0 & 0 & 0 & 0 \end{array}\right],$$

which gives solutions

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 4-2c_1 \\ c_1 \\ 5 \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \\ 5 \end{bmatrix} + c_1 \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix} = \mathbf{x}_p + c_1\mathbf{x}^{(1)},$$

where c_1 is an arbitrary constant. ◯

So, when solving systems of linear algebraic equations, we can solve for both $\mathbf{x}_p$ and $\mathbf{x}_h$ at the same time.

We will see that there is also a nonhomogeneous principle for solving ordinary differential equations. In fact, when solving ordinary differential equations, it *will* make sense to

www.TechnicalBooksPDF.com

solve separately for $\mathbf{x}_p$ and $\mathbf{x}_h$ and then add them together to have the general solution, as Theorem 1.18 suggests. So, while the *structure* of solutions of linear algebraic equations is the same as the *structure* of solutions of differential equations, the practical aspects of how you go about solving for the solutions is simpler for systems of linear algebraic equations.

Theorem 1.19

A nonhomogeneous system that has at least one solution has more than one solution if, and only if, the corresponding homogeneous system has a nontrivial solution.

1.4.1 Problems

In problems 1–3, solve the system and express the solutions in the form $\mathbf{x} = \mathbf{x}_p + \mathbf{x}_h$. Clearly label what is $\mathbf{x}_p$ and what is $\mathbf{x}_h$.

1. $\left\{\begin{array}{rrrr} x_1 & & +2x_3 & = -1 \\ x_1 & +x_2 & +4x_3 & = 1 \\ -x_1 & & -2x_3 & = 1 \end{array}\right\}$
2. $\left\{\begin{array}{rrrr} x_1 & +x_2 & +2x_3 & = 5 \\ & -x_2 & -3x_3 & = -4 \\ -2x_1 & +2x_2 & +8x_3 & = 6 \end{array}\right\}$
3. $\left\{\begin{array}{rrrrr} x_1 & +2x_2 & & +3x_4 & = 4 \\ & x_2 & +x_3 & -x_4 & = -5 \\ x_1 & & -2x_3 & +5x_4 & = -6 \\ & x_2 & +x_3 & -x_4 & = 5 \end{array}\right\}$
4. Suppose the system $A\mathbf{x} = \mathbf{b}^{(1)}$ has a solution $\mathbf{x}_{p,1}$, the system $A\mathbf{x} = \mathbf{b}^{(2)}$ has a solution $\mathbf{x}_{p,2}$, and the system $A\mathbf{x} = \mathbf{0}$ has general solution $\mathbf{x}_h$. Find the general solution of the system $A\mathbf{x} = \mathbf{b}^{(1)} + \mathbf{b}^{(2)}$.
5. Suppose the system $A\mathbf{x} = \begin{bmatrix} 2 \\ 5 \\ 8 \\ 0 \end{bmatrix}$ has general solution $\mathbf{x} = \begin{bmatrix} 4 + c_1 - c_2 \\ 1 - c_1 - c_2 \\ c_1 \\ c_2 \end{bmatrix}$, where c_1, c_2 are arbitrary constants.
   (a) Write the general solution in the form $\mathbf{x} = \mathbf{x}_p + \mathbf{x}_h$. Clearly label what is $\mathbf{x}_p$ and what is $\mathbf{x}_h$.
   (b) Use your result from part (a) to find the general solution of the system $A\mathbf{x} = \mathbf{0}$.
   (c) Use your result from part (b) to find *all* of the basic solutions of the system $A\mathbf{x} = \mathbf{0}$.
6. Assume A is an $m \times n$ matrix. Which of the following *possibly* may be true?
   (a) The equation $A\mathbf{x} = \mathbf{b}$ has a solution if the augmented matrix $[\,A \mid \mathbf{b}\,]$ has a pivot position in each row.
   (b) The equation $A\mathbf{x} = \mathbf{b}$ has a solution for all $\mathbf{b}$ in $\mathbb{R}^m$, if the matrix A has a pivot position in each row.

www.TechnicalBooksPDF.com

(c) Every linear combination of vectors in $\mathbb{R}^m$ can be written in the form $A\mathbf{x}$ for some matrix A and vector $\mathbf{x}$ in $\mathbb{R}^n$.

(d) The equation $A\mathbf{x} = \mathbf{b}$ can have a solution for all $\mathbf{b}$ in $\mathbb{R}^m$ and still the columns of A do not span $\mathbb{R}^m$.

7. [The Fredholm alternative] Suppose a nonhomogeneous problem $A\mathbf{x} = \mathbf{b}$ has a solution $\mathbf{x}$. Let $\mathbf{z}$ be any solution of the homogeneous system $A^T\mathbf{z} = \mathbf{0}$. Explain why $\mathbf{b}^T\mathbf{z} = 0$. [Hint: First, take the transpose of the equation $A\mathbf{x} = \mathbf{b}$.]

8. (a) Write down a 3×4 matrix A that is in RREF, has rank equal to two, and has at least two entries of "2" and at least one entry of "-3."

   (b) For the matrix that you wrote down in part (a), find all basic solutions of $A\mathbf{x} = \mathbf{0}$.

   (c) For the matrix that you wrote down in part (a), (i) find a vector $\mathbf{b}$ for which $A\mathbf{x} = \mathbf{b}$ has at least one solution, and (ii) find all solutions of $A\mathbf{x} = \mathbf{b}$.

9. Suppose A is a 6×5 matrix. Is it possible that for *all* vectors $\mathbf{b}$ in $\mathbb{R}^6$ the system $A\mathbf{x} = \mathbf{b}$ has at least one solution for $\mathbf{x}$? Why, or why not?

10. For the system

$$\left\{\begin{array}{rcrcrl} x_1 & + & x_2 & + & x_3 & = 0 \\ x_1 & - & x_2 & + & x_3 & = 6 \\ x_1 & & & + & 2x_3 & = 0 \end{array}\right\}$$

find the general solution as well as the general solution of the corresponding homogeneous system. Label which is which.

1.5 Inverse Matrix

We will see that the concept of an "inverse matrix" unifies many results concerning systems of equations.

Example 1.25

For $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ and $B = \begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix}$, we have $AB = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = BA$. ◯

Example 1.26

For $A = I_n$ and $B = I_n$, we have $AB = I_n = BA$. ◯

The following definition works *only* for square matrices.

Definition 1.23

If A is $n \times n$ and there is an $n \times n$ matrix B with

$$AB = I_n = BA, \tag{1.31}$$

www.TechnicalBooksPDF.com

then we say A is **invertible**, B is the **inverse of** A, and we write $B = A^{-1}$. We may also say "A is **non-singular**" or "A^{-1} exists." If a matrix does not have an inverse, we say the matrix is **singular**, also known as **non-invertible**.

So, in each of Examples 1.25 and 1.26, the given matrix A is invertible and the given $B = A^{-1}$.

Theorem 1.20

The inverse of A is unique, that is, there can be at most one distinct matrix B with $AB = I = BA$.

Why? Suppose B and C are both inverses of A, that is, $AB = I = BA$ and $AC = I = CA$. Then $CAB = (CA)B = (I)B = B$, and also $CAB = C(AB) = C(I) = C$, so $B = C$, that is, B and C are not distinct. □

So, we don't have to look for more than one inverse of the same matrix. Further, if two different numerical methods produce two distinct approximations of the inverse of a given matrix, then one or both of the approximations are incorrect.

I'm all in favor of working less, even if it sometimes requires thinking more. The next result does this.

Theorem 1.21

If A and B are $n \times n$ and *either* $AB = I_n$ *or* $BA = I_n$, then A^{-1} exists and equals B.

So, we don't have to check both $AB = I$ and $BA = I$. In fact, Theorem 1.26, as follows, will develop an algorithm that constructs B with $AB = I$, if possible.

The next result is among the most simple but useful results of mathematics for engineers.

Theorem 1.22

If 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ and $ad - bc \neq 0$, then A is invertible and

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}. \tag{1.32}$$

As we will see in Section 1.6, specifically Theorem 1.30, if A is 2×2 and $ad - bc = 0$, then A^{-1} does not exist. Further, in Section 1.6, we will see that the criterion $ad - bc \neq 0$ is generalized using the concept of the "determinant of a matrix."

www.TechnicalBooksPDF.com

Example 1.27

(Example 1.25 again) Find the inverse of the matrix $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$.

Method: A has $ad - bc = (1)(4) - (2)(3) = -2 \neq 0$, so there exists

$$A^{-1} = \frac{1}{-2}\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix}. \bigcirc$$

Example 1.28

Use the definition to explain why the matrix $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$ is not invertible.

Method: If there were a matrix $B = \begin{bmatrix} w & x \\ y & z \end{bmatrix}$ with $AB = I_2$, then

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = AB = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}\begin{bmatrix} w & x \\ y & z \end{bmatrix} = \begin{bmatrix} w + 2y & x + 2z \\ 2w + 4y & 2x + 4z \end{bmatrix};$$

hence, $1 = w + 2y$, $0 = 2w + 4y$, $0 = x + 2z$, and $1 = 2x + 4z$. Substituting the first of the four equations into the second gives $0 = 2w + 4y = 2(w + 2y) = 2(1) = 2$, giving a contradiction. So, A has no inverse. $\bigcirc$

In the notation A^{-1}, the $^{-1}$ indicates the operation of "taking the inverse of," the same notation as for "inverse function" that you saw in algebra and calculus courses. In fact, we have

$$A\mathbf{x} = \mathbf{b} \iff \mathbf{x} = A^{-1}\mathbf{b}. \tag{1.33}$$

This notation means both "if $A\mathbf{x} = \mathbf{b}$, then $\mathbf{x} = A^{-1}\mathbf{b}$" *and* "if $\mathbf{x} = A^{-1}\mathbf{b}$, then $A\mathbf{x} = \mathbf{b}$," that is, that $A\mathbf{x} = \mathbf{b}$ and $\mathbf{x} = A^{-1}\mathbf{b}$ are logically equivalent.

Note also that if $B = A^{-1}$, then $BA = I = AB$, so A satisfies the requirements in the definition to be B^{-1}. In fact,

$$B = A^{-1} \iff A = B^{-1}. \tag{1.34}$$

So, we often say "A and B are inverses of each other" rather than "$B = A^{-1}$" or "$A = B^{-1}$."

We have some general results:

Theorem 1.23

If A and C are invertible and of the same size, then

(a) $(A^{-1})^{-1} = A$.
(b) $(A^T)^{-1} = (A^{-1})^T$.
(c) $(AC)^{-1} = C^{-1}A^{-1}$.

Why? For (a), we have just seen in (1.34) that $A^{-1} = B$ yields $B^{-1} = A$; hence, $(A^{-1})^{-1} = (B)^{-1} = A$.

www.TechnicalBooksPDF.com

For (b), let $C = A^T$. Because $I = AA^{-1}$, taking the transpose of both sides and using Theorem 1.13 in Section 1.2 implies $I = I^T = (AA^{-1})^T = (A^{-1})^T A^T = (A^{-1})^T C$. It follows from Theorem 1.21 that $(A^{-1})^T = C^{-1}$, that is, that $(A^{-1})^T = (A^T)^{-1}$. □

For (c), explain this yourself in Problem 1.5.3.17.

Theorem 1.24

If $n \times n$ matrix A^{-1} exists, then for *any* n-vector $\mathbf{b}$, the system $A\mathbf{x} = \mathbf{b}$ has exactly one solution, namely, $\mathbf{x} = A^{-1}\mathbf{b}$.

Why? $\mathbf{b} = A\mathbf{x} \iff A^{-1}\mathbf{b} = A^{-1}(A\mathbf{x}) = (A^{-1}A)\mathbf{x} = (I)\mathbf{x} = \mathbf{x}$. □

Theorem 1.25

Suppose A is $n \times n$. Then the following are logically equivalent statements:

(a) A^{-1} exists.
(b) $rank(A) = n$.
(c) $\text{RREF}(A) = I_n$.
(d) The only solution of $A\mathbf{x} = \mathbf{0}$ is $\mathbf{x} = \mathbf{0}$.
(e) For every $\mathbf{b}$, there is exactly one solution of $A\mathbf{x} = \mathbf{b}$.

The reasons for these results will become apparent when we study the algorithm as follows for finding A^{-1}, if it exists.

1.5.1 Row Reduction Algorithm for Constructing the Inverse

We will see that to find a matrix B with $AB = I_n$, row reduce $[A \mid I_n] \sim [I_n \mid B]$, *if* A is invertible. This method is usually called the Gauss–Jordan method for finding A^{-1}. Here's why the method works: recall Theorem 1.9 in Section 1.2, that is, that $AB = \begin{bmatrix} A\mathbf{B}_{*1} & A\mathbf{B}_{*2} & \cdots & A\mathbf{B}_{*n} \end{bmatrix}$. Next, let's define the columns of I_n by $\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \ldots, \mathbf{e}^{(n)}$. So, solving $AB = I_n$ is the same as solving $[\, A\mathbf{B}_{*1} \quad A\mathbf{B}_{*2} \quad \cdots \quad A\mathbf{B}_{*n} \,] = \begin{bmatrix} \mathbf{e}^{(1)} & \mathbf{e}^{(2)} & \cdots & \mathbf{e}^{(n)} \end{bmatrix}$, that is, solving *all* of the systems

$$A\mathbf{B}_{*1} = \mathbf{e}^{(1)}, \quad A\mathbf{B}_{*2} = \mathbf{e}^{(2)}, \ldots, \quad A\mathbf{B}_{*n} = \mathbf{e}^{(n)}$$

for the columns of B. To solve for the columns of B, that is, $\mathbf{B}_{*1}, \ldots, \mathbf{B}_{*n}$, we could do row reduction individually on each of the corresponding augmented matrices $\left[A \mid \mathbf{e}^{(1)} \right], \ldots, \left[A \mid \mathbf{e}^{(n)} \right]$. Instead, we can row reduce the "mega"- augmented matrix $\left[A \mid \mathbf{e}^{(1)} \quad \cdots \quad \mathbf{e}^{(n)} \right]$:

www.TechnicalBooksPDF.com

$$[\,A \mid I_n\,] = \left[\, A \mid \mathbf{e}^{(1)} \quad \cdots \quad \mathbf{e}^{(n)} \,\right] \sim \cdots \sim \left[\, \mathrm{RREF}(A) \mid \mathbf{f}^{(1)} \quad \cdots \quad \mathbf{f}^{(n)} \,\right].$$

We can do row reduction instead on the "mega"- augmented matrix because the elementary row operations that reduce A to $\mathrm{RREF}(A)$ work just as well on the augmented matrices $\left[\, A \mid \mathbf{e}^{(j)} \,\right], j = 1, 2, \ldots$. They "work just as well" because all elementary row operations can be implemented by multiplication on the left by elementary matrices, as we mentioned at the end of Section 1.2.

Now, if $\mathrm{RREF}(A) = I_n$, then we have

$$\left[\, A \mid \mathbf{e}^{(1)} \quad \cdots \quad \mathbf{e}^{(n)} \,\right] \sim \cdots \sim \left[\, I_n \mid \mathbf{f}^{(1)} \quad \cdots \quad \mathbf{f}^{(n)} \,\right];$$

hence, we can immediately read off the solutions $\mathbf{B}_{*1} = \mathbf{f}^{(1)}$, $\mathbf{B}_{*2} = \mathbf{f}^{(2)}, \ldots, \mathbf{B}_{*n} = \mathbf{f}^{(n)}$. This gives $A^{-1} = B = \left[\, \mathbf{f}^{(1)} \quad \cdots \quad \mathbf{f}^{(n)} \,\right]$.

On the other hand, if $\mathrm{RREF}(A) \neq I_n$, then because of property (1.5) in Section 1.1, $\mathrm{RREF}(A)$ has at least its bottom row being an all zero row. But then, there are two cases concerning the bottom row of the matrix $\left[\, \mathrm{RREF}(A) \mid \mathbf{f}^{(1)} \quad \cdots \quad \mathbf{f}^{(n)} \right]$:

- If that bottom row is zero, then either there is no inverse matrix B or there are infinitely many choices for inverse matrix B, but the latter is impossible by Theorem 1.20.
- If that bottom row is nonzero, then by Theorem 1.2 in Section 1.1, at least one of the equations $A\mathbf{B}_{*1} = \mathbf{e}^{(1)}$, $A\mathbf{B}_{*2} = \mathbf{e}^{(2)}, \ldots, A\mathbf{B}_{*n} = \mathbf{e}^{(n)}$ cannot be solved for the corresponding column of the matrix B; hence, there is no matrix B satisfying $AB = I_n$.

In either case, A is not invertible.

The preceding text explains why we have the following:

Theorem 1.26

(a) If $\mathrm{RREF}(A) = I_n$ and $[\,A \mid I_n\,] \sim [\,I_n \mid B\,]$, then A^{-1} exists and equals B.

(b) If $\mathrm{RREF}(A) \neq I_n$, then A^{-1} does not exist.

(c) If at any point during the Gauss–Jordan row reduction process, there is a row of the form $[\,0 \cdots 0 \mid \alpha_1 \cdots \alpha_n\,]$ with at least one of the α_js being nonzero, then we can cut short the row reduction process and conclude that A^{-1} does not exist.

Example 1.29

(Example 1.25 again) For $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, use the Gauss–Jordan row reduction process to determine whether A is invertible, and if it is, find A^{-1}.

www.TechnicalBooksPDF.com

Method:

$$\left[\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 3 & 4 & 0 & 1 \end{array}\right] \underset{-3R_1+R_2\to R_2}{\sim} \left[\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 0 & -2 & -3 & 1 \end{array}\right]$$

$$\underset{\substack{R_2+R_1\to R_1 \\ -\frac{1}{2}R_2\to R_2}}{\sim} \left[\begin{array}{cc|cc} \textcircled{1} & 0 & -2 & 1 \\ 0 & \textcircled{1} & \frac{3}{2} & -\frac{1}{2} \end{array}\right].$$

So, by Theorem 1.26(a), there exists $A^{-1} = \begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix}$. ○

Example 1.30

For $A = \begin{bmatrix} 1 & -1 & -2 \\ 2 & -3 & -5 \\ -1 & 3 & 5 \end{bmatrix}$, use the Gauss–Jordan row reduction process to determine whether A is invertible, and if it is, find A^{-1}.

Method:

$$\left[\begin{array}{ccc|ccc} 1 & -1 & -2 & 1 & 0 & 0 \\ 2 & -3 & -5 & 0 & 1 & 0 \\ -1 & 3 & 5 & 0 & 0 & 1 \end{array}\right] \underset{\substack{-2R_1+R_2\to R_2 \\ R_1+R_3\to R_3}}{\sim} \left[\begin{array}{ccc|ccc} 1 & -1 & -2 & 1 & 0 & 0 \\ 0 & -1 & -1 & -2 & 1 & 0 \\ 0 & 2 & 3 & 1 & 0 & 1 \end{array}\right]$$

$$\underset{\substack{2R_2+R_3\to R_3 \\ -R_2+R_1\to R_1 \\ -R_2\to R_2}}{\sim} \left[\begin{array}{ccc|ccc} 1 & 0 & -1 & 3 & -1 & 0 \\ 0 & 1 & 1 & 2 & -1 & 0 \\ 0 & 0 & 1 & -3 & 2 & 1 \end{array}\right]$$

$$\underset{\substack{-R_3+R_2\to R_2 \\ R_3+R_1\to R_1}}{\sim} \left[\begin{array}{ccc|ccc} \textcircled{1} & 0 & 0 & 0 & 1 & 1 \\ 0 & \textcircled{1} & 0 & 5 & -3 & -1 \\ 0 & 0 & \textcircled{1} & -3 & 2 & 1 \end{array}\right]$$

So, by Theorem 1.26(a), there exists $A^{-1} = \begin{bmatrix} 0 & 1 & 1 \\ 5 & -3 & -1 \\ -3 & 2 & 1 \end{bmatrix}$. ○

Example 1.31

For $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 1 & -1 & -3 \end{bmatrix}$, use the Gauss–Jordan row reduction process to determine whether A is invertible, and if it is, find A^{-1}.

www.TechnicalBooksPDF.com

Method:

$$\left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 4 & 5 & 6 & 0 & 1 & 0 \\ 1 & -1 & -3 & 0 & 0 & 1 \end{array}\right] \underset{\substack{-4R_1+R_2\to R_2 \\ -R_1+R_3\to R_3}}{\sim} \left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & -3 & -6 & -4 & 1 & 0 \\ 0 & -3 & -6 & -1 & 0 & 1 \end{array}\right]$$

$$\underset{-R_2+R_3\to R_3}{\sim} \left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & -3 & -6 & -4 & 1 & 0 \\ 0 & 0 & 0 & 3 & -1 & 1 \end{array}\right].$$

The bottom row is nonzero and has the form $[0\ 0\ 0|\alpha_1\ \alpha_2\ \alpha_3]$, so Theorem 1.26(c) tells us that A^{-1} does not exist. ◯

1.5.2 Inverse of a Partitioned Matrix

In Section 1.2, we partitioned a matrix into its columns or rows. More generally, we can partition a matrix into "blocks of rows" or "blocks or columns"; for example, an $m \times n$ matrix A can be partitioned into two blocks containing its first r columns and its last $n - r$ columns:

$$A = \left[\begin{array}{ccc|ccc} \mathbf{A}_{*1} & \cdots & \mathbf{A}_{*r} & \mathbf{A}_{*r+1} & \cdots & \mathbf{A}_{*n} \end{array}\right]$$

Even more generally, we could **partition a matrix into blocks**, each of which holds part of the rows and part of the columns. For example,

$$A = \left[\begin{array}{cc|ccc} 2 & 0 & 1 & 2 & -4 \\ -1 & 3 & 0 & 1 & 5 \\ \hline 0 & 0 & 0 & 0 & 3 \\ 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 4 & -2 & 3 \end{array}\right] \triangleq \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline O & A_{22} \end{array}\right],$$

where A_{11} is 2×2, A_{12} is 2×3, O is the 3×2 zero matrix, and A_{22} is 3×3.

We can define multiplication of block matrices:

$$\left[\begin{array}{c|c} C_{11} & C_{12} \\ \hline C_{21} & C_{22} \\ \hline C_{31} & C_{32} \end{array}\right] \left[\begin{array}{c|c} E_{11} & E_{12} \\ \hline E_{21} & E_{22} \end{array}\right] \triangleq \left[\begin{array}{c|c} C_{11}E_{11} + C_{12}E_{21} & C_{11}E_{12} + C_{12}E_{22} \\ \hline C_{21}E_{11} + C_{22}E_{21} & C_{21}E_{12} + C_{22}E_{22} \\ \hline C_{31}E_{11} + C_{32}E_{21} & C_{31}E_{12} + C_{32}E_{22} \end{array}\right],$$

where the C_{ik}, E_{kj} are themselves matrices, assuming all of the products of blocks exist.

www.TechnicalBooksPDF.com

Example 1.32

Suppose a matrix A can be written in block form $\left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline O & I \end{array}\right]$, where A_{11} is square and invertible. Explain why A is invertible and find its inverse.

Method: We'll look for a matrix B in block form $B = \left[\begin{array}{c|c} B_{11} & B_{12} \\ \hline B_{21} & B_{22} \end{array}\right]$ that we want to satisfy $AB = I$. We calculate

$$\left[\begin{array}{c|c} I & O \\ \hline O & I \end{array}\right] = I \stackrel{?}{=} AB = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline O & I \end{array}\right]\left[\begin{array}{c|c} B_{11} & B_{12} \\ \hline B_{21} & B_{22} \end{array}\right]$$

$$= \left[\begin{array}{c|c} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ \hline O \cdot B_{11} + I \cdot B_{21} & O \cdot B_{12} + I \cdot B_{22} \end{array}\right] = \left[\begin{array}{c|c} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ \hline B_{21} & B_{22} \end{array}\right]; \tag{1.35}$$

hence, we need $O = B_{21}$ and $I = B_{22}$. Substitute those into (1.35) to get

$$\left[\begin{array}{c|c} I & O \\ \hline O & I \end{array}\right] = I_n \stackrel{?}{=} AB = \left[\begin{array}{c|c} A_{11}B_{11} & A_{11}B_{12} + A_{12} \\ \hline O & I \end{array}\right].$$

So, we need $I = A_{11}B_{11}$ and $O = A_{11}B_{12} + A_{12}$. The former is solvable because we assumed A_{11} is invertible: $B_{11} = A_{11}^{-1}$. The latter is solvable by the same reason: $O = A_{11}B_{12} + A_{12} \iff -A_{12} = A_{11}B_{12} \iff -A_{11}^{-1}A_{12} = A_{11}^{-1}A_{11}B_{12} = B_{12}$.

So, we conclude that A is invertible and A^{-1} has the partitioned form

$$A^{-1} = \left[\begin{array}{c|c} A_{11}^{-1} & -A_{11}^{-1}A_{12} \\ \hline O & I \end{array}\right]. \bigcirc$$

Note that we didn't specify the sizes of the matrix A or its blocks in Example 1.32. This was because we wanted to emphasize that our explanations did not need us to work with individual entries of the matrices, so we didn't need to know how many of them there were.

1.5.3 Problems

In problems 1–4, for each matrix, use the row reduction algorithm to either find the inverse matrix or explain why the inverse does not exist.

1. $\begin{bmatrix} 1 & 4 \\ -2 & 3 \end{bmatrix}$
2. $\begin{bmatrix} 1 & 1 & -1 \\ 2 & 0 & -1 \\ 3 & 0 & 2 \end{bmatrix}$

www.TechnicalBooksPDF.com

3. $\begin{bmatrix} 1 & -2 & 2 \\ -2 & 0 & -1 \\ 2 & -1 & 0 \end{bmatrix}$
4. $\begin{bmatrix} 1 & 1 & -1 \\ 3 & 0 & 2 \\ -2 & -1 & 0 \end{bmatrix}$
5. Given that

$$\left[\begin{array}{ccc|ccc} 1 & 0 & -1 & 1 & 0 & 0 \\ 2 & -2 & 1 & 0 & 1 & 0 \\ -3 & -2 & 1 & 0 & 0 & 1 \end{array}\right]$$

is row equivalent to

$$\left[\begin{array}{ccc|ccc} 1 & 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & -1.5 & 1 & -0.5 & 0 \\ 0 & 0 & 2.5 & -\frac{5}{2} & 0.5 & -0.5 \end{array}\right],$$

use the row reduction algorithm to find

$$\begin{bmatrix} 1 & 0 & -1 \\ 2 & -2 & 1 \\ -3 & -2 & 1 \end{bmatrix}^{-1}.$$

6. If $(A^T)^{-1} = \begin{bmatrix} 1 & 2 \\ -3 & 4 \end{bmatrix}$, find A^{-1}, A, and A^2.
7. If A is $n \times n$ and invertible and $B = A^{-1}$, find a formula for $(A^T B)^{-1}$ in terms of A and B^T.
8. Find an example of 2×2 matrices A, B such that A, B, and $A + B$ are all invertible but $(A + B)^{-1} \neq A^{-1} + B^{-1}$.
9. If A is upper triangular and A^{-1} exists, must A^{-1} also be upper triangular? If so, why? If not, give a specific counterexample.

In each of the problems 10–13 you are given information about $A\mathbf{y}^{(i)}$'s, which you will use to find a formula for A^{-1}. Recall that $\mathbf{e}^{(1)}, \mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(n)}$ are the columns of the $n \times n$ identity matrix I_n. [Hint: Try to find, for example, a vector $\mathbf{x}$ satisfying $A\mathbf{x} = \mathbf{e}^{(3)}$. Eventually, you will use (1.20), in Section 1.2 which is part of Theorem 1.9 in Section 1.2.]

10. If A is a 4×4 matrix that satisfies

$$A\mathbf{y}^{(1)} = \mathbf{e}^{(1)},\ A\mathbf{y}^{(2)} = -\mathbf{e}^{(3)},\ A\mathbf{y}^{(3)} = \mathbf{e}^{(4)},\ \text{and}\ A\mathbf{y}^{(4)} = -\mathbf{e}^{(2)}$$

for some vectors $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(4)}$, find a formula for A^{-1} in terms of the $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(4)}$.

www.TechnicalBooksPDF.com

11. If A is a 4×4 matrix that satisfies

$$A\mathbf{y}^{(1)} = \mathbf{e}^{(1)},\ A\mathbf{y}^{(2)} = -\mathbf{e}^{(3)},$$

$$A\mathbf{y}^{(3)} = \frac{1}{2}\left(\mathbf{e}^{(2)} + \mathbf{e}^{(4)}\right), \text{ and } A\mathbf{y}^{(4)} = \frac{1}{2}\left(\mathbf{e}^{(2)} - \mathbf{e}^{(4)}\right)$$

for some vectors $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(4)}$, find a formula for A^{-1} in terms of the $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(4)}$.

12. If A is a 4×4 matrix that satisfies

$$A\mathbf{y}^{(1)} = \mathbf{e}^{(2)},\ A\mathbf{y}^{(2)} = -\mathbf{e}^{(4)},\ A\mathbf{y}^{(3)} = \mathbf{e}^{(1)}, \text{ and } A\mathbf{y}^{(4)} = -\mathbf{e}^{(3)}$$

for some vectors $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(4)}$, find a formula for A^{-1} in terms of the $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(4)}$.

13. If A is a 3×3 matrix that satisfies

$$A\mathbf{y}^{(1)} = \mathbf{e}^{(1)} + \mathbf{e}^{(2)},\ A\mathbf{y}^{(2)} = -\mathbf{e}^{(1)}, \text{ and } A\mathbf{y}^{(3)} = \mathbf{e}^{(2)} + \mathbf{e}^{(3)}$$

for some vectors $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(3)}$, find a formula for A^{-1} in terms of the $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{y}^{(3)}$.

14. According to Theorem 1.21, if $AB = I_n$, then $BA = I_n$, also; hence, $B = A^{-1}$. Find an example of 2×2 matrices A and B for which $AB = O$ but $BA \neq O$.

15. Suppose C is an unspecified but invertible $m \times m$ matrix and A is an unspecified $m \times n$ matrix. Explain why every solution of the system $A\mathbf{x} = \mathbf{0}$ is a solution of the system $CA\mathbf{x} = \mathbf{0}$, and vice versa.

16. If $C\mathbf{y} = \mathbf{z}$, $B\mathbf{x} = \mathbf{y}$, and B and C are inverses of each other, must there be a relationship between $\mathbf{x}$ and $\mathbf{z}$?

17. Explain why AC is invertible and $(AC)^{-1} = C^{-1}A^{-1}$ if A and C are invertible and of the same size. [Hint: Let $D = AC$ and $B = C^{-1}A^{-1}$ and explain why $DB = I$.]

18. Suppose A and B are $n \times n$ invertible matrices.
    (a) If $AB = BA$, why must $A^{-1}B^{-1} = B^{-1}A^{-1}$?
    (b) If $AB = BA$, why must $AB^{-1} = B^{-1}A$ and $A^{-1}B = BA^{-1}$?
    (c) If $AB = BA$, why must

$$A + B = (A^{-1} + B^{-1})AB = AB(A^{-1} + B^{-1})?$$

19. Assume A is an $n \times n$ matrix. Which of the following must be true?
    If a system of equations $A\mathbf{x} = \mathbf{0}$ has a nontrivial solution, then
    (a) A is invertible.
    (b) A is singular.
    (c) The columns of A span $\mathbb{R}^n$.
    (d) The set $\mathcal{W} \triangleq \{\mathbf{x} :\ A\mathbf{x} = \mathbf{0}\}$ does not contain the zero vector.

20. Suppose that A is $m \times n$, $A\mathbf{x} = \mathbf{0}$ has a nontrivial solution $\mathbf{x}^{(1)}$, and $n \times n$ matrix B is invertible.
    (a) Find at least one nontrivial solution of $AB\mathbf{x} = \mathbf{0}$.

www.TechnicalBooksPDF.com

(b) Suppose that in addition to the aforementioned assumptions, we assume that $m = n$. What possible values can $rank(A)$ have, and why? What possible values can $rank(AB)$ have, and why?

21. Suppose A, B, and C are $n \times n$ and $ABC = I_n$.
    (a) Must A be invertible? If "yes," why? If "no," give an example.
    (b) Must C be invertible? If "yes," why? If "no," give an example.
    (c) Must B be invertible? If "yes," why? If "no," give an example.
22. If AC is invertible and C is invertible, must A be invertible? If so, find a formula for A^{-1} in terms of $(AC)^{-1}$ and C^{-1}. [Hint: You cannot assume that AC invertible implies that $(AC)^{-1} = C^{-1}A^{-1}$. Why not? Because Theorem 1.23(c) *assumes* that A and C are invertible, and you can't assume that A is invertible when trying to explain why A is invertible. *Nevertheless*, the formula $(AC)^{-1} = C^{-1}A^{-1}$ gives you a clue to find a formula for A^{-1} in terms of $(AC)^{-1}$ and C^{-1}.]
23. Suppose A, C are unspecified matrices and $B \triangleq (I - A)^{-1}$ exists. For each of (a), (b), (c), find a matrix X that satisfies the given equation. Give your conclusions in terms of B, C.
    (a) $X = AX + C$, (b) $AX = X + C$, and (c) $XA = X + C$.
24. Decide whether or not the partitioned matrix $A \triangleq \left[\begin{array}{c|c} I & O \\ \hline A_{21} & I \end{array}\right]$ is invertible; if so, find its inverse in terms of O, I, and A_{21}. Note that there is no reason to believe that A_{21} is square.*
25. Assume that A_{11} and A_{22} are invertible. Decide whether or not the partitioned matrix $A \triangleq \left[\begin{array}{c|c} A_{11} & O \\ \hline A_{21} & A_{22} \end{array}\right]$ is invertible; if so, find its inverse in terms of O, I, A_{11}, A_{21}, and A_{22}.
26. The matrix $A = \begin{bmatrix} 2 & 1 \\ -3 & -2 \end{bmatrix}$ satisfies $A^{-1} = A$. Find two other examples of 2×2 matrices, other than I, each of which equals its inverse.

1.6 Determinant, Adjugate Matrix, and Cramer's Rule

We saw in Theorem 1.22 in Section 1.5 that for 2×2 matrices $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, $ad - bc$ being nonzero versus zero completely determines whether A^{-1} exists or not. We will see that this generalizes to $n \times n$ matrices using the concept of "determinant." It turns out that this concept applies to many things, including to (a) solving systems of linear algebraic equations, later in this section; (b) eigenvalues and eigenvectors, in Section 2.1, which are some of the most important engineering tools; (c) area and volume in Chapter 6; and (d) multivariable integration, in Chapter 7.

* Problems 1.5.3.24 and 1.5.3.25 are from Carlson (1993).

www.TechnicalBooksPDF.com

Definition 1.24

The **determinant** of A, denoted by $\det(A)$ or $|A|$, is defined for 2×2 matrices by

$$\det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \triangleq a_{11}a_{22} - a_{12}a_{21}. \tag{1.36}$$

Definition 1.25

The determinant of A, denoted by $\det(A)$ or $|A|$, is defined for 3×3 matrices by

$$\det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \triangleq a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}. \tag{1.37}$$

The determinant of a 3×3 matrix is defined "recursively" in terms of determinants of 2×2 matrices.

The determinant of a 3×3 matrix is also given by

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - (a_{13}a_{22}a_{31} + a_{12}a_{21}a_{33} + a_{11}a_{23}a_{32}).$$

It can also be found by the so-called crisscross method shown in Figure 1.7. We caution that there is no crisscross method for matrices of size bigger than 3×3.

The determinant of $A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$ is defined using sub-matrix "minors"

$M_{ij} \triangleq$ the $(n-1) \times (n-1)$ matrix obtained from A by deleting its ith row and jth column

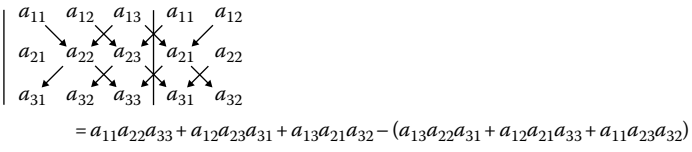


FIGURE 1.7
"Criss-cross" method for 3×3 determinant.

www.TechnicalBooksPDF.com

and

$$A_{ij} = (-1)^{i+j} \det(M_{ij}). \tag{1.38}$$

The A_{ij} are called "co-factors." The reason for this name will be given in Definition 1.26.

Example 1.33

For $A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 7 & 6 & 5 & 4 \\ 3 & 2 & -1 & 0 \end{bmatrix}$, find A_{11}, A_{12}, and A_{23}.

Method: We have $A_{11} = (-1)^{1+1} \begin{vmatrix} 6 & 7 & 8 \\ 6 & 5 & 4 \\ 2 & -1 & 0 \end{vmatrix} = -48$, $A_{12} = (-1)^{1+2} \begin{vmatrix} 5 & 7 & 8 \\ 7 & 5 & 4 \\ 3 & -1 & 0 \end{vmatrix} =$ 72, and $A_{23} = (-1)^{2+3} \begin{vmatrix} 1 & 2 & 4 \\ 7 & 6 & 4 \\ 3 & 2 & 0 \end{vmatrix} = 0$. ◯

Definition 1.26

The determinant of A, denoted by $\det(A)$ or $|A|$, is defined for $n \times n$ matrices by

$$\det(A) \triangleq a_{11}A_{11} + a_{12}A_{12} + \cdots + a_{1n}A_{1n}.$$

This definition is called "expansion of the determinant along the first row."

Theorem 1.27

$\det(A)$ can be calculated by expansion along any row or column, that is,

(a) $|A| = a_{i1}A_{i1} + a_{i2}A_{i2} + \cdots + a_{in}A_{in}$, any i for $1 \le i \le n$, and
(b) $|A| = a_{1j}A_{1j} + a_{2j}A_{2j} + \cdots + a_{nj}A_{nj}$, any j for $1 \le j \le n$.

So, we can expand a determinant along any row, as in (a), or along any column, as in (b). Theorem 1.27 is amazing because it says that $2n$ different ways of calculating $\det(A)$ give the same result.

Instead of Definition 1.26, there is an alternative way of defining $\det(A)$ using permutations of $\{1, \ldots, n\}$ and their parities. That alternative definition is very useful for theoretical results, including many of the theorems in this section; however, rather than lengthening the exposition considerably, we will refer the reader to other books that include this topic.

www.TechnicalBooksPDF.com

Example 1.34

Expand $\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 3 & -4 & 5 \end{vmatrix}$ along the first row, the third row, and the second column.

Method: Expansion along the first row, the third row, and the second column gives, respectively,

$$\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 3 & -4 & 5 \end{vmatrix} = 1 \cdot (-1)^{1+1} \begin{vmatrix} 5 & 6 \\ -4 & 5 \end{vmatrix} + 2 \cdot (-1)^{1+2} \begin{vmatrix} 4 & 6 \\ 3 & 5 \end{vmatrix} + 3 \cdot (-1)^{1+3} \begin{vmatrix} 4 & 5 \\ 3 & -4 \end{vmatrix}$$

$$= 49 - 2(2) + 3(-31) = -48,$$

$$\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 3 & -4 & 5 \end{vmatrix} = 3 \cdot (-1)^{3+1} \begin{vmatrix} 2 & 3 \\ 5 & 6 \end{vmatrix} + (-4) \cdot (-1)^{3+2} \begin{vmatrix} 1 & 3 \\ 4 & 6 \end{vmatrix} + 5 \cdot (-1)^{3+3} \begin{vmatrix} 1 & 2 \\ 4 & 5 \end{vmatrix}$$

$$= 3(-3) + 4(-6) + 5(-3) = -48, \text{ and}$$

$$\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 3 & -4 & 5 \end{vmatrix} = 2 \cdot (-1)^{1+2} \begin{vmatrix} 4 & 6 \\ 3 & 5 \end{vmatrix} + 5 \cdot (-1)^{2+2} \begin{vmatrix} 1 & 3 \\ 3 & 5 \end{vmatrix} + (-4) \cdot (-1)^{3+2} \begin{vmatrix} 1 & 3 \\ 4 & 6 \end{vmatrix}$$

$$= -2(2) + 5(-4) + 4(-6) = -48. \bigcirc$$

Here are a few important facts about determinants.

Theorem 1.28

If A and B are $n \times n$, then

(a) $|A^T| = |A|$.

(b) $|AB| = |A|\,|B|$.

(c) If A is an upper or lower triangular matrix, then $|A| = a_{11}a_{22}\cdots a_{nn}$, that is, the product of the diagonal entries.

In particular, if A is a diagonal matrix, then $|A| = a_{11}a_{22}\cdots a_{nn}$.

Using Theorem 1.27, calculation of a determinant of an $n \times n$ matrix requires, in general, calculation of n determinants of $(n-1) \times (n-1)$ matrices, each of which in turn requires calculation of $n-1$ determinants of $(n-2) \times (n-2)$ matrices, etc. This could require millennia of computational time, after which time, the accumulated round-off errors would invalidate the conclusion because the number $n!$, that is, $n(n-1)\cdots 2 \cdot 1$, quickly becomes huge* as $n \to \infty$, for example, $20! \approx 2.6 \times 10^{18}$. So, we want to have a better way to evaluate a determinant of a "large" matrix. Theorem 1.28(c) gives us a clue: maybe we can

* Stirling's approximation gives $n! \sim \sqrt{2\pi n}\, n^n\, e^{-n}$.

www.TechnicalBooksPDF.com

take the determinant of an upper triangular matrix that is row equivalent to a given matrix instead of taking the determinant of the given matrix.

Theorem 1.29

Suppose B is obtained from A by an elementary row operation:

(a) $|A| = |B|$, if $kR_i + R_j \to R_j$ on A yields B.
(b) $|A| = -|B|$, if $R_i \leftrightarrow R_j$ on A yields B.
(c) $|A| = k|B|$, if $\frac{1}{k}R_i \to R_i$ on A yields B.

Example 1.35

Find the elementary row operation that gives the equalities
(a) $\begin{vmatrix} a & b \\ c & d \end{vmatrix} = \begin{vmatrix} a & b \\ 2a+c & 2b+d \end{vmatrix}$, (b) $\begin{vmatrix} a & b \\ c & d \end{vmatrix} = -\begin{vmatrix} c & d \\ a & b \end{vmatrix}$, and (c) $\begin{vmatrix} a & b \\ c & d \end{vmatrix} = 2\begin{vmatrix} \frac{a}{2} & \frac{b}{2} \\ c & d \end{vmatrix}$.

Method:

(a) $2R_1 + R_2 \to R_2$ operates on $\begin{vmatrix} a & b \\ c & d \end{vmatrix}$ to produce $\begin{vmatrix} a & b \\ 2a+c & 2b+d \end{vmatrix}$.

(b) $R_1 \leftrightarrow R_2$ operates on $\begin{vmatrix} a & b \\ c & d \end{vmatrix}$ to produce $-\begin{vmatrix} c & d \\ a & b \end{vmatrix}$.

(c) This may look the strangest of the three examples. We can think of $\frac{1}{2}R_1 \to R_1$ as $2R_1 \leftarrow R_1$, so, for example, $\begin{vmatrix} 4 & 6 \\ 1 & 5 \end{vmatrix} = 2\begin{vmatrix} 2 & 3 \\ 1 & 5 \end{vmatrix}$, via $2R_1 \leftarrow R_1$, works by effectively "factoring out 2 from the first row" of $\begin{vmatrix} 4 & 6 \\ 1 & 5 \end{vmatrix}$. ○

If we combine Theorem 1.29 with Theorem 1.28(c), we have a row reduction algorithm to evaluate determinants.

Example 1.36

Use elementary row operations to evaluate $\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 3 & -4 & 5 \end{vmatrix}$.

Method:

$$\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 3 & -4 & 5 \end{vmatrix} \underset{\substack{-4R_1 + R_2 \to R_2 \\ -3R_1 + R_3 \to R_3}}{=} \begin{vmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & -10 & -4 \end{vmatrix} \underset{-\frac{10}{3}R_2 + R_3 \to R_3}{=} \begin{vmatrix} 1 & 2 & 3 \\ 0 & -3 & -3 \\ 0 & 0 & 16 \end{vmatrix}$$

$$= (1)(-3)(16) = -48. \text{ ○}$$

In Example 1.36 we did not use a row exchange or multiplication of a row during the row reduction process. A row exchange(s) may be needed to get to a row echelon form,

www.TechnicalBooksPDF.com

that is, a row equivalent matrix in upper triangular form. Also, a row exchange(s) may be useful to do "partial pivoting" to control numerical error.

Example 1.37

Use elementary row operations to evaluate $\begin{vmatrix} 0 & 2 & 3 \\ 0 & 4 & 5 \\ 1 & -2 & 8 \end{vmatrix}$.

Method:

$$\begin{vmatrix} 0 & 2 & 3 \\ 0 & 4 & 5 \\ 1 & -2 & 8 \end{vmatrix} \underset{R_1 \leftrightarrow R_3}{=} - \begin{vmatrix} 1 & -2 & 8 \\ 0 & 4 & 5 \\ 0 & 2 & 3 \end{vmatrix} \underset{-\frac{1}{2}R_2 + R_3 \to R_3}{=} - \begin{vmatrix} 1 & -2 & 8 \\ 0 & 4 & 5 \\ 0 & 0 & \frac{1}{2} \end{vmatrix} = -1 \cdot 4 \cdot \frac{1}{2} = -2. \bigcirc$$

This algorithm enables machines to find the determinant of a large matrix.
Because of the connection between row reduction and determinants, we have

Theorem 1.30

When A is $n \times n$, A^{-1} exists $\iff$ $rank(A) = n$ $\iff$ $|A| \neq 0$.

Theorem 1.30 is one of the mathematical results most often used by engineers and scientists.

Because of Theorem 1.30, we know the matrix A is invertible in each of Examples 1.36 and 1.37. Theorem 1.30 is an "existential result," that is, it tells us whether A^{-1} exists but does not tell us how to find A^{-1}. But Theorem 1.30 illustrates the concept that if we study the explanation of a theorem, we can discern an algorithm for getting numerical results. Here, the row reduction algorithm for determinants is what makes the explanation "tick."

Theorem 1.31

If A^{-1} exists, then $|A^{-1}| = \dfrac{1}{|A|}$.

Besides using elementary row operations to evaluate a determinant, we can also use "**elementary column operations**" because of the identity $|A| = |A^T|$. For example, the elementary column operation of interchanging two columns given by $C_i \leftrightarrow C_j$, multiplies a determinant by (-1) because the operation of interchanging two rows of A^T multiplies its determinant by (-1). Similarly, the operation of adding a multiple of one column into another column does not change the determinant, for example, $2C_1 + C_3 \to C_3$ doesn't affect the determinant.

www.TechnicalBooksPDF.com

1.6.1 Adjugate Matrix

Definition 1.27

The **adjugate** of A is defined by $adj(A) \triangleq \left[A_{ji}\right]_{\substack{1 \le j \le n \\ 1 \le i \le n}} = \left(\left[A_{ij}\right]_{\substack{1 \le i \le n \\ 1 \le j \le n}}\right)^T$, that is, $adj(A)$ is the *transpose* of the matrix of cofactors of A.

By the way, many people call $adj(A)$ the "adjoint of A" but use the same abbreviation, "adj." We prefer to not use the word adjoint because it means something entirely different in the subject of operator theory, which has many applications to matrix theory and differential equations.

Theorem 1.32

(a) $A\ adj(A) = |A|\ I_n = adj(A)\ A$.

(b) If $|A| \neq 0$, then $A^{-1} = \dfrac{1}{|A|} adj(A)$.

(c) If $|A| = 0$, then $A\ adj(A) = O = adj(A)\ A$.

Theorem 1.32(b) gives a formula for A^{-1} that can be useful for very small matrices and also for theoretical purposes. But we usually do not use that formula as a method for computing A^{-1} because computing $adj(A)$ requires computing n^2 determinants of $(n-1) \times (n-1)$ matrices.

Theorem 1.33

Given an $n \times n$ matrix A,

(a) $rank(adj(A)) = \begin{cases} 0, & \text{if } rank(A) < n-1 \\ 1, & \text{if } rank(A) = n-1 \\ n, & \text{if } rank(A) = n \end{cases}$.

(b) If the nullity of A is one, that is, $\nu(A) = 1$, then every nonzero column of $adj(A)$ is a nontrivial solution of $A\mathbf{x} = \mathbf{0}$.

Why? An explanation of Theorem 1.33(a) can be found in *Elementary Matrix Theory*, by H. Eves. As for Theorem 1.33(b), let $adj(A) \triangleq B = \left[\mathbf{B}_{*1} \quad \mathbf{B}_{*2} \quad \cdots \quad \mathbf{B}_{*n}\right]$ and note that $AB = \left[A\mathbf{B}_{*1} \quad A\mathbf{B}_{*2} \quad \cdots \quad A\mathbf{B}_{*n}\right]$ by Theorem 1.9 in Section 1.2. Recall from Theorem 1.17 in Section 1.3 that $rank(A) + \nu(A) = n$; hence, if $\nu(A) = 1$, then $rank(A) = n-1$ and $|A| = 0$. So, Theorem 1.32(c) gives $O = AB = \left[A\mathbf{B}_{*1} \quad A\mathbf{B}_{*2} \quad \cdots \quad A\mathbf{B}_{*n}\right]$. So, every nonzero column of $adj(A)$ is a nontrivial solution of $A\mathbf{x} = \mathbf{0}$. We call this the "adjugate matrix method" for finding a nontrivial solution of $A\mathbf{x} = \mathbf{0}$ when A is not invertible.

www.TechnicalBooksPDF.com

We will apply Theorem 1.33 to find an eigenvector in Section 2.1; this application was inspired by an example of "mode shape" (actually, an eigenvector of an associated 3×3 matrix!) of a three-mass, three-spring mechanical model in Timoshenko et al. (1990). We'll discuss an example like Timoshenko et al.'s example in Chapter 5.

Like anything else having to do with the adjugate matrix, this "adjugate matrix method" for solving $A\mathbf{x} = \mathbf{0}$ is practical only for very small matrices but *is* also useful for theoretical purposes.

Example 1.38

Find the adjugate of $A = \begin{bmatrix} -5 & 2 & 4 \\ 2 & -8 & 2 \\ 4 & 2 & -5 \end{bmatrix}$ and use the result of Theorem 1.33 to find a $\mathbf{x} \neq \mathbf{0}$ that solves $A\mathbf{x} = \mathbf{0}$.

Method:

$$adj(A) = \begin{bmatrix} \begin{vmatrix} -8 & 2 \\ 2 & -5 \end{vmatrix} & -\begin{vmatrix} 2 & 2 \\ 4 & -5 \end{vmatrix} & \begin{vmatrix} 2 & -8 \\ 4 & 2 \end{vmatrix} \\ -\begin{vmatrix} 2 & 4 \\ 2 & -5 \end{vmatrix} & \begin{vmatrix} -5 & 4 \\ 4 & -5 \end{vmatrix} & -\begin{vmatrix} -5 & 2 \\ 4 & 2 \end{vmatrix} \\ \begin{vmatrix} 2 & 4 \\ -8 & 2 \end{vmatrix} & -\begin{vmatrix} -5 & 4 \\ 2 & 2 \end{vmatrix} & \begin{vmatrix} -5 & 2 \\ 2 & -8 \end{vmatrix} \end{bmatrix}^T = \begin{bmatrix} 36 & 18 & 36 \\ 18 & 9 & 18 \\ 36 & 18 & 36 \end{bmatrix}.$$

$Adj(A)$ has rank equal to one, as we can see from doing the elementary row operations $-\frac{1}{2}R_1 + R_2 \to R_2$, $-R_1 + R_3 \to R_3$ on $adj(A)$.

Theorem 1.33(a) implies $rank(A) = 2$; hence, $\nu(A) = 1$. Illustrating Theorem 1.33(b), we see that $\begin{bmatrix} 36 \\ 18 \\ 36 \end{bmatrix}$ is a nonzero solution of $A\mathbf{x} = \mathbf{0}$. ◯

We can also define the determinant of a 1×1 matrix by $det([\, a_{11} \,]) = a_{11}$. Note that we should not use the $|\ |$ notation for the determinant of a 1×1 matrix because it would be confused with the absolute value notation. With this notation, the adjugate matrix of a 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is given by

$$adj(A) = \begin{bmatrix} det([\, d \,]) & -det([\, c \,]) \\ -det([\, b \,]) & det([\, a \,]) \end{bmatrix}^T = \begin{bmatrix} d & -c \\ -b & a \end{bmatrix}^T = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

So, the formula for the inverse of a 2×2 matrix given in (1.32) in Section 1.5 is actually the formula of Theorem 1.32(b) for $n = 2$.

1.6.2 Cramer's Rule

If A^{-1} exists, that is, if $|A| \neq 0$, then for every vector $\mathbf{b}$, the system $A\mathbf{x} = \mathbf{b}$ has unique solution given by

www.TechnicalBooksPDF.com

$$\mathbf{x} = A^{-1}\mathbf{b} = \frac{1}{|A|}\begin{bmatrix} A_{11} & . & . & . & A_{1n} \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ A_{n1} & . & . & . & A_{mn} \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix};$$

hence,

$$x_1 = \frac{1}{|A|}(A_{11}b_1 + A_{21}b_2 + \cdots + A_{n1}b_n) = \frac{1}{|A|}(b_1A_{11} + b_2A_{21} + \cdots + b_nA_{n1}).$$

But, $(b_1A_{11} + b_2A_{21} + \cdots + b_nA_{n1})$ is the determinant of a matrix expanded along its first column:

$$(b_1A_{11} + b_2A_{21} + \cdots + b_nA_{n1}) = \begin{vmatrix} b_1 & a_{12} & . & . & . & a_{1n} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ b_n & a_{n2} & . & . & . & a_{nn} \end{vmatrix}.$$

The latter determinant is the same as $|A|$ except for having the first column replaced by the vector $\mathbf{b}$. This suggests notations for new $n \times n$ matrices that mix A and $\mathbf{b}$:

$$\begin{aligned} A_1(\mathbf{b}) &\triangleq [\,\mathbf{b} \quad \mathbf{A}_{*2} \quad \mathbf{A}_{*3} \quad \cdots \quad \mathbf{A}_{*n}\,] \\ A_2(\mathbf{b}) &\triangleq [\,\mathbf{A}_{*1} \quad \mathbf{b} \quad \mathbf{A}_{*3} \quad \cdots \quad \mathbf{A}_{*n}\,] \\ &\vdots \\ A_n(\mathbf{b}) &\triangleq [\,\mathbf{A}_{*1} \quad \mathbf{A}_{*2} \quad \cdots \quad \mathbf{A}_{*(n-1)} \quad \mathbf{b}\,], \end{aligned}$$

that is, $A_j(\mathbf{b})$ is the matrix obtained from A by replacing its jth column by the vector $\mathbf{b}$. We have that the system $A\mathbf{x} = \mathbf{b}$ has unique solution given by

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} |A_1(\mathbf{b})| \,/\, |A| \\ |A_2(\mathbf{b})| \,/\, |A| \\ \vdots \\ |A_n(\mathbf{b})| \,/\, |A| \end{bmatrix}. \tag{1.39}$$

Theorem 1.34

(Cramer's rule) If $|A| \neq 0$, then for every n vector $\mathbf{b}$, the unique solution of $A\mathbf{x} = \mathbf{b}$ is given by (1.39).

Again, this method for solving $A\mathbf{x} = \mathbf{b}$ is useful only for very small systems but is useful for theoretical purposes.

www.TechnicalBooksPDF.com

Example 1.39

For the system

$$\mathbf{b} = A\mathbf{x} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 3 & -4 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix},$$

(a) Use Cramer's rule to find the unique solution in terms of the unspecified parameters b_1, b_2, b_3.

(b) Use your result for part (a) to find A^{-1}.

Method:

$$x_1 = \frac{\begin{vmatrix} b_1 & 2 & 3 \\ b_2 & 5 & 6 \\ b_3 & -4 & 5 \end{vmatrix}}{\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 3 & -4 & 5 \end{vmatrix}} = -\frac{1}{48}\left(b_1 \begin{vmatrix} 5 & 6 \\ -4 & 5 \end{vmatrix} - b_2 \begin{vmatrix} 2 & 3 \\ -4 & 5 \end{vmatrix} + b_3 \begin{vmatrix} 2 & 3 \\ 5 & 6 \end{vmatrix}\right),$$

$$x_2 = \frac{\begin{vmatrix} 1 & b_1 & 3 \\ 4 & b_2 & 6 \\ 3 & b_3 & 5 \end{vmatrix}}{\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 3 & -4 & 5 \end{vmatrix}} = -\frac{1}{48}\left(-b_1 \begin{vmatrix} 4 & 6 \\ 3 & 5 \end{vmatrix} + b_2 \begin{vmatrix} 1 & 3 \\ 3 & 5 \end{vmatrix} - b_3 \begin{vmatrix} 1 & 3 \\ 4 & 6 \end{vmatrix}\right),$$

and

$$x_3 = \frac{\begin{vmatrix} 1 & 2 & b_1 \\ 4 & 5 & b_2 \\ 3 & -4 & b_3 \end{vmatrix}}{\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 3 & -4 & 5 \end{vmatrix}} = -\frac{1}{48}\left(b_1 \begin{vmatrix} 4 & 5 \\ 3 & -4 \end{vmatrix} - b_2 \begin{vmatrix} 1 & 2 \\ 3 & -4 \end{vmatrix} + b_3 \begin{vmatrix} 1 & 2 \\ 4 & 5 \end{vmatrix}\right).$$

For each of the calculations of x_1, x_2, and x_3, we expanded along the column in which $\mathbf{b}$ is found.

The unique solution is

$$\mathbf{x} = \begin{bmatrix} -\frac{49}{48}b_1 + \frac{22}{48}b_2 + \frac{1}{16}b_3 \\ \frac{1}{24}b_1 + \frac{1}{12}b_2 - \frac{1}{8}b_3 \\ \frac{31}{48}b_1 - \frac{5}{24}b_2 + \frac{1}{16}b_3 \end{bmatrix}. \tag{1.40}$$

(c) We can rewrite (1.40) as

$$\mathbf{x} = \frac{1}{48}\begin{bmatrix} -49 & 22 & 3 \\ 2 & 4 & -6 \\ 31 & -10 & 3 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}, \tag{1.41}$$

www.TechnicalBooksPDF.com

and this is true for *all* vectors $\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$. Because (1.41) is really $\mathbf{x} = A^{-1}\mathbf{b}$, we can read A^{-1} from (1.41):

$$A^{-1} = \frac{1}{48}\begin{bmatrix} -49 & 22 & 3 \\ 2 & 4 & -6 \\ 31 & -10 & 3 \end{bmatrix}. \bigcirc$$

Learn More About It

A good discussion of the definition of determinants using permutations is in *Elementary Linear Algebra with Applications*, 7th ed., by Bernard Kolman and David Hill, Prentice-Hall, © 2000. A three-mass, three-spring mechanical model is found in *Vibration Problems in Engineering*, 4th edn. by S. Timoshenko, D. H. Young, and W. Weaver, Jr., John Wiley & Sons, © 1990.

1.6.3 Problems

In problems 1–2, for each system, find the determinant by (a) expanding along a row or column of your choice and (b) using the row reduction algorithm. In (a), state which row or column you are expanding along. In (b), use only elementary row operations and notate them.

1. $\begin{bmatrix} 0 & 1 & 4 \\ -1 & 3 & 2 \\ 2 & 0 & 1 \end{bmatrix}$
2. $\begin{bmatrix} 0 & -1 & 2 \\ 1 & 5 & 3 \\ -2 & -1 & 1 \end{bmatrix}$
3. Find an example of 2×2 matrices A, B for which $|A + B| \neq |A| + |B|$.
4. If A and B are 3×3 matrices, $|A| = -2$, and $|B| = 5$, find
   (a) $|AB|$, (b) $|A^T B^2|$, (c) $|-A^4 B|$, (d) $|AB^{-1}|$.
5. You may assume that $|A| = -132$, where

$$A = \begin{bmatrix} 1 & 5 & 1 & 10 \\ 2 & 7 & -1 & 8 \\ 3 & 9 & -2 & 6 \\ 4 & 0 & -3 & 4 \end{bmatrix}.$$

   (a) Use $|A|$ and an elementary row operation to find

$$\begin{vmatrix} 1 & 5 & 1 & 10 \\ 3 & 12 & 0 & 18 \\ 3 & 9 & -2 & 6 \\ 4 & 0 & -3 & 4 \end{vmatrix}.$$

www.TechnicalBooksPDF.com

(b) Use $|A|$ and an elementary row operation to find

$$\begin{vmatrix} 1 & 5 & 1 & 10 \\ 1 & 4 & 0 & 6 \\ 3 & 9 & -2 & 6 \\ 4 & 0 & -3 & 4 \end{vmatrix}.$$

6. Find $adj\left(\begin{bmatrix} 1 & 1 & 2 \\ -1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix}\right)$.
7. Suppose a, b, c are unspecified but distinct scalars. Explain why the determinant result

$$\begin{vmatrix} 1 & 1 & 1 \\ a & b & c \\ a^2 & b^2 & c^2 \end{vmatrix}$$

is $(c-a)(c-b)(b-a)$. [Note: This is an example of a **Vandermonde determinant**.]

8. Suppose a, b, c are unspecified but distinct scalars. For

$$A = \begin{bmatrix} a & c & 0 \\ 0 & b & a \\ b & 0 & c \end{bmatrix},$$

(a) Find $|A|$ in terms of a, b, c.
(b) Find $adj(A)$ in terms of a, b, c.
(c) Find $A\ adj(A)$.

9. Suppose the $n \times n$ matrix $B = \alpha A$ for some nonzero scalar α and A^{-1} exists.
(a) Why must B^{-1} exist? Find a formula for B^{-1} in terms of A^{-1} and α.
(b) Find a formula for $adj(B)$ in terms of $adj(A)$ and α.
10. Suppose A is 5×5 and $A^T = 2A$. Find all possible values of $|A|$.
11. You may assume that

$$\begin{vmatrix} 1 & 2 & -1 \\ 2 & 0 & 1 \\ -3 & -2 & 1 \end{vmatrix} = -4.$$

Use Cramer's rule to find the exact solution of the system

$$\left\{\begin{array}{rcrcrl} x_1 & + & 2x_2 & - & x_3 & = s \\ 2x_1 & & & + & x_3 & = 5 \\ -3x_1 & - & 2x_2 & + & x_3 & = t \end{array}\right\}$$

in terms of the unspecified parameters s and t.

www.TechnicalBooksPDF.com

12. Consider the system

$$\left\{\begin{array}{rcrcl} sx_1 & + & x_2 & = & 5 \\ -2x_1 & + & 3x_2 & = & 4 \end{array}\right\}.$$

(a) For which value(s) of the parameter "s" does the system have exactly one solution?

(b) For the value(s) of "s" you found in part (a), explain how to use Cramer's rule to find the solution $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. Your final conclusion should be in terms of "s."

13. Consider the system

$$\left\{\begin{array}{rcrcl} sx_1 & + & x_2 & = & 3 \\ 4x_1 & + & sx_2 & = & -6 \end{array}\right\}.$$

(a) For which value(s) of the parameter "s" does the system have exactly one solution?

(b) For the value(s) of "s" you found in part (a), explain how to use Cramer's rule to find the solution $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. Your final conclusion should be in terms of "s."

(c) For the all *other* value(s) of "s," decide whether the system has a solution(s), and if it does, find the solution(s).

14. Use Cramer's rule to find x_2 for the system

$$\left\{\begin{array}{rcrcrcl} x_1 & & & - & 2x_3 & = & 3 \\ 2x_1 & + & x_2 & - & x_3 & = & -2 \\ -x_1 & + & 2x_2 & + & x_3 & = & 1 \end{array}\right\}.$$

15. Use Cramer's rule to find x_2 in terms of b_1, b_2, and b_3 for the system

$$\left\{\begin{array}{rcrcrcl} x_1 & + & x_2 & + & x_3 & = & b_1 \\ & & x_2 & - & x_3 & = & b_2 \\ -x_1 & + & x_2 & - & x_3 & = & b_3 \end{array}\right\}.$$

Do not substitute in specific values for b_1, b_2, and b_3.

16. For the system as follows, use Cramer's rule to find x_3 in terms of b_1, b_2, and b_3. Do not substitute in specific values for b_1, b_2, and b_3.

$$\left\{\begin{array}{rcrcrcl} x_1 & + & x_2 & + & x_3 & = & b_1 \\ -x_1 & + & 2x_2 & - & x_3 & = & b_2 \\ & & x_2 & - & 2x_3 & = & b_3 \end{array}\right\}.$$

17. For what value(s) of k is $\begin{bmatrix} 4 & 0 & k \\ 1 & k & 0 \\ k & 0 & 9 \end{bmatrix}$ invertible?

www.TechnicalBooksPDF.com

18. Find all value(s) of k for which the matrix $A = \begin{bmatrix} k & 0 & k \\ 0 & 4 & 1 \\ k & 0 & 2 \end{bmatrix}$ is invertible. For those values of k, find A^{-1} *in terms of* k.
19. (a) Why does having two equal rows imply that the determinant of a matrix must be zero?
    (b) Why does having two equal columns imply that the determinant of a matrix must be zero?
    (c) Why are these "alien cofactor identities" true?
$$0 = a_{i1}A_{j1} + a_{i2}A_{j2} + \cdots + a_{in}A_{jn}, \text{ if } i \neq j,$$
$$0 = a_{1j}A_{1i} + a_{2j}A_{2i} + \cdots + a_{nj}A_{ni}, \text{ if } i \neq j.$$
20. Suppose A is invertible and
$$A^{-1} = \begin{bmatrix} 1 & -4 & 7 \\ -2 & 5 & -8 \\ 3 & -6 & 2 \end{bmatrix}.$$
    (a) If you had used Cramer's rule to find the solution of $A\mathbf{x} = \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$, what would be the formula for $|A_2(\mathbf{b})|/|A|$ in terms of b_1, b_2, and b_3? Do not substitute in specific values for b_1, b_2, and b_3.
    (b) Find $|A|$ and explain how you found it.
21. Suppose $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ and $B = \begin{bmatrix} e & f \\ g & h \end{bmatrix}$. Recall a theorem that tells when a 2×2 matrix is invertible. Explain why $(ae + bg)(cf + dh) - (af + bh)(ce + dg) \neq 0$ if $ad - bc \neq 0$ and $eh - fg \neq 0$.
22. Suppose B and C are invertible matrices and $AB = C$ and $CA = B$. Explain why $|A|$ is either 1 or -1.
23. Suppose A and B are $n \times n$ matrices and $AB = -I_n$. Which of the following must be true? For each one, explain why, or why not.
    (a) $|A| \neq 0$, (b) B is not invertible, (c) $A^{-1} = -B$, (d) $|A^T| = 0$.
24. Suppose A is a singular $n \times n$ matrix and $\mathbf{b}$ is in $\mathbb{R}^n$.
    (a) If $A\mathbf{x} = \mathbf{b}$ *does* have a solution $\mathbf{x}^*$, why must $adj(A)\mathbf{b} = \mathbf{0}$?
    (b) If $A\mathbf{x} = \mathbf{b}$ *does not* have a solution, must $adj(A)\mathbf{b} = \mathbf{0}$? If so, why? If not, give a specific counterexample.
25. (a) Suppose A and B are invertible matrices. Why must $adj(AB) = adj(B)adj(A)$?
    (b) If A and B are 2×2 matrices and either or both of A and B are singular, must $adj(AB) = adj(B)adj(A)$? If so, why? If not, give a specific counterexample.
26. Let $I_3 = [\, \mathbf{e}^{(1)} \quad \mathbf{e}^{(2)} \quad \mathbf{e}^{(3)} \,]$. Is it possible to have a 3×3 matrix A that satisfies
$$A(\mathbf{y}^{(1)} + \mathbf{y}^{(2)}) = \mathbf{e}^{(1)},\ A(2\mathbf{y}^{(1)} - \mathbf{y}^{(2)}) = \mathbf{e}^{(2)}, \text{ and } A\mathbf{y}^{(2)} = \mathbf{e}^{(3)}$$
for some vectors $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(3)}$? If so, find a specific example; if not, explain why not.

www.TechnicalBooksPDF.com

27. Suppose an $n \times n$ matrix A can be written in the partitioned form $A \triangleq \left[\begin{array}{c|c} A_{11} & O \\ \hline O & I \end{array}\right]$, where I is the $(n-r) \times (n-r)$ identity matrix. Explain why $|A| = |A_{11}|$. [Hints: The last $n-r$ columns of A are the last $n-r$ columns of the $n \times n$ identity matrix (why?), so expand the determinant of A along the last column, and then along the last column of a resulting $(n-1) \times (n-1)$ determinant, etc.]
28. Suppose an $n \times n$ matrix A can be written in the partitioned form $A \triangleq \left[\begin{array}{c|c} I & A_{12} \\ \hline O & A_{22} \end{array}\right]$, where I is the $r \times r$ identity matrix. Explain why $|A| = |A_{22}|$. [Hints: The first r columns of A are the first r columns of the $n \times n$ identity matrix (why?), so expand the determinant of A along the first column and then along the first column of a resulting $(n-1) \times (n-1)$ determinant, etc.]
29. Suppose an $n \times n$ matrix A can be written in the partitioned form $A \triangleq \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline O & A_{22} \end{array}\right]$, where A_{11} is $r \times r$. Suppose that A_{11} is invertible. Explain why $|A| = |A_{11}|\,|A_{22}|$. [Hints: First, verify that we can factor A as

$$A = \left[\begin{array}{c|c} A_{11} & O \\ \hline O & I \end{array}\right] \left[\begin{array}{c|c} I & A_{11}^{-1}A_{12} \\ \hline O & A_{22} \end{array}\right].$$

After that, use the results of Problems 1.6.3.27 and 1.6.3.28.]

1.7 Linear Independence, Basis and Dimension

Recall from Definition 1.18 in Section 1.3 that a set of vectors, $\mathcal{W}$, has a spanning set, $\{\mathbf{v}_1, \ldots, \mathbf{v}_\ell\}$, if *every* vector $\mathbf{v}$ in $\mathcal{W}$ can be written as a linear combination of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_\ell$, that is, there exists at least one choice of scalars $c_1, \ldots, c_\ell$ such that $\mathbf{v} = c_1\mathbf{v}_1 + \cdots + c_\ell\mathbf{v}_\ell$. The context that we are most concerned with is when $\mathcal{W}$ is the set of solutions of a homogeneous system. For example, in Example 1.18 in Section 1.3, the system $\left\{\begin{array}{rrrrr} x_1 & +x_2 & & -x_4 & = 0 \\ & & x_3 & +2x_4 & = 0 \end{array}\right\}$ has solution set spanned by

$$\left\{\begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ -2 \\ 1 \end{bmatrix}\right\}.$$

We recall from Theorem 1.16 in Section 1.3 that if $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\ell)}$ are *all* of the *basic* solutions of a homogeneous system $A\mathbf{x} = \mathbf{0}$, then $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\ell)}\}$ is a spanning set for the set of solutions, and no other spanning set has fewer than ℓ vectors.

www.TechnicalBooksPDF.com

There is one more underlying concept that we will need, that of "linear independence" of vectors. Intuitively, a set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_\ell\}$ is linearly independent if there are ℓ "truly different" directions expressed by that set. We need a precise definition in order to really pin down the concept of "truly different."

Before we do so, here is a widely useful calculation, which we will use often again in Chapter 5, as well as many times in this section.

Lemma 1.3

Given vectors $\mathbf{v}_1 = \begin{bmatrix} v_{11} \\ v_{21} \\ \vdots \\ v_{n1} \end{bmatrix}$, $\mathbf{v}_2 = \begin{bmatrix} v_{12} \\ v_{22} \\ \vdots \\ v_{n2} \end{bmatrix}$, $\ldots$, $\mathbf{v}_\ell = \begin{bmatrix} v_{1\ell} \\ v_{2\ell} \\ \vdots \\ v_{n\ell} \end{bmatrix}$, form the $n \times \ell$ partitioned matrix $A \triangleq \left[\mathbf{v}_1 \mid \mathbf{v}_2 \mid \cdots \mid \mathbf{v}_\ell \right]$ and define $\boldsymbol{\alpha} \triangleq \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_\ell \end{bmatrix}$. Then we have

$$A\boldsymbol{\alpha} = \alpha_1 \mathbf{v}_1 + \cdots + \alpha_\ell \mathbf{v}_\ell. \tag{1.42}$$

Why? This is really Lemma 1.2 in Section 1.2. □

1.7.1 Linear Independence

Definition 1.28

(a) In a vector space $\mathcal{V}$, a set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_\ell\}$ is **linearly independent** if there is only one choice of scalars $\alpha_1, \ldots, \alpha_\ell$ with $\mathbf{0} = \alpha_1 \mathbf{v}_1 + \cdots + \alpha_\ell \mathbf{v}_\ell$, namely, $\alpha_1 = \cdots = \alpha_\ell = 0$.

(b) A set of vectors is defined to be **linearly dependent** if it is not linearly independent.

Lemma 1.3 explains why a set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_\ell\}$ is linearly independent if, and only if, the homogeneous system $\left[\mathbf{v}_1 \mid \mathbf{v}_2 \mid \cdots \mid \mathbf{v}_\ell \right] \boldsymbol{\alpha} = \mathbf{0}$ has only the trivial solution for $\boldsymbol{\alpha}$.

Example 1.40

Determine whether the set of vectors

$$\left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 0 \\ 1 \end{bmatrix} \right\}$$

is linearly independent or linearly dependent.

www.TechnicalBooksPDF.com

Method: Using Lemma 1.2 in Section 1.2, we can rewrite $\mathbf{0} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \alpha_\ell\mathbf{v}_\ell$ as a system whose augmented matrix is

$$\begin{bmatrix} 1 & 2 & 1 & | 0 \\ 1 & 0 & -1 & | 0 \\ 0 & 0 & 0 & | 0 \\ 1 & 2 & 1 & | 0 \end{bmatrix} \begin{matrix} \\ \\ \\ -R_1 + R_2 \to R_2 \\ -R_1 + R_4 \to R_4 \\ R_2 + R_1 \to R_1 \end{matrix} \sim \begin{bmatrix} \textcircled{1} & 0 & -1 & | 0 \\ 0 & \textcircled{2} & -2 & | 0 \\ 0 & 0 & 0 & | 0 \\ 0 & 0 & 0 & | 0 \end{bmatrix}.$$

Because there is a free variable, this homogeneous system has infinitely many solutions; hence, there is a nontrivial solution for $\boldsymbol{\alpha}$, for example, $(\alpha_1, \alpha_2, \alpha_3) = (1, -1, 1)$, by setting the free variable $\alpha_3 = 1$, hence the given set of three vectors is linearly dependent. ◯

Theorem 1.35

A set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_\ell\}$ in a vector space $\mathcal{V}$ is *linearly dependent* if, and only if, there is a choice of the index j so that $\mathbf{v}_j$ is a linear combination of the other $(\ell - 1)$ vectors, that is, so that

$$\mathbf{v}_j \text{ is in the } \mathit{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_\ell\}.$$

Example 1.41

To illustrate Theorem 1.35, the set of three vectors

$$\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\} = \left\{ \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ -1 \\ -\frac{1}{2} \end{bmatrix} \right\}$$

is linearly independent: In Figure 1.8a, the dashed multiples of the vector $\mathbf{v}_3$ are not in the plane spanned by $\{\mathbf{v}_1, \mathbf{v}_2\}$. In Figure 1.8b, the dashed multiples of the vector $\mathbf{v}_1$ are not in the plane spanned by $\{\mathbf{v}_2, \mathbf{v}_3\}$. In Figure 1.8c, the dashed multiples of the vector $\mathbf{v}_2$ are not in the plane spanned by $\{\mathbf{v}_1, \mathbf{v}_3\}$.

Because there is no j for which $\mathbf{v}_j$ is a linear combination of the other two vectors, the set of three vectors is linearly independent. ◯

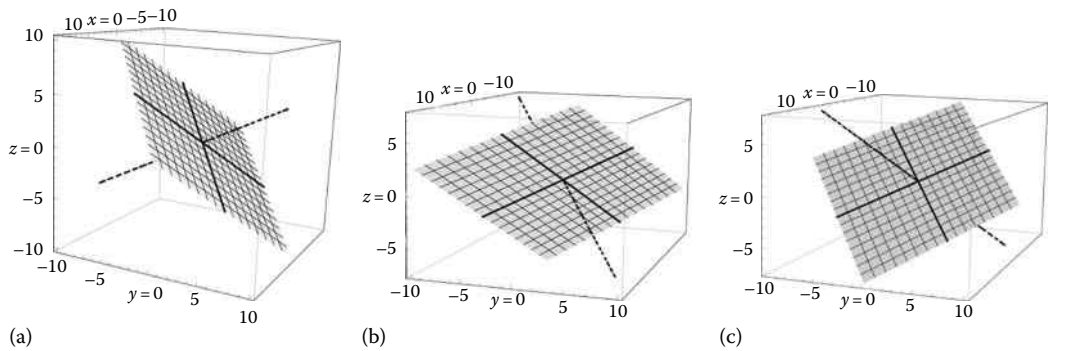


FIGURE 1.8
Linear independence. (a) $\mathbf{v}_3$ not in span $\{\mathbf{v}_1, \mathbf{v}_2\}$, (b) $\mathbf{v}_1$ not in span $\{\mathbf{v}_2, \mathbf{v}_3\}$, (c) $\mathbf{v}_2$ not in span $\{\mathbf{v}_1, \mathbf{v}_3\}$.

www.TechnicalBooksPDF.com

By the way, the planes in Figure 1.8 were plotted by *Mathematica*TM. For example, the plane in Figure 1.8a was drawn by the command

```
ParametricPlot3D[{u + v, u - v, -u + v}, {u, -5, 5}, {v, -5, 5}],
```

and then we rotated the picture using the mouse. Note that

$$\begin{bmatrix} u+v \\ u-v \\ -u+v \end{bmatrix} = u \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} + v \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$$

is a linear combination of the vectors $[1 \quad 1 \quad -1]^T, [1 \quad -1 \quad 1]^T$ for all scalars u, v.

Fortunately, if our original problem is to find a general solution of a homogeneous system and we want to decide if a set of solution vectors is a linearly independent set, the next theorem says that we don't have to do the work of row reduction as in Example 1.40. Once again, we see that a reason we learn theorems and their applications is so that we don't have to do the same, tedious work over and over again.

Theorem 1.36

If $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\ell)}$ are all of the basic solutions of a homogeneous system $A\mathbf{x} = \mathbf{0}$, then the set of vectors $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\ell)}\}$ is automatically linearly independent.

Corollary 1.2

If $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\ell)}$ are *all* of *the* basic solutions of a homogeneous system $A\mathbf{x} = \mathbf{0}$, then $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\ell)}\}$ is linearly independent and spans the set $\mathcal{W} = \{\mathbf{x} : A\mathbf{x} = \mathbf{0}\}$.

Basis

We begin with a definition that is a special case of a general definition we will state soon.

Definition 1.29

Suppose A is an $m \times n$ matrix and $\mathcal{W} = \{\mathbf{x} : A\mathbf{x} = \mathbf{0}\}$. A finite set of vectors $\mathcal{S} \triangleq \{\mathbf{v}_1, \ldots, \mathbf{v}_\ell\}$ is a **basis** for $\mathcal{W}$ if $\mathcal{S}$ is linearly independent and spans $\mathcal{W}$.

Example 1.42

Find a basis for $\mathcal{W} = \left\{ \mathbf{x} : \begin{bmatrix} 4 & 2 & 4 \\ 2 & 1 & 2 \\ 4 & 2 & 4 \end{bmatrix} \mathbf{x} = \mathbf{0} \right\}$.

www.TechnicalBooksPDF.com

Method: Row reduction gives $\left[\begin{array}{ccc|c} 4 & 2 & 4 & 0 \\ 2 & 1 & 2 & 0 \\ 4 & 2 & 4 & 0 \end{array}\right] \begin{array}{c} \sim \\ -\frac{1}{2}R_1 + R_2 \to R_2 \\ -R_1 + R_3 \to R_3 \\ \frac{1}{4}R_1 \to R_1 \end{array} \left[\begin{array}{ccc|c} \textcircled{1} & \frac{1}{2} & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$.

So, x_1 is the only basic variable, and $x_2 \triangleq c_1$, $x_3 \triangleq c_2$ are free variables. Solving for x_1 in terms of x_2, x_3 will enable us to find *all* of the basic solutions:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}x_2 - x_3 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}c_1 - c_2 \\ c_1 \\ c_2 \end{bmatrix} = c_1 \begin{bmatrix} -\frac{1}{2} \\ 1 \\ 0 \end{bmatrix} + c_2 \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \triangleq c_1\mathbf{x}^{(1)} + c_2\mathbf{x}^{(2)}.$$

By Corollary 1.2, $\left\{ \begin{bmatrix} -\frac{1}{2} \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \right\}$ is a basis for $\mathcal{W}$. ◯

1.7.2 Vector Spaces and Subspaces

A **vector space** is a nonempty set of objects, called vectors, equipped with two operations: vector addition and multiplication of a scalar times a vector. The operations have to satisfy certain "axioms."

Example 1.43

$\mathbb{R}^3$ is a vector space when given the usual operations of vector addition and multiplication of a scalar times a vector:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \triangleq \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ x_3 + y_3 \end{bmatrix} \text{ and } \alpha \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \triangleq \begin{bmatrix} \alpha x_1 \\ \alpha x_2 \\ \alpha x_3 \end{bmatrix}. \text{ ◯}$$

So, vector addition and multiplication by a scalar are defined entry by entry.

Example 1.44

$\mathbb{R}^3$ can be generalized. The set $\mathbb{R}^n$ consists of all vectors $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ and is a vector space when given the usual operations of vector addition and multiplication of a scalar times a vector:

$$\mathbf{x} + \mathbf{y} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \triangleq \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{bmatrix},$$

www.TechnicalBooksPDF.com

and for any scalar α,

$$\alpha\mathbf{x} = \alpha \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \triangleq \begin{bmatrix} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_n \end{bmatrix}. \bigcirc$$

The axioms are all very natural abstractions of properties of vector addition and scalar multiplication on $\mathbb{R}^3$, as in Example 1.43. For example, *commutativity of addition* says that for *all* vectors $\mathbf{x}, \mathbf{y}$,

$$\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x},$$

and *distributivity of vector addition* says that for all scalars α and vectors $\mathbf{x}, \mathbf{y}$,

$$\alpha(\mathbf{x} + \mathbf{y}) = (\alpha\mathbf{x}) + (\alpha\mathbf{y}).$$

There are many other axioms, but the easy way to remember them is that in a vector space, vector addition and scalar multiplication behave as they behave in $\mathbb{R}^3$. We will refer the reader to other books for a more complete exposition of the axioms.

Definition 1.30

Given a vector space $\mathcal{V}$, a **vector subspace** $\mathcal{W}$ is a nonempty subset of $\mathcal{V}$ such that when vector addition and multiplication by a scalar are restricted to only the vectors in $\mathcal{W}$, all of the vector space axioms are verified.

The definition allows $\mathcal{W} = \mathcal{V}$, that is, that $\mathcal{V}$ is a vector subspace of itself. This sounds silly, but allowing this will simplify certain results later.

Example 1.45

In $\mathbb{R}^3$, any plane or line that passes through the origin, that is, contains the point $(0, 0, 0)$, is a vector subspace. So, the line shown in Figure 1.4 and the planes shown in Figures 1.5 and 1.8 are vector subspaces of $\mathbb{R}^3$. The line in Figure 1.6 is not a vector subspace of $\mathbb{R}^3$ because that line does not pass through the origin. $\bigcirc$

Caution: The *only* sets in $\mathbb{R}^3$ that are vector subspaces are $\{\mathbf{0}\}$, lines through the origin, planes through the origin, and $\mathbb{R}^3$ itself. So, a set can contain the origin but still not be a vector subspace.

Using the definition of a vector subspace is, at best, tedious work, just as tedious as using the definition of a vector space. Fortunately, there are theorems that save time and effort.

Theorem 1.37

If $\mathcal{V}$ is a vector space, $\mathcal{W}$ is a subset of vectors in $\mathcal{V}$, and $\mathcal{W}$ satisfies the two "closure" axioms

www.TechnicalBooksPDF.com

(CloseAdd) If $\mathbf{x}$ and $\mathbf{y}$ are in $\mathcal{W}$, then $\mathbf{x}+\mathbf{y}$ is in $\mathcal{W}$,

(CloseMult) If $\mathbf{x}$ is in $\mathcal{W}$ and α is a scalar, then $\alpha\mathbf{x}$ is in $\mathcal{W}$,

then $\mathcal{W}$ is a vector subspace of $\mathcal{V}$.

Theorem 1.38

If $\mathcal{V}$ is a vector space, $\mathcal{W}$ is a subset of vectors in $\mathcal{V}$, and $\mathcal{W}$ satisfies the "closure" axiom

(CloseLin) If $\mathbf{x}$ and $\mathbf{y}$ are in $\mathcal{W}$ and α and β are scalars, then $\alpha\mathbf{x}+\beta\mathbf{y}$ is in $\mathcal{W}$,

then $\mathcal{W}$ is a vector subspace of $\mathcal{V}$.

Theorem 1.39

Let $\mathcal{V}$ be the vector space $\mathbb{R}^n$, that is, Example 1.44's vector space of all real n vectors, with its usual operations of addition and scalar multiplication. Suppose A is a real, $m\times n$ matrix. Define $\mathcal{W}\triangleq\{\mathbf{x}\text{ in }\mathbb{R}^n:\ A\mathbf{x}=\mathbf{0}\}$. Then $\mathcal{W}$ is a vector subspace of $\mathbb{R}^n$.

Why? Let $\mathbf{x}$ and $\mathbf{y}$ be any unspecified vectors in $\mathcal{W}$ and let α and β be any unspecified scalars, that is, real numbers. We will explain why $\alpha\mathbf{x}+\beta\mathbf{y}$ being in $\mathcal{W}$ follows from $\mathbf{x}$ and $\mathbf{y}$ being in $\mathcal{W}$: Because $\mathbf{x}$ is in $\mathcal{W}$, we have $A\mathbf{x}=\mathbf{0}$; similarly, $A\mathbf{y}=\mathbf{0}$. By linearity, that is, Theorem 1.15 in Section 1.3, we have

$$A(\alpha\mathbf{x}+\beta\mathbf{y})=\alpha A(\mathbf{x})+\beta A(\mathbf{y})=\alpha\mathbf{0}+\beta\mathbf{0}=\mathbf{0};$$

hence, $\alpha\mathbf{x}+\beta\mathbf{y}$ is in $\mathcal{W}$. Theorem 1.38 explains why $\mathcal{W}$ is a vector subspace of $\mathbb{R}^n$. □

Definition 1.31

(a) A subspace $\mathcal{W}$ has a **basis** $\{\mathbf{v}_1,\ldots,\mathbf{v}_\ell\}$ if $\{\mathbf{v}_1,\ldots,\mathbf{v}_\ell\}$ is linearly independent and $Span\{\mathbf{v}_1,\ldots,\mathbf{v}_\ell\}=\mathcal{W}$.

(b) The **dimension** of a vector subspace $\mathcal{W}$ is the number of vectors in a basis for $\mathcal{W}$.

Because of Theorem 1.39, we see that Definition 1.29 is a special case of Definition 1.31 because Definition 1.29 refers to a vector subspace that is defined to be the set of solutions of a homogeneous system.

Theorem 1.40

(a) For a given vector subspace $\mathcal{W}$, every basis has the same number of vectors.

(b) If $\mathcal{W}$ is a subspace of $\mathbb{R}^n$, then there exists a basis for $\mathcal{W}$.

www.TechnicalBooksPDF.com

Because of this theorem, the definition of dimension makes sense: Theorem 1.40(a) says we cannot get two different values for the dimension of a given vector subspace by using two bases having different numbers of vectors. Theorem 1.40(b) says we can find a basis; therefore, we can calculate the dimension of a given vector subspace.

Theorem 1.41

If a vector subspace $\mathcal{W}$ in $\mathbb{R}^n$ is spanned by a set of vectors $\{\mathbf{a}_1, \ldots, \mathbf{a}_\ell\}$, form the $n \times \ell$ matrix $A = [\mathbf{a}_1 \quad \cdots \quad \mathbf{a}_\ell]$ and row reduce it to a row echelon form. If the pivot columns are the $i_1, \ldots, i_\ell$ columns, then $\{\mathbf{a}_{i_1}, \ldots, \mathbf{a}_{i_\ell}\}$ is a basis for $\mathcal{W}$.

Theorem 1.42

(Goldilocks and the sets of vectors): Suppose $\mathcal{W}$ is a vector subspace and $dim(\mathcal{W}) = \ell$. Suppose $\mathcal{E}$ is a set of $\ell - 1$ vectors in $\mathcal{W}$ and $\mathcal{G}$ is a set of $\ell + 1$ vectors in $\mathcal{W}$. Then we know for sure that neither $\mathcal{E}$ nor $\mathcal{G}$ can be a basis for $\mathcal{W}$.

Why? Because every basis for $\mathcal{W}$ should have exactly ℓ vectors. □

Intuitively, $\mathcal{E}$, a set of $\ell - 1$ vectors, is not a basis for $\mathcal{W}$ because $\mathcal{E}$ does not have enough directions, that is, $\mathcal{E}$ cannot span $\mathcal{W}$. Intuitively, the reason why $\mathcal{G}$, a set of $\ell + 1$ vectors, is not a basis for $\mathcal{W}$ is that $\mathcal{G}$ has too many directions, that is, $\mathcal{G}$ is not linearly independent. It's as if Goldilocks tried $\mathcal{E}$, but that was too small, and then tried $\mathcal{G}$, but that was too big. In the fairy tale, "one bed was just right."

But note that if $\mathcal{F}$ is a set of ℓ vectors in $\mathcal{W}$, $\mathcal{F}$ may or may not be a basis for $\mathcal{W}$; even though $\mathcal{F}$ has the right *number* of vectors, they may not be linearly independent or may not span $\mathcal{W}$.

As a special case, consider $\mathcal{W} = \mathbb{R}^n = \mathcal{V}$, that is, the case where the vector subspace is the whole space. We have two useful results. The first one is particularly beloved by students because it can be easy to apply. Indeed, we will use it in Section 2.2 and often in Chapter 5.

Theorem 1.43

Suppose $\mathcal{S} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a set of n vectors in $\mathbb{R}^n$. Then the following are logically equivalent:

(a) $\mathcal{S}$ is a basis for $\mathbb{R}^n$.
(b) $\mathcal{S}$ is linearly independent.
(c) $\mathcal{S}$ spans $\mathbb{R}^n$.
(d) $|\,\mathbf{v}_1 \quad \cdots \quad \mathbf{v}_n\,| \neq 0$.

Why? From the given vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$, form the matrix $A \triangleq \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_n\end{bmatrix}$. Recall that Theorem 1.30 says that $|A| \neq 0 \iff rank(A) = n \iff A^{-1}$ exists.

www.TechnicalBooksPDF.com

As to linear independence of the columns of A, Lemma 1.1 tells us that

$$\mathbf{0} = [\,\mathbf{v}_1 \quad \cdots \quad \mathbf{v}_n\,]\,\boldsymbol{\alpha} = A\boldsymbol{\alpha}$$

has its only solution being $\boldsymbol{\alpha} = \mathbf{0}$ if, and only if, $rank(A) = n$, if and only if $|A| \neq 0$. The set of vectors $\mathcal{S}$ is linearly independent if, and only if, $|A| \neq 0$.

As to whether the set of vectors $\mathcal{S}$ spans $\mathbb{R}^n$, that is, whether for every $\mathbf{b}$, there exist scalars $c_1, \ldots, c_n$ so that $\mathbf{b} = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n = A\mathbf{c}$, we know there is a solution, $\mathbf{c} = A^{-1}\mathbf{b}$, when A^{-1} exists, that is, when $|A| \neq 0$. But if $|A| = 0$, that is, $rank(A) < n$, then* the system $A\mathbf{c} = \mathbf{b}$ fails to have a solution, $\mathbf{c}$, for at least one choice of the vector $\mathbf{b}$; hence, $\mathcal{S}$ fails to span $\mathbb{R}^n$. □

Corollary 1.3

A set of vectors $\mathcal{S} = \{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ in $\mathbb{R}^n$

(a) Is linearly *dependent* if $p > n$
(b) Fails to span $\mathbb{R}^n$ if $p < n$

Note that Corollary 1.3 gives more specific information than the (Goldilocks) Theorem 1.42.

Theorem 1.44

For any $m \times n$ matrix A, $rank(A^T) = rank(A)$.

Learn More About It

A good discussion of Theorem 1.44 is in Section 1.15 of *Theory of Matrices*, by Peter Lancaster, Academic Press, © 1969. Chapter 4 of *Linear Algebra and Its Applications*, 3rd ed., by David C. Lay, Pearson, © 2003, has a very good discussion of the axioms of vector spaces, and the whole book is an excellent resource for matrix algebra.

1.7.3 Problems

1. Find a criterion satisfied by all vectors $\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$ that are in *Span* $\left\{ \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -2 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \\ 3 \end{bmatrix} \right\}$.

* Suppose $E_q, E_{q-1}, \ldots, E_2, E_1$ are elementary matrices that row reduce A to its RREF, that is, $E_qE_{q-1}\cdots E_2E_1A = \text{RREF}(A)$. Let $\mathbf{b} = E_q^{-1}\cdots E_1^{-1}\mathbf{e}^{(n)}$, so $[\,A\,|\,\mathbf{b}\,] \sim \left[\,\text{RREF}(A)\,|\,\mathbf{e}^{(n)}\,\right]$ explains why $A\mathbf{x} = \mathbf{b}$ has no solution for this vector $\mathbf{b}$ because $\text{RREF}(A) \neq I_n$ implies that at least the bottom row of RREF(A) is all zeros.

www.TechnicalBooksPDF.com

2. For what value(s) of k do the vectors

$$\left\{\begin{bmatrix}1\\0\\-1\end{bmatrix},\begin{bmatrix}1\\k\\0\end{bmatrix},\begin{bmatrix}0\\-2\\k\end{bmatrix}\right\}$$

span $\mathbb{R}^3$?

3. Find the exact value(s) of t for which the set of vectors

$$\left\{\begin{bmatrix}1\\-1\\2\end{bmatrix},\begin{bmatrix}t\\2\\3\end{bmatrix},\begin{bmatrix}0\\1\\t\end{bmatrix}\right\}$$

is linearly dependent.

4. Are the vectors

$$\begin{bmatrix}1\\-1\\1\end{bmatrix},\begin{bmatrix}-2\\3\\-1\end{bmatrix},\begin{bmatrix}0\\1\\1\end{bmatrix}$$

linearly independent? If so, why? If not, explicitly write one of the three vectors as a linear combination of the other two vectors.

5. For each of the sets of vectors as follows, determine whether the set is linearly independent, spans $\mathbb{R}^3$, and/or is a basis for $\mathbb{R}^3$, and justify your conclusions.
   (a) $\left\{\begin{bmatrix}1\\1\\1\end{bmatrix},\begin{bmatrix}3\\1\\1\end{bmatrix},\begin{bmatrix}1\\-1\\-1\end{bmatrix}\right\}$
   (b) $\left\{\begin{bmatrix}1\\1\\0\end{bmatrix},\begin{bmatrix}1\\0\\1\end{bmatrix},\begin{bmatrix}0\\1\\1\end{bmatrix}\right\}$
   (c) $\left\{\begin{bmatrix}1\\1\\0\end{bmatrix},\begin{bmatrix}1\\0\\1\end{bmatrix}\right\}$
   (d) $\left\{\begin{bmatrix}1\\2\\3\end{bmatrix},\begin{bmatrix}0\\1\\4\end{bmatrix},\begin{bmatrix}-1\\0\\5\end{bmatrix},\begin{bmatrix}1\\-3\\0\end{bmatrix}\right\}$

6. Suppose $A = \begin{bmatrix}\mathbf{v}^{(1)} & \mathbf{v}^{(2)} & \mathbf{v}^{(3)} & \mathbf{v}^{(4)}\end{bmatrix}$ is a 6×4 matrix and $\left\{\mathbf{v}^{(2)}, \mathbf{v}^{(3)}\right\}$ is a linearly independent set. Determine all possible values of $rank(A)$ and state your conclusions in the form $\_\_ \leq rank(A) \leq \_\_$. Explain how you arrived at your conclusions.

7. Suppose $A = \begin{bmatrix}\mathbf{v}^{(1)} & \mathbf{v}^{(2)} & \mathbf{v}^{(3)} & \mathbf{v}^{(4)} & \mathbf{v}^{(5)}\end{bmatrix}$ is a 4×5 matrix and $\left\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}\right\}$ is a linearly independent set. Fill in the blanks as follows and explain how you arrived at your conclusions.
   (a) Determine all possible values of $rank(A)$ and state your conclusions in the form $\_\_ \leq rank(A) \leq \_\_$. Explain how you arrived at your conclusions.
   (b) Determine all possible values of the $nullity(A)$ and state your conclusions in the form $\_\_ \leq \nu(A) \leq \_\_$. Explain how you arrived at your conclusions.

www.TechnicalBooksPDF.com

8. Suppose A is an invertible 3×3 matrix and $\{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$ is a basis for $\mathbb{R}^3$.
   (a) Why is $B \triangleq [\mathbf{b}_1 \quad \mathbf{b}_2 \quad \mathbf{b}_3]$ invertible?
   (b) Why is $|AB| \neq 0$?
   (c) Use parts (a) and (b) and Theorem 1.9 in Section 1.2 to explain why $\{A\mathbf{b}_1, A\mathbf{b}_2, A\mathbf{b}_3\}$ is a basis for $\mathbb{R}^3$.
9. For the matrix $A = \begin{bmatrix} 1 & -1 & 2 \\ 3 & 1 & 0 \end{bmatrix}$, find
   (a) $rank(A)$, (b) $\nu(A)$, and (c) a basis for $\mathcal{W} \triangleq \{\mathbf{x} \text{ in } \mathbb{R}^3 : A\mathbf{x} = \mathbf{0}\}$.
10. Assume A is an 4×4 matrix that is not invertible. Which of the following *must* be true?
   (a) The columns of A are linearly independent.
   (b) There is no 4×4 matrix B such that $AB = I_4$.
   (c) The system $A\mathbf{x} = \mathbf{0}$ has no solution.
   (d) Two columns of A are identical.
   (e) There is a vector $\mathbf{x}$ in $\mathbb{R}^4$ with $\mathbf{x} \neq \mathbf{0}$ and $A\mathbf{x} = \mathbf{0}$.
   (f) A has four pivot positions.
11. (a) Find an example of a 2×3 matrix A such that the nullity of A^T does not equal the nullity of A.
   (b) Why does A being square imply that the nullity of A^T must equal the nullity of A? [Hint: Theorem 1.44 is relevant.]
12. Explain why Theorem 1.35 is true.

Key Terms

adjugate: Definition 1.27 in Section 1.6
augmented matrix: before Lemma 1.1
basic solution: Definition 1.19 in Section 1.3
basic variable: Definition 1.5 in Section 1.1.3
basis: Definitions 1.29 and 1.31 in Section 1.7
complete set of basic solutions: Definition 1.19 in Section 1.3
Cramer's rule: Theorem 1.34 in Section 1.6.2
determinant: Definitions 1.24 through 1.26 in Section 1.6
diagonal matrix: Definition 1.11 in Section 1.2
dimension: Definition 1.31 in Section 1.7
elementary column operations: after Theorem 1.31 in Section 1.6
elementary row operations: Definition 1.3 in Section 1.1
equivalent system: after (1.1) in Section 1.1
free variable: Definition 1.5 in Section 1.1.3
Gaussian elimination method: before Example 1.2 in Section 1.1
Gauss–Jordan method: after Definition 1.4 in Section 1.1
general linear combination: Definition 1.16 in Section 1.3
general solution: Definition 1.17 in Section 1.3, Definition 1.22 in Section 1.4
homogeneous system: Definition 1.15 in Section 1.3
identity matrix: Definition 1.9 in Section 1.2

www.TechnicalBooksPDF.com

inverse, invertible: Definition 1.23 in Section 1.5
leading entry: before Definition 1.1 in Section 1.1
leading 1: before Definition 1.1 in Section 1.1
linearly dependent: Definition 1.28 in Section 1.7
linearly independent: Definition 1.28 in Section 1.7
lower triangular: Definition 1.12 in Section 1.2
matrix: before Definition 1.1 in Section 1.1
matrix addition: beginning of Section 1.2
matrix multiplication: Definition 1.8 in Section 1.2
nonhomogeneous system: Section 1.4
non-invertible: Definition 1.23 in Section 1.5
non-singular: Definition 1.23 in Section 1.5
nonzero row: before Definition 1.1 in Section 1.1
nullity: Definition 1.20 in Section 1.3
particular solution: Definition 1.21 in Section 1.4
partition a matrix into blocks: Section 1.5.2; Example 1.32 in Section 1.5
partitioned matrix: before Lemma 1.2 in Section 1.2
pivot column: before Definition 1.2 in Section 1.1
pivot positions: Definition 1.1 in Section 1.1
powers: Definition 1.13 in Section 1.2
$\mathbb{R}^n$: before Definition 1.7 in Section 1.2
rank: Definition 1.2 in Section 1.1
row echelon form: Definition 1.1 in Section 1.1
row echelon form of *C*: after Definition 1.3 in Section 1.1
row reduced echelon form (RREF): Definition 1.4 in Section 1.1.3
singular: Definition 1.23 in Section 1.5
solution: after (1.1) in Section 1.1
spanning set: Definition 1.18 in Section 1.3
square: before Definition 1.1 in Section 1.1
transpose: Definition 1.14 in Section 1.2
trivial solution: Theorem 1.14 in Section 1.3
upper triangular: Definition 1.12 in Section 1.2
Vandermonde determinant: Problem 1.6.3.7
vector space: Section 1.7.2
vector subspace: Definition 1.30 in Section 1.7
zero matrix: Definition 1.10 in Section 1.2.1
zero row: before Definition 1.1 in Section 1.1

MATLAB® Commands

RREF: Definition 1.4 in Section 1.1

Mathematica Commands

ParametricPlot3D`[{u + v, u - v, -u + v}, {u, -5, 5}, {v, -5, 5}]`:
`after Ex. 1.41`

www.TechnicalBooksPDF.com

References

Carlson, D. Teaching linear algebra: Must the fog always roll in? *College Mathematics Journal* 24, 29–40, 1993.

Lederer, E.M. (ed.) *Linear Algebra Exam File*. Engineering Process, Inc., San Jose, CA, 1989.

Timoshenko, S., Young, D.H., and Kleaver, W., *Vibration Problems in Engineering*, 4th edn. John Wiley & Sons, New York, 1990.

Yuster, T. The reduced row echelon form of a matrix is unique: A short proof. *Mathematics Magazine* 57, 93–94, 1984.

2

Matrix Theory

2.1 Eigenvalues and Eigenvectors

This topic is both at the heart of matrix theory and essential to the study of many topics in engineering and science. We will see that Chapter 5 relies on this topic to solve systems of linear ordinary differential equations. In Chapter 11, we will see how the same concept applies to solving linear partial differential equations.

Example 2.1

Let $A = \begin{bmatrix} -4 & -2 \\ 6 & 3 \end{bmatrix}$, $\mathbf{x}^{(1)} = \begin{bmatrix} 2 \\ -3 \end{bmatrix}$, and $\mathbf{x}^{(2)} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$. We have

$$A\mathbf{x}^{(1)} = \begin{bmatrix} -4 & -2 \\ 6 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ -3 \end{bmatrix} = \begin{bmatrix} -2 \\ 3 \end{bmatrix} = (-1) \begin{bmatrix} 2 \\ -3 \end{bmatrix} = (-1)\mathbf{x}^{(1)}, \text{ and}$$

$$A\mathbf{x}^{(2)} = \begin{bmatrix} -4 & -2 \\ 6 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = (0) \begin{bmatrix} 1 \\ -2 \end{bmatrix} = (0)\mathbf{x}^{(2)}.$$

If we denote $\lambda_1 = -1$ and $\lambda_2 = 0$, we have

$$A\mathbf{x}^{(1)} = \lambda_1 \mathbf{x}^{(1)} \quad \text{and} \quad A\mathbf{x}^{(2)} = \lambda_2 \mathbf{x}^{(2)}. \bigcirc$$

These fit the following definition, which involves a universal concept and "tool" of engineering, science, and mathematics:

Definition 2.1

A square matrix A has **eigenvalue** λ if the system $A\mathbf{x} = \lambda\mathbf{x}$ has at least one non-trivial solution for $\mathbf{x}$, in which case we call such an $\mathbf{x} \neq \mathbf{0}$ a corresponding **eigenvector**.

So, in Example 2.1, $A = \begin{bmatrix} -4 & -2 \\ 6 & 3 \end{bmatrix}$ has eigenvalue $\lambda_1 = -1$, with corresponding eigenvector $\begin{bmatrix} 2 \\ -3 \end{bmatrix}$, and eigenvalue $\lambda_2 = 0$, with corresponding eigenvector $\begin{bmatrix} 1 \\ -2 \end{bmatrix}$.

Geometrically, A has an eigenvector $\mathbf{x}$ if $A\mathbf{x}$ is either parallel to $\mathbf{x}$ or is $\mathbf{0}$. If λ is the corresponding eigenvalue, then $|\lambda|$ is the magnification of $A\mathbf{x}$ versus $\mathbf{x}$; if λ is negative then $A\mathbf{x}$ points in the direction opposite to that of $\mathbf{x}$. The results of Example 2.1 are summarized pictorially in Figure 2.1.

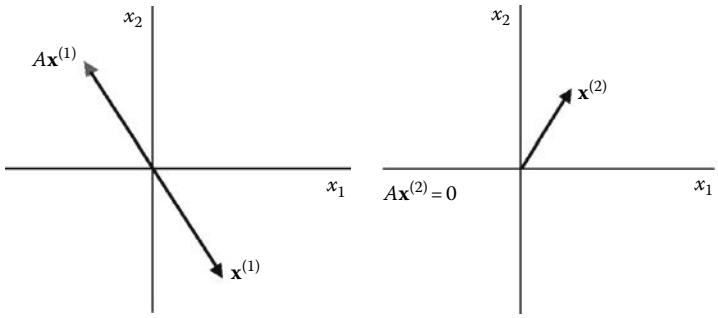


FIGURE 2.1
Geometry of eigenvectors.

Theorem 2.1

λ is an eigenvalue of A if, and only if, $0 = |A - \lambda I|$.

Why? First, let's rewrite $A\mathbf{x} = \lambda\mathbf{x}$:

$$A\mathbf{x} = \lambda\mathbf{x} \iff \mathbf{0} = A\mathbf{x} - \lambda\mathbf{x} = A\mathbf{x} - \lambda I_n\mathbf{x} = (A - \lambda I_n)\mathbf{x}.$$

So,

$$\lambda \text{ is an eigenvalue of } A \iff (A - \lambda I)\mathbf{x} = \mathbf{0} \text{ has a solution } \mathbf{x} \neq \mathbf{0}. \tag{2.1}$$

But, by Theorems 1.30 in Section 1.6 and 1.25 in Section 1.5, the homogeneous system $(A - \lambda I)\mathbf{x} = \mathbf{0}$ has a solution $\mathbf{x} \neq \mathbf{0}$ if, and only if, the matrix $(A - \lambda I)$ has zero determinant. □

Corollary 2.1

A is invertible if, and only if, 0 is not an eigenvalue of A.

Why? A is invertible $\iff |A| \neq 0 \iff |A - 0 \cdot I| \neq 0 \iff \lambda = 0$ is not an eigenvalue of A. □

Definition 2.2

The **characteristic equation** for an $n \times n$ matrix A is $0 = |A - \lambda I_n|$, and the **characteristic polynomial** for A is the function $\mathcal{P}(\lambda) \triangleq |A - \lambda I_n|$.

Example 2.2

(Example 2.1 again) Find all of the eigenvalues of $A = \begin{bmatrix} -4 & -2 \\ 6 & 3 \end{bmatrix}$.

Method:

$$0 = |A - \lambda I_2| = \left| \begin{bmatrix} -4 & -2 \\ 6 & 3 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = \begin{vmatrix} -4-\lambda & -2 \\ 6 & 3-\lambda \end{vmatrix}$$

$$= (-4-\lambda)(3-\lambda) - (-2)(6) = \lambda^2 + \lambda = \lambda(\lambda+1).$$

The eigenvalues λ satisfy $0 = \lambda$ or $0 = \lambda + 1$. The eigenvalues are $\lambda = 0$ and $\lambda = -1$. ◯

Example 2.3

Find *all* of the eigenvalues of $A = \begin{bmatrix} 2 & 2 & 4 \\ 2 & -1 & 2 \\ 4 & 2 & 2 \end{bmatrix}$.

Method:

$$0 = |A - \lambda I_3| = \begin{vmatrix} 2-\lambda & 2 & 4 \\ 2 & -1-\lambda & 2 \\ 4 & 2 & 2-\lambda \end{vmatrix}$$

$$\underset{-R_1 + R_3 \to R_3}{=} \begin{vmatrix} 2-\lambda & 2 & 4 \\ 2 & -1-\lambda & 2 \\ 2+\lambda & 0 & -2-\lambda \end{vmatrix}$$

$$\underset{R_3 \leftarrow (2+\lambda)R_3}{=} (2+\lambda) \begin{vmatrix} 2-\lambda & 2 & 4 \\ 2 & -1-\lambda & 2 \\ 1 & 0 & -1 \end{vmatrix}$$

$$= (2+\lambda)\left(1 \cdot \begin{vmatrix} 2 & 4 \\ -1-\lambda & 2 \end{vmatrix} + (-1) \cdot \begin{vmatrix} 2-\lambda & 2 \\ 2 & -1-\lambda \end{vmatrix}\right), \text{ by expanding along } R_3$$

$$= (2+\lambda)(4 - 4(-1-\lambda) - ((2-\lambda)(-1-\lambda) - 4)) = (2+\lambda)(-\lambda^2 + 5\lambda + 14)$$

$$= (2+\lambda)(2+\lambda)(7-\lambda).$$

The eigenvalues are $\lambda_1 = -2$, $\lambda_2 = -2$, $\lambda_3 = 7$. ◯

You may wonder why we bothered to list -2 twice in Example 2.3. It will turn out that we need the repetition of the eigenvalue both for a practical reason when solving differential equations and for theoretical reasons.

Note that we used elementary row operations to simplify the determinant. There are two primary reasons for this: expanding the determinant along a row or column gives the cubic equation $0 = -\lambda^3 + 3\lambda^2 + 24\lambda + 28$, for which factoring could be a chore. (Admittedly, we could use technology to find the roots of the characteristic equation pretty easily.) Second, if we can simplify an algebraic expression, we usually make fewer algebraic errors. You can test this yourself by seeing if you get the same characteristic equation by another method.

How did we know that $-R_1 + R_3 \to R_3$ would be so useful? We tried it because it created a 0 in the $(3,2)$ entry, which we thought might help. To tell the truth, using that particular elementary row operation turned out to be even more helpful than we imagined.

Similarly, if instead we had used the column operation $C_1 - C_2 \to C_1$, it also would have been very helpful.

The beauty of the characteristic equation is that, in theory, it finds the *exact* eigenvalues. But for "large" matrices, solving the characteristic equation can be a monster of a problem. We will study numerical methods for finding eigenvalues in Section 8.5. Certainly, we should solve the characteristic equation by hand only for *very* small matrices.

Nevertheless, exploring eigenvalues and eigenvectors of small matrices by hand is useful for all but the most abstract thinkers. This is a case where, in the long run, it helps to spend some effort "opening up the hood to see what's really going on" and "getting your hands dirty" with the details. There are quite a few concepts concerning eigenvalues and eigenvectors that are more easily understood in the context of simple examples.

Given a matrix and its eigenvalues, how do we find the corresponding eigenvectors? The answer is found in (2.1): Individually for each eigenvalue λ, solve the homogeneous system $(A - \lambda I_n)\mathbf{x} = \mathbf{0}$ for $\mathbf{x}$.

Example 2.4

Find all of the eigenvectors of the matrix $A = \begin{bmatrix} -4 & -2 \\ 6 & 3 \end{bmatrix}$:

Method: In Example 2.2, we found that $\lambda_1 = -1$, $\lambda_2 = 0$ gave all of the eigenvalues.

$$\left[A - \lambda_1 I_2 \mid \mathbf{0}\right] = \left[A - (-1)I_2 \mid \mathbf{0}\right] = \left[\begin{array}{cc|c} -4-(-1) & -2 & 0 \\ 6 & 3-(-1) & 0 \end{array}\right] = \left[\begin{array}{cc|c} -3 & -2 & 0 \\ 6 & 4 & 0 \end{array}\right]$$

$$\begin{matrix} \sim \\ -\frac{1}{3}R_1 \to R_1 \\ -6R_1 + R_2 \to R_2 \end{matrix} \left[\begin{array}{cc|c} \textcircled{1} & \frac{2}{3} & 0 \\ 0 & 0 & 0 \end{array}\right].$$

The solutions are $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -\frac{2}{3}c_1 \\ c_1 \end{bmatrix} = c_1 \begin{bmatrix} -\frac{2}{3} \\ 1 \end{bmatrix}$; this gives an eigenvector for any $c_1 \neq 0$.

Likewise,

$$\left[A - \lambda_2 I_2 \mid \mathbf{0}\right] = \left[A - (0)I_2 \mid \mathbf{0}\right] = \left[\begin{array}{cc|c} -4 & -2 & 0 \\ 6 & 3 & 0 \end{array}\right] \begin{matrix} \sim \\ -\frac{1}{4}R_1 \to R_1 \\ -6R_1 + R_2 \to R_2 \end{matrix} \left[\begin{array}{cc|c} \textcircled{1} & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{array}\right].$$

The solutions are $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}c_1 \\ c_1 \end{bmatrix} = c_1 \begin{bmatrix} -\frac{1}{2} \\ 1 \end{bmatrix}$.
This gives an eigenvector for any $c_1 \neq 0$. ◯

Some people prefer that eigenvectors be written as multiples of vectors without fractions and with as few negative signs as possible. So, in Example 2.4, we might summarize the eigenvectors as being

$$\mathbf{x} = \hat{c}_1\mathbf{x}^{(1)}, \text{where } \hat{c}_1 \neq 0 \text{ and } \mathbf{x}^{(1)} = \begin{bmatrix} 2 \\ -3 \end{bmatrix}, \text{ corresponding to } \lambda_1 = -1, \quad \text{and}$$

$$\mathbf{x} = \hat{c}_1\mathbf{x}^{(2)}, \text{where } \hat{c}_1 \neq 0 \text{ and } \mathbf{x}^{(2)} = \begin{bmatrix} 1 \\ -2 \end{bmatrix} \text{ corresponding to } \lambda_2 = 0.$$

For each eigenvalue λ_j, RREF$(A-\lambda_j I)$ *must* have a row of zeros. Indeed, if RREF$(A-\lambda_j I)$ does not have a row of zeros, then an "alarm bell" should ring in our heads because the system $(A-\lambda_j I)\mathbf{x}=\mathbf{0}$ would have only the trivial solution for $\mathbf{x}$, that is, there would be no eigenvector, hence λ_j would not be an eigenvalue.

So, if we fail to get an eigenvector, what went wrong? Perhaps we made an error in finding the eigenvalues, or in substituting them into $(A-\lambda_j I)\mathbf{x}=\mathbf{0}$, or in row reducing $\left[A-\lambda_j I \mid \mathbf{0}\right]$.

2.1.1 The Adjugate Matrix Method for Finding an Eigenvector

Theorem 2.2

(Adjugate matrix method) Suppose an $n\times n$ matrix A has an eigenvalue λ and $adj(A-\lambda I)$ is not the zero matrix. Then each nonzero column of $adj(A-\lambda I)$ is an eigenvector of A corresponding to eigenvalue λ.

Why? Because A has an eigenvalue λ, the $n\times n$ matrix $B \triangleq A-\lambda I$ has $|B|=0$. By Theorem 1.30 in Section 1.6, $rank(B)<n$. By Theorem 1.33b in Section 1.6, $rank(adj(B))\le 1$. Because $adj(B)=adj(A-\lambda I)$ is not the zero matrix, $rank(adj(B))\neq 0$. It follows that $rank(adj(B))=1$, so Theorem 1.33(b) in Section 1.6 implies that every nonzero column of $adj(B)$ is a nontrivial solution of $(A-\lambda I)\mathbf{x}=B\mathbf{x}=\mathbf{0}$, hence is an eigenvector of A corresponding to eigenvalue λ. □

Example 2.5

Find all of the eigenvectors of the matrix $A=\begin{bmatrix}2 & 2 & 4\\ 2 & -1 & 2\\ 4 & 2 & 2\end{bmatrix}$.

Method: In Example 2.3, we found that $\lambda_1=-2$, $\lambda_2=-2$, $\lambda_3=7$ gave all of the eigenvalues.

$$\left[A-\lambda_1 I_3 \mid \mathbf{0}\right]=\left[A-(-2)I_3 \mid \mathbf{0}\right]$$
$$=\begin{bmatrix}2-(-2) & 2 & 4 & \mid 0\\ 2 & -1-(-2) & 2 & \mid 0\\ 4 & 2 & 2-(-2) & \mid 0\end{bmatrix}=\begin{bmatrix}4 & 2 & 4 & \mid 0\\ 2 & 1 & 2 & \mid 0\\ 4 & 2 & 4 & \mid 0\end{bmatrix}.$$

We studied this in Example 1.42 in Section 1.7 and found all solutions. The eigenvectors are

$$\mathbf{x}=\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}=\begin{bmatrix}-\frac{1}{2}c_1-c_2\\ c_1\\ c_2\end{bmatrix}=c_1\begin{bmatrix}-\frac{1}{2}\\ 1\\ 0\end{bmatrix}+c_2\begin{bmatrix}-1\\ 0\\ 1\end{bmatrix},\ \text{any } c_1,c_2 \text{ with } |c_1|+|c_2|\neq 0.$$

Note that we don't have to repeat the above for λ_2 because $\lambda_2=-2=\lambda_1$.

As for the last eigenvalue,

$$\left[A-\lambda_3 I_3 \mid \mathbf{0}\right]=\left[A-(7)I_3 \mid \mathbf{0}\right]=\begin{bmatrix}-5 & 2 & 4 & \mid 0\\ 2 & -8 & 2 & \mid 0\\ 4 & 2 & -5 & \mid 0\end{bmatrix}.$$

Rather than do row reduction, we will use the "adjugate matrix method" of Theorem 2.2: because $\lambda_3 = 7$ is an eigenvalue of A, each nonzero column of $adj(A - 7I)$ is an eigenvector of A corresponding to eigenvalue $\lambda = 7$. We have from Example 1.38 in Section 1.6 that $adj(A - 7I) = \begin{bmatrix} 36 & 18 & 36 \\ 18 & 9 & 18 \\ 36 & 18 & 36 \end{bmatrix}$, so $\mathbf{x} = \begin{bmatrix} 36 \\ 18 \\ 36 \end{bmatrix}$ is an eigenvector of A corresponding to eigenvalue $\lambda = 7$. We can conclude that there are eigenvectors $\mathbf{x} = c_1 \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}$, any $c_1 \neq 0$. ◯

How did we know that the adjugate matrix method would work, that is, that $adj(A - 7I)$ would have a nonzero column? The truth is that we didn't know it would work for sure but we gave it a try. But, in Example 2.7 we will use a little more theory to be sure of ourselves. In this textbook, the role of theory is to organize the results, provide methods for solving things, know what to calculate, and know when it's o.k. to go ahead and start calculating.

Alternatively, we could have found an eigenvector corresponding to $\lambda = 7$ by using the usual row reduction:

$$\left[A - \lambda_3 I_3 \mid \mathbf{0}\right] = \left[A - (7)I_3 \mid \mathbf{0}\right] = \begin{bmatrix} -5 & 2 & 4 & \mid 0 \\ 2 & -8 & 2 & \mid 0 \\ 4 & 2 & -5 & \mid 0 \end{bmatrix} \sim \cdots \sim \begin{bmatrix} \textcircled{1} & 0 & -1 & \mid 0 \\ 0 & \textcircled{1} & -\frac{1}{2} & \mid 0 \\ 0 & 0 & 0 & \mid 0 \end{bmatrix}.$$

Multiplicities: First, we note that the characteristic polynomial of $n \times n$ matrix A,

$$\mathcal{P}(\lambda) \triangleq |\, A - \lambda I_n \,| - \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix},$$

is an n-th degree polynomial in λ. So,

> $\mathcal{P}(\lambda) = 0$ has exactly n roots $\lambda_1, \lambda_2, \ldots, \lambda_n$, possibly complex and possibly including repetition(s)

hence

$$\mathcal{P}(\lambda) = (\lambda_1 - \lambda)(\lambda_2 - \lambda) \cdots (\lambda_n - \lambda). \tag{2.2}$$

For example, in Example 2.3 we saw that $A = \begin{bmatrix} 2 & 2 & 4 \\ 2 & -1 & 2 \\ 4 & 2 & 2 \end{bmatrix}$ has

$$\mathcal{P}(\lambda) = (-2 - \lambda)(-2 - \lambda)(7 - \lambda).$$

Definition 2.3

If λ_j is an eigenvalue of A, the **algebraic multiplicity** of λ_j is the number of times $(\lambda_j - \lambda)$ appears as a factor of $\mathcal{P}(\lambda)$ in (2.2). We will denote by α_j the algebraic multiplicity of λ_j. An eigenvalue whose algebraic multiplicity is one is called **simple**. Some people refer to an eigenvalue whose algebraic multiplicity is greater than one as **degenerate**.

If $\mu_1, \mu_2, \ldots, \mu_p$ are the *distinct* eigenvalues of A, then

$$\mathcal{P}(\lambda) = (\mu_1 - \lambda)^{\alpha_1}(\mu_2 - \lambda)^{\alpha_2} \cdots (\mu_p - \lambda)^{\alpha_p}. \tag{2.3}$$

This way of writing the characteristic polynomial displays all of the eigenvalues *and* their algebraic multiplicities.

Example 2.6

For the matrix A of Examples 2.3 and 2.5, $\mathcal{P}(\lambda) = (-2 - \lambda)^2(7 - \lambda)^1$. ◯

Definition 2.4

If μ_j is an eigenvalue of A, the **geometric multiplicity** of μ_j is the nullity of the matrix $A - \mu_j I$, that is, $\nu(A - \mu_j I)$, namely, the number of basic solutions of the homogeneous system $(A - \mu_j I)\mathbf{x} = \mathbf{0}$. The geometric multiplicity of μ_j is denoted by m_j.

Theorem 2.3

Suppose $n \times n$ matrix A has distinct eigenvalues $\mu_1, \mu_2, \ldots, \mu_p$ and corresponding algebraic multiplicities $\alpha_1, \alpha_2, \ldots, \alpha_p$ and geometric multiplicities $m_1, m_2, \ldots, m_p$.
Then,

(a) $\alpha_1 + \alpha_2 + \ldots + \alpha_p = n$.
(b) $1 \leq m_j \leq \alpha_j$ for each $1 \leq j \leq p$.
(c) if $\alpha_j = 1$, then $m_j = 1$.

We recall from Theorem 1.17 in Section 1.3 that for each $1 \leq j \leq p$, $rank(A - \mu_j I) + \nu(A - \mu_j I) = n$. So, $rank(A - \mu_j I) = n - \nu(A - \mu_j I) =$ (the number of pivot columns of $A - \mu_j I$).

Example 2.7

For $A = \begin{bmatrix} 2 & 2 & 4 \\ 2 & -1 & 2 \\ 4 & 2 & 2 \end{bmatrix}$, that is, the matrix of Examples 2.3 and 2.5, find the algebraic and geometric multiplicities of the eigenvalues.

Method: From Example 2.6, we know that the distinct eigenvalues are $\mu_1 = -2$ and $\mu_2 = 7$ and that their algebraic multiplicities are $\alpha_1 = 2, \alpha_2 = 1$, respectively. From Theorem 2.3(c), we have $m_2 = 1$, that is, the geometric multiplicity of the eigenvalue $\lambda = 7$ must be one, hence the rank of $(A - 7I)$ is $3 - 1 = 2$. It follows from Theorem 2.2 that the adjugate matrix method works here, that is, that $adj(A - 7I)$ has a nonzero column of that is an eigenvector of A corresponding to eigenvalue $\lambda = 7$. ◯

Example 2.8

Suppose $\mathbf{x}$ is an eigenvector for each of the matrices A, B, C, corresponding to eigenvalues λ, β, γ, respectively. If $A^{-1}B = C$, find a relationship among λ, β, and γ.

Method: First, writing $C = A^{-1}B$ implicitly tells us that A must be invertible. From Corollary 2.1, it follows that $\lambda \neq 0$. Next, by definition of eigenvalue and eigenvector, we were given that $\mathbf{x} \neq \mathbf{0}$ and $A\mathbf{x} = \lambda\mathbf{x}$, $B\mathbf{x} = \beta\mathbf{x}$, and $C\mathbf{x} = \gamma\mathbf{x}$. But, $A\mathbf{x} = \lambda\mathbf{x}$ yields $\mathbf{x} = A^{-1}A\mathbf{x} = \lambda A^{-1}\mathbf{x}$, hence $A^{-1}\mathbf{x} = \lambda^{-1}\mathbf{x}$. So,

$$\gamma\mathbf{x} = C\mathbf{x} = (A^{-1}B)\mathbf{x} = A^{-1}(B\mathbf{x}) = A^{-1}(\beta\mathbf{x}) = \beta(A^{-1}\mathbf{x}) = \beta(\lambda)^{-1}\mathbf{x}.$$

Since $\mathbf{x} \neq \mathbf{0}$, it follows that $\gamma = \frac{\beta}{\lambda}$, which is a relationship among λ, β, and γ. ◯

2.1.2 Complex Numbers

First, let's note some facts about complex numbers: a number z is **complex** if $z = x + iy$, where x and y are real and $i \triangleq \sqrt{-1}$. [If you're an electrical engineer, use j instead of i.] For example, both $-1 + i2$ and $-1 + i0$ are complex. So, all real numbers are also complex.

Given a complex number $z = x + iy$, where x and y are real, we call x the **real part** of z and write $x = \mathcal{R}e(z)$, and we call y the **imaginary part** of z and write $y = \mathcal{I}m(z)$. If $\mathcal{I}m(z) = 0$, we say z is **real**. We denote $\overline{z} = x - iy$ and call it the **complex conjugate** of z. We have a basic fact: $\overline{wz} = \overline{w}\,\overline{z}$ for all complex numbers w, z. Complex numbers are illustrated in Figure 2.2.

We have two useful facts:

$$\mathcal{R}e(z) = \frac{z + \overline{z}}{2} \quad \text{and} \quad \mathcal{I}m(z) = \frac{z - \overline{z}}{2i}.$$

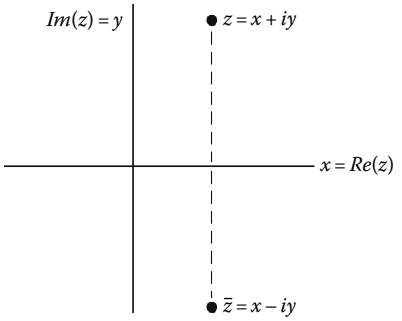


FIGURE 2.2
Complex numbers.

When we divide by a complex number, it helps to use the complex conjugate to "rationalize the denominator." For example,

$$\frac{1-i2}{3+i4}=\frac{1-i2}{3+i4}\cdot\frac{3-i4}{3-i4}=\frac{(1-i2)(3-i4)}{(3+i4)(3-i4)}=\frac{-5-i10}{3^2-(i4)^2}=\frac{-5-i10}{25}=-\frac{1}{5}-i\frac{2}{5}.$$

It helps to multiply by the complex conjugate of the denominator because if $z=x+iy$, then $z\bar{z}=(x+iy)(x-iy)=x^2-(iy)^2=x^2+y^2$ is real.

2.1.3 Complex Eigenvalues and Eigenvectors

Example 2.9

For $A=\begin{bmatrix}1 & -2\\ 4 & -3\end{bmatrix}$, find all of the eigenvalues and the corresponding eigenvectors.

Method: $0=|\,A-\lambda I_2\,|=\begin{vmatrix}1-\lambda & -2\\ 4 & -3-\lambda\end{vmatrix}=(1-\lambda)(-3-\lambda)-(-2)(4)=\lambda^2+2\lambda+5$, so by the quadratic formula the eigenvalues are

$$\lambda=\frac{-2\pm\sqrt{2^2-4\cdot1\cdot5}}{2\cdot1}=\frac{-2\pm i4}{2}=\frac{-2}{2}\pm\frac{i4}{2}=-1\pm i2,$$

To find the eigenvectors, use row reduction on $\left[A-\lambda I\mid\mathbf{0}\right]$:

$$\left[A-(-1+i2)I_2\mid\mathbf{0}\right]$$

$$=\begin{bmatrix}2-i2 & -2 & |\,0\\ 4 & -2-i2 & |\,0\end{bmatrix}\underset{\substack{R_1\leftrightarrow R_2\\ \frac{1}{4}R_1\to R_1\\ -(2-i2)R_1+R_2\to R_2}}{\sim}\begin{bmatrix}\textcircled{1} & -\frac{1}{2}-i\frac{1}{2} & |\,0\\ 0 & 0 & |\,0\end{bmatrix}.$$

Note that the last $(2,2)$ entry is $-(2-i2)\left(-\frac{1}{2}-i\frac{1}{2}\right)+(-2)=-(-1-1)-2=0$. The eigenvectors are $\mathbf{x}=c_1\begin{bmatrix}\frac{1}{2}+i\frac{1}{2}\\ 1\end{bmatrix}$, for any complex number $c_1\neq0$.

Likewise,

$$\left[A-(-1-i2)I_2\mid\mathbf{0}\right]$$

$$=\begin{bmatrix}2+i2 & -2 & |\,0\\ 4 & -2+i2 & |\,0\end{bmatrix}\underset{\substack{R_1\leftrightarrow R_2\\ \frac{1}{4}R_1\to R_1\\ -(2+i2)R_1+R_2\to R_2}}{\sim}\begin{bmatrix}\textcircled{1} & -\frac{1}{2}+i\frac{1}{2} & |\,0\\ 0 & 0 & |\,0\end{bmatrix}.$$

The eigenvectors are $\mathbf{x}=c_1\begin{bmatrix}\frac{1}{2}-i\frac{1}{2}\\ 1\end{bmatrix}$, for any complex number $c_1\neq0$. ◯

We notice that in Example 2.9, the eigenvalues came in a "complex conjugate pair": $\lambda_1 = -1 + i2$, $\lambda_2 = -1 - i2$. Not only that, the eigenvectors come in a "complex conjugate pair":

$$\begin{bmatrix} \frac{1}{2} + i\frac{1}{2} \\ 1 \end{bmatrix} \text{ and } \begin{bmatrix} \frac{1}{2} - i\frac{1}{2} \\ 1 \end{bmatrix}, \text{ that is,}$$

$$\begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix} + i \begin{bmatrix} \frac{1}{2} \\ 0 \end{bmatrix} \text{ and } \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix} - i \begin{bmatrix} \frac{1}{2} \\ 0 \end{bmatrix}.$$

Here we define complex conjugation of a vector or matrix "entry by entry," for example,

$$\overline{\begin{bmatrix} \frac{1}{2} + i\frac{1}{2} \\ 1 \end{bmatrix}} \triangleq \begin{bmatrix} \overline{\frac{1}{2} + i\frac{1}{2}} \\ \overline{1} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} - i\frac{1}{2} \\ 1 \end{bmatrix}.$$

This can be generalized:

Theorem 2.4

If A is real, then its eigenvalues and eigenvectors come in complex conjugate pairs. That is, if λ is an eigenvalue of A with corresponding eigenvector $\mathbf{z}$, then $\overline{\lambda}$ is an eigenvalue of A and $\overline{\mathbf{z}}$ is a corresponding eigenvector.

Why? First, A being real means $\overline{A} = A$, so

$$\overline{A - \lambda I} = \overline{A} - \overline{\lambda I} = A - \overline{\lambda} \overline{I} = A - \overline{\lambda} I.$$

If $\mathbf{z}$ is an eigenvector corresponding to A's eigenvalue λ, then $\mathbf{z} \neq \mathbf{0}$ and $\mathbf{0} = (A - \lambda I)\mathbf{z}$. It follows that

$$\overline{\mathbf{z}} \neq \mathbf{0} \text{ and } \mathbf{0} = \overline{\mathbf{0}} = \overline{(A - \lambda I)\mathbf{z}} = \overline{(A - \lambda I)}\, \overline{\mathbf{z}} = (A - \overline{\lambda} I)\overline{\mathbf{z}},$$

hence $\overline{\mathbf{z}}$ is an eigenvector of A corresponding to eigenvalue $\overline{\lambda}$. □

So, strangely enough, finding eigenvectors corresponding to complex eigenvalues of a real matrix A needs about half of the work of finding eigenvectors corresponding to real eigenvalues. That's because Theorem 2.4 tells us that once we have found an eigenvector corresponding to eigenvalue λ, we don't need to do row reduction to find an eigenvector corresponding to eigenvalue $\overline{\lambda}$. We will take advantage of this in Example 2.12 in Section 2.2.

2.1.4 Eigenvalues and Eigenvectors of Triangular and Diagonal Matrices

It's easy to find the eigenvalues and eigenvectors of a diagonal matrix. We'll start with an example.

Example 2.10

Find all eigenvalues and eigenvectors of $A = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$.

Method: $0 = |\ A - \lambda I_4\ | = \begin{vmatrix} 4-\lambda & 0 & 0 & 0 \\ 0 & 3-\lambda & 0 & 0 \\ 0 & 0 & 2-\lambda & 0 \\ 0 & 0 & 0 & 1-\lambda \end{vmatrix} = (4 - \lambda)(3 - \lambda)(2 - \lambda)(1 - \lambda)$ by Theorem 1.28(c) in Section 1.6, the fact that the determinant of a triangular matrix is the product of the diagonal entries. To find corresponding eigenvectors, row reduction is easy:

$$0 = \left[A - 4I_4 \mid \mathbf{0}\right] = \left[\begin{array}{cccc|c} 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & -3 & 0 \end{array}\right] \sim \left[\begin{array}{cccc|c} 0 & \textcircled{1} & 0 & 0 & 0 \\ 0 & 0 & \textcircled{1} & 0 & 0 \\ 0 & 0 & 0 & \textcircled{1} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right],$$

so it's easy to see that the eigenvectors corresponding to eigenvalue 4 are

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} c_1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = c_1 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \text{ for any } c_1 \neq 0.$$

Similarly, the eigenvectors corresponding to eigenvalue 3 are $c_1 \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$, for any $c_1 \neq 0$; the eigenvectors corresponding to eigenvalue 2 are $c_1 \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$, for any $c_1 \neq 0$; and the eigenvectors corresponding to eigenvalue 1 are $c_1 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$, for any $c_1 \neq 0$. ◯

By the way, for *any* $n \times n$ diagonal matrix, the eigenvectors are the columns of the identity matrix, that is, the vectors $\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \ldots, \mathbf{e}^{(n)}$ used in the discussion of the row reduction algorithm for constructing the inverse in Section 1.5.

Theorem 2.5

If A is an upper or lower triangular matrix, then the eigenvalues of A are the diagonal entries of A.

Additionally, in the special case when A is an $n \times n$ upper or lower triangular matrix and has n distinct eigenvalues, we can say a lot about the eigenvectors of A, as we will see in Problem 2.1.6.21.

Caution: These results sometimes mislead people. It would be natural to ask, "If we row reduce a matrix to a row echelon form, which is upper triangular, can we use that to find the eigenvalues of the original matrix?" Unfortunately, the answer is "No!" The eigenvalues of a matrix in RREF are all ones and zeros, so this idea of using a row echelon form would seem to say that every matrix has all of its eigenvalues being only ones and zeros, which is clearly nonsense. So, it's good to have a degree of skepticism about things that sound good but need further study and explanation.

2.1.5 MATLAB® and *Mathematica*™

The software package MATLAB was originally developed to apply robust, industrial-strength algorithms for working with matrices, which is where the MAT part of its name came from. Since then, the software package has developed in many directions.

The eig commands in MATLAB give approximate eigenvalues and eigenvectors. For example, for the matrix in Problem 2.1.6.14, we entered

```
>A=[-10.07106781,-5.656854249,0;5.656854249, 4.071067812,0;0,0,3]
```

and `>[V,D]=eig(A)` and got

$$\mathtt{V} = \begin{matrix} -0.8944 & 0.4472 & 0 \\ 0.4472 & -0.8944 & 0 \\ 0 & 0 & 1.0000 \end{matrix}$$

$$\mathtt{D} = \begin{matrix} -7.2426 & 0 & 0 \\ 0 & 1.2426 & 0 \\ 0 & 0 & 3.0000 \end{matrix}$$

This says that the approximate eigenvalues of A are `-7.2426,1.2426,3.0000`, and corresponding approximate eigenvectors are

$$\begin{bmatrix} -0.8944 \\ 0.4472 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.4472 \\ -0.8944 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1.0000 \end{bmatrix}.$$

Eigenvectors produced by MATLAB are all **normalized** to have length one, for example, $(-0.8944)^2 + 0.4472^2 + 0^2 \approx 0.9999392$. We can ask for more *displayed* accuracy by earlier giving the command `>format long`, so

$$\mathtt{V} = \begin{matrix} -0.89442719098937 & 0.44721359552104 & 0 \\ 0.44721359552104 & -0.89442719098937 & 0 \\ 0 & 0 & 1.00000000000000 \end{matrix}$$

$$\mathtt{D} = \begin{matrix} -7.24264068533334 & 0 & 0 \\ 0 & 1.24264068733334 & 0 \\ 0 & 0 & 3.00000000000000 \end{matrix}$$

The normalization we see is much more accurate:

$$(-0.89442719098937)^2 + 0.44721359552104^2 + 0^2 \approx 1.00000000000000.$$

Some versions of MATLAB also come with a symbolic manipulation toolbox, for example, MAPLE, in which case we can get exact eigenvalues and eigenvectors.

In *Mathematica*, we entered $\mathbf{A} = \{\{-3-5\sqrt{2}, -4\sqrt{2}, 0\}, 4\sqrt{2}, -3+5\sqrt{2}, 0\}, \{0, 0, 3\}\}$ and then **`Eigenvalues[A]`**, and got

$$\left\{3(-1-\sqrt{2}), 3, 3(-1+\sqrt{2})\right\}.$$

We entered **`Eigenvectors[A]`** and got

$$\left\{\{-2, 1, 0\}, \{0, 0, 1\}, \left\{\frac{-1}{2}, 1, 0\right\}\right\}.$$

The *Mathematica* command **`Eigenvalues[A,k]`** produces the first k eigenvalues of A, where "first" means the largest in absolute value. For example, if a 4×4 matrix has eigenvalues $-3, 2, 2, 1$ then the first two eigenvalues are -3 and 2. The *Mathematica* command Eigenvectors[A, k] produces the corresponding eigenvectors.

Each of the *Mathematica* commands **`Eigensystem[A]`** and **`Eigensystem[A,k]`** gives both eigenvalues and eigenvectors.

Many calculators these days will produce approximate eigenvalues and eigenvectors, and some more expensive models give exact results, that is, do symbolic manipulation.

2.1.6 Problems

In problems 1–9, for each matrix, find all *exact* eigenvalues and eigenvectors.

1. $\begin{bmatrix} -2 & 7 \\ 1 & 4 \end{bmatrix}$
2. $\begin{bmatrix} 1 & 2 \\ 3 & 2 \end{bmatrix}$
3. $\begin{bmatrix} -1 & 4 \\ 1 & 1 \end{bmatrix}$
4. $\begin{bmatrix} -2 & -5 \\ 1 & 0 \end{bmatrix}$
5. $\begin{bmatrix} 1 & 1 & 2 \\ -1 & 3 & 2 \\ 1 & 1 & 2 \end{bmatrix}$

[Hint: the eigenvalues are $0, 2$, and 4.]

6. $\begin{bmatrix} 1 & -3 & 6 \\ 0 & -2 & 0 \\ 2 & 0 & 0 \end{bmatrix}$

[Hint: -2 is an eigenvalue]

7. $\begin{bmatrix} -3 & 0 & 0 \\ 4 & -4 & -3 \\ -1 & 1 & 0 \end{bmatrix}$
8. $\begin{bmatrix} 6 & 1 & 4 \\ -4 & 1 & -4 \\ -1 & -1 & 0 \end{bmatrix}$
9. $\begin{bmatrix} 1 & 0 & 0 \\ 2 & 3 & 1 \\ -1 & 2 & 5 \end{bmatrix}$
10. Suppose a, b, c are unspecified but *distinct, nonzero* scalars. For $\begin{bmatrix} a & b & 0 \\ 0 & a & b \\ 0 & 0 & c \end{bmatrix}$,
    a. Find the characteristic equation.
    b. Find the eigenvalues and their corresponding algebraic and geometric multiplicities.
11. For which values of λ does $A\mathbf{x} = \lambda B\mathbf{x}$ have a nontrivial solution $\mathbf{x}$, where

$$A = \begin{bmatrix} -1 & 0 & 5 \\ 2 & 1 & 4 \\ 3 & -2 & 3 \end{bmatrix} \text{ and } B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}?$$

12. Suppose $A = \begin{bmatrix} 4 & -1 & 2 \\ 2 & 1 & 4 \\ -1 & 0 & 5 \end{bmatrix}$ has an eigenvector $\begin{bmatrix} 2 \\ 4 \\ 1 \end{bmatrix}$ and has characteristic polynomial $\mathcal{P}(\lambda) = -\lambda^3 + 10\lambda^2 - 33\lambda + 36$. Use all of the aforementioned information to find *all* of the eigenvalues of A. [Hint: use the definition of the word eigenvector.]
13. Suppose $A = \begin{bmatrix} 4 & 0 & 10 \\ -5 & -6 & -5 \\ 5 & 0 & -1 \end{bmatrix}$ has eigenvectors $\begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$. Use these eigenvectors to find three, not necessarily distinct, eigenvalues of A. State their geometric and algebraic multiplicities and explain how you arrived at those multiplicities. Also, can A have another eigenvalue other than the ones you found? Why or why not?
14. Suppose

$$A = \begin{bmatrix} -3 - 5\sqrt{2} & -4\sqrt{2} & 0 \\ 4\sqrt{2} & -3 + 5\sqrt{2} & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

Find all exact eigenvalues and eigenvectors of A.

15. a. Can a vector $\mathbf{x}$ be an eigenvector for two unequal eigenvalues λ_1 and λ_2 for the same matrix A? Answer "Yes" or "No" and justify that conclusion. How? If your answer is "Yes," give a specific example. If your answer is "No," explain why.
    b. Can a nonzero vector $\mathbf{x}$ be an eigenvector for two unequal eigenvalues λ_1 and λ_2 corresponding to two different matrices A and B, respectively? Answer "Yes" or "No" and justify that conclusion. How? If your answer is "Yes," give a specific example. If your answer is "No," explain why.
16. Suppose an $n \times n$ matrix A is invertible, $B = 2I_n - A$, B is invertible, $C = B^{-1}$, λ is an eigenvalue of A, and $\lambda \neq 2$.
    a. Find an eigenvalue γ for C, in terms of λ, and explain how you found it.
    b. Can $\gamma = \frac{1}{2}$? Why, or why not?
17. Suppose A and B are $n \times n$ matrices and $C = A + B$ and $\mathbf{x}$ is an eigenvector for *both* A and B, corresponding to λ and β, respectively.
    a. Use $\mathbf{x}$ to find an eigenvalue γ for C, in terms of λ and β.
    b. If, in addition, $C\mathbf{x} = A^2\mathbf{x}$, use part (a) to explain why either $\lambda = \frac{1}{2}(1 + \sqrt{1 + 4\beta})$ or $\lambda = \frac{1}{2}(1 - \sqrt{1 + 4\beta})$.
18. Suppose $\mathbf{x}$ is an eigenvector for both of the $n \times n$ matrices A and B, corresponding to eigenvalues λ and β, respectively. Must $\mathbf{x}$ be an eigenvector of AB? If so, corresponding to what eigenvalue?
19. Suppose A is square and for some vector $\mathbf{b}$, $A\mathbf{x} = \mathbf{b}$ has two distinct solutions. (a) Must $\lambda = 0$ be an eigenvalue of A? Why, or why not? (b) If, in addition, $\mathbf{b} = \mathbf{0}$, must $A\mathbf{x} = \mathbf{0}$ have infinitely many solutions? Why, or why not?
20. Suppose A is an unspecified $n \times n$ matrix whose eigenvalues are $\lambda_1, \lambda_2, \ldots, \lambda_n$. Use (2.2) and the definition of the characteristic polynomial to explain why $|A| = \lambda_1 \cdot \lambda_2 \cdot \cdots \cdot \lambda_n$, that is, the determinant is the product of the eigenvalues.
21. Suppose A is $n \times n$, has n distinct eigenvalues, and is an upper or lower triangular matrix. You'll explain why you can say a lot about the eigenvectors of A: if

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdot & \cdot & \cdot & a_{1n} \\ 0 & a_{22} & \cdot & \cdot & \cdot & a_{2n} \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & a_{nn} \end{bmatrix},$$

first find all of the eigenvalues of A. After that, use the given information that all of the eigenvalues are distinct to imply that

$$A - a_{11}I = \begin{bmatrix} 0 & a_{12} & \cdot & \cdot & \cdot & a_{1n} \\ 0 & a_{22} - a_{11} & \cdot & \cdot & \cdot & a_{2n} \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & a_{nn} - a_{11} \end{bmatrix}$$

is row equivalent to $B \triangleq \begin{bmatrix} 0 & a_{12} & . & . & . & a_{1n} \\ 0 & 1 & . & . & . & 0 \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ 0 & 0 & . & . & . & 1 \end{bmatrix}$.

Use that to explain why $\mathbf{e}^{(1)}$ is an eigenvector corresponding to eigenvalue $\lambda_1 = a_{11}$ is $\mathbf{e}^{(1)}$.

Continue by working with all of the other $n-1$ eigenvalues.

22. Suppose A is $n \times n$, $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\}$ is linearly independent, $A\mathbf{x}^{(1)} = -3\mathbf{x}^{(2)}$, and $A\mathbf{x}^{(2)} = 2\mathbf{x}^{(1)}$. Without using a specific matrix A, specific value of n, or specific vector $\mathbf{x}^{(1)}$ or $\mathbf{x}^{(2)}$,
    a. Find two eigenvectors of A^2.
    b. Explain why $(-6-\lambda)^2$ is a factor of the characteristic polynomial of A^2.
    c. Give a specific example of 2×2 matrix A and vectors $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ which satisfy all of the given hypotheses in the beginning of the problem, and verify that conclusions (a) and (b) hold true for your example.
23. Suppose A is an $n \times n$ invertible matrix such that $\lambda = 1$ is not an eigenvalue. Explain why
    a. $(I - A^{-1})$ is invertible.
    b. $A^{-1}(I - A^{-1})^{-1} = (A - I)^{-1}$.
    c. $(I - A^{-1})^{-1}A^{-1} = (A - I)^{-1}$.
24. Suppose A, B, C are $n \times n$ matrices, $B = C^T AC$, C is invertible, and λ is an eigenvalue of A with corresponding eigenvector $\mathbf{x}$.
    a. Why is $|B - \lambda C^T C| = 0$?
    b. Find a nontrivial solution of the system $(B - \lambda C^T C)\mathbf{y} = \mathbf{0}$.
25. Suppose A and B are invertible $n \times n$ matrices. $AB = BA$, λ is an eigenvalue of A with corresponding eigenvector $\mathbf{x}$, and $\mathbf{y} = B\mathbf{x}$.
    a. Explain why $\mathbf{y} \neq \mathbf{0}$.
    b. Why is $\mathbf{y}$ an eigenvector of A? [Hint: Multiply $AB = BA$ on the right by $\mathbf{x}$.]
    c. If, in addition to all of the aforementioned assumptions, A's eigenvalue λ has geometric multiplicity equal to one, why does it follow that $\mathbf{x}$ is also an eigenvector of B?
26. Find all of the eigenvalues of

$$\begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}$$

and their algebraic and geometric multiplicities.

27. Suppose A is an unspecified $n \times n$ matrix.
    a. Why is it true that A is invertible if, and only if, 0 is not an eigenvalue of A?
    b. Suppose that all eigenvalues of A are greater than 2. Why must $I - A$ be invertible?

c. Suppose that all eigenvalues of A are greater than 2. Why must $I - A^{-1}$ be invertible?

28. Find an example of a 2×2 matrix A with such that the eigenvectors of A^T are not eigenvectors of A, and vice-versa. Why must your example have $A \neq A^T$?

29. Suppose $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$.

a. Find all values of λ for which $(A - \lambda B)\mathbf{x} = \mathbf{0}$ has a nontrivial solution.

b. For each such value of λ, find a nontrivial solution for $\mathbf{x}$.

Note: $(A - \lambda B)\mathbf{x} = \mathbf{0}$ is called a **generalized eigenvalue problem**. The *Mathematica* commands `Eigenvalues[A,B]`, `Eigen-vectors[A,B]`, etc. can find generalized eigenvalues and corresponding eigenvectors.

30. (Project) Investigate whether knowing that A can be partitioned into blocks can help determined its eigenvalues. You may start with studying the eigenvalues of matrices of the form

$$\left[\begin{array}{c|c} D_{11} & O \\ \hline O & A_{22} \end{array}\right],$$

where D_{11} is a diagonal matrix, or an upper triangular matrix, and A_{22} is square.

31. (Project) The paper Bryan and Leise (2006) discusses some of the mathematics behind Google's Page Rank Algorithm. Read the paper and do at least 8 of the 17 exercises contained in the paper.

The paper may be available at ⟨http://148.85.1.57/∼ tleise/Math22Spring2007/GoogleMath22.pdf⟩.

By the way, the $\$2.5 \times 10^{10}$ in the title was the approximate "market value," that is, the value of all of the stock shares owned, of Google when the company went public in 2004; more recently, the market value has been between $\$1 \times 10^{11}$ and $\$2 \times 10^{11}$.

32. Find an example of 2×2 matrices A and B for which no eigenvector of B is an eigenvector of AB.

2.2 Basis of Eigenvectors and Diagonalization

Now that we've seen how a little theory can help us, we can use a little more abstraction. When learning a theory, it's always a good idea to refer back to a concrete example, for example, Example 2.5 in Section 2.1, to see what the definitions and results would say about an example.

Definition 2.5

If A has an eigenvalue μ_j whose geometric multiplicity is m_j, then we call $\{\mathbf{x} : (A - \mu_j I)\mathbf{x} = \mathbf{0}\}$ the **eigenspace** of A corresponding to eigenvalue μ_j and denote it by $\mathcal{E}_j$ or $\mathcal{E}_{\lambda=\mu_j}$.

Example 2.11

Find the eigenspaces for the matrix A of Example 2.5 in Section 2.1.

Method: From the results of Example 2.5 in Section 2.1, we know that A has distinct eigenvalues $\mu_1 = -2$, with corresponding eigenvectors $\mathbf{x} = c_1 \begin{bmatrix} -\frac{1}{2} \\ 1 \\ 0 \end{bmatrix} + c_2 \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$, any c_1, c_2 with $|c_1| + |c_2| \neq 0$, and $\mu_2 = 7$, with corresponding eigenvectors $\mathbf{x} = c_1 \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}$, any $c_1 \neq 0$. The eigenspaces are $\mathcal{E}_{\lambda=-2} = \text{Span}\left\{ \begin{bmatrix} -\frac{1}{2} \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \right\}$ and $\mathcal{E}_{\lambda=7} = \text{Span}\left\{ \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} \right\}$. ○

Theorem 2.6

If $n \times n$ matrix A has distinct eigenvalues $\mu_1, \mu_2, \ldots, \mu_p$ and corresponding eigenspaces $\mathcal{E}_1, \ldots, \mathcal{E}_p$, then for each $j = 1, \ldots, p$

(a) $\mathcal{E}_j$ is a vector subspace of $\mathbb{R}^n$, if A and μ_j are real.
(b) $\mathcal{E}_j$ is a vector subspace of $\mathbb{C}^n$, if A and/or μ_j is *not* real.
(c) $dim(\mathcal{E}_j) = m_j$ and $1 \leq m_j \leq \alpha_j \leq n$.
(d) $\mathcal{E}_j$ has basis consisting of the basic solutions, $\{\mathbf{x}^{j,1}, \ldots, \mathbf{x}^{j,m_j}\}$, of the homogeneous system $(A - \mu_j I)\mathbf{x} = \mathbf{0}$.

Why? Theorem 1.39 in Section 1.7 explains both (a) and (b). Definition 2.4 in Section 2.1 explains most of (c). Theorem 2.3 in Section 2.1 explains the $1 \leq m_j \leq \alpha_j \leq n$ part of (c). Corollary 1.2 in Section 1.7 explains (d). □

So, adopting the notation found in Theorem 2.6(d), the eigenspaces $\mathcal{E}_1, \ldots, \mathcal{E}_p$ have basis vectors

$$\mathbf{x}^{1,1}, \ldots, \mathbf{x}^{1,m_1}; \mathbf{x}^{2,1}, \ldots, \mathbf{x}^{2,m_2}; \ldots; \mathbf{x}^{p,1}, \ldots, \mathbf{x}^{p,m_p}. \tag{2.4}$$

Theorem 2.7

Suppose $n \times n$ matrix A has distinct eigenvalues $\mu_1, \mu_2, \ldots, \mu_p$ whose geometric multiplicities are $m_1, m_2, \ldots, m_p$:

(a) If $m_1 + m_2 + \cdots + m_p = n$, then $\{\mathbf{x}^{1,1}, \ldots, \mathbf{x}^{1,m_1}, \mathbf{x}^{2,1}, \ldots, \mathbf{x}^{2,m_2}, \ldots, \mathbf{x}^{p,1}, \ldots, \mathbf{x}^{p,m_p}\}$ is a basis for $\mathbb{R}^n$ when A and all its eigenvalues are real (respectively, $\mathbb{C}^n$, if A and/or at least one of its eigenvalues is not real);

(b) Any set of vectors chosen from $\{\mathbf{x}^{1,1},\ldots,\mathbf{x}^{1,m_1},\mathbf{x}^{2,1},\ldots,\mathbf{x}^{2,m_2},\ldots,\mathbf{x}^{p,1},\ldots,\mathbf{x}^{p,m_p}\}$ is linearly independent.

(c) If $p=n$, that is, A has n distinct eigenvalues, rewrite the corresponding eigenvectors as $\mathbf{x}^{(1)},\mathbf{x}^{(2)},\ldots,\mathbf{x}^{(n)}$. Then $\{\mathbf{x}^{(1)},\mathbf{x}^{(2)},\ldots,\mathbf{x}^{(n)}\}$ is a basis for $\mathbb{R}^n$(or $\mathbb{C}^n$).

Why? We will explain only why (b) is true: by Theorem 1.43 in Section 1.7, it will suffice to explain why $\{\mathbf{x}^{(1)},\mathbf{x}^{(2)},\ldots,\mathbf{x}^{(n)}\}$ is linearly independent. First, $A\mathbf{x}^{(j)}=\lambda_j\mathbf{x}^{(j)}$ implies $(A-\lambda I)\mathbf{x}^{(j)}=(\lambda_j-\lambda)\mathbf{x}^{(j)}$. Now, suppose

$$\mathbf{0}=c_1\mathbf{x}^{(1)}+c_2\mathbf{x}^{(2)}+\cdots+c_n\mathbf{x}^{(n)}.$$

If we multiply on the left by the matrix $(A-\lambda_2 I)(A-\lambda_3 I)\cdots(A-\lambda_n I)$, we have

$$\begin{aligned}\mathbf{0}&=c_1(A-\lambda_2 I)(A-\lambda_3 I)\cdots(A-\lambda_n I)\mathbf{x}^{(1)}+c_2(A-\lambda_2 I)(A-\lambda_3 I)\cdots(A-\lambda_n I)\mathbf{x}^{(2)}+\cdots+\\&\qquad+c_n(A-\lambda_2 I)(A-\lambda_3 I)\cdots(A-\lambda_n I)\mathbf{x}^{(n)}\\&=c_1(\lambda_1-\lambda_2)(\lambda_1-\lambda_3)\cdots(\lambda_1-\lambda_n)\mathbf{x}^{(1)}+c_2(\lambda_2-\lambda_2)(\lambda_2-\lambda_3)\cdots(\lambda_2-\lambda_n)\mathbf{x}^{(2)}+\cdots+\\&\qquad+c_n(\lambda_n-\lambda_2)(\lambda_n-\lambda_3)\cdots(\lambda_n-\lambda_n)\mathbf{x}^{(n)}\\&=c_1(\lambda_1-\lambda_2)(\lambda_1-\lambda_3)\cdots(\lambda_1-\lambda_n)\mathbf{x}^{(1)}+\mathbf{0}+\cdots+\mathbf{0}.\end{aligned}$$

Because the eigenvalues were assumed to be distinct, and $\mathbf{x}^{(1)}\neq\mathbf{0}$, it follows that $c_1=0$. In a similar way (see Problem 2.2.3.24), we can conclude $c_2=\cdots=c_n=0$. By the way, we may call what we did an "annihilator method," and we will use something like this when solving nonhomogeneous differential equations in Section 4.1. □

Example 2.12

For $A=\begin{bmatrix}-1 & 2\\ -2 & -1\end{bmatrix}$, find a basis for $\mathbb{R}^2$, or $\mathbb{C}^2$ consisting of eigenvectors of A.

Method: $0=|\,A-\lambda I_2\,|=\begin{vmatrix}-1-\lambda & 2\\ -2 & -1-\lambda\end{vmatrix}=(-1-\lambda)^2+4$, so the eigenvalues λ satisfy $(-1-\lambda)^2=-4$, hence $(-1-\lambda)=\pm i2$, where $i\triangleq\sqrt{-1}$. The eigenvalues are $\lambda=-1\pm i2$. We may call $\lambda_1=-1+i2$ and $\lambda_2=-1-i2$, although we could just as well reverse the roles as long as we are consistent.

Because the 2×2 matrix A has two distinct eigenvalues, Theorem 2.7(b) guarantees that $\mathbb{C}^2$ has a basis consisting of eigenvectors of A: $\{\mathbf{x}^{(1)},\mathbf{x}^{(2)}\}$. Let's find them explicitly, using row reduction on $\left[A-\lambda I\mid\mathbf{0}\right]$:

$$\left[A-(-1+i2)I_2\mid\mathbf{0}\right]=\begin{bmatrix}-i2 & 2 & \mid 0\\ -2 & -i2 & \mid 0\end{bmatrix}\quad\begin{matrix}\sim\\ R_1\leftrightarrow R_2\\ -\frac{1}{2}R_1\to R_1\\ i2R_1+R_2\to R_2\end{matrix}\quad\begin{bmatrix}\textcircled{1} & i & \mid 0\\ 0 & 0 & \mid 0\end{bmatrix}.$$

The eigenvectors corresponding to eigenvalue $-1+i2$ are $\mathbf{x}=c_1\begin{bmatrix}-i\\ 1\end{bmatrix}$, for any $c_1\neq 0$.

Using Theorem 2.4 in Section 2.1 about the "complex conjugate pair" nature of eigenvalues and eigenvectors, the eigenvectors corresponding to eigenvalue $-1-i2$ are $\mathbf{x}=c_1\begin{bmatrix} i \\ 1 \end{bmatrix}$, for any $c_1 \neq 0$.

Because we have two distinct eigenvalues, Theorem 2.7(b) implies that

$$\left\{\begin{bmatrix} -i \\ 1 \end{bmatrix}, \begin{bmatrix} i \\ 1 \end{bmatrix}\right\}$$

is a basis, consisting of eigenvectors of A, for $\mathbb{C}^2$. ◯

Example 2.13

For $A=\begin{bmatrix} 2 & 2 & 4 \\ 2 & -1 & 2 \\ 4 & 2 & 2 \end{bmatrix}$, that is, the matrix of Examples 2.3, 2.5, and 2.7 in Section 2.1, find a basis for $\mathbb{R}^3$, or $\mathbb{C}^3$, consisting of eigenvectors of A.

Method: We can use Theorem 2.7(a), because in Example 2.5 in Section 2.1, we found two basic solutions, $\begin{bmatrix} -\frac{1}{2} \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$, corresponding to eigenvalue $\mu_1=-2$, and one basic solution, $\begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}$, corresponding to eigenvalue $\mu_2=7$; in effect, these conclusions were summarized in Example 2.11. It follows that

$$\left\{\begin{bmatrix} -\frac{1}{2} \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}\right\}$$

is a basis for $\mathbb{R}^3$ consisting of eigenvectors of A. ◯

Here are further useful results.

Theorem 2.8

Suppose A has distinct eigenvalues $\mu_1, \mu_2, \ldots, \mu_p$ and corresponding eigenspaces $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_p$:

(a) If $B=A+\gamma I$ for some nonzero scalar γ, then B has distinct eigenvalues

$$\mu_1+\gamma,\ \mu_2+\gamma,\ \ldots,\ \mu_p+\gamma$$

and corresponding eigenspaces $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_p$, that is, the same eigenspaces as A.

(b) If $B = \gamma A$ for some nonzero scalar γ, then B has distinct eigenvalues $\gamma\mu_1, \gamma\mu_2, \ldots, \gamma\mu_p$ and corresponding eigenspaces $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_p$, that is, the same eigenspaces as A.

(c) If $B = A^T$, then B has distinct eigenvalues $\mu_1, \mu_2, \ldots, \mu_p$, that is, the same eigenvalues as A, but not necessarily the same eigenspaces as A.

(d) If $B = A^{-1}$ exists, then B has distinct eigenvalues ${\mu_1}^{-1}, {\mu_2}^{-1}, \ldots, {\mu_p}^{-1}$, that is, the reciprocals of the eigenvalues of A, and the same eigenspaces as A.

(e) If $B = f(A)$ for some polynomial function, then B has eigenvalues

$$f(\mu_1), f(\mu_2), \ldots, f(\mu_p),$$

which are *not* necessarily distinct, and the same eigenspaces as A.

Parts (a) and (b) are special cases of part (e) but with more informative conclusions than part (e) because the distinctiveness of the eigenvalues is maintained. As for part (e), the reason why we must allow for the eigenvalues losing distinctiveness can be seen if 2×2 matrix A has eigenvalues -1 and 1 and $f(A) = A^2$: By Theorem 2.8(e), A^2 has eigenvalues $(-1)^2$ and 1, which are not distinct. In Problem 2.2.3.12, you will be asked to come up with a specific example of such an A. □

2.2.1 Diagonalizing a Matrix

Given a matrix A, we will see that it is sometimes useful to express it in terms of a diagonal matrix: $A = PDP^{-1}$, for some diagonal matrix D and invertible matrix P. But

$$\begin{aligned} A = PDP^{-1} &\iff AP = (PDP^{-1})P = (PD)(P^{-1}P) = (PD)(I) = PD \\ &\iff P^{-1}(AP) = P^{-1}(PD) = (P^{-1}P)D = (I)D = D. \end{aligned}$$

So, $A = PDP^{-1}$ is logically equivalent to both $AP = PD$ and to $D = P^{-1}AP$.

Definition 2.6

Matrix A is **diagonalizable** if there is an invertible matrix P with $P^{-1}AP$ being a diagonal matrix. In this case, we say that P **diagonalizes** A.

Theorem 2.9

A is diagonalizable if, and only if, $AP = PD$ for some diagonal matrix D and invertible matrix P.

Theorem 2.10

If $\mathbb{R}^n$ (respectively, $\mathbb{C}^n$) has a basis $\{\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(n)}\}$ consisting of eigenvectors of A, then

(a) A is diagonalizable.
(b) the $n \times n$ matrix $P = \begin{bmatrix} \mathbf{p}^{(1)} & \mathbf{p}^{(2)} & \dots & \mathbf{p}^{(n)} \end{bmatrix}$ diagonalizes A.
(c) $P^{-1}AP = D = diag(\lambda_1, \lambda_2, \dots, \lambda_n)$, where the λ_j's are the n eigenvalues of A, possibly including repetitions, corresponding to eigenvectors $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \dots, \mathbf{p}^{(n)}$.

Why? By Theorem 1.9 in Section 1.2,

$$\begin{aligned} AP &= A\begin{bmatrix} \mathbf{p}^{(1)} & \mathbf{p}^{(2)} & \dots & \mathbf{p}^{(n)} \end{bmatrix} \\ &= \begin{bmatrix} A\mathbf{p}^{(1)} & A\mathbf{p}^{(2)} & \dots & A\mathbf{p}^{(n)} \end{bmatrix} = \begin{bmatrix} \lambda_1\mathbf{p}^{(1)} & \lambda_2\mathbf{p}^{(2)} & \dots & \lambda_n\mathbf{p}^{(n)} \end{bmatrix}. \\ &= \begin{bmatrix} \mathbf{p}^{(1)} & \mathbf{p}^{(2)} & \dots & \mathbf{p}^{(n)} \end{bmatrix} diag(\lambda_1, \lambda_2, \dots, \lambda_n) = PD. \end{aligned}$$

by Theorem 1.10 in Section 1.2. □

Example 2.14

Find an exact matrix that diagonalizes $A = \begin{bmatrix} -1 & 2 \\ 1 & 1 \end{bmatrix}$.

Method: First, let's find the eigenvalues of A and then a basis for $\mathbb{R}^2$ or $\mathbb{C}^2$ that consists of eigenvectors of A: $0 = |A - \lambda I_2| = \begin{vmatrix} -1-\lambda & 2 \\ 1 & 1-\lambda \end{vmatrix} = \lambda^2 - 3$, so the eigenvalues are $\lambda = \pm\sqrt{3}$. We have

$$\begin{aligned} &\left[A - \sqrt{3}I_2 \mid \mathbf{0}\right] \\ &= \begin{bmatrix} -1-\sqrt{3} & 2 & \mid 0 \\ 1 & 1-\sqrt{3} & \mid 0 \end{bmatrix} \underset{\substack{R_1 \leftrightarrow R_2 \\ (1+\sqrt{3})R_1 + R_2 \to R_2}}{\sim} \begin{bmatrix} \textcircled{1} & 1-\sqrt{3} & \mid 0 \\ 0 & 0 & \mid 0 \end{bmatrix}. \end{aligned}$$

Corresponding to eigenvalue $\lambda_1 = \sqrt{3}$, we have eigenvectors $\mathbf{x} = c_1\begin{bmatrix} -1+\sqrt{3} \\ 1 \end{bmatrix}$, where $c_1 \neq 0$. Similarly,

$$\left[A - (-\sqrt{3})I_2 \mid \mathbf{0}\right] = \begin{bmatrix} -1+\sqrt{3} & 2 & \mid 0 \\ 1 & 1+\sqrt{3} & \mid 0 \end{bmatrix} \sim \begin{bmatrix} \textcircled{1} & 1+\sqrt{3} & \mid 0 \\ 0 & 0 & \mid 0 \end{bmatrix}.$$

Corresponding to eigenvalue $\lambda_2 = -\sqrt{3}$, we have eigenvectors $\mathbf{x} = c_1\begin{bmatrix} -1-\sqrt{3} \\ 1 \end{bmatrix}$, where $c_1 \neq 0$.

So, we can take $\mathbf{p}^{(1)} = \begin{bmatrix} -1+\sqrt{3} \\ 1 \end{bmatrix}$ and $\mathbf{p}^{(2)} = \begin{bmatrix} -1-\sqrt{3} \\ 1 \end{bmatrix}$ as the columns of the diagonalizing matrix P, that is,

$$P = \begin{bmatrix} \mathbf{p}^{(1)} & \mathbf{p}^{(2)} \end{bmatrix} = \begin{bmatrix} -1+\sqrt{3} & -1-\sqrt{3} \\ 1 & 1 \end{bmatrix}.$$

We only have one more thing to check: because $|P| = (-1+\sqrt{3})(1) - (-1-\sqrt{3})(1) = 2\sqrt{3} \neq 0$, P is invertible. According to Theorem 2.10(c), this matrix P should diagonalize A. ○

It doesn't hurt to *check* that this works out correctly, that is, that $P^{-1}AP = D$, the diagonal matrix whose diagonal entries are the eigenvalues of A. First, we need the inverse of the 2×2 matrix P:

$$P^{-1} = \frac{1}{|P|}\begin{bmatrix} 1 & 1+\sqrt{3} \\ -1 & -1+\sqrt{3} \end{bmatrix} = \frac{1}{2\sqrt{3}}\begin{bmatrix} 1 & 1+\sqrt{3} \\ -1 & -1+\sqrt{3} \end{bmatrix}.$$

So,

$$\begin{aligned} P^{-1}AP &= P^{-1}\begin{bmatrix} -1 & 2 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} -1+\sqrt{3} & -1-\sqrt{3} \\ 1 & 1 \end{bmatrix} \\ &= \frac{1}{2\sqrt{3}}\begin{bmatrix} 1 & 1+\sqrt{3} \\ -1 & -1+\sqrt{3} \end{bmatrix}\begin{bmatrix} 3-\sqrt{3} & 3+\sqrt{3} \\ \sqrt{3} & -\sqrt{3} \end{bmatrix} \\ &= \frac{1}{2\sqrt{3}}\begin{bmatrix} 6 & 0 \\ 0 & -6 \end{bmatrix} = \begin{bmatrix} \sqrt{3} & 0 \\ 0 & -\sqrt{3} \end{bmatrix} = D = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \end{aligned}$$

the diagonal matrix predicted by Theorem 2.10(c). ◯

Example 2.15

Find an exact matrix that diagonalizes $A = \begin{bmatrix} 2 & 2 & 4 \\ 2 & -1 & 2 \\ 4 & 2 & 2 \end{bmatrix}$, and find the corresponding diagonal matrix D.

Method: By Theorem 2.10(b), if we have a basis for $\mathbb{R}^3$ consisting of eigenvectors of A, then we can use those eigenvectors as the columns of a matrix, P, that diagonalizes A. But, we found such a basis in Example 2.13, so

$$P = \begin{bmatrix} -\frac{1}{2} & -1 & 2 \\ 1 & 0 & 1 \\ 0 & 1 & 2 \end{bmatrix},$$

diagonalizes A. As to the matrix D, we can either note that the three columns of P correspond to eigenvalues $-2, -2, 7$ of A, hence according to Theorem 2.10(c), we should have

$$D = \begin{bmatrix} -2 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 7 \end{bmatrix},$$

or we could calculate that

$$\begin{aligned} P^{-1}AP &= \frac{1}{9}\begin{bmatrix} -2 & 8 & -2 \\ -4 & -2 & 5 \\ 2 & 1 & 2 \end{bmatrix}\begin{bmatrix} 2 & 2 & 4 \\ 2 & -1 & 2 \\ 4 & 2 & 2 \end{bmatrix}\begin{bmatrix} -\frac{1}{2} & -1 & 2 \\ 1 & 0 & 1 \\ 0 & 1 & 2 \end{bmatrix} \\ &= \frac{1}{9}\begin{bmatrix} -2 & 8 & -2 \\ -4 & -2 & 5 \\ 2 & 1 & 2 \end{bmatrix}\begin{bmatrix} 1 & 2 & 14 \\ -2 & 0 & 7 \\ 0 & -2 & 14 \end{bmatrix} = \begin{bmatrix} -2 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 7 \end{bmatrix} = D. \text{ ◯} \end{aligned}$$

In this example, it made sense to check that $P^{-1}AP = D$ even though the theorem predicted what D should be.

Definition 2.7

Two matrices A and B are **similar** if there is an invertible matrix P with $B = P^{-1}AP$.

So, a matrix is diagonalizable if it is similar to a diagonal matrix. Notice that the word "similar" has nothing to do with the idea of "row equivalence." We mention this because some books write $A \sim B$ to mean "A is similar to B," while we have reserved the symbol $\sim$ to mean "is row equivalent to."

In the definition of similar matrices, either, both, or neither A and B may be diagonal matrices.

Theorem 2.11

If A and B are similar, then B's eigenvalues and their algebraic and geometric multiplicities exactly equal those of A. Further, if $B = P^{-1}AP$, then $\mathbf{y}$ is an eigenvector of B if, and only if, $P\mathbf{y}$ is an eigenvector of A.

2.2.2 Deficient Eigenvalues

Does every matrix have a set of eigenvectors that is a basis for $\mathbb{R}^n$ or $\mathbb{C}^n$? Unfortunately, no.

Example 2.16

$A = \begin{bmatrix} 29 & 18 \\ -50 & -31 \end{bmatrix}$ has only one linearly independent eigenvector.

Why? First, we find the eigenvalues: $0 = |\ A - \lambda I_2\ | = \begin{vmatrix} 29-\lambda & 18 \\ -50 & -31-\lambda \end{vmatrix} = \lambda^2 + 2\lambda + 1 = (\lambda + 1)^2$, so the only eigenvalue is $\lambda = -1$. So, to find *all* the corresponding eigenvectors, we use

$$\left[A - (-1)I_2 \mid \mathbf{0}\right] = \begin{bmatrix} 30 & 18 & \mid 0 \\ -50 & -30 & \mid 0 \end{bmatrix} \sim \begin{bmatrix} \textcircled{1} & \frac{3}{5} & \mid 0 \\ 0 & 0 & \mid 0 \end{bmatrix},$$

so the eigenvectors are $\mathbf{x} = c_1 \begin{bmatrix} -\frac{3}{5} \\ 1 \end{bmatrix}$, $c_1 \neq 0$. So, the dimension of the only eigenspace is one, and A does not have a set of eigenvectors that is a basis for $\mathbb{R}^2$. ○

Definition 2.8

If an eigenvalue has its geometric multiplicity strictly less than its algebraic multiplicity, we call that eigenvalue **deficient**.

So far, our only example of a deficient eigenvalue is $\lambda = -1$ in Example 2.16.

2.2.3 Problems

In problems 1–7, for each matrix, find an exact matrix that diagonalizes the given matrix, if possible.

1. $\begin{bmatrix} 5 & -1 \\ 3 & 1 \end{bmatrix}$
2. $\begin{bmatrix} 5 & -2 \\ -2 & 2 \end{bmatrix}$
3. $\begin{bmatrix} 2 & 0 \\ -1 & -1 \end{bmatrix}$
4. $\begin{bmatrix} -3 & \sqrt{3} \\ -\sqrt{3} & 1 \end{bmatrix}$
5. $\begin{bmatrix} -2 & \sqrt{2} \\ -\sqrt{2} & 2 \end{bmatrix}$
6. $\begin{bmatrix} -3 & -1 & 2 \\ 0 & -2 & 0 \\ -1 & -1 & 0 \end{bmatrix}$
7. $\begin{bmatrix} 6 & 7 & 7 \\ -7 & -8 & -7 \\ 7 & 7 & 6 \end{bmatrix}$

In problems 8–10, a matrix A and information about one or more of its eigenvalues is given. Find an exact matrix, P, that diagonalizes A. Also, find P^{-1}, for example by using a calculator, and explicitly calculate $D = P^{-1}AP$ to check your work.

8. $A = \begin{bmatrix} 1 & 5 & -10 \\ 5 & 1 & -10 \\ 0 & 0 & -4 \end{bmatrix}$, $\lambda = 6$
9. $A = \begin{bmatrix} 3 & 0 & -12 \\ 4 & -1 & -12 \\ 0 & 0 & -1 \end{bmatrix}$, $\lambda = -1, 3$
10. $A = \begin{bmatrix} -3 & 2 & 2 \\ 0 & -1 & 0 \\ -1 & 1 & 0 \end{bmatrix}$, $\lambda = -1$
11. (a) Write down a 3×3 matrix A that satisfies all of the following properties: (i) A is in RREF, (ii) A has exactly two pivot positions, and (iii) at least one entry of A is 2 and at least one entry of A is -1.
    (b) For the matrix A you wrote down in part (a), find a basis for $\mathbb{R}^3$ consisting of eigenvectors of A.
12. Find an example of a 2×2 matrix A that has eigenvalues -1 and 1 and define $f(A) = A^2$. For your matrix A, explain why A^2 has eigenvalues $(-1)^2$ and 1, which are not distinct. This illustrates Theorem 2.8.
13. Assume A is an $n \times n$ matrix, $n \geq 2$, and $\mathbf{x}$ is an eigenvector of A. For each of (a) through (g), decide whether it must be true, must be false, or may be true and may be false.

a. $\mathbf{x}$ is in $\mathcal{W} \triangleq \{\mathbf{x} : \ (A - \lambda I_n)\mathbf{x} = \mathbf{0}\}$ for some scalar λ.
b. $\mathbf{x}$ is an eigenvector of A^T.
c. $\mathbf{x} = \mathbf{0}$
d. $\mathbf{x}$ is a factor of the characteristic polynomial.
e. $\mathbf{x}$ is an eigenvector of A^2.
f. $\{\mathbf{x}\}$ is a basis of an eigenspace of A.
g. $\mathbf{x}$ is an eigenvector of B, if B is similar to A.

14. For the matrix A of Example 2.16, explain why there is no invertible matrix P that diagonalizes A. [Hint: Consider the equation $AP = PD$; additionally, use Theorems 2.5 in Section 2.1 and 2.11 to decide in advance what the diagonal entries of D must be.]
15. Suppose A is a 3×3 matrix for which

$$[\,A + 2I \,|\, \mathbf{0}\,] = \left[\begin{array}{ccc|c} -2 & -2 & 2 & 0 \\ 1 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \text{ and } [\,A + 3I \,|\, \mathbf{0}] = \left[\begin{array}{ccc|c} -1 & -2 & 2 & 0 \\ 1 & 2 & -1 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right].$$

Find a set of three eigenvectors of A that is a basis for $\mathbb{R}^3$.

16. (a) Explain why matrices that are similar have equal determinant.
(b) Is the "converse" true? That is, if $|A| = |B|$, does it follow that A and B are similar? If so, why? If not, give a specific counter-example.
17. Suppose A is an unspecified 3×3 matrix for which $rank(A - 2I_3) = 2$ and $rank(A - 3I_3) = 1$. Find $|A|$ either by using the result of Problem 2.1.6.20 or by using the result of Problem 2.2.3.16.
18. Suppose $A + I$ is row equivalent to

$$\begin{bmatrix} 1 & -2 & 3 & 5 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Find a basis for the eigenspace $\mathcal{E}_{\mu=-1}$.

19. Suppose A is an $n \times n$ real matrix, all of whose eigenvalues are *positive* real numbers. Suppose A has a set of eigenvectors $\{\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(n)}\}$ that is a basis for $\mathbb{R}^n$. Define, as usual, $P = \left[\mathbf{p}^{(1)} \ \mathbf{p}^{(2)} \ \ldots \ \mathbf{p}^{(n)}\right]$, so $P^{-1} A P = D \triangleq diag(\lambda_1, \ldots, \lambda_n)$. Define $S \triangleq P \, diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}) \, P^{-1}$. Explain why $S^2 = A$, hence we may refer to S as the "square root of the matrix" A.
20. If $B = P^{-1}AP$, does it follow that $A = PBP^{-1}$? If so, why? If not, give a specific counterexample using 2×2 matrices.
21. (Designing a desired matrix) Find a 2×2 matrix A that has two distinct eigenvalues and has eigenvectors

$$\begin{bmatrix} -5 \\ 1 \end{bmatrix}, \ \begin{bmatrix} 4 \\ 1 \end{bmatrix}.$$

22. For the matrix

$$A = \begin{bmatrix} 2 & 0 & 1 \\ 0 & -5 & 0 \\ 3 & 0 & 2 \end{bmatrix},$$

(a) Find all of the *exact* eigenvalues and eigenvectors and (b) explain why A has a set of eigenvectors that is a basis for $\mathbb{R}^3$.

23. Suppose A is an unspecified 3×3 matrix that has eigenvalues $2, -2, \sqrt{3}$ and corresponding eigenvectors $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$.
    a. Find the characteristic polynomial of A.
    b. Find a set of three linearly independent eigenvectors of A^2.
    c. Find the characteristic polynomial of A^2.
24. In explaining why Theorem 2.7(c) was true, we used an "annihilator" to explain why c_1 must equal zero. What annihilator would you use to explain why c_2 must equal zero? What annihilator would you use to explain why c_n must equal zero?
25. In explaining why Theorem 2.7(c) was true, we used the "annihilator" $(A - \lambda_2 I)(A - \lambda_3 I) \cdots (A - \lambda_n I)$ to explain why c_1 must equal zero. In an effort to explain why Theorem 2.7(a) is true, with a similar annihilator using $\mu_2, \ldots, \mu_p$, explain why

$$\mathbf{0} = c_{1,1}\mathbf{x}^{1,1} + \cdots + c_{1,m_1}\mathbf{x}^{1,m_1} + c_{2,1}\mathbf{x}^{2,1} + \cdots +$$

$$+ c_{2,m_2}\mathbf{x}^{2,m_2} + \cdots + c_{p,1}\mathbf{x}^{p,1} + \cdots + c_{p,m_p}\mathbf{x}^{p,m_p}$$

would imply that $c_{1,1} = \cdots = c_{1,m_1} = 0$. Here, we're using the notation of Theorem 2.7 that $\{\mathbf{x}^{1,1}, \ldots, \mathbf{x}^{1,m_1}\}$ is the set of basic solutions that span the eigenspace $\mathcal{E}_1$.

26. Can $\mathbb{R}^3$ have a basis of eigenvectors all of which have 0 in their first components?
27. For $A = \begin{bmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ 0 & 1 & 8 \end{bmatrix}$, find all of the eigenvalues and use that information to explain why A has a set of three linearly independent eigenvectors.

2.3 Inner Product and Orthogonal Sets of Vectors

We're familiar with the dot product in $\mathbb{R}^3$ given by

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \bullet \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \triangleq x_1y_1 + x_2y_2 + x_3y_3.$$

For example, in physics, *Work* = **Force** • **displacement**, assuming **Force** is constant. The dot product generalizes to $\mathbb{R}^n$.

Definition 2.9

If $\mathbf{x}, \mathbf{y}$ are in $\mathbb{R}^n$, then

$$\langle \mathbf{x}, \mathbf{y} \rangle \triangleq x_1 y_1 + x_2 y_2 + \cdots + x_n y_n, \tag{2.5}$$

which we call the **inner product** of $\mathbf{x}$ and $\mathbf{y}$.

So, in $\mathbb{R}^3$, $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x} \bullet \mathbf{y}$.

Theorem 2.12

(Properties of inner products) For all $\mathbf{x}, \mathbf{y}, \mathbf{x}_1, \mathbf{x}_2$ in $\mathbb{R}^n$ and α in $\mathbb{R}$,

$$\left\{ \begin{array}{l} \text{(a)} \ \ \langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle \\ \text{(b)} \ \ \langle \alpha \mathbf{x}, \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle \\ \text{(c)} \ \ \langle \mathbf{x}_1 + \mathbf{x}_2, \mathbf{y} \rangle = \langle \mathbf{x}_1, \mathbf{y} \rangle + \langle \mathbf{x}_2, \mathbf{y} \rangle \\ \text{(d)} \ \ \langle \mathbf{x}, \mathbf{x} \rangle \geq 0, \text{with equality only if } \mathbf{x} = \mathbf{0} \end{array} \right\}. \tag{2.6}$$

By the way, from properties (2.6)(a) and (b) it follows that

$$\langle \mathbf{x}, \alpha \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle.$$

Sometimes we will write

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y}, \tag{2.7}$$

that is, that the real number $\langle \mathbf{x}, \mathbf{y} \rangle$ equals the real, 1×1 matrix $\mathbf{x}^T \mathbf{y}$, in an abuse of notation.

Definition 2.10

If $\mathbf{x}, \mathbf{y}$ are in $\mathbb{R}^n$

$$|| \mathbf{x} || \triangleq \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}, \tag{2.8}$$

which we call the **norm** or **magnitude** or **length** of $\mathbf{x}$.

The "direction vector" from the point $(0, 0, \ldots, 0)$, to the point $(x_1, x_2, \ldots, x_n)$ is given by

$$\mathbf{x} \triangleq \overrightarrow{OP} \triangleq \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},$$

so $||\,\mathbf{x}\,||$ measures the distance from the origin, $(0, 0, \ldots, 0)$, to the point $(x_1, x_2, \ldots, x_n)$.

Theorem 2.13

(Properties of norms) For all $\mathbf{x}, \mathbf{y}$ in $\mathbb{R}^n$ and α in $\mathbb{R}$,

$$\left\{\begin{array}{l} \text{(a)}\ \ ||\,\mathbf{x}\,||^2 = \langle \mathbf{x}, \mathbf{x} \rangle \\ \text{(b)}\ \ ||\mathbf{x} + \mathbf{y}||^2 = ||\,\mathbf{x}\,||^2 + ||\,\mathbf{y}\,||^2 + 2\langle \mathbf{x}, \mathbf{y} \rangle \\ \text{(c)}\ \ ||\,\mathbf{x}\,|| \geq 0, \text{with equality only if } \mathbf{x} = \mathbf{0} \\ \text{(d)}\ \ ||\alpha\mathbf{x}|| = |\alpha|\ ||\,\mathbf{x}\,|| \\ \text{(e)}\ \ |\langle \mathbf{x}, \mathbf{y} \rangle| \leq ||\,\mathbf{x}\,||\ ||\,\mathbf{y}\,|| \\ \text{(f)}\ \ ||\mathbf{x} + \mathbf{y}|| \leq ||\,\mathbf{x}\,|| + ||\,\mathbf{y}\,|| \end{array}\right\}. \tag{2.9}$$

Why? (a) and (c) are easily explained. As for (b), from (a) and then properties of $\langle\ ,\ \rangle$ we have

$$||\mathbf{x} + \mathbf{y}||^2 = \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{x} + \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle,$$

from which (2.9)(b) follows.

As for (d), you will explain this in Problem 2.3.4.14, using the fact that for real numbers α, we have $\sqrt{\alpha^2} = |\alpha|$, for example, $\sqrt{(-5)^2} = 5 = |(-5)|$.

Part (e) is called the **Cauchy–Schwarz** inequality. We will explain this result in the appendix at the end of this section.

Part (f) is known as the **triangle inequality** and is illustrated in Figure 2.3: In the plane spanned by the vectors $\mathbf{x}$ and $\mathbf{y}$, here visualized as the plane $\mathbb{R}^2$, we know that $||\mathbf{x} + \mathbf{y}||$, the distance from O to P, is the length of the shortest curve from O to P, hence is less than or equal to $||\,\mathbf{x}\,|| + ||\,\mathbf{y}\,||$, the sum of the distance from O to Q and the distance from Q to P. To summarize, the triangle OQP illustrates that

$$||\mathbf{x} + \mathbf{y}|| \leq ||\,\mathbf{x}\,|| + ||\,\mathbf{y}\,||.$$

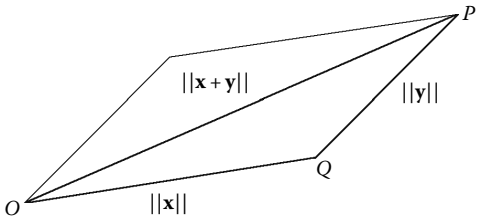


FIGURE 2.3
Triangle inequality.

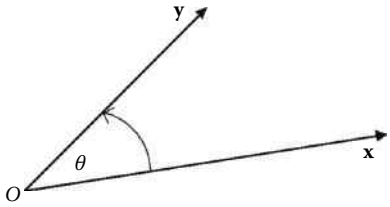


FIGURE 2.4
Angle between vectors.

To explain why (f) follows from (e), note that

$$\begin{aligned} ||\mathbf{x}+\mathbf{y}||^2 &= ||\,\mathbf{x}\,||^2 + ||\,\mathbf{y}\,||^2 + 2\langle \mathbf{x},\mathbf{y}\rangle \le ||\,\mathbf{x}\,||^2 + ||\,\mathbf{y}\,||^2 + 2|\langle \mathbf{x},\mathbf{y}\rangle| \\ &\le ||\,\mathbf{x}\,||^2 + ||\,\mathbf{y}\,||^2 + 2||\,\mathbf{x}\,||\,||\,\mathbf{y}\,|| = \big(||\,\mathbf{x}\,|| + ||\,\mathbf{y}\,||\big)^2, \end{aligned}$$

by the Cauchy–Schwarz inequality. It follows that $||\mathbf{x}+\mathbf{y}|| \le ||\,\mathbf{x}\,|| + ||\,\mathbf{y}\,||$. □

Definition 2.11

The angle θ, with $0 \le \theta \le \pi$, between two nonzero vectors $\mathbf{x}, \mathbf{y}$ is defined implicitly by

$$\cos\theta \triangleq \frac{\langle \mathbf{x},\mathbf{y}\rangle}{||\,\mathbf{x}\,||\,||\,\mathbf{y}\,||}. \tag{2.10}$$

and the requirement that $0 \le \theta \le \pi$. This is illustrated in Figure 2.4.

Note that the Cauchy inequality guarantees that $\left|\dfrac{\langle \mathbf{x},\mathbf{y}\rangle}{||\,\mathbf{x}\,||\,||\,\mathbf{y}\,||}\right| \le 1$, that is, that $|\cos\theta| \le 1$. We can rewrite (2.10) as

$$\langle \mathbf{x},\mathbf{y}\rangle = ||\,\mathbf{x}\,||\,||\,\mathbf{y}\,||\cos\theta. \tag{2.11}$$

Some people prefer to define $\langle \mathbf{x},\mathbf{y}\rangle$ via (2.11) and then explain why $\langle \mathbf{x},\mathbf{y}\rangle$ satisfies the equality in Definition 2.9 as a result rather than as a definition.

Theorem 2.14

Equality in the Cauchy–Schwarz inequality, that is, $|\langle \mathbf{x}, \mathbf{y} \rangle| = ||\,\mathbf{x}\,||\;||\,\mathbf{y}\,||$, holds if and only if the set of vectors $\{\mathbf{x}, \mathbf{y}\}$ is linearly dependent.

This result makes sense because when $\mathbf{x}$ and $\mathbf{y}$ are nonzero, $\cos\theta = \pm 1$ is equivalent to $\theta = 0$ or $\theta = \pi$, that is, equivalent to the vectors $\mathbf{x}, \mathbf{y}$ being parallel. Note also that if $\mathbf{x} = \mathbf{0}$, then $\{\mathbf{0}, \mathbf{y}\}$ is linearly dependent.

By the way, given two nonzero vectors $\mathbf{x}, \mathbf{y}$, the quantity

$$r \triangleq \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{||\mathbf{x}||\;||\mathbf{y}||}$$

is called the **correlation coefficient** in statistics. When $r = \pm 1$, the nonzero vectors $\mathbf{x}, \mathbf{y}$ are linearly dependent, hence there is a scalar k with $\mathbf{y} = k\mathbf{x}$. If we think of the components of the vectors $\mathbf{x}, \mathbf{y}$ as pairing off in data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ in the plane, then $r = \pm 1$ implies that those data points lie on a common line. We'll see more about fitting a line to data points in Section 2.5.

Definition 2.12

$\mathbf{x}, \mathbf{y}$ are **orthogonal** if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$.

The notation "$\mathbf{x} \perp \mathbf{y}$" means that "$\mathbf{x}, \mathbf{y}$ are orthogonal."

Theorem 2.15

(Pythagoras) $||\mathbf{x} + \mathbf{y}||^2 = ||\,\mathbf{x}\,||^2 + ||\,\mathbf{y}\,||^2$ holds if, and only if, $\mathbf{x} \perp \mathbf{y}$.

2.3.1 Orthogonal Set of Vectors

Definition 2.13

(a) $\{\mathbf{x}, \mathbf{y}\}$ is an orthogonal set of vectors if $\mathbf{x} \perp \mathbf{y}$.

(b) $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is an **orthogonal set of vectors** if $\mathbf{x}_i \perp \mathbf{x}_j$ for all $i \neq j$.

Example 2.17

Here are three examples of orthogonal sets of vectors:

(a) $\left\{ \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ -4 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} \right\}$.

(b) $\left\{ \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \frac{1}{3\sqrt{2}} \begin{bmatrix} 1 \\ -4 \\ 1 \end{bmatrix}, \frac{1}{3} \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} \right\}$.

(c) $\left\{ \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} \right\}$.

Example 2.17(b) is very similar to Example 2.17(a). Each of the vectors in Example 2.17(b) has length one, that is, is a **unit vector** and is a **normalized** version of a vector in Example 2.17(a). Given a nonzero vector $\mathbf{x}$, we can "divide" it by its length to get a unit vector, $\widehat{\mathbf{x}}$, that points in the same direction:

$$\widehat{\mathbf{x}} \triangleq \frac{1}{||\,\mathbf{x}\,||}\mathbf{x}. \tag{2.12}$$

We use a conventional notation that a hat, $\widehat{\ }$, over a vector signifies that it is a unit vector.

Definition 2.14

$\mathcal{S}$ is an **orthonormal set of vectors** if $\mathcal{S}$ is an orthogonal set of vectors and each vector in $\mathcal{S}$ is a unit vector. We may abbreviate "orthonormal set of vectors" by writing "o.n. set."

So, Example 2.17(b) gives an o.n. set of three vectors.

Based on Example 2.17(a) and (b), we note that an orthogonal set of nonzero vectors will give us an o.n. set of vectors after normalizing each of the vectors. If $\mathbf{0}$ is in an orthogonal set of vectors, then we can't normalize to get an o.n. set of vectors.

A common notation is the **Kronecker delta** defined by

$$\delta_{ij} \triangleq \begin{Bmatrix} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{Bmatrix}.$$

Using this symbol, we note that $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is an o.n. set if $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \delta_{ij}$ for all i, j. We also note that the $n \times n$ identity matrix

$$I_n = \left[\delta_{ij}\right]_{\substack{1 \le i \le n \\ 1 \le j \le n}}.$$

2.3.2 The Gram–Schmidt Process

This is a very useful technique both for producing o.n. sets and, as we will see in Section 2.7, factoring matrices.

Theorem 2.16

(The Gram–Schmidt process) From a linearly independent set of vectors $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ in $\mathbb{R}^m$ we can *construct* vectors $\mathbf{q}_1, \ldots, \mathbf{q}_n$ in $\mathbb{R}^m$ with the properties that

(a) $\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$ is an o.n. set.
(b) $Span\{\mathbf{q}_1, \ldots, \mathbf{q}_n\} = Span\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$.
(c) For each $i \leq n$, $\mathbf{q}_i$ is a linear combination of $\mathbf{a}_1, \ldots, \mathbf{a}_i$.

Before explaining the process in general, it's useful to do an example.

Example 2.18

Demonstrate the Gram–Schmidt process for the set of vectors

$$\left\{ \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ -2 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 5 \end{bmatrix} \right\}. \tag{2.13}$$

Method: Denote the vectors listed in (2.13) as $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$. To start a "recursive" process, let

$$\text{(1a)} \quad \mathbf{v}_1 \triangleq \mathbf{a}_1, \quad r_{11} \triangleq ||\mathbf{v}_1|| = \sqrt{2}, \quad \text{and} \quad \text{(1b)} \quad \mathbf{q}_1 = r_{11}^{-1}\mathbf{v}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}.$$

Next, let

$$\text{(2a)} \quad \mathbf{v}_2 \triangleq \mathbf{a}_2 - (\mathbf{a}_2 \bullet \mathbf{q}_1)\mathbf{q}_1 = \begin{bmatrix} 1 \\ -2 \\ 0 \end{bmatrix} - \left(\begin{bmatrix} 1 \\ -2 \\ 0 \end{bmatrix} \bullet \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \right) \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 \\ -2 \\ 0 \end{bmatrix} - \left(\frac{1}{\sqrt{2}} \right) \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0.5 \\ -2 \\ 0.5 \end{bmatrix}, \quad r_{22} \triangleq ||\mathbf{v}_2|| = \frac{3}{\sqrt{2}},$$

$$\text{and} \quad \text{(2b)} \quad \mathbf{q}_2 = r_{22}^{-1}\mathbf{v}_2 = \frac{1}{3\sqrt{2}} \begin{bmatrix} 1 \\ -4 \\ 1 \end{bmatrix}.$$

Finally, let

$$\text{(3a)} \quad \mathbf{v}_3 \triangleq \mathbf{a}_3 - (\mathbf{a}_3 \bullet \mathbf{q}_1)\mathbf{q}_1 - (\mathbf{a}_3 \bullet \mathbf{q}_2)\mathbf{q}_2$$

$$= \begin{bmatrix} 0 \\ 1 \\ 5 \end{bmatrix} - \left(\begin{bmatrix} 0 \\ 1 \\ 5 \end{bmatrix} \bullet \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \right) \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} - \left(\begin{bmatrix} 0 \\ 1 \\ 5 \end{bmatrix} \bullet \frac{1}{3\sqrt{2}} \begin{bmatrix} 1 \\ -4 \\ 1 \end{bmatrix} \right) \frac{1}{3\sqrt{2}} \begin{bmatrix} 1 \\ -4 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ 1 \\ 5 \end{bmatrix} - \left(\frac{-5}{\sqrt{2}} \right) \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} - \left(\frac{1}{3\sqrt{2}} \right) \frac{1}{3\sqrt{2}} \begin{bmatrix} 1 \\ -4 \\ 1 \end{bmatrix} = \cdots = \frac{11}{9} \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix},$$

$$r_{33} \triangleq ||\mathbf{v}_3|| = \frac{11}{3}, \quad \text{and} \quad \text{(3b)} \quad \mathbf{q}_3 = r_{33}^{-1}\mathbf{v}_3 = \frac{1}{3} \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}.$$

It is clear by construction that the conclusion in Theorem 2.16 (b) is true and that the $\mathbf{q}_i$'s are unit vectors; to see that they are orthogonal, we have, for example, that

$$\langle \mathbf{q}_2, \mathbf{q}_1 \rangle = r_{22}^{-1} \langle \mathbf{a}_2 - (\mathbf{a}_2 \bullet \mathbf{q}_1)\mathbf{q}_1, \mathbf{q}_1 \rangle = r_{22}^{-1} (\langle \mathbf{a}_2, \mathbf{q}_1 \rangle - (\mathbf{a}_2 \bullet \mathbf{q}_1) \langle \mathbf{q}_1, \mathbf{q}_1 \rangle)$$

$$= r_{22}^{-1} (\langle \mathbf{a}_2, \mathbf{q}_1 \rangle - \langle \mathbf{a}_2, \mathbf{q}_1 \rangle \cdot 1) = 0.$$

To summarize the conclusions, the o.n. set

$$\{\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3\} = \left\{ \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \frac{1}{3\sqrt{2}} \begin{bmatrix} 1 \\ -4 \\ 1 \end{bmatrix}, \frac{1}{3} \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} \right\}$$

satisfies the conclusions of Theorem 2.16 for the given set of three vectors $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$. ○

As an aside, the order in which we list the given vectors may affect what vectors are produced by the Gram–Schmidt process. Also, in Section 2.7, we will see why we denote the "r's" as r_{ii} instead of more simply r_i.

The general description of the Gram–Schmidt process is an **algorithm**, that is, a computational procedure: given a linearly independent set of vectors $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$, the initial step is to let

$$\mathbf{v}_1 \triangleq \mathbf{a}_1, r_{11} \triangleq ||\mathbf{v}_1||, \text{ and } \mathbf{q}_1 = r_{11}^{-1}\mathbf{v}_1.$$

Here's the "inductive step": Having already constructed an o.n. set $\{\mathbf{q}_1, \ldots, \mathbf{q}_{j-1}\}$ that has the property that

$$\text{for each } i \leq j-1, \mathbf{q}_i \text{ is a linear combination of } \mathbf{a}_1, \ldots, \mathbf{a}_i$$

and the property that

$$Span\{\mathbf{q}_1, \ldots, \mathbf{q}_{j-1}\} = Span\{\mathbf{a}_1, \ldots, \mathbf{a}_{j-1}\},$$

we let

$$\mathbf{v}_j \triangleq \mathbf{a}_j - (\mathbf{a}_j \bullet \mathbf{q}_1)\mathbf{q}_1 - \cdots - (\mathbf{a}_j \bullet \mathbf{q}_{j-1})\mathbf{q}_{j-1}, \quad r_{jj} \triangleq ||\mathbf{v}_j||, \quad \text{and} \quad \mathbf{q}_j \triangleq r_{jj}^{-1}\mathbf{v}_j.$$

By construction, $\mathbf{q}_j$ is a unit vector. To see that $\mathbf{q}_j$ is orthogonal to $\mathbf{q}_1, \ldots, \mathbf{q}_{j-1}$, we note that for $i \leq j-1$,

$$\begin{aligned} \langle \mathbf{q}_j, \mathbf{q}_i \rangle &= r_{jj}^{-1} \langle \mathbf{a}_j - (\mathbf{a}_j \bullet \mathbf{q}_1)\mathbf{q}_1 - \cdots - (\mathbf{a}_j \bullet \mathbf{q}_{j-1})\mathbf{q}_{j-1}, \mathbf{q}_i \rangle \\ &= r_{jj}^{-1} \big(\langle \mathbf{a}_j, \mathbf{q}_i \rangle - \langle \mathbf{a}_j, \mathbf{q}_1 \rangle \langle \mathbf{q}_1, \mathbf{q}_i \rangle - \cdots - \langle \mathbf{a}_j, \mathbf{q}_{j-1} \rangle \langle \mathbf{q}_{j-1}, \mathbf{q}_i \rangle \big). \end{aligned}$$

By orthonormality, all of the *subtracted* terms are zero except for the ith term, which is

$$-\langle \mathbf{a}_j, \mathbf{q}_i \rangle \langle \mathbf{q}_i, \mathbf{q}_i \rangle = -\langle \mathbf{a}_j, \mathbf{q}_i \rangle \cdot 1.$$

So,

$$\langle \mathbf{q}_j, \mathbf{q}_i \rangle = r_{jj}^{-1} (\langle \mathbf{a}_j, \mathbf{q}_i \rangle - \langle \mathbf{a}_j, \mathbf{q}_i \rangle \cdot 1) = r_{jj}^{-1} \cdot 0 = 0.$$

Because $\{\mathbf{a}_1, \ldots, \mathbf{a}_j\}$ is a linearly independent set of vectors, we can't write $\mathbf{a}_j$ as a linear combination of $\mathbf{a}_1, \ldots, \mathbf{a}_{j-1}$. After a little further reasoning, it follows that $\mathbf{v}_j \neq \mathbf{0}$.

As for conclusion (c) of Theorem 2.16, $\mathbf{q}_1$ is a linear combination of $\mathbf{a}_1$, and $\mathbf{q}_2$ is a linear combination of $\mathbf{a}_1, \mathbf{a}_2$. Continuing in this way, $\mathbf{q}_j$ is a linear combination of $\mathbf{a}_1, \ldots, \mathbf{a}_{j-1}$.

It follows that $\mathbf{v}_j$, and hence $\mathbf{q}_j$, is a linear combination of $\mathbf{a}_j$ and $\mathbf{a}_1, \ldots, \mathbf{a}_{j-1}$, hence is a linear combination of $\mathbf{a}_1, \ldots, \mathbf{a}_j$. This explains (*c*) and also explains why

$$Span\{\mathbf{q}_1, \ldots, \mathbf{q}_n\} \subseteq Span\{\mathbf{a}_1, \ldots, \mathbf{a}_n\},$$

that is, every vector in $Span\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$ must also be in $Span\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$.

As for conclusion (b) of Theorem 2.16, $\mathbf{a}_1$ is a linear combination of $\mathbf{q}_1$ and $\mathbf{a}_2 = \mathbf{v}_2 + (\mathbf{a}_2 \bullet \mathbf{q}_1)\mathbf{q}_1 = r_{22}\mathbf{q}_2 + (\mathbf{a}_2 \bullet \mathbf{q}_1)\mathbf{q}_1$ is a linear combination of $\mathbf{q}_1, \mathbf{q}_2$. Continuing in this way, the inductive step explains why

$$\mathbf{a}_j = r_{jj}\mathbf{q}_j + (\mathbf{a}_j \bullet \mathbf{q}_1)\mathbf{q}_1 + \cdots + (\mathbf{a}_j \bullet \mathbf{q}_{j-1})\mathbf{q}_{j-1} \tag{2.14}$$

so $\mathbf{a}_j$ is a linear combination of $\mathbf{q}_1, \ldots, \mathbf{q}_{j-1}, \mathbf{q}_j$. It follows that

$$Span\{\mathbf{a}_1, \ldots, \mathbf{a}_n\} \subseteq Span\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}.$$

This, along with the other subset relationship at the end of the previous paragraph, explains why $Span\{\mathbf{a}_1, \ldots, \mathbf{a}_n\} = Span\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$. □

We will use (2.14) again when we study the "*QR* factorization" in Section 2.7.

2.3.3 Orthogonal Projections

In physics, work is defined as $\overrightarrow{Force} \bullet \overrightarrow{displacement}$, assuming $\overrightarrow{Force}$ is constant. In effect, we are interested in the "component" of force in the direction of the displacement vector. This concept is formalized in the concept of "orthogonal projection."

Definition 2.15

An **orthogonal projection** on $\mathbb{R}^n$ is a real, $n \times n$, matrix P that satisfies both $P^2 = P$ and $P^T = P$.

Theorem 2.17

If P is an orthogonal projection on $\mathbb{R}^n$ then (a) for every $\mathbf{x}, \mathbf{y}$ in $\mathbb{R}^n$

$$P\mathbf{x} \perp (I - P)\mathbf{y}, \tag{2.15}$$

and (b) every $\mathbf{x}$ in $\mathbb{R}^n$ can be decomposed into a sum of orthogonal vectors:

$$\mathbf{x} = I\mathbf{x} = (P + I - P)\mathbf{x} = P\mathbf{x} + (I - P)\mathbf{x}.$$

Why? Concerning (a), for any $\mathbf{x}$ in $\mathbb{R}^n$,

$$\begin{aligned}\langle P\mathbf{x}, (I - P)\mathbf{y}\rangle &= (P\mathbf{x})^T(I - P)\mathbf{y} = \mathbf{x}^T P^T(I - P)\mathbf{y} = \mathbf{x}^T P(I - P)\mathbf{y} = \mathbf{x}^T(P - P^2)\mathbf{y} \\ &= \mathbf{x}^T O\mathbf{y} = 0. \ \square\end{aligned}$$

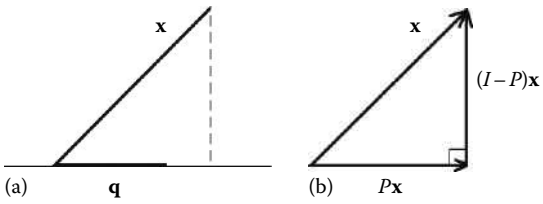


FIGURE 2.5
Projection onto a line: (a) $\mathbf{x}$ is to be projected onto $\mathbf{q}$, (b) Decompose $\mathbf{x} = P\mathbf{x} + (I - P)\mathbf{x}$.

Theorem 2.17(b) is illustrated in Figure 2.5b.

Lemma 2.1

If $\mathbf{q}$ is a unit vector in $\mathbb{R}^n$ then

$$P \triangleq \mathbf{q}\mathbf{q}^T$$

is a projection.

You will explain why $P \triangleq \mathbf{q}\mathbf{q}^T$ satisfies the two requirements of an orthogonal projection in Problem 2.3.4.17.

Theorem 2.18

If $\mathbf{q}$ is a unit vector and $P = \mathbf{q}\mathbf{q}^T$, then for every $\mathbf{x}$ in $\mathbb{R}^n$

$$P\mathbf{x} = \langle \mathbf{x}, \mathbf{q} \rangle \mathbf{q}. \tag{2.16}$$

Why?

$$P\mathbf{x} = (\mathbf{q}\mathbf{q}^T)\mathbf{x} = \mathbf{q}(\mathbf{q}^T\mathbf{x}) = (\mathbf{q}^T\mathbf{x})\mathbf{q} = \langle \mathbf{q}, \mathbf{x} \rangle \mathbf{q} = \langle \mathbf{x}, \mathbf{q} \rangle \mathbf{q}. \square$$

In the latter calculation, we again used the "abuse of notation" of considering $(\mathbf{q}^T\mathbf{x})$ to be both a 1×1 matrix and a scalar.

Geometrically, $P = \mathbf{q}\mathbf{q}^T$ is a projection onto the vector subspace spanned by the (unit) vector $\mathbf{q}$, that is, a line through the origin, as illustrated in Figure 2.5a.

Theorem 2.19

Suppose $\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$ is an o.n. set in $\mathbb{R}^m$ and we define

$$P_n = \mathbf{q}_1\mathbf{q}_1^T + \cdots + \mathbf{q}_n\mathbf{q}_n^T.$$

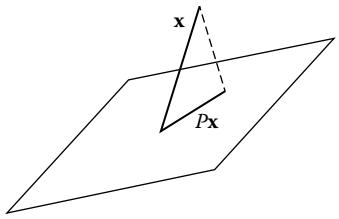


FIGURE 2.6
Projection onto a plane.

Then, (a) P_n is an orthogonal projection and (b) $P_n\mathbf{q}_i = \mathbf{q}_i$ for every $i \leq n$.

Geometrically, P_n is a projection onto the vector subspace $\mathcal{V}_n = Span\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$. For the case $n = 2$, $\mathcal{V}_2$ is a plane, and the projection P_2 is illustrated in Figure 2.6.

Using the concept of orthogonal projection and the notation for P_n we can restate the inductive step in the Gram–Schmidt process as

$$\mathbf{v}_j \triangleq (I - P_{j-1})\mathbf{a}_j, \qquad r_{jj} = ||\mathbf{v}_j||, \quad \text{and} \quad \mathbf{q}_j \triangleq r_{jj}^{-1}\mathbf{v}_j.$$

In this clean notation, it's even easier to see that $\mathbf{q}_j$ is orthogonal to each of $\mathbf{q}_1, \ldots, \mathbf{q}_{j-1}$, using Theorem 2.19(b).

Appendix

To explain Theorem 2.13(e), that is, (2.9)(e), the "Cauchy–Schwarz" inequality, let's consider two cases: if $\mathbf{y} = \mathbf{0}$, then (2.9)(e) would follow from

$$|\langle \mathbf{x}, \mathbf{0}\rangle| = |\ 0\ | = 0 = ||\,\mathbf{x}\,|| \cdot 0 = ||\,\mathbf{x}\,||\ ||\mathbf{0}||.$$

On the other hand, suppose $\mathbf{y} \neq \mathbf{0}$. Our subsequent work will essentially be a calculation. Let α be a scalar; later we will choose a useful value for it. We calculate

$$0 \leq ||\mathbf{x} - \alpha\mathbf{y}||^2 = ||\,\mathbf{x}\,||^2 + ||\alpha\mathbf{y}||^2 - 2\langle \mathbf{x}, \alpha\mathbf{y}\rangle = ||\,\mathbf{x}\,||^2 + |\alpha|^2||\,\mathbf{y}\,||^2 - 2\alpha\langle \mathbf{x}, \mathbf{y}\rangle.$$

If we think of the RHS as a polynomial in α, we can complete the square:

$$0 \leq \alpha^2||\,\mathbf{y}\,||^2 - 2\alpha\langle \mathbf{x}, \mathbf{y}\rangle + ||\,\mathbf{x}\,||^2 = ||\,\mathbf{y}\,||^2\left(\alpha^2 - \left(2\,\frac{\langle \mathbf{x}, \mathbf{y}\rangle}{||\,\mathbf{y}\,||^2}\right)\alpha\right) + ||\,\mathbf{x}\,||^2$$

$$= ||\,\mathbf{y}\,||^2\left(\alpha^2 - \left(2\,\frac{\langle \mathbf{x}, \mathbf{y}\rangle}{||\,\mathbf{y}\,||^2}\right)\alpha + \left(\frac{\langle \mathbf{x}, \mathbf{y}\rangle}{||\,\mathbf{y}\,||^2}\right)^2 - \left(\frac{\langle \mathbf{x}, \mathbf{y}\rangle}{||\,\mathbf{y}\,||^2}\right)^2\right) + ||\,\mathbf{x}\,||^2$$

$$= ||\,\mathbf{y}\,||^2\left(\alpha^2 - \left(2\,\frac{\langle \mathbf{x}, \mathbf{y}\rangle}{||\,\mathbf{y}\,||^2}\right)\alpha + \left(\frac{\langle \mathbf{x}, \mathbf{y}\rangle}{||\,\mathbf{y}\,||^2}\right)^2\right) - ||\,\mathbf{y}\,||^2\left(\frac{\langle \mathbf{x}, \mathbf{y}\rangle}{||\,\mathbf{y}\,||^2}\right)^2 + ||\,\mathbf{x}\,||^2$$

that is,

$$0 \leq ||\,\mathbf{y}\,||^2 \left(\alpha - \frac{\langle \mathbf{x}, \mathbf{y}\rangle}{||\,\mathbf{y}\,||^2}\right)^2 - \frac{\left(\langle \mathbf{x}, \mathbf{y}\rangle\right)^2}{||\,\mathbf{y}\,||^2} + ||\,\mathbf{x}\,||^2 \tag{2.17}$$

Now, if we choose

$$\alpha = \frac{\langle \mathbf{x}, \mathbf{y}\rangle}{||\,\mathbf{y}\,||^2},$$

then (2.17) becomes

$$0 \leq 0 - \frac{\left(\langle \mathbf{x}, \mathbf{y}\rangle\right)^2}{||\,\mathbf{y}\,||^2} + ||\,\mathbf{x}\,||^2.$$

that is,

$$0 \leq \frac{\left|\langle \mathbf{x}, \mathbf{y}\rangle\right|^2}{||\,\mathbf{y}\,||^2} \leq ||\,\mathbf{x}\,||^2.$$

Multiply both sides of the inequality by $||\,\mathbf{y}\,||^2$ to get

$$\left|\langle \mathbf{x}, \mathbf{y}\rangle\right|^2 \leq ||\,\mathbf{x}\,||^2 ||\,\mathbf{y}\,||^2$$

and take the square root of both sides to get $|\langle \mathbf{x}, \mathbf{y}\rangle| \leq ||\,\mathbf{x}\,||\;||\,\mathbf{y}\,||$. □

2.3.4 Problems

In problems 1–4, find an o.n. set whose span equals the span of the given set of vectors. Use exact values, that is, do not make decimal approximations.

1. $\left\{\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}2\\0\end{bmatrix}\right\}$
2. $\left\{\begin{bmatrix}1\\0\\1\end{bmatrix}, \begin{bmatrix}0\\1\\1\end{bmatrix}, \begin{bmatrix}1\\1\\1\end{bmatrix}\right\}$
3. $\left\{\begin{bmatrix}-1\\2\\0\end{bmatrix}, \begin{bmatrix}2\\1\\-1\end{bmatrix}, \begin{bmatrix}1\\0\\2\end{bmatrix}\right\}$
4. $\left\{\begin{bmatrix}1\\-1\\1\end{bmatrix}, \begin{bmatrix}1\\1\\-1\end{bmatrix}, \begin{bmatrix}-1\\1\\1\end{bmatrix}\right\}$
5. Suppose $\mathbf{a}_1 = \begin{bmatrix}3\\-4\end{bmatrix}$ and $\mathbf{a}_2 = \begin{bmatrix}1\\\sqrt{2}\end{bmatrix}$. Find an o.n. set $\mathcal{S} = \{\mathbf{q}_1, \mathbf{q}_2\}$ such that $\text{Span}(\{\mathbf{a}_1, \mathbf{a}_2\}) = \text{Span}(\mathcal{S})$.

6. Let

$$\mathbf{a}_1 = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}, \ \mathbf{a}_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \ \mathbf{a}_3 = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}.$$

   a. Use the Gram–Schmidt (G.–S.) process on $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$.

   b. Let $\mathbf{w}_1 = \mathbf{a}_2, \mathbf{w}_2 = \mathbf{a}_3, \mathbf{w}_3 = \mathbf{a}_1$. Use the Gram–Schmidt (G.–S.) process on $\{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3\}$.

7. If you use the Gram–Schmidt (G.–S.) process on a set of vectors, the conclusion may depend on the order in which you listed the original vectors. Find an example of two vectors $\mathbf{x}, \mathbf{y}$ in $\mathbb{R}^2$ so that G.–S. on $\{\mathbf{x}, \mathbf{y}\}$ and G.–S. on $\{\mathbf{y}, \mathbf{x}\}$ give two different o.n. sets.

8. Suppose $\{\mathbf{a}_1, \mathbf{a}_2\}$ is a linearly independent set of vectors in $\mathbb{R}^{137}$ and $\langle \mathbf{a}_1, \mathbf{a}_2 \rangle = \frac{1}{3}$. Find, in terms of $\{\mathbf{a}_1, \mathbf{a}_2\}$, the set of vectors $\{\mathbf{q}_1, \mathbf{q}_2\}$ produced by the Gram–Schmidt process.

9. Is it true that for all vectors $\mathbf{x}, \mathbf{y}, \mathbf{z}$

$$\langle \mathbf{x}, \mathbf{x} + \mathbf{y} - \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{y} + \mathbf{z} - \mathbf{x} \rangle + \langle \mathbf{z}, \mathbf{z} + \mathbf{x} - \mathbf{y} \rangle = ||\mathbf{x}||^2 + ||\mathbf{y}||^2 + ||\mathbf{z}||^2?$$

   If so, why? If not, give a specific counterexample.

10. Suppose $\mathbf{x}, \mathbf{y}, \mathbf{z}$ are nonzero vectors in $\mathbb{R}^n$, $\mathbf{x} \perp \mathbf{y}$, and $\mathbf{y} \perp \mathbf{z}$. For two different values of n and for each of the three conclusions (i), (ii), (iii), state whether it is true or false. If true, explain why; if false, give a specific counterexample.
    (i) $(\mathbf{x} + \mathbf{z}) \perp \mathbf{y}$, (ii) $\mathbf{x} \parallel \mathbf{z}$, (iii) $\mathbf{x} \perp (-\mathbf{y})$.
    a. if $n = 2$
    b. if $n = 3$

11. Suppose $\mathbf{a}_i$, $i = 1, 2, 3$ are unspecified vectors in $\mathbb{R}^3$. Suppose that all we know about them is that

$$\langle \mathbf{a}_1, \mathbf{a}_1 \rangle = 2, \langle \mathbf{a}_1, \mathbf{a}_2 \rangle = 3, \langle \mathbf{a}_1, \mathbf{a}_3 \rangle = 4,$$

$$\langle \mathbf{a}_2, \mathbf{a}_2 \rangle = 5, \langle \mathbf{a}_2, \mathbf{a}_3 \rangle = 6, \langle \mathbf{a}_3, \mathbf{a}_3 \rangle = 9.$$

    a. Construct from $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$ a set of three vectors that is a basis for $\mathbb{R}^3$ and is an orthogonal (not necessarily o.n.) set.

    b. Construct from $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$ an o.n. basis for $\mathbb{R}^3$. [Hint: it may help to use the identity (2.9)(b).]

12. Explain why for all vectors $\mathbf{x}, \mathbf{y}$

$$||\mathbf{x} + \mathbf{y}||^2 + ||\mathbf{x} - \mathbf{y}||^2 = 2(||\mathbf{x}||^2 + ||\mathbf{y}||^2).$$

    [Hint: Use (2.9)(b).]

13. Use Theorem 2.13 (b) to explain why Theorem 2.15, that is, the Pythagorean Theorem, is true.

14. Explain why $||\alpha \mathbf{x}|| = |\alpha| \, ||\mathbf{x}||$ for all scalars α and vectors $\mathbf{x}$ in $\mathbb{R}^n$.

15. Suppose that $\{\mathbf{a}_1, \ldots, \mathbf{a}_{n-1}\}$ is a linearly independent set of vectors in $\mathbb{R}^m$ and we have used the Gram–Schmidt process to construct an o.n. set $\{\mathbf{q}_1, \ldots, \mathbf{q}_{n-1}\}$. Suppose $\mathbf{a}_n$ is a linear combination of $\{\mathbf{a}_1, \ldots, \mathbf{a}_{n-1}\}$. What happens if we try to continue the Gram–Schmidt process, that is, we try to construct $\mathbf{q}_n$?
16. Suppose A has a real eigenvalue λ and corresponding unit eigenvector $\mathbf{u}$. In terms of λ, find the exact values of (a) $\langle \mathbf{u}, A\mathbf{u} \rangle$, (b) $\langle \mathbf{u}, A^2\mathbf{u} \rangle$, (c) $||A\mathbf{u}||^2$.
17. Assume $\mathbf{q}$ is a unit vector and define $P \triangleq \mathbf{q}\mathbf{q}^T$. Explain why it is an orthogonal projection, that is, satisfies the two properties of an orthogonal projection.
18. Suppose $\{\mathbf{q}_1, \mathbf{q}_2\}$ is an o.n. set of vectors in $\mathbb{R}^m$. Let $P_i \triangleq \mathbf{q}_i\mathbf{q}_i^T$ for $i = 1, 2$. Let $A = \gamma_1 P_1 + \gamma_2 P_2$ where γ_i are unspecified scalars γ_1, γ_2.
    a. If $\mathbf{w} = \alpha_1\mathbf{q}_1 + \alpha_2\mathbf{q}_2$ for unspecified scalars α_1, α_2, find $A\mathbf{w}$ in terms of $\alpha_1, \alpha_2, \gamma_1, \gamma_2, \mathbf{q}_1, \mathbf{q}_2$.
    b. If, in addition, $m = 2$ and $\gamma_1 \neq \gamma_2$, find all of the eigenvalues and eigenvectors of A in terms of γ_1, γ_2.
19. Suppose P_1 and P_2 are orthogonal projections and $P_1P_2 = P_2P_1$. Explain why P_1P_2 is an orthogonal projection, too.

2.4 Orthonormal Bases and Orthogonal Matrices

2.4.1 Orthogonal Sets and Bases

Theorem 2.20

Every o.n. set is linearly independent.

Why? Suppose $\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$ is an o.n. set. To explain why it is linearly independent, we will explain why the equation

$$\mathbf{0} = c_1\mathbf{q}_1 + \cdots + c_n\mathbf{q}_n \tag{2.18}$$

has only the trivial solution for $c_1, \ldots, c_n$. To see this, for any i with $1 \leq i \leq n$, operate with $P \triangleq \mathbf{q}_i\mathbf{q}_i^T$ on both sides of (2.18) to get

$$\mathbf{0} = P\mathbf{0} = P(c_1\mathbf{q}_1 + \cdots + c_n\mathbf{q}_n) = c_1P\mathbf{q}_1 + \cdots + c_nP\mathbf{q}_n = c_i\mathbf{q}_i,$$

hence, $c_i = 0$. Since this is true for each i, (2.18) has only the trivial solution, and so $\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$ is linearly independent. □

Corollary 2.2

If $\mathcal{S}$ is an orthogonal set of nonzero vectors, then $\mathcal{S}$ is a basis for $Span(\mathcal{S})$.

Corollary 2.3

If $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ is a linearly independent set of vectors in $\mathbb{R}^m$ and $\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$ is the o.n. set produced by the Gram–Schmidt process (and thus satisfies the conclusions of Theorem 2.16) in Section 2.3, then $\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$ is an o.n. basis for $Span\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$.

Given a real, $m \times n$ matrix, define

$$Col(A) \triangleq \{A\mathbf{x} : \ \mathbf{x} \text{ in } \mathbb{R}^n\},$$

which is called the **column space** of A or **range** of A. The MATLAB command **`colspace(A)`** gives a basis for the column space of A.

Write A in terms of its columns, that is,

$$A = \left[\begin{array}{c:c:c:c} \mathbf{a}_1 & & \cdots & \mathbf{a}_n \end{array}\right].$$

By Lemma 1.3 in Section 1.7, for every

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}.$$

we have

$$A\mathbf{x} = x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n,$$

so every vector in $Col(A)$ is a linear combination of the columns of A, that is,

$$Col(A) = Span(\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}).$$

Corollary 2.4

If the columns of a real, $m \times n$ matrix A are linearly independent then the following are bases for $Col(A)$:

(a) $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$, and
(b) $\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$, the o.n. set produced by the Gram–Schmidt process.

Corollary 2.5

If a real, $m \times n$ matrix A has $rank(A) = \ell$ then use Theorem 1.41 in Section 1.7 to construct a basis $\{\mathbf{a}_{i_1}, \ldots, \mathbf{a}_{i_\ell}\}$ for $Col(A)$. Then use the Gram–Schmidt process in Theorem 2.16 in Section 2.3 to construct an o.n. basis $\{\mathbf{q}_1, \ldots, \mathbf{q}_\ell\}$ for $Col(A)$. From that o.n. basis we can use Theorem 2.19 in Section 2.3 to construct P_A, an orthogonal projection onto $Col(A)$.

Corollary 2.6

If $\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$ is an o.n. set in $\mathbb{R}^n$ and $\mathbf{x}, \mathbf{y}$ are in $\mathbb{R}^n$ then

$$\left\{\begin{array}{l} \text{(a)}\ \ \mathbf{x} = \langle \mathbf{x}, \mathbf{q}_1\rangle \mathbf{q}_1 + \cdots + \langle \mathbf{x}, \mathbf{q}_n\rangle \mathbf{q}_n \\ \text{(b)}\ \ ||\,\mathbf{x}\,||^2 = |\langle \mathbf{q}_1, \mathbf{x}\rangle|^2 + \cdots + |\langle \mathbf{q}_n, \mathbf{x}\rangle|^2 \\ \text{(c)}\ \ \langle \mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{q}_1\rangle\langle \mathbf{q}_1, \mathbf{y}\rangle + \cdots + \langle \mathbf{x}, \mathbf{q}_n\rangle\langle \mathbf{q}_n, \mathbf{y}\rangle \end{array}\right\}. \tag{2.19}$$

Note that these results discuss the special case of an o.n. set of n vectors in $\mathbb{R}^n$.

Why? By Theorems 1.43 in Section 1.7 and 2.20 in Section 2.4, $\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$ is a basis for $\mathbb{R}^n$, so there are scalars $c_1, \ldots, c_n$ such that

$$\mathbf{x} = c_1\mathbf{q}_1 + \cdots + c_n\mathbf{q}_n. \tag{2.20}$$

If we take the inner product of both sides of (2.20) with $\mathbf{q}_j$ we get

$$\langle \mathbf{x}, \mathbf{q}_j\rangle = c_1\langle \mathbf{q}_1, \mathbf{q}_j\rangle + \cdots + c_n\langle \mathbf{q}_n, \mathbf{q}_j\rangle = c_1\delta_{1j} + \cdots + c_n\delta_{nj} = c_j,$$

from which (2.19)(a) follows. Because the Pythagorean theorem explains why

$$||\,\mathbf{x}\,||^2 = ||c_1\mathbf{q}_1||^2 + \cdots + ||c_n\mathbf{q}_n||^2 = |c_1|^2||\mathbf{q}_1||^2 + \cdots + |c_n|^2||\mathbf{q}_n||^2 = |c_1|^2 + \cdots + |c_n|^2,$$

(2.19)(a) implies (2.19)(b). As for (2.19)(c), you will do that calculation in Problem 2.4.4.15. □

By the way, (2.19)(b) and (c) are known as **Parseval identities**. Both (2.19)(a) and the conclusion of the next result are known as "Expansion" theorems.

Corollary 2.7

Suppose $\mathbf{v}_1, \ldots, \mathbf{v}_\ell$ is an orthogonal set of nonzero vectors in $\mathbb{R}^m$ and $\mathbf{y}$ is a vector in $Span\{\mathbf{v}_1, \ldots, \mathbf{v}_\ell\}$. Then

$$\mathbf{y} = \sum_{i=1}^{k} \frac{1}{||\mathbf{v}_i||^2}\langle \mathbf{y},\ \mathbf{v}_i\rangle \mathbf{v}_i.$$

2.4.2 Orthogonal Matrices

Definition 2.16

A square matrix Q is an **orthogonal matrix** if it satisfies $Q^TQ = I$.

Partitioning *any* matrix Q in terms of its columns, that is, $Q = \left[\begin{array}{c:c:c:c} \mathbf{q}_1 & \cdots & \mathbf{q}_n \end{array}\right]$, we have

$$Q^TQ = \left[\begin{array}{c} \mathbf{q}_1^T \\ \hdashline \\ \vdots \\ \hdashline \\ \mathbf{q}_n^T \end{array}\right] \left[\begin{array}{c:c:c:c} \mathbf{q}_1 & \cdots & \mathbf{q}_n \end{array}\right] = \left[\mathbf{q}_i^T\mathbf{q}_j\right]_{\substack{i=1,\ldots,n \\ j=1,\ldots,n}}.$$

Remarks

(1) For an $n \times n$ matrix to be a real, orthogonal matrix, its columns must be an o.n. set in $\mathbb{R}^n$, hence must be an o.n. basis for $\mathbb{R}^n$.
(2) If Q is an orthogonal matrix then it is invertible and $Q^{-1} = Q^T$. You will explain the following related result in Problem 2.4.4.22.

Corollary 2.8

Suppose $Q = \left[\begin{array}{c:c:c:c} \mathbf{q}_1 & \cdots & \mathbf{q}_n \end{array}\right]$ is an orthogonal matrix. Then the unique solution of $Q\mathbf{x} = \mathbf{b}$ is

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \langle \mathbf{b}, \mathbf{q}_1 \rangle \\ \vdots \\ \langle \mathbf{b}, \mathbf{q}_n \rangle \end{bmatrix} = \sum_{i=1}^{n} \langle \mathbf{b}, \ \mathbf{q}_j \rangle \mathbf{e}^{(j)}.$$

Theorem 2.21

If $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ is an o.n. set in $\mathbb{R}^m$ and $r < m$, then we can "complete the o.n. set to get an o.n. basis for $\mathbb{R}^m$," that is, construct $(m - r)$ vectors $\{\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m\}$ in $\mathbb{R}^m$ so that

$$\{\mathbf{u}_1, \ldots, \mathbf{u}_r; \mathbf{u}_{r+1}, \ldots, \mathbf{u}_m\}$$

is an o.n. set in $\mathbb{R}^m$.

We will explain this theorem in the appendix later.

Corollary 2.9

Suppose $\mathbf{x}^{(1)}$ is a unit vector in $\mathbb{R}^n$ and $\mathbf{z}$ is a vector in $\mathbb{R}^n$ that has the property that every vector that is orthogonal to $\mathbf{x}^{(1)}$ must also be orthogonal to $\mathbf{z}$. Then $\mathbf{z}$ is a scalar multiple of $\mathbf{x}^{(1)}$.

Why? In $\mathbb{R}^3$ this seems to be geometrically obvious: A unit vector $\mathbf{x}^{(1)}$ is, by definition, normal to the plane $\Pi \triangleq \{\mathbf{x}^{(1)}\}^{\perp}$, so if $\mathbf{z} \neq \mathbf{0}$, then $\mathbf{z}$ being orthogonal to every vector that is orthogonal to $\mathbf{x}^{(1)}$ says that $\mathbf{z}$ is also normal to the plane Π. Because $\mathbf{x}^{(1)}$ and $\mathbf{z}$ are both normal to the plane Π, geometrical intuition tells us that $\mathbf{z}$ must be a scalar multiple of $\mathbf{x}^{(1)}$.

To explain the result in general and without using intuition, we apply Theorem 2.21: because $\mathbf{x}^{(1)}$ is a unit vector in $\mathbb{R}^n$, there must be vectors $\mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}$ so that $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ is an o.n. basis for $\mathbb{R}^n$. It follows from Corollary 2.6 (a) that

$$\mathbf{z} = \langle \mathbf{z}, \mathbf{x}^{(1)} \rangle \mathbf{x}^{(1)} + \langle \mathbf{z}, \mathbf{x}^{(2)} \rangle \mathbf{x}^{(2)} + \cdots + \langle \mathbf{z}, \mathbf{x}^{(n)} \rangle \mathbf{x}^{(n)}.$$

But $\mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}$, being orthogonal to $\mathbf{x}^{(1)}$, must be orthogonal to $\mathbf{z}$, as we assumed. So,

$$\mathbf{z} = \langle \mathbf{z}, \mathbf{x}^{(1)} \rangle \mathbf{x}^{(1)} + \mathbf{0} + \cdots + \mathbf{0} = (\langle \mathbf{z}, \mathbf{x}^{(1)} \rangle) \mathbf{x}^{(1)},$$

that is, $\mathbf{z}$ is a multiple of $\mathbf{x}^{(1)}$. □

2.4.3 Appendix

We will explain Theorem 2.21, that is, why a given o.n. set $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ in $\mathbb{R}^m$ can be "completed" to give an o.n. basis for $\mathbb{R}^m$. We will use a version of the Gram–Schmidt process on the set of vectors

$$\left\{ \mathbf{u}_1, \ldots, \mathbf{u}_r; \mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(m)} \right\},$$

where $\mathbf{e}^{(j)}, j = 1, .., m$ are the columns of the $m \times m$ identity matrix. The vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ are assumed to be an o.n. set, so the Gram–Schmidt process starts with the next vector, $\mathbf{e}^{(1)}$: We calculate

$$\mathbf{v}_{r+1} = \mathbf{e}^{(1)} - \sum_{j=1}^{r} (\mathbf{e}^{(1)} \bullet \mathbf{u}_j) \mathbf{u}_j. \tag{2.21}$$

There are two possibilities regarding (2.21):

1. First, suppose $\mathbf{v}_{r+1} \neq \mathbf{0}$, that is, that $\mathbf{e}^{(1)}$ is not in the $Span\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$. In this case, we let $\mathbf{u}_{r+1} = \frac{1}{||\mathbf{v}_{r+1}||} \mathbf{v}_{r+1}$, so now we have an o.n. set of $r+1$ vectors. For example, in Example 2.30 in Section 2.7, we will have $r = 2 < 3 = m$ and we will be able to construct $\mathbf{u}_3$ using the Gram–Schmidt process on $\mathbf{e}^{(1)}$.

   The next step would be to try to find the $(r+2)$th vector, using

$$\mathbf{v}_{r+2} = \mathbf{e}^{(2)} - \sum_{j=1}^{r+1} (\mathbf{e}^{(2)} \bullet \mathbf{u}_j) \mathbf{u}_j$$

2. Second, suppose instead that $\mathbf{v}_{r+1} = \mathbf{0}$. Then our o.n. set is stuck at still having only the original r vectors, $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$. We try to find the $(r+1)$th vector in the o.n. set using

$$\mathbf{v}_{r+1} = \mathbf{e}^{(2)} - \sum_{j=1}^{r} (\mathbf{e}^{(2)} \bullet \mathbf{u}_j)\mathbf{u}_j.$$

Continuing in this way, eventually we will find an $(r+1)$th vector. Why? Because it cannot be true that each of $\mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(m)}$ are in $Span\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$, because that would say that $Span\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ is $\mathbb{R}^m$, which would violate the initial assumption that $r < m$ and the fact that the dimension of $\mathbb{R}^m$ is m.

So, eventually we will find a vector $\mathbf{u}_{r+1}$ to put in our o.n. set. Likewise, eventually we will find an $(r+2)$th vector $\mathbf{u}_{r+1}$ to put in our o.n. set, etc. Continuing in this way, we use as many of $\mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(m)}$ as we need to construct $\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m$. □

To increase our intuition about the aforementioned method, consider a special situation when all of the vectors $\mathbf{u}_1, \ldots, \mathbf{u}_r$ have zeros in their bottom $m-r$ rows. Then clearly each of $\mathbf{e}^{(r+1)}, \ldots, \mathbf{e}^{(m)}$ would be in $\big(\text{Span}\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}\big)^{\perp}$ so we could take $\mathbf{u}_\ell = \mathbf{e}^{(\ell)}$, $\ell = r+1, \ldots, m$. This would give us a shortcut to completing the basis.

2.4.4 Problems

1. Can we use the set of vectors

$$\left\{ \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ 0 \\ -1 \end{bmatrix} \right\}$$

   to construct an o.n. basis for $\mathbb{R}^3$ using the G.-S. process? If so, do that; if not, why not?
2. Find three different real, 2×2, orthogonal matrices.
3. Find all values of a, b for which

$$A = \begin{bmatrix} a & 0 & 1/\sqrt{2} \\ 1/\sqrt{2} & 0 & b \\ 0 & 1 & 0 \end{bmatrix}$$

   is a real, orthogonal matrix.
4. Find a 3×3 real, orthogonal matrix Q that has *no* zero entry, that is, $q_{ij} \neq 0$ for all i, j. Show *how* you found your Q.
5. Find a 3×3 real, orthogonal matrix Q whose first column is $\frac{1}{\sqrt{14}} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$.
6. Is it possible to find a 3×3 real, orthogonal matrix Q that has exactly three entries of zero? Why, or why not?
7. Find a 3×3 real, orthogonal matrix Q that has exactly four entries of zero.

8. Explain why for all angles ϕ, θ the matrix

$$A = \begin{bmatrix} \sin\phi\cos\theta & -\sin\theta & \cos\phi\cos\theta \\ \sin\phi\sin\theta & \cos\theta & \cos\phi\sin\theta \\ \cos\phi & 0 & -\sin\phi \end{bmatrix}$$

is orthogonal.

9. Suppose Q_1 and Q_2 are both orthogonal matrices. Explain why $Q \triangleq Q_1 Q_2$ is also an orthogonal matrix.
10. Suppose Q is a real, $m \times n$, matrix whose set of columns is an o.n. set.
    a. If $m = n$, explain why $rank(Q) = n$.
    b. If $m > n$, explain why $rank(Q) = n$.
    c. Why do we not need to discuss the case $m < n$?
11. If Q is a real, orthogonal matrix, explain why $\langle Q\mathbf{x}, Q\mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{y}\rangle$ for all $\mathbf{x}, \mathbf{y}$. [Hint: Rewrite the left-hand side as a matrix multiplication involving a transpose.] By the way, sometimes this property of real, orthogonal matrices is stated as "(multiplication by) a real, orthogonal matrix preserves angles between vectors."
12. If Q is a real, orthogonal matrix, explain why
    a. $||Q\mathbf{x}|| = ||\mathbf{x}||$ for all $\mathbf{x}$. [Hint: First, rewrite $||Q\mathbf{x}||^2 = \langle Q\mathbf{x}, Q\mathbf{x}\rangle$.]
    b. Every eigenvalue λ of Q satisfies $|\lambda| = 1$. [Hint: If $\mathbf{x}$ is a unit eigenvector of Q corresponding to λ, then $1 = ||\mathbf{x}||^2 = ||Q\mathbf{x}||^2 = ||\lambda\mathbf{x}||^2$.] By the way, sometimes the property of orthogonal matrices in part (a) is stated as "(multiplication by) an orthogonal matrix preserves lengths of vectors."
13. Let $B = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\frac{2}{\sqrt{6}} \end{bmatrix}$. Is it true that $||B\mathbf{x}||^2 = ||\mathbf{x}||^2$ for all $\mathbf{x}$ in $\mathbb{R}^3$?
14. Let $A = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{\sqrt{2}} \end{bmatrix}$. Is it true that $||A\mathbf{x}||^2 = ||\mathbf{x}||^2$ for all $\mathbf{x}$ in $\mathbb{R}^3$?
15. Use a calculation, that is, derivation, to explain why (2.19)(c) is correct.
16. Explain why the determinant of every real, orthogonal matrix is ± 1. [Hints: Use $Q^T Q = I$ and Theorem 1.28(a) and (b) in Section 1.6.]
17. A **Householder matrix** is of the form $Q = I - 2\mathbf{q}\mathbf{q}^T$, where $\mathbf{q}$ is *any* unit vector. Explain why each such matrix satisfies $Q^T = Q$ and is an orthogonal matrix.
18. Suppose $A = I - 2\mathbf{q}_1\mathbf{q}_1^T - 2\mathbf{q}_2\mathbf{q}_2^T$, where $\{\mathbf{q}_1, \mathbf{q}_2\}$ is *any* o.n. set of vectors. Explain why each such matrix satisfies $A^T = A$ and is an orthogonal matrix.

In problems 19–21, find an orthogonal projection onto the column space of the given matrix.

19. $A = \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix}$
20. $A = \begin{bmatrix} 1 & -1 \\ 2 & -2 \end{bmatrix}$
21. $A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \\ 2 & 0 & 2 \end{bmatrix}$
22. Explain why Corollary 2.8 is true.

2.5 Least Squares Solutions

If A is an invertible square matrix, then the system of linear algebraic equations $A\mathbf{x} = \mathbf{b}$ has a unique solution $\mathbf{x} = A^{-1}\mathbf{b}$. This is a good theoretical result.

But what if A is not invertible or not square? In the real world, we might still need to find some kind of approximate solution, or, if there are infinitely many solutions, we might need to find the "best" solution. Now we will study these situations, which go back in history at least as far back as Gauss's "day job" of using data from geodesic surveys.

Suppose A is a real, $m \times n$ matrix and $\mathbf{b}$ is in $\mathbb{R}^m$. The first fundamental problem we want to solve is to find an $\mathbf{x}^\star$ that *minimizes* the squared error

$$E(\mathbf{x}) \triangleq ||A\mathbf{x} - \mathbf{b}||^2 = \sum_{i=1}^{m} \left((A\mathbf{x})_i - b_i\right)^2. \tag{2.22}$$

$E(\mathbf{x})$ is the square of the distance between $A\mathbf{x}$, an "arrow", and $\mathbf{b}$, our "target." A perfect "bullseye" would be when $E(\mathbf{x}) = 0$, that is, $A\mathbf{x} = \mathbf{b}$. Note that $A\mathbf{x})_i = (A\mathbf{x}) \bullet \mathbf{e}^{(i)}$ is the ith component of the vector $A\mathbf{x}$.

We call such a minimizer $\mathbf{x}^\star$ a **least squares solution** (l.s.s.) of $A\mathbf{x} = \mathbf{b}$. The vector

$$\mathbf{r} \triangleq \mathbf{b} - A\mathbf{x}$$

is called the **residual**, so the goal of finding a least squares solution is the goal of minimizing the norm of the residual.

Geometrically, the goal of this problem is to find a vector $\mathbf{b}_A$ in the vector subspace $Col(A) \triangleq \{A\mathbf{x} : \ \mathbf{x} \text{ in } \mathbb{R}^n\}$ such that $\mathbf{b}_A$ comes closest to $\mathbf{b}$, as shown in Figure 2.7.

We recall that this is basically the same picture as the illustration of projection onto a plane in Figure 2.6. This picture suggests that if P_A is the orthogonal projection (recall Corollary 2.5) in Section 2.4 for the vector subspace $Col(A)$, then $\mathbf{b}_A$ will be the vector in $Col(A)$ that comes closest to $\mathbf{b}$. Let's explain why this is true: Given any vector subspace $\mathcal{W}$ contained in $\mathbb{R}^m$, we can define

$$\mathcal{W}^\perp = \{\mathbf{y} \text{ in } \mathbb{R}^m : \ \mathbf{y} \perp \mathbf{v} \text{ for all } \mathbf{v} \text{ in } \mathcal{W}\}, \tag{2.23}$$

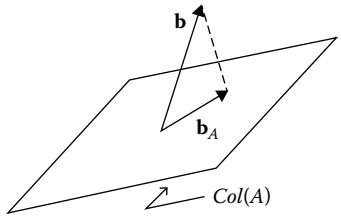


FIGURE 2.7
Least squares problems: Projection P_A.

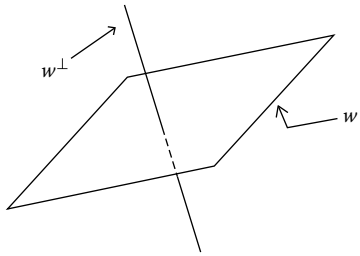


FIGURE 2.8
$\mathcal{W}^{\perp}$.

as illustrated in Figure 2.8, and define P to be the projection onto $\mathcal{W}$. Then every vector $\mathbf{b}$ in $\mathbb{R}^m$ can be written uniquely as a sum of a vector in $\mathcal{W}$ and a vector in $\mathcal{W}^{\perp}$, namely,

$$\mathbf{b} = P\mathbf{b} + (I - P)\mathbf{b},$$

according to Theorem 2.17(b) in Section 2.3.

If we define P_A to be the projection onto the vector subspace $Col(A)$, then indeed $P_A\mathbf{b}$ will be the vector in $Col(A)$ that is closest to $\mathbf{b}$. Why? By the Pythagorean Theorem, if $\mathbf{v}$ is *any* vector in $Col(A)$, then

$$\begin{aligned} ||\mathbf{b} - \mathbf{v}||^2 &= ||(P_A\mathbf{b} + (I - P_A)\mathbf{b}) - \mathbf{v}||^2 \\ &= ||(P_A\mathbf{b} - \mathbf{v}) + (I - P_A)\mathbf{b}||^2 = ||P_A\mathbf{b} - \mathbf{v}||^2 + ||(I - P_A)\mathbf{b}||^2, \end{aligned} \tag{2.24}$$

by the facts that (1) $Col(A)$ is a vector subspace, hence closed under vector addition and scalar multiplication and (2) both $P_A\mathbf{b}$ and $\mathbf{v}$ are in $Col(A)$. Clearly, (2.24) tells us that the closest $\mathbf{v}$ can get to $\mathbf{b}$ is to have $\mathbf{v} = P_A\mathbf{b} \triangleq \mathbf{b}_A$, in order to have $||P_A\mathbf{b} - \mathbf{v}|| = 0$.

By the way, the aforementioned says that $\mathbf{b}_A = P_A\mathbf{b}$ solves the **best approximation** problem for the vector subspace $Col(A)$.

That potentially leaves open at least one more theoretical issue: How do we find a solution(s) for $\mathbf{x}^\star$ so that $A\mathbf{x}^\star = \mathbf{b}_A$? We'll see that there is a very satisfying theoretical answer to this: "By solving the 'normal equations.' "

Because $\mathbf{b}_A$ is in $Col(A)$, we are certain to be able to find at least one least squares solution, that is, an $\mathbf{x}^\star$ with $A\mathbf{x}^\star = \mathbf{b}_A$. Later, in this section, we will discuss the second fundamental problem: If there are infinitely many least squares solutions, which should we choose?

2.5.1 The Normal Equations

Recall that we denote the columns of I_n by $\mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(n)}$, each of which is a vector in $\mathbb{R}^n$.

Because $\mathbf{b}_A = P_A\mathbf{b}$,

$$\mathbf{b} - \mathbf{b}_A = (I - P_A)\mathbf{b}$$

is automatically orthogonal to every vector in $Col(A)$, that is, $(\mathbf{b} - \mathbf{b}_A)$ is in $Col(A)^{\perp}$, by Theorem 2.17(b) in Section 2.3. In particular, for $i = 1, \ldots, n$, $A\mathbf{e}^{(i)}$ is in $Col(A)$, so

$$(\mathbf{b} - \mathbf{b}_A) \perp A\mathbf{e}^{(i)},\ i = 1, \ldots, n,$$

that is,

$$0 = \left(A\mathbf{e}^{(i)}\right)^T (\mathbf{b} - \mathbf{b}_A) = (\mathbf{e}^{(i)})^T \left(A^T(\mathbf{b} - \mathbf{b}_A)\right),\ i = 1, \ldots, n.$$

This being true for every $i = 1, \ldots, n$, we know that

$$A^T(\mathbf{b} - \mathbf{b}_A) = \mathbf{0},$$

that is,

$$A^T\mathbf{b} = A^T\mathbf{b}_A = A^T\left(A\mathbf{x}^\star\right).$$

So, we have derived the basic theoretical method for finding least squares solutions:

Theorem 2.22

Every least squares solution $\mathbf{x}^\star$ for $A\mathbf{x} = \mathbf{b}$ must satisfy **the normal equations**

$$A^TA\mathbf{x} = A^T\mathbf{b}. \tag{2.25}$$

The normal equations may have exactly one solution or infinitely many solutions.

Corollary 2.10

If A^TA is invertible, then there is exactly one least squares solution,

$$\mathbf{x}^\star = \left(A^TA\right)^{-1}A^T\mathbf{b}. \tag{2.26}$$

Example 2.19

Find a l.s.s. of $\begin{bmatrix} 4 & 0 \\ -1 & 1 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 3 \\ 5 \end{bmatrix}$.

Method: With $A = \begin{bmatrix} 4 & 0 \\ -1 & 1 \\ 0 & 2 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 0 \\ 3 \\ 5 \end{bmatrix}$, we have that

$$A^T A = \begin{bmatrix} 4 & -1 & 0 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ -1 & 1 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 17 & -1 \\ -1 & 5 \end{bmatrix}$$

is invertible. There is exactly one l.s.s.:

$$\mathbf{x}^\star = \left(A^T A\right)^{-1} A^T \mathbf{b} = \frac{1}{84} \begin{bmatrix} 5 & 1 \\ 1 & 17 \end{bmatrix} \begin{bmatrix} 4 & -1 & 0 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 3 \\ 5 \end{bmatrix} = \cdots = \frac{1}{42} \begin{bmatrix} -1 \\ 109 \end{bmatrix}. \bigcirc$$

Example 2.20

For the system $\begin{bmatrix} 1 & -1 \\ -2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$, (a) find (all of) the least squares solution(s), and (b) find the l.s.s. whose norm is minimized.

Method: With $A = \begin{bmatrix} 1 & -1 \\ -2 & 2 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$, we have that

$$A^T A = \begin{bmatrix} 5 & -5 \\ -5 & 5 \end{bmatrix}$$

is not invertible. The normal equations are

$$\begin{bmatrix} 5 & -5 \\ -5 & 5 \end{bmatrix} \mathbf{x} = A^T A \mathbf{x} = A^T \mathbf{b} = \begin{bmatrix} 6 \\ -6 \end{bmatrix}.$$

The Gauss–Jordan method gives

$$\left[\begin{array}{cc|c} 5 & -5 & 6 \\ -5 & 5 & -6 \end{array}\right] \sim \left[\begin{array}{cc|c} \textcircled{1} & -1 & 6/5 \\ 0 & 0 & 0 \end{array}\right]$$

after $R_1 + R_2 \to R_2$, $\frac{1}{5} R_1 \to R_1$. So, there are infinitely many l.s.s.

$$\mathbf{x}^\star = \begin{bmatrix} 1.2 \\ 0 \end{bmatrix} + c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1.2 + c_1 \\ c_1 \end{bmatrix}, \tag{2.27}$$

where c_1 is an arbitrary scalar.

(b) To find the minimum norm solution, we complete the square to calculate

$$\begin{aligned} ||\mathbf{x}^\star||^2 &= (1.2 + c_1)^2 + c_1^2 = 2c_1^2 + 2.4c_1 + 1.44 = 2(c_1^2 + 1.2c_1) + 1.44 \\ &= 2(c_1^2 + 1.2c_1 + 0.36 - 0.36) + 1.44 = 2(c_1 + 0.6)^2 + 0.72. \end{aligned}$$

So, $||\mathbf{x}^\star||$ is minimized by taking $c_1 = -0.6$. The minimum norm l.s.s. is

$$\mathbf{x}^{\star\star} \triangleq \begin{bmatrix} 0.6 \\ -0.6 \end{bmatrix}. \bigcirc$$

Geometrically, in Example 2.20, we have that

$$Col(A) = \left\{ \begin{bmatrix} t \\ -2t \end{bmatrix} : \ -\infty < t < \infty \right\}$$

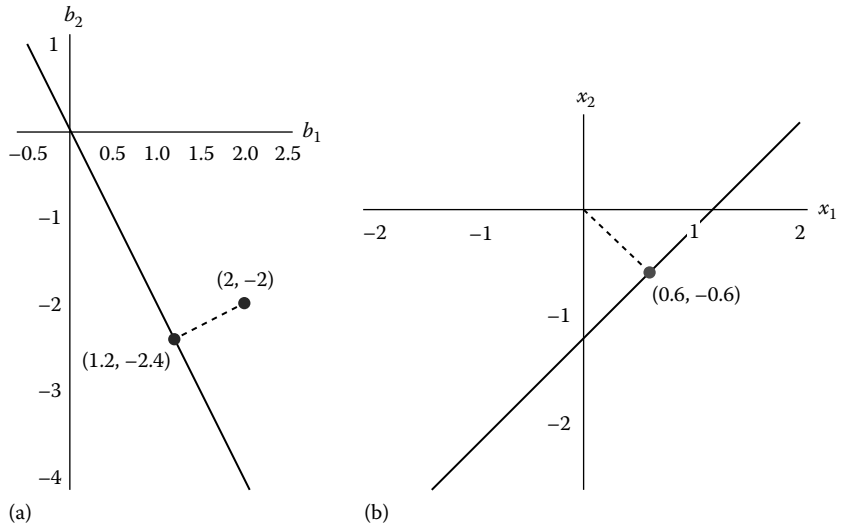


FIGURE 2.9
Example 2.20. (a) In **b** space and (b) in **x** space.

is a line, as depicted in Figure 2.9a, which illustrates what is happening in the space where $\mathbf{b} = [b_1 \quad b_2]^T$ and $A\mathbf{x}$ live. Figure 2.9b shows a picture of *all* of the least squares solutions $\mathbf{x}^\star$ and the l.s.s. of minimum norm, $\mathbf{x}^{\star\star}$, in particular.

Write the matrix $A = [\,\mathbf{a}_1 \;\vdots\; \ldots \;\vdots\; \mathbf{a}_n]$ in terms of its columns. Then the $n \times n$ matrix

$$A^T A = \begin{bmatrix} \mathbf{a}_1^T \\ - - \\ \vdots \\ - - \\ \mathbf{a}_n \end{bmatrix} [\,\mathbf{a}_1 \;\vdots\; \ldots \;\vdots\; \mathbf{a}_n\,] = \begin{bmatrix} \mathbf{a}_1 \bullet \mathbf{a}_1 & . & . & . & \mathbf{a}_1 \bullet \mathbf{a}_n \\ . & & . & & . \\ . & & & . & . \\ . & & . & & . \\ \mathbf{a}_n \bullet \mathbf{a}_1 & . & . & . & \mathbf{a}_n \bullet \mathbf{a}_n \end{bmatrix} = [\mathbf{a}_i \cdot \mathbf{a}_j]_{\substack{1 \le i \le m \\ 1 \le j \le n}}$$

is called the **Gramian** or **Gram matrix**.

Example 2.21

Find the **regression line** $y = f(x) \triangleq \mu x + \beta$ for the data given in Table 2.1.

Method: Denote by (x_i, y_i) the data points, for $i = 1, 2, 3$. Our linear fit for the data will be in the form $y = f(x) \triangleq \mu x + \beta$, where constants μ and β are to be chosen to minimize the total squared error

TABLE 2.1
Data for Example 2.21

x	1	2	3
y	0.5	1.0	2.0

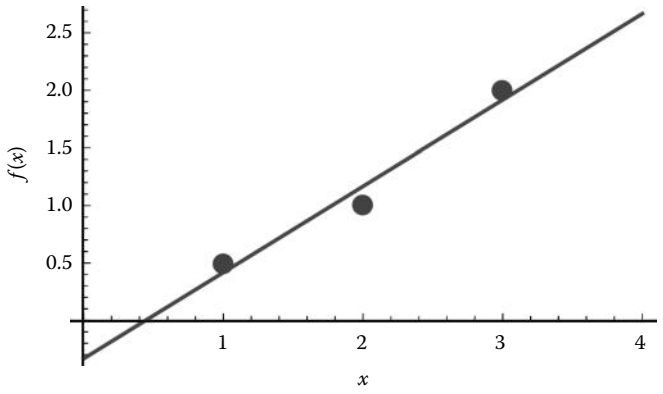


FIGURE 2.10
Regression line for Example 2.21.

$$E \triangleq \sum_{i=1}^{3} \big(y_i - f(x_i)\big)^2 = \sum_{i=1}^{3} (y_i - \mu x_i - \beta)^2.$$

To understand the concept of regression line, it helps to look at Figure 2.10. Ideally, we could find μ, β that satisfy

$$\mu x_i + \beta = y_i,\ i = 1, 2, 3,$$

which can be written as a system of three linear equations in the two unknowns μ, β:

$$\left\{\begin{array}{rcl} \mu \cdot 1 + \beta & = & 0.5 \\ \mu \cdot 2 + \beta & = & 1.0 \\ \mu \cdot 3 + \beta & = & 2.0 \end{array}\right\},$$

that is,

$$A \begin{bmatrix} \mu \\ \beta \end{bmatrix} \triangleq \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1.0 \\ 2.0 \end{bmatrix}.$$

With $A = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix}$, the normal equations,

$$\begin{bmatrix} 14 & 6 \\ 6 & 3 \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix} = A^T A \begin{bmatrix} \mu \\ \beta \end{bmatrix} = A^T A \mathbf{x} = A^T \mathbf{b} = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0.5 \\ 1.0 \\ 2.0 \end{bmatrix} = \begin{bmatrix} 8.5 \\ 3.5 \end{bmatrix},$$

have unique solution

$$\begin{bmatrix} \mu \\ \beta \end{bmatrix} = \begin{bmatrix} 14 & 6 \\ 6 & 3 \end{bmatrix}^{-1} \begin{bmatrix} 8.5 \\ 3.5 \end{bmatrix} = \frac{1}{6} \begin{bmatrix} 3 & -6 \\ -6 & 14 \end{bmatrix} \begin{bmatrix} 8.5 \\ 3.5 \end{bmatrix} = \begin{bmatrix} \frac{3}{4} \\ -\frac{1}{3} \end{bmatrix}.$$

The regression line is

$$y = \frac{3}{4} x - \frac{1}{3}. \bigcirc$$

Figure 2.10 shows the data and the regression line for Example 2.21.

In general, we want to find a regression line $y = \mu x + \beta$ for data points $(x, y) = (x_i, y_i)$, $i = 1, \ldots, m$, where at least two of the x_i's are distinct. Trying to solve $y_i = \mu x_i + \beta$, $i = 1, \ldots, m$ can be restated as $A \begin{bmatrix} \mu \\ \beta \end{bmatrix} = \mathbf{y}$. The normal equations are

$$\begin{bmatrix} \sum_{i=1}^{m} x_i^2 & \sum_{i=1}^{m} x_i \\ \sum_{i=1}^{m} x_i & m \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m} x_i y_i \\ \sum_{i=1}^{m} y_i \end{bmatrix}. \tag{2.28}$$

Some averages commonly used in statistics and mechanics are defined by

$$\overline{x} \triangleq \frac{1}{m} \sum_{i=1}^{m} x_i, \quad \overline{x^2} \triangleq \frac{1}{m} \sum_{i=1}^{m} x_i^2, \quad \overline{xy} \triangleq \frac{1}{m} \sum_{i=1}^{m} x_i y_i, \quad \text{and} \quad \overline{y} \triangleq \frac{1}{m} \sum_{i=1}^{m} y_i.$$

System (2.28) is

$$\begin{bmatrix} m\,\overline{x^2} & m\,\overline{x} \\ m\overline{x} & m \end{bmatrix} = \begin{bmatrix} \mu \\ \beta \end{bmatrix} = \begin{bmatrix} m\,\overline{xy} \\ m\,\overline{y} \end{bmatrix}.$$

Divide both sides by m to see that this is

$$\begin{bmatrix} \overline{x^2} & \overline{x} \\ \overline{x} & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix} = \begin{bmatrix} \overline{xy} \\ \overline{y} \end{bmatrix}. \tag{2.29}$$

The matrix $A^T A$ is invertible because its determinant, $\overline{x^2} - (\overline{x})^2$, is guaranteed to be positive. (In Problem 2.5.2.9 you will explain why.) The l.s.s. is

$$\begin{bmatrix} \mu \\ \beta \end{bmatrix} = \left(A^T A\right)^{-1} \begin{bmatrix} \overline{xy} \\ \overline{y} \end{bmatrix} = \frac{1}{\overline{x^2} - (\overline{x})^2} \begin{bmatrix} 1 & -\overline{x} \\ -\overline{x} & \overline{x^2} \end{bmatrix} \begin{bmatrix} \overline{xy} \\ \overline{y} \end{bmatrix}. \tag{2.30}$$

The regression line is given by

$$y = \frac{1}{\overline{x^2} - (\overline{x})^2} \left((\,\overline{xy} - \overline{x}\,\overline{y}\,)x + (-\overline{x}\,\overline{xy} + \overline{x^2}\,\overline{y}) \right). \tag{2.31}$$

2.5.1.1 Least Squares Solution and Orthogonal Matrices

Suppose the set of columns of A is already an o.n. set $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ in $\mathbb{R}^m$. By using Lemma 1.3 in Section 1.7, we can rewrite the system satisfied by least squares solutions, that is, $A\mathbf{x}^\star = \mathbf{b}_A$, as

$$x_1^\star \mathbf{a}_1 + \cdots + x_n^\star \mathbf{a}_n = \mathbf{b}_A. \tag{2.32}$$

Because $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ is an o.n. set, the projection onto $Col(A)$ is given by

$$P_A\mathbf{b} = \langle \mathbf{b}, \mathbf{a}_1 \rangle \mathbf{a}_1 + \cdots + \langle \mathbf{b}, \mathbf{a}_n \rangle \mathbf{a}_n.$$

Because $\mathbf{b}_A = P_A\mathbf{b}$, we can rewrite (2.32) as

$$x_1^\star \mathbf{a}_1 + \cdots + x_n^\star \mathbf{a}_n = (\mathbf{a}_1^T\mathbf{b})\mathbf{a}_1 + \cdots + (\mathbf{a}_n^T\mathbf{b})\mathbf{a}_n.$$

So, if the set of columns of A is an o.n. set, then there is exactly one l.s.s.:

$$\mathbf{x}^\star = \begin{bmatrix} x_1^\star \\ \vdots \\ x_n^\star \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1^T\mathbf{b} \\ \vdots \\ \mathbf{a}_n^T\mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_n^T \end{bmatrix} \mathbf{b} = A^T\mathbf{b}. \tag{2.33}$$

What if the set of columns of the given $m \times n$ matrix A is not o.n.? Can we use the Gram–Schmidt process to get a general result, as long as the columns of A are linearly independent? We will see the answers when we discuss the QR factorization in Section 2.7.

So, what if we're trying to solve a system $A\mathbf{x} = \mathbf{b}$ and it happens that A is a real, orthogonal matrix? Then $A^{-1} = A^T$, so the unique solution of $A\mathbf{x} = \mathbf{b}$ is given by $\mathbf{x} = A^T\mathbf{b}$, which equals the unique l.s.s. given by (2.33). Actually, this should not be a great surprise: Equation (2.33) holds when the set of columns of the $m \times n$ matrix A is an o.n. set, which is a generalization of the special case when A is square and is a real, orthogonal matrix.

Suppose the set of columns of A is an o.n. set but A is $m \times n$ with $m \neq n$. Because the columns of A are linearly independent, we must have $m > n$, by the (Goldilocks) Theorem 1.42 in Section 1.7. So, what are A^TA and AA^T in this case? Because the set of columns of A is an o.n. set, we are guaranteed to have $A^TA = I_n$. But we will not be able to say much about AA^T.

Example 2.22

(a) Let $A = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 1 & -1 \end{bmatrix}$. Then

$$AA^T = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 1 & -1 \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

which would seem to suggest that we could get something like "part of the identity matrix."

(b) Let

$$A = \frac{1}{\sqrt{6}} \begin{bmatrix} \sqrt{2} & -2 \\ \sqrt{2} & 1 \\ \sqrt{2} & 1 \end{bmatrix}.$$

Then

$$AA^T = \frac{1}{\sqrt{6}}\begin{bmatrix}\sqrt{2} & -2\\ \sqrt{2} & 1\\ \sqrt{2} & 1\end{bmatrix}\frac{1}{\sqrt{6}}\begin{bmatrix}\sqrt{2} & \sqrt{2} & \sqrt{2}\\ -2 & 1 & 1\end{bmatrix} = \begin{bmatrix}1 & 0 & 0\\ 0 & 0.5 & 0.5\\ 0 & 0.5 & 0.5\end{bmatrix}.$$

2.5.2 Problems

In problems 1–6, find all least squares solutions.

1. $\begin{Bmatrix} x_1 - x_2 = 0\\ x_1 + 2x_2 = 1\\ x_2 = 2\end{Bmatrix}$
2. $\begin{Bmatrix} x_1 + 2x_2 = 3\\ 3x_1 + 4x_2 = 5\\ 5x_1 - x_2 = 1\end{Bmatrix}$
3. $\begin{bmatrix}4 & -1\\ 4 & -3\\ 2 & -1\end{bmatrix}\mathbf{x} = \begin{bmatrix}2\\ -1\\ 1\end{bmatrix}$
4. $\begin{bmatrix}2 & -1\\ 2 & -2\\ -1 & 3\end{bmatrix}\mathbf{x} = \begin{bmatrix}2\\ -1\\ 1\end{bmatrix}$
5. $\begin{bmatrix}1 & 1\\ 1 & -2\\ 0 & 1\end{bmatrix}\mathbf{x} = \begin{bmatrix}3\\ 1\\ 4\end{bmatrix}$
6. $\begin{bmatrix}1 & 0 & -1\\ 1 & -1 & 0\\ 0 & 1 & -1\end{bmatrix}\mathbf{x} = \begin{bmatrix}1\\ 2\\ 0\end{bmatrix}$
7. If an $n \times n$ matrix Q satisfies $Q^TQ = I_n$, explain why $QQ^T = I_n$.
8. If the set of columns of an $m \times n$ matrix Q is an o.n. set, explain why QQ^T is an orthogonal projection.
9. Assuming that at least two of the x_i's are distinct, explain why $\overline{x^2} - (\bar{x})^2$ is guaranteed to be positive. [Hint: Apply the Cauchy–Schwarz inequality in $\mathbb{R}^m$ using $\mathbf{y} \triangleq \begin{bmatrix}\frac{1}{m} & \frac{1}{m} & \dots & \frac{1}{m}\end{bmatrix}^T$. As part of the explanation, explain why $\mathbf{x}, \mathbf{y}$ are linearly independent.]

In problems 10–12, find the regression line for the data given in the table.

10. Table 2.2
11. Table 2.3
12. Table 2.4
13. Table 2.5 has course grades on a 4.0 scale for students of Dynamics and Circuits. Using a least squares regression line for the data,

TABLE 2.2

Data for Problem 2.5.2.10

x	1	2	3
y	1.5	2.0	3.0

TABLE 2.3

Data for Problem 2.5.2.11

x	0.9	2.1	2.8
y	1.0	2.0	3.0

TABLE 2.4

Data for Problem 2.5.2.12

x	0.6	0.7	1.3	1.8
y	0.9	1.1	2.0	3.0

TABLE 2.5

Data for Problem 2.5.2.13

Student#	Dynamics	Circuits
1	3	2.3
2	3.7	2.7
3	2.3	3
4	2.3	2.7
5	1	0
6	2.7	3.3

a. Predict the grade in Circuits for students who receive a B^-, that is, 2.7, in Dynamics, and

b. Predict the grade in Dynamics for students who receive a B^-, that is, 2.7, in Circuits.

14. We wish to fit data (x_i, y_i), for $i = 1, \ldots, m$, to a function of the form $f(x) = \alpha x + \beta x^2$. Find the normal equations and find the best such function $f(x)$.

15. If $m > n$, that is, there are more equations than unknowns, can there be infinitely many l.s.s.'s?

16. Suppose A is invertible. Explain why there is exactly one least squares solution for $A\mathbf{x} = \mathbf{b}$ and it is given by $\mathbf{x}^\star = A^{-1}\mathbf{b}$. Explain also why this is the only solution of the normal equations.

17. We wish to fit a function of the form $f(x) = Ae^{\alpha x}$ to the data in Table 2.6. Define $z = \ln y$ and fit the data in $z = \ln f(x)$.

TABLE 2.6
Data for Problem 2.5.2.17

x	1.00	2.00	3.00	4.00
y	1.65	2.70	4.50	7.35

2.6 Symmetric Matrices, Definite Matrices, and Applications

We will see that if A is a real matrix and $A^T = A$, then many wonderful things are true, particularly concerning eigenvalues and eigenvectors. Not only that, but these results help greatly in using all matrices, as we will see in the next section.

Definition 2.17

A real, square matrix A is **symmetric** if $A = A^T$.

Note that a real matrix is symmetric if and only if $a_{ij} = a_{ji}$ for all i, j.

Example 2.23

$$A = \begin{bmatrix} 2 & 2 & 4 \\ 2 & -1 & 2 \\ 4 & 2 & 2 \end{bmatrix} \text{ is symmetric.}$$

Definition 2.18

Two square matrices A and B are **orthogonally similar** if there is an orthogonal matrix Q such that $B = Q^T A Q$.

2.6.1 A Spectral Theorem

Another name for eigenvalues is **spectra**.

Theorem 2.23

(Spectral theorem) If A is a real, symmetric, $n \times n$ matrix, then A has real eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ and there is a choice of corresponding eigenvectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ satisfying (a) $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ is an o.n. set, and (b) the matrix $Q = \left[\mathbf{x}^{(1)} \mid \ldots \mid \mathbf{x}^{(n)} \right]$ is orthogonal and diagonalizes A.

Why? Here we will explain all of these conclusions except that we will omit the explanation why there are no deficient eigenvalues, that is, why each eigenvalue's geometric

multiplicity equals its algebraic multiplicity. (The latter property will be discussed in Section 13.6.) It follows that we can find a linearly independent set of exactly n eigenvectors.

We have a choice concerning the order in which the eigenvalues are listed, so there are many choices for the diagonal matrix $D = Q^T AQ$. Nevertheless, the diagonal entries of D are the eigenvalues of A, including multiplicities. And, even if we specify D, we may have many choices for the matrix Q, as we will see in the explanation of Theorem 2.23 which follows a lemma.

First, Lemma 2.2 will establish orthogonality for eigenvectors corresponding to distinct eigenvalues. After that we will explain why we can choose a real, orthogonal set of eigenvectors corresponding to a repeated eigenvalue.

Lemma 2.2

If A is a real, symmetric matrix, then (a) eigenvectors corresponding to distinct eigenvalues are *automatically* orthogonal and (b) all of its eigenvalues are real.

Why? First, notice that we didn't bother to state the size of A because we want to emphasize that the derivation doesn't need to be explicit about that size.

(a) If A has distinct eigenvalues λ, μ with corresponding eigenvectors $\mathbf{x}, \mathbf{y}$, then we calculate

$$\mathbf{y}^T(A\mathbf{x}) = \mathbf{y}^T(\lambda\mathbf{x}) = \lambda\mathbf{y}^T\mathbf{x}.$$

Also, using the assumption that $A = A^T$ and the fact that every 1×1 matrix is symmetric,

$$\mathbf{y}^T A\mathbf{x} = (\mathbf{y}^T A\mathbf{x})^T = \mathbf{x}^T A^T\mathbf{y} = \mathbf{x}^T(A\mathbf{y}) = \mathbf{x}^T(\mu\mathbf{y}) = \mu\mathbf{x}^T\mathbf{y} = \mu\mathbf{y}^T\mathbf{x}.$$

Putting these two results together, we get

$$\mu\mathbf{y}^T\mathbf{x} = \mathbf{y}^T(A\mathbf{x}) = \lambda\mathbf{y}^T\mathbf{x},$$

hence

$$(\mu - \lambda)\mathbf{y}^T\mathbf{x} = 0.$$

Because μ and λ are distinct, we can divide both sides by $(\mu - \lambda) \neq 0$ to get $\mathbf{y}^T\mathbf{x} = 0$, that is, that $\mathbf{x}, \mathbf{y}$ are orthogonal.

(b) If A has an eigenvalue λ with corresponding eigenvector $\mathbf{x}$, then we can explain why λ must be real: let $^-$ denote complex conjugation, as in Section 2.1. We saw in Section 2.1 that we may need complex eigenvectors for real matrices, so we should not (accidentally) assume that $\mathbf{x}$ is real. Define a scalar z by

$$z \triangleq \mathbf{x}^T\,\overline{(A\mathbf{x})} = \mathbf{x}^T\,\overline{A}\,\overline{\mathbf{x}}.$$

Using the given information that A is real and the fact that the complex conjugate applied twice cancels out, we calculate that

$$\overline{z} = \overline{\mathbf{x}^T A\,\overline{\mathbf{x}}} = \overline{\mathbf{x}^T}\,\overline{A}\,\overline{\overline{\mathbf{x}}} = \overline{\mathbf{x}^T}\,A\,\mathbf{x}.$$

Also, the latter is a 1×1 matrix, so it equals its transpose. It follows that

$$\overline{z} = \left(\overline{\mathbf{x}^T}\,A\,\mathbf{x}\right)^T = \mathbf{x}^T A^T\,\overline{\mathbf{x}^T}^T$$

Using the assumption that A is symmetric and the fact that the transpose applied twice cancels out, we have

$$\overline{z} = \mathbf{x}^T A\,\overline{\mathbf{x}} = z.$$

To summarize, $z = \overline{z}$. It follows that z is real. But $z = \mathbf{x}^T\,\overline{(A\mathbf{x})} = \mathbf{x}^T\overline{(\lambda\mathbf{x})} = \overline{\lambda}\,\mathbf{x}^T\,\overline{\mathbf{x}}$, and in Problem 2.6.3.11 you will explain why $\mathbf{x}^T\overline{\mathbf{x}}$ is real. It follows that λ must be real. □

Explanation for Theorem 2.23:

Lemma 2.2 tells us that the real, symmetric, $n \times n$ matrix A has only real eigenvalues. Because the eigenvectors $\mathbf{x}$ satisfy real, homogeneous systems of the form $(A - \lambda I)\mathbf{x} = \mathbf{0}$, we may choose real eigenvectors. (We could choose them to be complex, for example, if $[1 \;\; 3]^T$ is an eigenvector, then $[(2+i) \;\; 3(2+i)]^T$ is also an eigenvector, too, but let's keep things as simple and real as possible.)

Are eigenvectors corresponding to the same eigenvalue automatically orthogonal? No. But, we can choose them to be: If, for example, $\lambda_1 = \cdots = \lambda_\ell > \lambda_{\ell+1} \geq \cdots \geq \lambda_n$ and we can find a complete set of basic solutions $\mathbf{v}_1, \ldots, \mathbf{v}_\ell$ for the system $(A - \lambda_1 I)\mathbf{x} = \mathbf{0}$, then $\{\mathbf{v}_1, \ldots, \mathbf{v}_\ell\}$ would be a basis for the ℓ dimensional eigenspace $\mathcal{E}_{\lambda=\lambda_1}$. Being a basis, it's linearly independent, so the Gram–Schmidt process (Theorem 2.16) in Section 2.3 can be used to produce an o.n. set $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\ell)}\}$ with

$$Span(\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\ell)}\}) = Span(\{\mathbf{v}_1, \ldots, \mathbf{v}_\ell\}) = \mathcal{E}_{\lambda=\lambda_1}.$$

So, if we assume that no eigenvalues of a real, symmetric matrix are deficient, then we can put together o.n. sets of eigenvectors that span the eigenspaces that correspond to the distinct eigenvalues of A. By Lemma 2.2, eigenvectors corresponding to distinct eigenvalues are automatically orthogonal, so this explains why Theorem 2.23(a) is true, that is, there is an o.n. set of eigenvectors $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ of A. By Theorem 2.20 in Section 2.4, that set is linearly independent, hence is a basis for $\mathbb{R}^n$, by Theorem 1.43 in Section 1.7. By Theorem 2.10 in Section 2.2, the matrix $P = \begin{bmatrix}\mathbf{x}^{(1)} & \cdots & \mathbf{x}^{(n)}\end{bmatrix}$ is orthogonal and diagonalizes A. This completes the explanations for Theorem 2.23, except we omit the explanation why real, symmetric matrices do not have deficient eigenvalues. □

If there is an eigenvalue of multiplicity $\ell \geq 2$, we have many choices for the order in which we list $\mathbf{v}_1, \ldots, \mathbf{v}_\ell$. The vectors $\mathbf{q}_1, \ldots, \mathbf{q}_\ell$ produced by the Gram–Schmidt process may depend on that order. That is why there may be many choices for the orthogonal matrix Q even for the same diagonal matrix $D = Q^T A Q$.

Example 2.24

Find a real, orthogonal matrix Q that diagonalizes $A = \begin{bmatrix} 2 & 2 & 4 \\ 2 & -1 & 2 \\ 4 & 2 & 2 \end{bmatrix}$.

Method: From Example 2.5 in Section 2.1 (and work done in examples that led up to it), the three eigenvalues of this matrix are $7 > -2 = -2$ and the corresponding eigenspaces have bases

$$\mathcal{E}_{\lambda_1=7} = \text{Span}\left(\left\{\begin{bmatrix}2\\1\\2\end{bmatrix}\right\}\right),\ E_{\lambda_2=\lambda_3=-2} = \text{Span}\left(\left\{\begin{bmatrix}-0.5\\1\\0\end{bmatrix}, \begin{bmatrix}-1\\0\\1\end{bmatrix}\right\}\right).$$

The G.–S. process for the linearly independent set $\left\{\begin{bmatrix}2\\1\\2\end{bmatrix}\right\}$ is easy: let $\mathbf{x}^{(1)} = \frac{1}{3}\begin{bmatrix}2\\1\\2\end{bmatrix}$.

Here is the G.–S. process for the linearly independent set $\left\{\begin{bmatrix}-0.5\\1\\0\end{bmatrix}, \begin{bmatrix}-1\\0\\1\end{bmatrix}\right\}$:

$$\mathbf{x}^{(2)} = \frac{1}{\sqrt{1.25}}\begin{bmatrix}-0.5\\1\\0\end{bmatrix} = \frac{1}{\sqrt{5}}\begin{bmatrix}-1\\2\\0\end{bmatrix},$$

$$\mathbf{v}_3 = \begin{bmatrix}-1\\0\\1\end{bmatrix} - \left(\begin{bmatrix}-1\\0\\1\end{bmatrix} \bullet \frac{1}{\sqrt{5}}\begin{bmatrix}-1\\2\\0\end{bmatrix}\right)\frac{1}{\sqrt{5}}\begin{bmatrix}-1\\2\\0\end{bmatrix} = \cdots = \frac{1}{5}\begin{bmatrix}-4\\-2\\5\end{bmatrix}, \text{ and}$$

$$r_{33} = \frac{\sqrt{45}}{5},\ \mathbf{x}^{(3)} = \frac{5}{\sqrt{45}} \cdot \frac{1}{5}\begin{bmatrix}-4\\-2\\5\end{bmatrix}.$$

The matrix

$$Q = \left[\frac{1}{3}\begin{bmatrix}2\\1\\2\end{bmatrix} \quad \frac{1}{\sqrt{5}}\begin{bmatrix}-1\\2\\0\end{bmatrix} \quad \frac{1}{\sqrt{45}}\begin{bmatrix}-4\\-2\\5\end{bmatrix}\right],$$

that is,

$$Q = \frac{1}{3\sqrt{5}}\begin{bmatrix}2\sqrt{5} & -3 & -4\\ \sqrt{5} & 6 & -2\\ 2\sqrt{5} & 0 & 5\end{bmatrix},$$

is a real, orthogonal matrix that diagonalizes A. ◯

Because Q is an orthogonal matrix, $Q^{-1} = Q^T$, so the diagonalization of A is given by

$$Q^T A Q = D.$$

It is very good to avoid needing to find the inverse of a matrix, even if it means relying on theory. That's because, as we will see better in Chapter 8, finding an inverse is likely to introduce numerical error.

As usual, it is good to check that $Q^T A Q = D$ actually is true.

2.6.1.1 A Spectral Formula

Suppose A is a real, symmetric matrix and Q is a real, orthogonal matrix that diagonalizes A, specifically $A = QDQ^{-1} = QDQ^T$, where $D = diag(\lambda_1, \ldots, \lambda_n)$. We calculate that

$$A = QDQ^T = \left[\begin{array}{c|c|c|c} \mathbf{q}^{(1)} & \cdots & \mathbf{q}^{(n)} \end{array}\right] diag(\lambda_1, \ldots, \lambda_n) \begin{bmatrix} \mathbf{q}_1^T \\ \vdots \\ \mathbf{q}_n^T \end{bmatrix} = \left[\begin{array}{c|c|c|c} \mathbf{q}^{(1)} & \cdots & \mathbf{q}^{(n)} \end{array}\right] \begin{bmatrix} \lambda_1\mathbf{q}_1^T \\ \vdots \\ \lambda_n\mathbf{q}_n^T \end{bmatrix}$$

$$= \mathbf{q}_1\lambda_1\mathbf{q}_1^T + \cdots + \mathbf{q}_n\lambda_n\mathbf{q}_n^T$$

that is,

Theorem 2.24

(Spectral decomposition) If A is a real, symmetric, $n \times n$ invertible matrix whose eigenvalues, including multiplicities, are $\lambda_1, \lambda_2, \ldots, \lambda_n$ and whose corresponding set of real eigenvectors $\{\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(n)}\}$ is an o.n. basis for $\mathbb{R}^n$, then

$$A = \sum_{i=1}^{n} \lambda_i \mathbf{q}_i \mathbf{q}_i^T. \tag{2.34}$$

This says that a real, symmetric matrix can be decomposed into a "weighted sum" of projections $\mathbf{q}_i\mathbf{q}_i^T$. We saw this in action in Problem 2.3.4.18. We will see something similar when we discuss the Singular Value Decomposition in Section 2.7.

Example 2.25

For the matrix $A = \begin{bmatrix} 2 & 2 & 4 \\ 2 & -1 & 2 \\ 4 & 2 & 2 \end{bmatrix}$ of Example 2.5 in Section 2.1 state explicitly the spectral decomposition.

Method: Using the eigenvalues of Example 2.5 in Section 2.1, and set of corresponding eigenvectors that is an o.n. basis for $\mathbb{R}^3$, the spectral decomposition of A is

$$A = \lambda_1\mathbf{q}_1\mathbf{q}_1^T + \lambda_2\mathbf{q}_2\mathbf{q}_2^T + \lambda_3\mathbf{q}_3\mathbf{q}_3^T$$

$$= 7\begin{bmatrix} \frac{2}{3} \\ \frac{1}{3} \\ \frac{2}{3} \end{bmatrix}\begin{bmatrix} \frac{2}{3} \\ \frac{1}{3} \\ \frac{2}{3} \end{bmatrix}^T + (-2)\begin{bmatrix} \frac{-1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \\ 0 \end{bmatrix}\begin{bmatrix} \frac{-1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \\ 0 \end{bmatrix}^T + (-2)\begin{bmatrix} -\frac{4}{\sqrt{45}} \\ -\frac{2}{\sqrt{45}} \\ \frac{5}{\sqrt{45}} \end{bmatrix}\begin{bmatrix} -\frac{4}{\sqrt{45}} \\ -\frac{2}{\sqrt{45}} \\ \frac{5}{\sqrt{45}} \end{bmatrix}^T$$

$$= \frac{7}{9}\begin{bmatrix} 4 & 2 & 4 \\ 2 & 1 & 2 \\ 4 & 2 & 4 \end{bmatrix} - \frac{2}{5}\begin{bmatrix} 1 & -2 & 0 \\ -2 & 4 & 0 \\ 0 & 0 & 0 \end{bmatrix} - \frac{2}{45}\begin{bmatrix} 16 & 8 & -20 \\ 8 & 4 & -10 \\ -20 & -10 & 25 \end{bmatrix}. \bigcirc$$

2.6.1.2 Positive Definite and Positive Semi-Definite Matrices

Definition 2.19

A real, square matrix A is

(a) **Positive definite** if $\mathbf{x}^T A\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$
(b) **Positive semi-definite** if $\mathbf{x}^T A\mathbf{x} \geq 0$ for all $\mathbf{x}$
(c) **Negative definite** if $\mathbf{x}^T A\mathbf{x} < 0$ for all $\mathbf{x} \neq \mathbf{0}$, that is, if $-A$ is positive definite

Example 2.26

Suppose A is a positive definite matrix and P is invertible. Explain why $P^T AP$ is also positive definite.

Method: While the problem did not say so, implicitly we should assume that the sizes of A and P agree so that the matrix product $P^T AP$ exists. For all $\mathbf{x} \neq \mathbf{0}$,

$$\mathbf{x}^T \left(P^T AP\right) \mathbf{x} = \left(\mathbf{x}^T P^T\right) A\,(P\mathbf{x}) = (P\mathbf{x})^T A\,(P\mathbf{x}) > 0,$$

because $\mathbf{x} \neq \mathbf{0}$ and P invertible together imply $P\mathbf{x} \neq \mathbf{0}$. (Why?) ◯

Theorem 2.25

Suppose A is a real, symmetric matrix and $\lambda_1, \lambda_2, \ldots, \lambda_n$ and $\{\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(n)}\}$ are as in Theorem 2.24. Then,

(a) All of its eigenvalues are positive if, and only if, A is positive definite.
(b) All of its eigenvalues are nonnegative if, and only if, A is positive semi-definite.

Why? By Corollary 2.6(a) and (b) in Section 2.4, we have

$$\mathbf{x} = c_1\mathbf{q}_1 + \cdots + c_n\mathbf{q}_n \text{ and } ||\,\mathbf{x}\,||^2 = |c_1|^2 + \cdots + |c_n|^2.$$

Using the spectral decomposition of A, we have

$$A\mathbf{x} = \left(\sum_{i=1}^{n} \lambda_i \mathbf{q}_i \mathbf{q}_i^T\right)(c_1\mathbf{q}_1 + \cdots + c_n\mathbf{q}_n) = \sum_{i=1}^{n} \lambda_i c_i \mathbf{q}_i.$$

Corollary 2.6(c) in Section 2.4, a Parseval identity, implies

$$\mathbf{x}^T A\mathbf{x} = \langle \mathbf{x}, A\mathbf{x} \rangle = \left\langle \sum_{i=1}^{n} c_i\mathbf{q}_i, \sum_{j=1}^{n} \lambda_j c_j \mathbf{q}_j \right\rangle = \sum_{i=1}^{n} \lambda_i |c_i|^2. \tag{2.35}$$

The desired conclusions (a) and (b) follow from (2.35) and Definition 2.19. □

2.6.1.3 Application to $A^2, A^{-1}, \sqrt{A}$

Not only is (2.34) a nice formula, it will give similar, nice formulas for A^2 and A^{-1} and even to a concept of $\sqrt{A}$, if the latter two exist.

If A is a real, symmetric, $n \times n$, and invertible (the latter is equivalent to all of its eigenvalues being nonzero), we can calculate A^{-1}: if Q is a real, orthogonal matrix that diagonalizes A, then the set of columns of Q is a basis for $\mathbb{R}^n$. To solve $A\mathbf{x} = \mathbf{b}$, we use

$$\sum_{i=1}^{n} (\mathbf{q}_i^T \mathbf{b}) \mathbf{q}_i = \mathbf{b} = A\mathbf{x} = \left(\sum_{i=1}^{n} \lambda_i \mathbf{q}_i \mathbf{q}_i^T \right) \mathbf{x} = \sum_{i=1}^{n} \lambda_i (\mathbf{q}_i^T \mathbf{x}) \mathbf{q}_i.$$

The solution $\mathbf{x}$ is easily found to have

$$\mathbf{q}_i^T \mathbf{b} = \lambda_i \mathbf{q}_i^T \mathbf{x}, \; i = 1, \ldots, n,$$

so

$$A^{-1}\mathbf{b} = \mathbf{x} = \sum_{i=1}^{n} (\mathbf{q}_i^T \mathbf{x}) \mathbf{q}_i = \sum_{i=1}^{n} \lambda_i^{-1} (\mathbf{q}_i^T \mathbf{b}) \mathbf{q}_i = \left(\sum_{i=1}^{n} \lambda_i^{-1} \mathbf{q}_i \mathbf{q}_i^T \right) \mathbf{b}. \tag{2.36}$$

This gives us

Theorem 2.26

If A is a real, symmetric, $n \times n$ invertible matrix whose set of eigenvectors $\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(n)}$ is an o.n. basis for $\mathbb{R}^n$, then

$$A^{-1} = \sum_{i=1}^{n} \lambda_i^{-1} \mathbf{q}_i \mathbf{q}_i^T. \tag{2.37}$$

and, similarly,

$$A^2 = \sum_{i=1}^{n} \lambda_i^2 \mathbf{q}_i \mathbf{q}_i^T. \tag{2.38}$$

Theorem 2.27

If A is a real, symmetric, $n \times n$ matrix all of whose eigenvalues are nonnegative and whose set of eigenvectors $\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(n)}$ is an o.n. basis for $\mathbb{R}^n$, then

$$\sqrt{A} \triangleq \sum_{i=1}^{n} \sqrt{\lambda_i} \; \mathbf{q}_i \mathbf{q}_i^T \tag{2.39}$$

satisfies $\left(\sqrt{A}\right)^2 = A$.

It would be interesting for you to compare this result with that of Problem 2.2.3.19.

2.6.1.4 Application to Least Squares Solutions

If A is a real, symmetric, $n \times n$ invertible matrix, then (2.36) gives a nice formula for $\mathbf{x}$, the unique solution of $A\mathbf{x} = \mathbf{b}$. In these circumstances, this is also a formula for the unique least squares solution (l.s.s.), by Problem 2.5.2.16.

But what *if* A is a real, symmetric, $n \times n$ matrix that is *not* invertible? Say that $\lambda = 0$ is an eigenvalue of algebraic multiplicity $(n - r)$ for A and the other eigenvalues, $\lambda_1, \ldots, \lambda_r$, are nonzero; for convenience, we are *no longer* ordering the eigenvalues as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Using the assumption that $A = A^T$, the normal equations become

$$A^2\mathbf{x} = A^T A\mathbf{x} = A^T\mathbf{b} = A\mathbf{b}. \tag{2.40}$$

The spectral decomposition (2.34) and the assumption that $\lambda_{r+1} = \cdots = \lambda_n = 0$ give

$$A = \sum_{i=1}^{n} \lambda_i \mathbf{q}_i \mathbf{q}_i^T = \sum_{i=1}^{r} \lambda_i \mathbf{q}_i \mathbf{q}_i^T + \sum_{i=r+1}^{n} 0 \cdot \mathbf{q}_i \mathbf{q}_i^T = \sum_{i=1}^{r} \lambda_i \mathbf{q}_i \mathbf{q}_i^T$$

and (2.38) gives

$$A^2 = \sum_{i=1}^{r} \lambda_i^2 \mathbf{q}_i \mathbf{q}_i^T.$$

The normal equations, that is, (2.40), become

$$\left(\sum_{i=1}^{r} \lambda_i^2 \mathbf{q}_i \mathbf{q}_i^T \right) \mathbf{x} = \left(\sum_{i=1}^{r} \lambda_i \mathbf{q}_i \mathbf{q}_i^T \right) \mathbf{b}.$$

By the usual sort of manipulations, for example, $(\mathbf{q}_i \mathbf{q}_i^T)\mathbf{x} = (\mathbf{q}_i^T \mathbf{x})\mathbf{q}_i$, the solutions of the normal equations, that is, all least squares solutions, are given by

$$\mathbf{x}^\star = \left(\sum_{i=1}^{r} \lambda_i^{-1} \mathbf{q}_i \mathbf{q}_i^T \right) \mathbf{b} + \sum_{i=r+1}^{n} c_i \mathbf{q}_i,$$

where the c_i are arbitrary constants.

If we ask for the l.s.s. of minimum norm, Parseval identity (2.19)(b) in Section 2.4 implies that this is

$$\mathbf{x}^{\star\star} = \left(\sum_{i=1}^{r} \lambda_i^{-1} \mathbf{q}_i \mathbf{q}_i^T \right) \mathbf{b}.$$

Definition 2.20

Suppose A is a real, symmetric, $n \times n$ matrix having $\lambda = 0$ as an eigenvalue of algebraic multiplicity $(n - r)$ and all other eigenvalues of A being $\lambda_1, \ldots, \lambda_r$. Then we define

$$A^+ \triangleq \sum_{i=1}^{r} \lambda_i^{-1} \mathbf{q}_i \mathbf{q}_i^T, \tag{2.41}$$

called a "partial inverse" or the "Moore–Penrose generalized inverse."

In these special circumstances, the minimum norm l.s.s. of $A\mathbf{x} = \mathbf{b}$ is given by

$$\mathbf{x}^{\star\star} = A^+ \mathbf{b}.$$

2.6.2 Further Study of Positive Definite Matrices

Theorem 2.28

Suppose A is a real, symmetric matrix. Then scalars $\eta < \infty$ and $\xi > -\infty$ can be chosen so that $\eta \geq \mathbf{x}^T A \mathbf{x} \geq \xi$ for all unit vectors $\mathbf{x}$. The values of η, ξ may depend upon A.

We postpone the explanation to the appendix at the end of this section. □

Example 2.27

Explain why for any given any real matrix A we can choose a scalar γ so that $(A + \gamma I)$ is positive definite.

Method: Given a real, $n \times n$, matrix A, choose ξ as in conclusion (b) of Theorem 2.28 and any $\gamma > -\xi$. Then, for all unit vectors $\mathbf{x}$,

$$\mathbf{x}^T (A + \gamma I)\mathbf{x} = \mathbf{x}^T A \mathbf{x} + \gamma \mathbf{x}^T \mathbf{x} = \mathbf{x}^T A \mathbf{x} + \gamma ||\mathbf{x}||^2 \geq \xi + \gamma \cdot 1 > 0. \bigcirc$$

Theorem 2.29

Suppose W is a real, symmetric positive definite $n \times n$ matrix. Define

$$\langle \mathbf{x}, \mathbf{y} \rangle_W \triangleq \langle \mathbf{x}, W\mathbf{y} \rangle = \mathbf{x}^T W^T \mathbf{y}$$

and

$$||\mathbf{x}||_W = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_W}.$$

Then all of the conclusions of Theorems 2.12 and 2.13 in Section 2.3 are true for $\langle \mathbf{x}, \mathbf{y} \rangle_W$ and $||\mathbf{x}||_W$ replacing $\langle \mathbf{x}, \mathbf{y} \rangle$ and $||\mathbf{x}||$, respectively.

You will explain the results of Theorem 2.29 in Problem 2.6.3.17.

2.6.2.1 Vibrations and the Generalized Eigenvalue Problem

In Section 5.3, we will study vibrations of spring–mass systems modeled by systems of differential equations;

$$\ddot{\mathbf{x}} = A\mathbf{x}.$$

In fact, those systems can also be written in the form

$$M\ddot{\mathbf{x}} + K\mathbf{x} = \mathbf{0}, \tag{2.42}$$

where the $n \times n$ **mass matrix** (also known as an **inertia matrix**) M is real, symmetric, and positive definite and the **stiffness matrix** K is also real, symmetric, and positive definite. If we try solutions of linear constant coefficients homogeneous system of ODEs (LCCHS) (2.42) in the form $\mathbf{x} = e^{\lambda t}\mathbf{v}$ where $\mathbf{v}$ is a constant vector, then we would get the **generalized eigenvalue problem**

$$(K + \mu M)\mathbf{x} = \mathbf{0}, \tag{2.43}$$

where $\mu = \lambda^2$.

Theorem 2.30

For the generalized eigenvalue problem (2.43), all eigenvalues μ_j are negative and there is a set eigenvectors that is an o.n. basis for $\mathbb{R}^n$. Correspondingly, system (2.42) has frequencies of vibration $\omega \triangleq \sqrt{-\mu_j}$.

Why? We can rewrite (2.43) as

$$\mathbf{0} = (K + \mu\sqrt{M}^2)\,\mathbf{x} = \sqrt{M}\,(\sqrt{M}^{-1}K\sqrt{M}^{-1} + \mu I)\,\sqrt{M}\,\mathbf{x}.$$

Defining $\mathbf{y} = \sqrt{M}\mathbf{x}$, (2.43) is equivalent to

$$(\sqrt{M}^{-1}K\sqrt{M}^{-1} + \mu I)\,\mathbf{y} = \mathbf{0}.$$

Defining $C \triangleq \sqrt{M}^{-1}K\sqrt{M}^{-1}$ and $\gamma = -\mu$, (2.43) is equivalent to

$$C\mathbf{x} = \gamma\mathbf{x}. \tag{2.44}$$

In problem 2.6.3.20, you will explain why the facts that both M and K are real, symmetric, and positive definite imply that C is also real, symmetric, and positive definite. It follows from Theorems 2.23 and 2.25 that all eigenvalues $\gamma = -\mu$ are positive and $\mathbb{R}^n$ has an o.n. basis of eigenvectors. This implies the conclusions of this theorem. □.

2.6.2.2 Positive Definiteness and Determinants

Suppose $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ is a real, 2×2, symmetric, positive definite matrix. Then for *all* $\mathbf{x} = [x_1 \quad x_2]^T \neq \mathbf{0}$,

$$0 < \mathbf{x}^T A \mathbf{x} = [x_1 \quad x_2]^T \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \cdots = ax_1^2 + 2bx_1x_2 + cx_2^2.$$

In particular, for $\mathbf{x} = [1 \quad 0]^T$, we get $0 < ax_1^2$, so necessarily $a > 0$. Similarly, $\mathbf{x} = [0 \quad 1]^T$ yields $c > 0$. So, the diagonal entries of A must be positive.

Further, by completing a square, we get

$$\begin{aligned} 0 < \mathbf{x}^T A \mathbf{x} &= ax_1^2 + 2bx_1x_2 + cx_2^2 = a\left(x_1^2 + 2\left(\frac{bx_2}{a}\right)x_1\right) + cx_2^2 \\ &= a\left(x_1^2 + 2\left(\frac{bx_2}{a}\right)x_1 + \left(\frac{bx_2}{a}\right)^2 - \left(\frac{bx_2}{a}\right)^2\right) + cx_2^2 = a\left(x_1 + \frac{bx_2}{a}\right)^2 - \frac{b^2x_2^2}{a} + cx_2^2 \\ &= a\left(x_1 + \frac{bx_2}{a}\right)^2 + \left(ac - b^2\right)x_2^2. \end{aligned}$$

Positivity for all $\mathbf{x} = [x_1 \quad x_2]^T \neq \mathbf{0}$ yields $a > 0$, which we already knew, and $ac - b^2 > 0$, that is, $|A| > 0$.

Now, if $A = \left[a_{ij}\right]_{\substack{1 \le i \le n \\ 1 \le j \le n}}$ is $n \times n$ a real, symmetric, positive definite matrix, then for all $\mathbf{x} \neq \mathbf{0}$ of the special form

$$\mathbf{x} = [x_1 \quad x_2 \quad \ldots \quad x_n]^T = [x_1 \quad x_2 \quad 0 \quad \ldots \quad 0]^T,$$

we have, using $a_{ij} = a_{ji}$,

$$0 < \mathbf{x}^T A \mathbf{x} = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 = [x_1 \quad x_2] \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

By the result in the previous paragraph and again noting that $a_{ij} = a_{ji}$, we conclude that necessarily

$$a_{11} > 0, a_{22} > 0, \text{ and } \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0.$$

This motivates.

Definition 2.21

If A is $n \times n$, then its **principal minors** are defined by $A_1 \triangleq det([a_{11}])$, $A_2 \triangleq \begin{vmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{vmatrix}, \ldots,$ $A_j \triangleq \begin{vmatrix} a_{11} & . & . & . & a_{1j} \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ a_{j1} & . & . & . & a_{jj} \end{vmatrix}, \ldots.$

Theorem 2.31

Suppose A is a real, symmetric $n \times n$ matrix. If A is positive definite, then its principal minors $A_1, A_2, \ldots, A_j, \ldots, A_n$ all must be positive. In particular, $a_{11} > 0$ and $|A| > 0$. Conversely, if all of the principal minors are positive, then A is positive definite.

Note that A_j is the determinant of the matrix obtained from A by deleting all *except* the rows $1, 2, \ldots, j$ and the columns $1, 2, \ldots, j$. This corresponds to only considering $x_1, \ldots, x_j$ to be allowed to be nonzero in $\mathbf{x}^T A \mathbf{x}$. Now, if we change the order in which the variables $x_1, \ldots, x_n$ are listed, then we can define a corresponding minor

$$A_{i_1,\ldots,i_j}$$

to be the determinant of the matrix obtained from A by deleting all *except* the rows $i_1, \ldots, i_j$ and columns $i_1, \ldots, i_j$. Comparing this notation with the principal minors notation, we have

$$A_{1,\ldots,j} = A_j$$

Theorem 2.32

Suppose A is a real, symmetric, positive definite $n \times n$ matrix. Then for all $1 \leq j \leq n$ and choices of $i_1, \ldots, i_j$, we must have $A_{i_1,\ldots,i_j} > 0$; in particular, all diagonal entries of A must be positive.

Example 2.28

Suppose A is 3×3 and has the form $A = \begin{bmatrix} a & -1 & b \\ -1 & c & 2 \\ b & 2 & 4 \end{bmatrix}$. If A is required to be positive definite, reach as many conclusions as possible about the parameters a, b, c, d.

Method: First, the diagonal entries must be positive, so $a > 0$ and $c > 0$. Using 2×2 minors, all of the following must be positive:

$$A_{1,2} = \begin{vmatrix} a & -1 \\ -1 & c \end{vmatrix} = ac - 1, A_{1,3} = \begin{vmatrix} a & b \\ b & 4 \end{vmatrix} = 4a - b^2, A_{2,3} = \begin{vmatrix} c & 2 \\ 2 & 4 \end{vmatrix} = 4c - 4 = 4(c - 1).$$

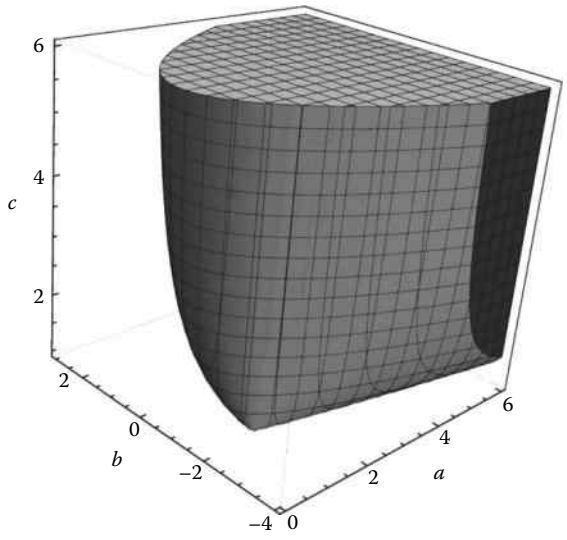


FIGURE 2.11
Example 2.28: Solid in parameter space.

Further,

$$|A| = A_{1,2,3} = 4ac - 4a - 4 - 4b - b^2c.$$

The conclusions we reach are $a > 0, c > 1, ac > 1, 4a - b^2 > 0$, and $(2a - b)c - 2(a + b) > 2$. ◯

Figure 2.11 shows part of the region in parameter space for which the matrix A of Example 2.28 is positive definite, specifically those (a, b, c) satisfying the inequalities and contained in the box $\{(a, b, c) : 0 < a < 6, -5 < b < 2, 1 < c < 5\}$. We used the *Mathematica* command

```
RegionPlot3D[ac > 1 && 4a − b^2 > 0 && 4ac − 4a − 4b − b^2c > 4, {a, 0, 6}, {b, −4, 3},
  {c, 1, 6}]
```

and then rotated the picture using the mouse.

Appendix

First, we need a basic calculation:

Theorem 2.33

For any $\mathbf{x} \neq \mathbf{0}$ and any scalar $\beta \neq 0$,

$$\frac{\mathbf{x}^T A \mathbf{x}}{\mathbf{x}^T \mathbf{x}} = \frac{(\beta \mathbf{x})^T A (\beta \mathbf{x})}{(\beta \mathbf{x})^T (\beta \mathbf{x})}. \tag{2.45}$$

Why?

$$\frac{(\beta\mathbf{x})^T A(\beta\mathbf{x})}{(\beta\mathbf{x})^T(\beta\mathbf{x})} = \frac{\cancel{\beta^2}(\mathbf{x}^T A\mathbf{x})}{\cancel{\beta^2}(\mathbf{x})^T\mathbf{x})} = \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T\mathbf{x}}. \square$$

Corollary 2.11

$$\left\{ \frac{\mathbf{x}^T A\mathbf{x}}{||\,\mathbf{x}\,||^2} :\ \mathbf{x} \neq \mathbf{0} \right\} = \{\mathbf{x}^T A\mathbf{x} :\ ||\,\mathbf{x}\,|| = 1\}.$$

Why? If $||\,\mathbf{x}\,|| = 1$, then

$$\mathbf{x}^T A\mathbf{x} = \frac{\mathbf{x}^T A\mathbf{x}}{1} = \frac{\mathbf{x}^T A\mathbf{x}}{||\,\mathbf{x}\,||^2}. \tag{2.46}$$

If α is *any* number in the set

$$S_1 \triangleq \{\mathbf{x}^T A\mathbf{x} :\ ||\,\mathbf{x}\,|| = 1\}$$

that is, $\alpha = \mathbf{x}^T A\mathbf{x}$, for some $\mathbf{x}$ satisfying $||\,\mathbf{x}\,|| = 1$, then

$$\alpha = \frac{\mathbf{x}^T A\mathbf{x}}{1} = \frac{\mathbf{x}^T A\mathbf{x}}{||\,\mathbf{x}\,||^2}$$

is also in the set

$$S_2 \triangleq \left\{ \frac{\mathbf{x}^T A\mathbf{x}}{||\,\mathbf{x}\,||^2} :\ \mathbf{x} \neq \mathbf{0} \right\}.$$

This being true for any number α in S_1 explains why $S_1 \subseteq S_2$.

Now, suppose α is *any* number in the set S_2, that is,

$$\alpha = \frac{\mathbf{x}^T A\mathbf{x}}{||\,\mathbf{x}\,||^2}$$

for some $\mathbf{x} \neq \mathbf{0}$. If we denote

$$\widehat{\mathbf{x}} = \frac{1}{||\,\mathbf{x}\,||}\,\mathbf{x},$$

then $||\widehat{\mathbf{x}}|| = 1$. Theorem 2.33, with $\beta = \dfrac{1}{||\,\mathbf{x}\,||}$, explains why

$$\alpha = \frac{\mathbf{x}^T A\mathbf{x}}{||\,\mathbf{x}\,||^2} = \frac{(\beta\mathbf{x})^T A(\beta\mathbf{x})}{||\beta\mathbf{x}||^2} = \frac{\widehat{\mathbf{x}}^T A\widehat{\mathbf{x}}}{||\widehat{\mathbf{x}}||^2} = \frac{\widehat{\mathbf{x}}^T A\widehat{\mathbf{x}}}{1} = \widehat{\mathbf{x}}^T A\widehat{\mathbf{x}}.$$

This explains why α is in the set S_1. This being true for any number α in S_2 explains why $S_2 \subseteq S_1$.

Because $S_1 \subseteq S_2$ and $S_2 \subseteq S_2$, the sets are equal, that is, $S_1 = S_2$. □.

The explanation for Theorem 2.28 is a little bit sophisticated: although it involves calculations, often it uses logical explanations. Also, using set inclusions adds to the sophistication of the explanations.

Here we explain the lower inequality in Theorem 2.28, that is, the existence of ξ as desired; the explanation of the upper inequality, that is, the existence of η, is similar. By Corollary 2.11,

$$\breve{m} \triangleq \min\left\{\frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T\mathbf{x}} : \mathbf{x} \text{ satisfying } \mathbf{x} \neq \mathbf{0}\right\} = \min\{\mathbf{x}^T A\mathbf{x} : \mathbf{x} \text{ satisfying } ||\,\mathbf{x}\,|| = 1\} \triangleq \widehat{m}.$$

The explanations for why these minimums are "achieved," for example, why there is a specific vector $\widehat{\mathbf{x}} \neq \mathbf{0}$ for which the minimum defined to be $\widehat{m} = \widehat{\mathbf{x}}^T A\widehat{\mathbf{x}}$, is more sophisticated than we choose to discuss here.*

Given a real, $n \times n$, matrix A, of the n^2 numbers a_{ij}, there is a positive number υ such that $|a_{ij}| \leq \upsilon$. For any $\mathbf{x}$ satisfying $||\,\mathbf{x}\,|| = 1$, we have

$$\mathbf{x}^T A\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j.$$

To get a lower bound on $\mathbf{x}^T A\mathbf{x}$, use the triangle inequality to see that

$$\left|\mathbf{x}^T A\mathbf{x}\right| = \left|\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j\right| \leq \sum_{i=1}^{n}\sum_{j=1}^{n}\left|a_{ij}x_i x_j\right| = \sum_{i=1}^{n}\sum_{j=1}^{n}\left|a_{ij}\right|\,|x_i|\,\left|x_j\right| \leq \sum_{i=1}^{n}\sum_{j=1}^{n}\upsilon\,|x_i|\,\left|x_j\right|.$$

Now, for any two real numbers x, y, we have $0 \leq \frac{1}{2}\,(|x| - |y|\,)^2 = \frac{1}{2}\,x^2 - |x|\,|y| + \frac{1}{2}\,y^2$, hence,

$$|x|\,|y| \leq \frac{1}{2}\left(x^2 + y^2\right).$$

Replacing x by x_i and y by x_j gives

$$\left|\mathbf{x}^T A\mathbf{x}\right| \leq \sum_{i=1}^{n}\sum_{j=1}^{n}\upsilon\cdot\frac{1}{2}\left(x_i^2 + x_j^2\right) = \frac{\upsilon}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_i^2 + x_j^2\right).$$

Noting that $\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n} x_i^2 = n(x_1^2 + \cdots + x_n^2) = \sum_{i=1}^{n}\sum_{j=1}^{n} x_j^2$, we have

$$\left|\mathbf{x}^T A\mathbf{x}\right| \leq \frac{\upsilon}{2}\cdot 2n(x_1^2 + \cdots + x_n^2) = n\upsilon||\,\mathbf{x}\,||^2 = n\upsilon\cdot 1 = n\upsilon.$$

* See Theorem 6.10.13 of Naylor and Sell (1982) mentioned in the "Read more about it" at the end of Section 2.10.

If $b > 0$ and a is real, then $|a| \leq b$ is equivalent to $b \geq a \geq -b$. So,

$$n\upsilon \geq \mathbf{x}^T A\mathbf{x} \geq -n\upsilon \triangleq \xi. \square$$

2.6.3 Problems

In problems 1–10, find a real, orthogonal matrix that diagonalizes the given matrix. Use exact values, that is, do not make decimal approximations.

1. $\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$
2. $\begin{bmatrix} 0 & -\sqrt{3} \\ -\sqrt{3} & 0 \end{bmatrix}$
3. $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}$
4. $\begin{bmatrix} 1 & 0 & -5 \\ 0 & -2 & 0 \\ -5 & 0 & 1 \end{bmatrix}$ [Hint: 6 is an eigenvalue.]
5. $\begin{bmatrix} -2 & 4 & 4 \\ 4 & 7 & -5 \\ 4 & -5 & 7 \end{bmatrix}$ [Hint: 12 is an eigenvalue.]
6. $\begin{bmatrix} 1 & -3 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix}$
7. $\begin{bmatrix} -2 & 0 & 0 \\ 0 & 3 & -1 \\ 0 & -1 & 3 \end{bmatrix}$
8. $\begin{bmatrix} 5 & 3 & 0 \\ 3 & 5 & 0 \\ 0 & 0 & 5 \end{bmatrix}$
9. $\begin{bmatrix} 7 & 4 & -4 \\ 4 & -8 & -1 \\ -4 & -1 & -8 \end{bmatrix}$
10. $\begin{bmatrix} 4 & 2 & -2 \\ 2 & 1 & -1 \\ -2 & -1 & 1 \end{bmatrix}$ [Hint: 0 and 6 are eigenvalues.]
11. Explain why $\mathbf{x}^T\overline{\mathbf{x}}$ is real for every vector $\mathbf{x}$ in $\mathbb{C}^n$.
12. Suppose A is a real, $m \times n$ matrix, A^TA is invertible, $m > n$, and we define $B = A(A^TA)^{-1}A^T$. Explain why B is symmetric.
13. Suppose A is a real, $m \times n$ matrix, A^TA is invertible, $m > n$, and we define $B = A(A^TA)^{-1}A^T$. Explain why $B^2 = B$.

14. Explain why (2.38) is correct.
15. Explain why (2.39) is correct.
16. Find all values of the scalar α for which $A \triangleq \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}$ is (a) positive definite and (b) positive semi-definite.
17. Suppose W is a real, symmetric, positive definite $n \times n$ matrix. Define

$$\langle \mathbf{x}, \mathbf{y} \rangle_W \triangleq \langle W\mathbf{x}, \mathbf{y} \rangle \triangleq \mathbf{x}^T W^T \mathbf{y}$$

and

$$|| \mathbf{x} ||_W \triangleq \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_W}.$$

Explain why all of the conclusions of Theorem 2.29 are true. [Hint: Since $\langle \mathbf{x}, \mathbf{y} \rangle_W$ is defined using $\langle \mathbf{x}, \mathbf{y} \rangle$, you can use the results of Theorem 2.12 in Section 2.3 giving conclusions about $\langle \mathbf{x}, \mathbf{y} \rangle$ to get conclusions about $\langle \mathbf{x}, \mathbf{y} \rangle_W$.]

18. (Small Project) Suppose W is a real, symmetric, positive definite $n \times n$ matrix and $\langle \mathbf{x}, \mathbf{y} \rangle_W$ and $|| \mathbf{x} ||_W$ are defined as in Problem 2.6.3.17. Suppose real, $m \times n$ matrix A and vector $\mathbf{b}$ in $\mathbb{R}^m$ are given. Consider the problem of finding $\mathbf{x}^\star$ that minimizes $||A\mathbf{x} - \mathbf{b}||_W^2$. We call this a **generalized**, or **weighted least squares** problem, with the letter "W" standing for "weighted." This problem allows us to study situations where some errors are more important than others. [for example, if $W = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$, then the total squared error is $E(\mathbf{x}) = 2|(A\mathbf{x})_1 - b_1|^2 + |(A\mathbf{x})_2 - b_2|^2$.]
Explain why $\mathbf{x}^\star$ should satisfy the **generalized normal equations**

$$A^T W^T A\mathbf{x} = A^T W^T \mathbf{b}.$$

[Hints: We can still think of $Col(A)$ as a vector subspace of $\mathbb{R}^m$. The orthogonal projection P_A should now be understood as being with respect to the inner product $\langle \mathbf{x}, \mathbf{y} \rangle_W$ and the concept of orthogonality similarly understood. For example, $(\mathbf{b} - \mathbf{b}_A) \perp A\mathbf{e}^{(i)}$ means

$$0 = \langle A\mathbf{e}^{(i)}, \mathbf{b} - \mathbf{b}_A \rangle_W = \left(A\mathbf{e}^{(i)}\right)^T W^T(\mathbf{b} - \mathbf{b}_A) = (\mathbf{e}^{(i)})^T A^T W^T(\mathbf{b} - \mathbf{b}_A).]$$

19. Suppose that instead of wanting to minimize the squared error, $||A\mathbf{x} - \mathbf{b}||^2$, that is, $\sum_{i=1}^m \left((A\mathbf{x})_i - b_i \right)^2$, we want to minimize the **relative squared error**,

$$\sum_{i=1}^m \left(\frac{(A\mathbf{x})_i - b_i}{b_i} \right)^2,$$

assuming all of the $b_i \neq 0$. Explain why this can be written as a weighted least squares problem.

20. Suppose both M and K are real, symmetric, and positive definite. Explain why $C = \sqrt{M}^{-1} K \sqrt{M}^{-1}$ is also real, symmetric, and positive definite.
21. Find two different examples of real, 2×2 matrices that are both symmetric and orthogonal.
22. Suppose Q is a real, symmetric, orthogonal $n \times n$ matrix. Use the result of Problem 2.4.4.12(b) to explain why only ± 1 can be eigenvalues of Q.

2.7 Factorizations: QR and SVD

In this section and the next we will explain why we can factor a matrix A in various ways that are useful foundations for practical, "industrial strength" methods for solving systems and least squares problems. These factorizations are

$$A = QR, \quad A = U\Sigma V^T, \quad EA = LU, \quad \text{and} \quad A = LL^T \tag{2.47}$$

2.7.1 QR Factorization

Theorem 2.34

If $m \geq n$, A is $m \times n$, and its set of columns is a linearly independent set of vectors in $\mathbf{R}^m$, then $A = QR$ where Q is an $m \times n$ matrix whose set of columns is an o.n. set and R is an invertible, upper triangular $n \times n$ matrix. In addition, if $m = n$, then Q is a real, orthogonal matrix.

Note that if $m < n$, then the columns of A cannot be linearly independent.

Why is Theorem 2.34 true? The Gram–Schmidt process used in explaining Theorem 2.16 in Section 2.3 can be summarized by (2.14) in Section 2.3, that is, for $j = 1, \ldots, n$,

$$\mathbf{a}_j = r_{jj}\mathbf{q}_j + (\mathbf{a}_j \bullet \mathbf{q}_1)\mathbf{q}_1 + \cdots + (\mathbf{a}_j \bullet \mathbf{q}_{j-1})\mathbf{q}_{j-1},$$

where $r_{jj} = ||\mathbf{v}_j||$ and $\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$ is an o.n. set. For example, for $j = 1, 2, 3$, (2.14) in Section 2.3 says

$$\mathbf{a}_1 = r_{11}\mathbf{q}_1,$$

$$\mathbf{a}_2 = r_{22}\mathbf{q}_2 + (\mathbf{a}_2 \bullet \mathbf{q}_1)\mathbf{q}_1,$$

and

$$\mathbf{a}_3 = r_{33}\mathbf{q}_3 + (\mathbf{a}_3 \bullet \mathbf{q}_1)\mathbf{q}_1 + (\mathbf{a}_3 \bullet \mathbf{q}_2)\mathbf{q}_2.$$

Define $r_{ij} = \mathbf{a}_j \bullet \mathbf{q}_i$ for $j > i$. From (2.14) in Section 2.3, we know that

$$A = [\mathbf{a}_1 \,\vdots\, \mathbf{a}_2 \,\vdots\, \cdots \,\vdots\, \mathbf{a}_n] = [r_{11}\mathbf{q}_1 \,\vdots\, r_{22}\mathbf{q}_2 + r_{12}\mathbf{q}_1 \,\vdots\, \cdots \,\vdots\, r_{nn}\mathbf{q}_n + r_{1n}\mathbf{q}_1 + \cdots + r_{n-1,n}\mathbf{q}_{n-1}]$$
$$= [r_{11}\mathbf{q}_1 \,\vdots\, r_{12}\mathbf{q}_1 + r_{22}\mathbf{q}_2 \,\vdots\, \cdots \,\vdots\, r_{1n}\mathbf{q}_1 + \cdots + r_{n-1,n}\mathbf{q}_{n-1} + r_{nn}\mathbf{q}_n].$$

By a calculation similar to Theorem 1.10 in Section 1.2 (see Problem 1.2.5.13),

$$A = [\mathbf{q}_1 \,\vdots\, \mathbf{q}_2 \,\vdots\, \cdots \,\vdots\, \mathbf{q}_n] \begin{bmatrix} r_{11} & r_{12} & . & . & . & r_{1n} \\ 0 & r_{22} & . & . & . & r_{2n} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ 0 & 0 & . & . & . & r_{nn} \end{bmatrix} \triangleq QR. \tag{2.48}$$

The properties of Q and R follow from the Gram–Schmidt process, that is, from the conclusions of Theorem 2.16 in Section 2.3. ◯

If $m > n$ then the matrix Q is not square and thus is not a real, orthogonal matrix, but it is still true that $Q^T Q = I_n$.

Example 2.29

(Example 2.18 in Section 2.3 again) Find the QR factorization of $A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & -2 & 1 \\ -1 & 0 & 5 \end{bmatrix}$.

Method: Denoting the columns of A by $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$, the calculations we did when solving Example 2.18 in Section 2.3 explain why

$$\mathbf{q}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix},\ r_{11} = \sqrt{2}\,.$$

After that, we had

$$\mathbf{q}_2 = \frac{1}{3\sqrt{2}} \begin{bmatrix} 1 \\ -4 \\ 1 \end{bmatrix},\ r_{22} = \frac{3}{\sqrt{2}},\ r_{12} = \mathbf{a}_2 \bullet \mathbf{q}_1 = \frac{1}{\sqrt{2}},$$

and then

$$\mathbf{q}_3 = \frac{1}{3} \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix},\ r_{33} = \frac{11}{3},\ r_{13} = \mathbf{a}_3 \bullet \mathbf{q}_1 = \frac{-5}{\sqrt{2}},\ r_{23} = \mathbf{a}_3 \bullet \mathbf{q}_2 = \frac{1}{3\sqrt{2}}\,.$$

It follows that

$$Q = \left[\mathbf{q}_1 \,\vdots\, \mathbf{q}_2 \,\vdots\, \mathbf{q}_3 \right] = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{3\sqrt{2}} & \frac{2}{3} \\ 0 & -\frac{4}{3\sqrt{2}} & \frac{1}{3} \\ -\frac{1}{\sqrt{2}} & \frac{1}{3\sqrt{2}} & \frac{2}{3} \end{bmatrix}$$

and

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{bmatrix} = \begin{bmatrix} \sqrt{2} & \frac{1}{\sqrt{2}} & -\frac{5}{\sqrt{2}} \\ 0 & \frac{3}{\sqrt{2}} & \frac{1}{3\sqrt{2}} \\ 0 & 0 & \frac{11}{3} \end{bmatrix}. \bigcirc$$

2.7.2 QR and Solving Systems

Suppose an $m \times n$ matrix $A = QR$ where Q, R are as in the conclusions of Theorem 2.34. Then we have a convenient method for solving the system $A\mathbf{x} = \mathbf{b}$, that is,

$$QR\mathbf{x} = \mathbf{b}. \tag{2.49}$$

Note that the set of columns of Q is an o.n. set of vectors in $\mathbb{R}^m$, so $Q^TQ = I_n$.

If (2.49) has a solution, then multiplication on the left by Q^T gives

$$R\mathbf{x} = RI\mathbf{x} = (Q^TQ)R\mathbf{x} = Q^T(QR\mathbf{x}) = Q^T\mathbf{b}. \tag{2.50}$$

Because R is upper triangular and invertible, we *can* solve the augmented matrix

$$\left[R \mid Q^T\mathbf{b} \right]$$

by back substitution: $x_n = r_{nn}^{-1}(Q^T\mathbf{b})_n$, $x_{n-1} = r_{n-1,n-1}^{-1}\left((Q^T\mathbf{b})_{n-1} - r_{n-1,n}x_n\right)$, etc. So, *if* $\mathbf{x}$ is a solution, then it satisfies (2.50) and can be computed using back substitution.

MATLAB can do a very good job of using back substitution to solve (2.50) using the command

$$\mathbf{x} = R\backslash Q^T\mathbf{b}.$$

A natural question is the converse: "If $\mathbf{x}$ satisfies (2.50), will it solve the original system, $A\mathbf{x} = \mathbf{b}$?" Let's substitute it in and see: Because R is invertible, $\mathbf{x} = R^{-1}Q^T\mathbf{b}$ follows from (2.50). Does

$$\mathbf{b} = A\mathbf{x} = QR\mathbf{x} = QRR^{-1}Q^T\mathbf{b} = QQ^T\mathbf{b}\,? \tag{2.51}$$

Unfortunately, if $m > n$ then the $m \times m$ matrix $QQ^T \neq I_m$, as we saw in Example 2.22 in Section 2.5 and will see in Problem 2.7.7.13. So, no, in answer to the question posed in (2.51), we have

Theorem 2.35

If A is $m \times n$ and its columns are a linearly independent set of vectors, let $A = QR$ where Q is an $m \times n$ matrix whose set of columns is an o.n. set and R is an invertible, upper triangular $n \times n$ matrix. Suppose $\mathbf{x}$ satisfies (2.50), that is, $R\mathbf{x} = Q^T\mathbf{b}$. Then $\mathbf{x}$ satisfies $A\mathbf{x} = \mathbf{b}$ if $m = n$.

2.7.3 QR and Least Squares Solutions

So, what if we used the QR factorization of A, as in the conclusions of Theorem 2.34, and $\mathbf{x}$ satisfies (2.50), that is, $R\mathbf{x} = Q^T\mathbf{b}$. If $m > n$, then what good does that do for us?

Theorem 2.36

If A is $m \times n$ and its columns are a linearly independent set of vectors, let $A = QR$ where Q is an $m \times n$ matrix whose set of columns is an o.n. set and R is an invertible, upper triangular $n \times n$ matrix. Suppose $\mathbf{x}^\star$ satisfies (2.50), that is, $R\mathbf{x} = Q^T\mathbf{b}$. Then $\mathbf{x}$ is a l.s.s. for $A\mathbf{x} = \mathbf{b}$.

Why? We will explain why $\mathbf{x}^\star$ satisfies the normal equations. We know that R is invertible and $Q^TQ = I_n$, so $\mathbf{x}^\star = R^{-1}Q^T\mathbf{b}$. We substitute this into the normal equations:

$$A^TA\mathbf{x}^\star = (QR)^T(QR)\mathbf{x}^\star = R^TQ^TQR\mathbf{x}^\star = \big(R^TI_nR\big)\mathbf{x}^\star = \big(R^TR\big)R^{-1}Q^T\mathbf{b} = R^TI_nQ^T\mathbf{b}$$
$$= (QR)^T\mathbf{b} = A^T\mathbf{b}.$$

So, yes, $\mathbf{x}^\star = R^{-1}Q^T\mathbf{b}$ is a l.s.s. of $A\mathbf{x} = \mathbf{b}$. □

The explanation of Theorem 2.36 is a nice example of a derivation because it is a string of equalities that review our knowledge of both least squares solutions and the QR factorization.

2.7.4 SVD

Definition 2.22

A matrix Σ is **pseudo-diagonal** if it can be written in one of the forms D, $[\,D \;\vdots\; O\,]$, or $\begin{bmatrix} D \\ \vdots \\ O \end{bmatrix}$ for some diagonal matrix D and some zero matrix O.

Equivalently, Σ is pseudo-diagonal if its entries σ_{ij} are zero for all $i \neq j$.

Theorem 2.37

If A is $m \times n$, then it has a **singular value decomposition** (SVD)

$$A = U\Sigma V^T,$$

where

U is a real, $m \times m$, orthogonal matrix
V is a real, $n \times n$ orthogonal matrix
Σ is an $m \times n$ pseudo-diagonal matrix.

Also, Σ is in the form D if $m = n$, $[\,D \;\vdots\; O\,]$ if $n > m$, or $\begin{bmatrix} D \\ \vdots \\ O \end{bmatrix}$ if $m > n$.

Note that we do *not* have to assume that the columns of A are linearly independent. Here is an outline of the steps used to construct the SVD:

1. Define the real, symmetric, positive semi-definite $n \times n$ matrix $B \triangleq A^T A$ and find its eigenvalues and an o.n. set of eigenvectors.
2. Use the r eigenvalues of B that are positive to construct an invertible, diagonal $r \times r$ matrix S, and use the corresponding eigenvectors of B to construct an $n \times r$ matrix V_1 whose columns are an o.n. set of vectors. The columns of the $n \times n$ orthogonal matrix $V = [\, V_1 \;\vdots\; V_2 \,]$ consist of all of the eigenvectors.
3. Construct the $m \times r$ matrix $U_1 \triangleq AV_1S^{-1}$, whose columns are an o.n. set of vectors.
4. If $r < m$, then the set of r columns of U_1 can be completed to give an o.n. basis for $\mathbb{R}^m$, for example, using the method of the appendix to Section 2.4. That basis will be the columns of the $m \times m$ orthogonal matrix $U = [\, U_1 \;\vdots\; U_2 \,]$.

Before explaining why an SVD exists by "construction," we need some further background that is itself useful. Also, we will apply this factorization to least squares problem.

Recall from Section 2.6 that if a real matrix is symmetric (hence square), then it has only real eigenvalues and has a basis that is an o.n. set of real eigenvectors. Given *any* real, $m \times n$ matrix A, define

$$B = A^T A.$$

Then B is real, $n \times n$, *and* symmetric. In addition, B is positive semi-definite because for all $\mathbf{x}$,

$$\mathbf{x}^T(B\mathbf{x}) = \mathbf{x}^T(A^T A\mathbf{x}) = (A\mathbf{x})^T(A\mathbf{x}) = ||\, A\mathbf{x}\,||^2 \geq 0.$$

By Theorem 2.25 in Section 2.6, all of B's eigenvalues $\mu_1, \ldots, \mu_n$ are automatically nonnegative. By Theorem 2.23 in Section 2.6, we may choose corresponding real eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ that form an o.n. basis for $\mathbb{R}^n$.

Even though A may not be symmetric and/or A may not be positive semi-definite, by constructing $B = A^T A$, we can take advantage of the results of Section 2.6 applied to B. Also, recall from Section 2.5 that $A^T A$ played a significant role in studying least squares solutions.

It follows that the eigenvalues of B can be denoted by

$$\sigma_1^2, \ldots, \sigma_r^2, 0, \ldots, 0$$

assuming that $\mu_1 \geq \ldots \geq \mu_r > 0 = \mu_{r+1} = \cdots = \mu_n$.

Our motivation for separating the positive eigenvalues of B from the zero eigenvalues is our recollection that when studying nonhomogeneous systems of linear algebraic equations, we used the solutions of the corresponding homogeneous system, that is, solutions of $B\mathbf{x} = \mathbf{0}$, among which are all of the eigenvectors of B corresponding to eigenvalue $\mu = 0$. Also, as we will see later, the zero eigenvalues of B will play a role in solving least squares problems.

To discover *how* it is possible to factor A as $U\Sigma V^T$ for some real, orthogonal matrices U, V and pseudo-diagonal matrix Σ, let's assume that we could and see what that would tells us about Σ and V:

$$B = A^TA = (U\Sigma V^T)^T(U\Sigma V^T) = (V\Sigma^TU^T)(U\Sigma V^T) = V\Sigma(U^TU)\Sigma V^T = V\Sigma^2V^T. \quad (2.52)$$

But we know from Theorem 2.23 in Section 2.6 that a real, symmetric matrix such as B can be diagonalized by the real, orthogonal matrix

$$P = \left[\begin{array}{c|c|c|c} \mathbf{v}_1 & & \cdots & \mathbf{v}_n \end{array}\right],$$

because $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is an o.n. set of eigenvectors for B. So,

$$B = P\, diag(\sigma_1^2, \ldots, \sigma_r^2, 0, \ldots, 0)\, P^T. \quad (2.53)$$

Comparing (2.52) and (2.53), we have

$$V\big(\Sigma^2\big)V^T = B = P\, diag(\sigma_1^2, \ldots, \sigma_r^2, 0, \ldots, 0)\, P^T.$$

So, choose

$$V = P = \left[\begin{array}{c|c|c|c} \mathbf{v}_1 & & \cdots & \mathbf{v}_n \end{array}\right]$$

and pseudo-diagonal matrix

$$\Sigma = [\,\sigma_{ij}\,]_{\substack{1 \le i \le m \\ 1 \le j \le n}} \text{with } \sigma_{ii} \triangleq \sigma_i \text{ for } i = 1, \ldots, r, \text{ and all other entries of } \Sigma \text{ being zero.}$$

After that, it's a little more work to discover what U should be. To do that, it helps to use block matrices introduced in Sections 1.5 and 1.6. For example, we can partition Σ by writing it as

$$\Sigma = \left[\begin{array}{cccccc|cccc} \sigma_1 & 0 & . & . & . & 0 & 0 & . & . & . & 0 \\ 0 & \sigma_2 & . & . & . & 0 & 0 & . & . & . & 0 \\ . & . & . & & & . & . & . & & & . \\ . & . & & . & & . & . & & . & & . \\ . & . & & & . & . & . & & & . & . \\ 0 & 0 & . & . & . & \sigma_r & 0 & . & . & . & 0 \\ \hline 0 & 0 & . & . & . & 0 & 0 & . & . & . & 0 \\ . & 0 & & . & & . & . & & . & & . \\ . & 0 & & & . & . & . & & & . & . \\ 0 & 0 & . & . & . & 0 & 0 & . & . & . & 0 \end{array}\right] = \left[\begin{array}{c|c} S_{r\times r} & O_{r\times(n-r)} \\ \hline O_{(m-r)\times r} & O_{(m-r)\times(n-r)} \end{array}\right]. \quad (2.54)$$

In (2.54), $S = diag(\sigma_1, \ldots, \sigma_r)$ is $r \times r$, the upper right zero matrix, O, is $r \times (n-r)$, the bottom left O is $(m-r) \times r$, and the bottom right O is $(m-r) \times (n-r)$. Note that S is invertible because it is a diagonal matrix and all of its diagonal entries are nonzero.

If $m = r$ then $\Sigma = [\, D \;\vdots\; O \,]$ or $\Sigma = D$. If $n = r$ then $\Sigma = \begin{bmatrix} D \\ O \end{bmatrix}$ or $\Sigma = D$.

We recall from Section 1.5 the definition of multiplication of block matrices:

$$\begin{bmatrix} C_{11} & | & C_{12} \\ -- & | & -- \\ C_{21} & | & C_{22} \end{bmatrix} \begin{bmatrix} E_{11} & | & E_{12} \\ -- & | & -- \\ E_{21} & | & E_{22} \end{bmatrix} \triangleq \begin{bmatrix} C_{11}E_{11} + C_{12}E_{21} & | & C_{11}E_{12} + C_{12}E_{22} \\ - - - & | & - - - \\ C_{21}E_{11} + C_{22}E_{21} & | & C_{21}E_{12} + C_{22}E_{22} \end{bmatrix},$$

where the C_{ik}, E_{kj} are themselves matrices.

Next, we want to find a real, orthogonal matrix U satisfying $A = U\Sigma V^T$. Write U and V as block matrices appropriate for multiplications involving block matrix Σ as in (2.54):

$$A = U\Sigma V^T = \begin{bmatrix} U_1 \;\vdots\; U_2 \end{bmatrix} \begin{bmatrix} S & | & O \\ -- & | & -- \\ O & | & O \end{bmatrix} \begin{bmatrix} V_1^T \\ -- \\ V_2^T \end{bmatrix} = \begin{bmatrix} U_1 \;\vdots\; U_2 \end{bmatrix} \begin{bmatrix} SV_1^T \\ -- \\ O \end{bmatrix} = U_1 S V_1^T. \tag{2.55}$$

The fact that

$$A = U_1 S V_1^T \tag{2.56}$$

is sometimes called the **thin SVD** or **reduced SVD** factorization.

Recall that $V = \begin{bmatrix} \mathbf{v}_1 \;\vdots\; \ldots \;\vdots\; \mathbf{v}_n \end{bmatrix}$, where $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is an o.n. set of eigenvectors for $A^T A$. So,

$$V_1 = \begin{bmatrix} \mathbf{v}_1 \;\vdots\; \ldots \;\vdots\; \mathbf{v}_r \end{bmatrix}, \quad V_2 = \begin{bmatrix} \mathbf{v}_{r+1} \;\vdots\; \ldots \;\vdots\; \mathbf{v}_n \end{bmatrix}.$$

As an aside, *if* $r = n$, then we don't need to write V in blocks, that is, $V = V_1$.

Because $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is an o.n. set, so is its subset $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$, thus the $n \times r$ matrix V_1 satisfies $V_1^T V_1 = I_r$. Also, to be of the correct sizes for the block multiplications to make sense, U_1 must be $m \times r$ and U_2 must be $m \times (m - r)$. As an aside, note that if $r = m$ then we don't need to write U in blocks, that is, $U = U_1$.

Because of (2.56) and the fact that $V_1^T V_1 = I_r$, we want

$$AV_1 = (U_1 S V_1^T)V_1 = U_1 S(V_1^T V_1) = U_1 S I = U_1 S. \tag{2.57}$$

It follows that we want

$$U_1 \triangleq AV_1 S^{-1} \tag{2.58}$$

because we began by assuming we could find U, Σ, V to satisfy $A = U\Sigma V^T$ and then we proceeded to discover what they should be. If we choose U, Σ, V such that (2.54) and (2.58) are true, then we have almost all of the desired conclusions about the SVD.

The remaining thing to discuss is why U is a real, orthogonal matrix. First, we can explain why the set of columns of the block matrix U_1 is an o.n. set in $\mathbb{R}^m$:

$$U_1^T U_1 = (AV_1 S^{-1})^T (AV_1 S^{-1}) = (S^{-1})^T V_1^T (A^T A) V_1 S^{-1} = (S^{-1})^T V_1^T B V_1 S^{-1}$$

But S is a diagonal matrix so $(S^{-1})^T = S^{-1}$ so

$$U_1^T U_1 = S^{-1} V_1^T B V_1 S^{-1},$$

But, by (2.52) and (2.54),

$$B = V\Sigma^2 V^T = \left[\begin{array}{c|c} V_1 & V_2 \end{array}\right] \left[\begin{array}{c|c} S & O \\ \hline O & O \end{array}\right]^2 \left[\begin{array}{c|c} V_1 & V_2 \end{array}\right] = \left[\begin{array}{c|c} V_1 & V_2 \end{array}\right] \left[\begin{array}{c} S^2 V_1^T \\ \hline O \end{array}\right] = V_1 S^2 V_1^T.$$

It follows that

$$\begin{aligned} U_1^T U_1 &= S^{-1} V_1^T B V_1 S^{-1} = S^{-1} V_1^T \left(V_1 S^2 V_1^T\right) V_1 S^{-1} = S^{-1} \left(V_1^T V_1\right) S^2 \left(V_1^T V_1\right) S^{-1} \\ &= S^{-1} I_r S^2 I_r S^{-1} = I_r, \end{aligned}$$

hence the set of columns of U_1 is an o.n. set in $\mathbb{R}^m$. In the appendix at the end of Section 2.4 we explained why having defined $U_1 \triangleq AV_1S^{-1}$ and having explained why the set of columns of U_1 is an o.n. set, it follows that we *can* find an $m \times (m-r)$ matrix U_2 so that $U \triangleq \left[\begin{array}{c|c} U_1 & U_2 \end{array}\right]$ is a real, orthogonal matrix.

This completes the explanation of why we can find the SVD of any real, $m \times n$ matrix A. □

On uniqueness of SVD, just by choosing a different order of listing the eigenvalues of B, we can get a different SVD. We can get more than one *significantly* different SVD if $B = A^T A$ has an eigenvalue with multiplicity greater than one. Also, if $r < m$, then there can be more than one choice for the matrix U_2 and thus for U.

Example 2.30

Find an SVD factorization of $A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$.

Method: Here $m = 3$ and $n = 2$, so U should be 3×3, Σ should be 3×2, and V should be 2×2.

$$B = A^T A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

has eigenvalues satisfying $0 = |B - \mu I| = (\mu - 2)^2 - 1$. So $\mu_1 = 3, \mu_2 = 1$, so $\sigma_1 = \sqrt{3}, \sigma_2 = 1$. Because B has two positive eigenvalues, $r = 2$ and we don't need to decompose V into blocks, that is, $V = V_1$.

Note that it is μ_js that are the eigenvalues of B, not the σ_js. The eigenvectors of B are found by row reduction on $\left[\begin{array}{c|c} B - \mu_j I & \mathbf{0} \end{array}\right]$, for example,

$$\left[\begin{array}{c|c} B - \mu_1 I & \mathbf{0} \end{array}\right] = \left[\begin{array}{cc|c} -1 & 1 & 0 \\ 1 & -1 & 0 \end{array}\right] \sim \left[\begin{array}{cc|c} \textcircled{1} & -1 & 0 \\ 0 & 0 & 0 \end{array}\right],$$

so we can use

$$\mathbf{v}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

as a unit eigenvector in the eigenspace of B corresponding to eigenvalue $\mu_1 = 3$. We will omit the similar details which find that

$$\mathbf{v}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

as a unit eigenvector in the eigenspace of B corresponding to eigenvalue $\mu_2 = 1$.

So, we may take the orthogonal matrix V to be

$$V = V_1 = [\,\mathbf{v}_1 \;\vdots\; \mathbf{v}_2\,] = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Because B has no zero eigenvalue, it is not necessary to write V in block form.

Also, because $m = 3$ and $n = 2$,

$$\Sigma = \left[\begin{array}{cc} \sigma_1 & 0 \\ 0 & \sigma_2 \\ \hline 0 & 0 \end{array}\right] = \left[\begin{array}{c} S \\ \hline O \end{array}\right], \text{ where } S = \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \end{bmatrix}.$$

Because $m - r = 3 - 2 > 0$, it is necessary to write U in block form. To find U_1, we use $A = U_1 S V_1^T = U_1 S V^T$ to imply

$$U_1 = AV_1S^{-1} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1/\sqrt{3} & 0 \\ 0 & 1 \end{bmatrix} = \cdots = \frac{1}{\sqrt{6}}\begin{bmatrix} 2 & 0 \\ 1 & \sqrt{3} \\ 1 & -\sqrt{3} \end{bmatrix}.$$

To find the third column of $U = \left[\begin{array}{c:c} U_1 & U_2 \end{array}\right] = \left[\begin{array}{cc:c} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{array}\right]$, that is, to find $U_2 = [\mathbf{u}_3]$, we need to find a unit vector $\mathbf{u}_3$ that is orthogonal to both of the columns of U_1. In the appendix of Section 2.4 we explained, using a process akin to the Gram–Schmidt process, how to "complete" an o.n. set to get an o.n. basis. Based on that appendix, using

$$\mathbf{e}^{(1)} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

calculate

$$\mathbf{w}_3 = \mathbf{e}^{(1)} - \langle \mathbf{e}^{(1)}, \mathbf{u}_1 \rangle \mathbf{u}_1 - \langle \mathbf{e}^{(1)}, \mathbf{u}_2 \rangle \mathbf{u}_2 = \cdots = \frac{1}{3}\begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix},$$

$||\mathbf{w}_3|| = 1/\sqrt{3}$, and finally

$$\mathbf{u}_3 = \frac{1}{\sqrt{3}}\begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}.$$

So, putting it all together, an SVD for A is given by

$$A = U\Sigma V^T = \left(\frac{1}{\sqrt{6}}\left[\begin{array}{cc:c} 2 & 0 & \sqrt{2} \\ 1 & \sqrt{3} & -\sqrt{2} \\ 1 & -\sqrt{3} & -\sqrt{2} \end{array}\right]\right)\left(\left[\begin{array}{cc} \sqrt{3} & 0 \\ 0 & 1 \\ \hline 0 & 0 \end{array}\right]\right)\left(\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\right). \bigcirc$$

Example 2.31

Find an SVD factorization of $A = \begin{bmatrix} 1 & \frac{1}{2} & -\frac{1}{2} \\ \sqrt{2} & 0 & 0 \\ -1 & \frac{1}{2} & -\frac{1}{2} \end{bmatrix}$.

Method: We will skim over routine details in order to highlight the steps of the method. In this example, as opposed to Example 2.30, it will turn out that the matrix V must be written in blocks.

$$B = A^T A = \cdots = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 0.5 & -0.5 \\ 0 & -0.5 & 0.5 \end{bmatrix}$$

has eigenvalues $\mu_1 = 4, \mu_2 = 1, \mu_3 = 0$, so $\sigma_1 = 2, \sigma_2 = 1, \sigma_3 = 0$. Because B has two positive eigenvalues, $r = 2$. The eigenvectors of B are found by row reduction on $\left[\, B - \mu_j I \,|\, \mathbf{0} \,\right]$, for example,

$$[\, B - \mu_1 I \,|\, \mathbf{0} \,] = \begin{bmatrix} 0 & 0 & 0 & |\, 0 \\ 0 & -3.5 & -0.5 & |\, 0 \\ 0 & -0.5 & -3.5 & |\, 0 \end{bmatrix} \sim \cdots \sim \begin{bmatrix} 0 & \textcircled{1} & 0 & |\, 0 \\ 0 & 0 & \textcircled{1} & |\, 0 \\ 0 & 0 & 0 & |\, 0 \end{bmatrix}$$

so we can use

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

as a unit eigenvector in the eigenspace of B corresponding to eigenvalue $\mu_1 = 4$. We will omit the similar details which find that

$$\mathbf{v}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$$

and

$$\mathbf{v}_3 = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

are unit eigenvectors in the eigenspaces of B corresponding to eigenvalues $\mu_2 = 1, \mu_3 = 0$.

So, we have the orthogonal matrix

$$V = [\, \mathbf{v}_1 \,\vdots\, \mathbf{v}_2 \,\vdots\, \mathbf{v}_3 \,] = \frac{1}{\sqrt{2}} \begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & 1 & 1 \\ 0 & -1 & 1 \end{bmatrix}$$

and the corresponding diagonal matrix is

$$\Sigma = diag(\sigma_1, \sigma_2, \sigma_3) = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \left[\begin{array}{cc|c} S & & O \\ \hline O & & 0 \end{array}\right].$$

Because $n - r = 3 - 2 > 0$, it *is* necessary to write V in block form. Because $m - r = 3 - 2 > 0$, it is necessary to write U in block form. Here,

$$V_1 = [\, \mathbf{v}_1 \,\vdots\, \mathbf{v}_2 \,] = \frac{1}{\sqrt{2}} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix},$$

$$V_2 = [\, \mathbf{v}_3 \,] = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \text{ and } S = diag(\sigma_1, \sigma_2) = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}.$$

To find U_1, we use $A = U_1 S V_1^T$ to find

$$U_1 = AV_1S^{-1} = \begin{bmatrix} 1 & \frac{1}{2} & -\frac{1}{2} \\ \sqrt{2} & 0 & 0 \\ -1 & \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix} = \cdots = \frac{1}{2} \begin{bmatrix} 1 & \sqrt{2} \\ \sqrt{2} & 0 \\ -1 & \sqrt{2} \end{bmatrix}.$$

To find the third column of U, that is, $U_2 = [\mathbf{u}_3]$, we need to find a unit vector $\mathbf{u}_3$ that is orthogonal to both of the columns of U_1. As in Example 2.30, we use the method of the appendix of Section 2.4: First, calculate

$$\mathbf{w}_3 = \mathbf{e}^{(1)} - \langle \mathbf{e}^{(1)}, \mathbf{u}_1 \rangle \mathbf{u}_1 - \langle \mathbf{e}^{(1)}, \mathbf{u}_2 \rangle \mathbf{u}_2 = \cdots = \frac{1}{4} \begin{bmatrix} 1 \\ -\sqrt{2} \\ -1 \end{bmatrix},$$

$||\mathbf{w}_3|| = 1/2$, and finally

$$\mathbf{u}_3 = \frac{1}{2} \begin{bmatrix} 1 \\ -\sqrt{2} \\ -1 \end{bmatrix}.$$

So, putting it all together, a SVD for A is given by

$$A = U\Sigma V^T = \left(\frac{1}{2} \begin{bmatrix} 1 & \sqrt{2} & 1 \\ \sqrt{2} & 0 & -\sqrt{2} \\ -1 & \sqrt{2} & -1 \end{bmatrix} \right) \left(\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \right) \left(\frac{1}{\sqrt{2}} \begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 1 & 1 \end{bmatrix} \right). \bigcirc$$

2.7.5 SVD and L.S.S.

By (2.56), that is, $A = U_1 S V_1^T$, the normal equations are

$$V_1 S^2 V_1^T \mathbf{x} = \cdots = A^T A \mathbf{x} = A^T \mathbf{b} = V_1 S U_1^T \mathbf{b}. \tag{2.59}$$

Example 2.32

Use (2.59) to find a l.s.s. of $A\mathbf{x} = \mathbf{b}$.

Method: Multiply (2.59) on the left by V_1^T and recall that $V_1^T V_1 = I_r$. After that, multiply on the left by S^{-1} to get

$$SV_1^T \mathbf{x} = U_1^T \mathbf{b}.$$

Taking advantage of the fact that $V_1^T V_1 = I$, we try to find a solution of this in the form $\mathbf{x} = V_1 \mathbf{y}$. When we substitute this in, after exchanging the right and left sides, we get

$$U_1^T \mathbf{b} = SV_1^T V_1 \mathbf{y} = S(V_1^T V_1)\mathbf{y} = SI_r \mathbf{y},$$

whose solution is $\mathbf{y} = S^{-1} U_1^T \mathbf{b}$, hence a l.s.s. is given by

$$\mathbf{x}^\star = V_1 S^{-1} U_1^T \mathbf{b}. \bigcirc$$

If we had any doubts before, this tells us that

Theorem 2.38

$A\mathbf{x} = \mathbf{b}$ always has at least one l.s.s.

Analogous to the spectral decomposition of (2.34) in Section 2.6, we have

Theorem 2.39

If $A = U\Sigma V^T$ is an SVD, then

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T, \tag{2.60}$$

where $\mathbf{u}_i, i = 1, \ldots, r$, are the columns of U_1.

Why? By the thin SVD (2.56), that is, $A = U_1 S V_1^T$, Theorem 1.10 in Section 1.2 implies

$$A = \begin{bmatrix} \mathbf{u}_1 & \vdots & \cdots & \vdots & \mathbf{u}_r \end{bmatrix} diag(\sigma_1, \ldots, \sigma_r) \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_r^T \end{bmatrix} = \begin{bmatrix} \sigma_1 \mathbf{u}_1 & \vdots & \cdots & \vdots & \sigma_r \mathbf{u}_r \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_r^T \end{bmatrix}$$

$$= \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^T. \square$$

2.7.6 Moore–Penrose Generalized Inverse

We saw a formula for A^+, where $\mathbf{x}^{\star\star} = A^+\mathbf{b}$ gives the minimum norm l.s.s. of $A\mathbf{x} = \mathbf{b}$, as long as A is real and symmetric. Let's get a formula for A^+ in general.

We saw, using the SVD of any matrix A, that

$$\mathbf{x}^\star = V_1 S^{-1} U_1^T \mathbf{b}$$

is a l.s.s. of $A\mathbf{x} = \mathbf{b}$. In Problem 2.7.7.21, you will explain why (2.60) explains why we can rewrite this as

$$\mathbf{x}^\star = \left(\sum_{i=1}^{r} \sigma_i^{-1} \mathbf{v}_i \mathbf{u}_i^T \right) \mathbf{b}. \tag{2.61}$$

We will see why (2.61) gives the l.s.s. of minimum norm: rewriting this as

$$\mathbf{x}^\star = \sum_{i=1}^{r} (\sigma_i^{-1} \mathbf{u}_i^T \mathbf{b}) \mathbf{v}_i$$

shows that $\mathbf{x}^\star$ is a linear combination of the columns of V_1. But $\{\mathbf{v}_1, \ldots, \mathbf{v}_r, \mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$ is an o.n. basis for $\mathbb{R}^n$, so *every* l.s.s. has the form

$$\mathbf{x} = \sum_{i=1}^{r} \alpha_i \mathbf{v}_i + \sum_{i=r+1}^{n} c_i \mathbf{v}_i \tag{2.62}$$

for some scalars $\alpha_i, i = 1, \ldots, r$ and scalars $c_i, i = r+1, \ldots, n$. We substitute (2.62) into the normal equations by first calculating

$$A^T A\mathbf{x} = (V_1 S^2 V_1^T)\left(\sum_{i=1}^{r} \alpha_i \mathbf{v}_i + \sum_{i=r+1}^{n} c_i \mathbf{v}_i\right) = \cdots = \left(\sum_{i=1}^{r} \sigma_i^2 \alpha_i \mathbf{v}_i + \sum_{i=r+1}^{n} 0 \cdot c_i \mathbf{v}_i\right) = \sum_{i=1}^{r} \sigma_i^2 \alpha_i \mathbf{v}_i \tag{2.63}$$

and

$$A^T \mathbf{b} = V_1^T S U_1^T \mathbf{b} = \sum_{i=1}^{r} (\sigma_i \mathbf{u}_i^T \mathbf{b}) \mathbf{v}_i. \tag{2.64}$$

Equating (2.63) and (2.64), we need to have

$$\alpha_i = \sigma_i^{-1} \mathbf{u}_i^T \mathbf{b},$$

so in order to be a l.s.s. we must have $\mathbf{x}$ to be of the form

$$\mathbf{x} = \sum_{i=1}^{r} (\sigma_i^{-1} \mathbf{u}_i^T \mathbf{b}) \mathbf{v}_i + \sum_{i=r+1}^{n} c_i \mathbf{v}_i = \mathbf{x}^\star + \sum_{i=r+1}^{n} c_i \mathbf{v}_i \tag{2.65}$$

for some scalars $c_i, i = r+1, \ldots, n$. By Parseval identity (2.19)(b) in Section 2.4 and the orthonormality of $\{\mathbf{v}_1, \ldots, \mathbf{v}_r, \mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$,

$$||\mathbf{x}||^2 = ||\mathbf{x}^\star||^2 + \sum_{i=r+1}^{n} |c_i|^2.$$

The l.s.s. of minimum norm has $c_{r+1} = \cdots = c_n = 0$ and thus is $\mathbf{x}^\star$, as we wanted to explain. □

Because of this, we have a formula for the **Moore–Penrose generalized inverse** for any matrix A:

$$A^+ = \sum_{i=1}^{r} \sigma_i^{-1} \mathbf{v}_i \mathbf{u}_i^T = V_1 S^{-1} U_1^T. \tag{2.66}$$

The Moore–Penrose generalized inverse satisfies many properties:

Theorem 2.40

If A is real, $X \triangleq A^+$ satisfies

$$\left\{\begin{array}{l}\text{(a)}\ \ AXA = A \\ \text{(b)}\ \ XAX = X \\ \text{(c)}\ \ (AX)^T = AX \\ \text{(d)}\ \ (XA)^T = XA\end{array}\right\}. \tag{2.67}$$

Example 2.33

For the system of Example 2.20 in Section 2.5, that is, $A\mathbf{x} \triangleq \begin{bmatrix} 1 & -1 \\ -2 & 2 \end{bmatrix}\mathbf{x} = \begin{bmatrix} 2 \\ -2 \end{bmatrix} = \mathbf{b}$,

a. Find the Moore–Penrose generalized inverse, A^+, and
b. Use it to find the l.s.s. of minimum norm.

Method: (a) We will skim over routine details in order to make clear the method. Begin to calculate the SVD of A by finding the eigenvalues σ_j^2 and corresponding eigenvectors $\mathbf{v}_j$ of $B \triangleq A^TA = \begin{bmatrix} 5 & -5 \\ -5 & 5 \end{bmatrix}$: $\sigma_1 = \sqrt{10} > 0 = \sigma_2$, so $r = 1$, with corresponding o.n. basis of eigenvectors

$$\{\mathbf{v}_1, \mathbf{v}_2\} = \left\{\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}, \mathbf{v}_2\right\}.$$

After that, we have $V = [V_1 \ \vdots \ V_2]$ with $V_1 = \mathbf{v}_1$, the 1×1 diagonal matrix $S = diag(\sigma_1) = [\sqrt{10}]$, and so $U = [U_1 \ \vdots \ U_2]$ with

$$U_1 \triangleq AV_1S^{-1} = \cdots = \frac{1}{\sqrt{5}}\begin{bmatrix} 1 \\ -2 \end{bmatrix} = [\mathbf{u}_1].$$

In order to construct the Moore–Penrose inverse, A^+, it is not necessary to know U_2 or V_2:

$$\begin{aligned} A^+ &= \sum_{i=1}^{1} \sigma_i^{-1}\mathbf{v}_i\mathbf{u}_i^T = (\sqrt{10})^{-1}\left(\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}\right)\left(\frac{1}{\sqrt{5}}\begin{bmatrix} 1 \\ -2 \end{bmatrix}^T\right) \\ &= \frac{1}{10}\begin{bmatrix} 1 \\ -1 \end{bmatrix}[1 \ \ -2] = \frac{1}{10}\begin{bmatrix} 1 & -2 \\ -1 & 2 \end{bmatrix}. \end{aligned}$$

(b) The l.s.s. solution of minimum norm is

$$\mathbf{x}^{\star\star} = A^+\mathbf{b} = \frac{1}{10}\begin{bmatrix} 1 & -1 \\ -2 & 2 \end{bmatrix}\begin{bmatrix} 2 \\ -2 \end{bmatrix} = \begin{bmatrix} 0.6 \\ -0.6 \end{bmatrix}. \ \bigcirc$$

It is a relief to find that this agrees with conclusion (b) of Example 2.20 in Section 2.5.

Example 2.34

For the system

$$A\mathbf{x} \triangleq \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \mathbf{x} = \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} = \mathbf{b},$$

a. Find all least squares solutions.
b. Find the l.s.s. of minimum norm.
c. Find the Moore–Penrose generalized inverse of A.

Method: (a) Writing, as in (2.56),

$$A = U\Sigma V^T = \left[\mathbf{u}_1 \mid \mathbf{u}_2 \mid \mathbf{u}_3\right] \left[\begin{array}{cc|c} 3 & 0 & 0 \\ 0 & 2 & 0 \\ \hline 0 & 0 & 0 \end{array}\right] \left[\begin{array}{c} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \hline \mathbf{v}_3^T \end{array}\right],$$

(2.65) implies that all least squares solutions are of the form

$$\mathbf{x} = (\sigma_1^{-1}\mathbf{u}_1^T\mathbf{b})\mathbf{v}_1 + (\sigma_2^{-1}\mathbf{u}_2^T\mathbf{b})\mathbf{v}_2 + c_3\mathbf{v}_3,$$

where c_3 is an arbitrary scalar. Note that $\sigma_1 = 3$, $\sigma_2 = 2$, and $\sigma_3 = 0$. Here we have that all least squares solutions are given by

$$\mathbf{x} = 3^{-1}\left(\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix} \bullet \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}\right)\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 2^{-1}\left(\begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix} \bullet \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}\right)\begin{bmatrix} 0 \\ \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} + c_3\begin{bmatrix} 0 \\ -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{3\sqrt{2}} \\ -\frac{3}{4} - c_3 \cdot \frac{1}{\sqrt{2}} \\ -\frac{3}{4} + c_3 \cdot \frac{1}{\sqrt{2}} \end{bmatrix},$$

where c_3 is an arbitrary scalar.
(b) Because $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is an o.n. set, Parseval identity (2.19)(b) in Section 2.4 implies that

$$||\mathbf{x}||^2 = ||(\sigma_1^{-1}\mathbf{u}_1^T\mathbf{b})\mathbf{v}_1 + (\sigma_2^{-1}\mathbf{u}_2^T\mathbf{b})\mathbf{v}_2 + c_3\mathbf{v}_3||^2 = |\sigma_1^{-1}(\mathbf{u}_1^T\mathbf{b})|^2 + |\sigma_2^{-1}(\mathbf{u}_2^T\mathbf{b})|^2 + |c_3|^2$$

is minimized by taking $c_3 = 0$. The l.s.s. of minimum norm is

$$\mathbf{x}^{\star\star} = \frac{1}{12}\begin{bmatrix} 2\sqrt{2} \\ -9 \\ -9 \end{bmatrix}.$$

(c) By (2.66)

$$A^+ = \sigma_1^{-1}\mathbf{v}_1\mathbf{u}_1^T + \sigma_2^{-1}\mathbf{v}_2\mathbf{u}_2^T = 3^{-1}\begin{bmatrix}1\\0\\0\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0\end{bmatrix} + 2^{-1}\begin{bmatrix}0\\\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\end{bmatrix}\begin{bmatrix}-\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0\end{bmatrix}$$

$$= \cdots = \begin{bmatrix}\frac{1}{3\sqrt{2}} & \frac{1}{3\sqrt{2}} & 0\\ -\frac{1}{4} & \frac{1}{4} & 0\\ -\frac{1}{4} & \frac{1}{4} & 0\end{bmatrix}. \bigcirc$$

Learn More About It

A good reference for much of the material in this chapter is *Matrix Analysis for Scientists and Engineers*, by Alan J. Laub, SIAM (Society for Industrial and Applied Mathematics), © 2005. Indeed, we use the notations of that book as much as possible.

2.7.7 Problems

In problems 1–4, find the QR factorization of the given matrix. Use exact values, that is, do not make decimal approximations. Check that $A = QR$.

1. $\begin{bmatrix}1 & 0\\1 & 1\\0 & 1\end{bmatrix}$
2. $\begin{bmatrix}1 & 0\\-1 & 1\\1 & 2\\0 & 1\end{bmatrix}$
3. $\begin{bmatrix}1 & 0\\1 & -2\\1 & 1\end{bmatrix}$
4. $\begin{bmatrix}1 & 0 & 0\\-1 & 1 & 0\\0 & -1 & 1\\0 & 0 & -1\end{bmatrix}$
5. Suppose A is a real, $m \times n$ matrix whose set of columns is an orthogonal set of nonzero vectors. Find the QR factorization of A.
6. Suppose A is a real, $m \times n$ matrix whose set of columns is an o.n. set. Find the QR factorization of A.
7. Suppose A is a real, invertible, upper triangular matrix. What is the QR factorization of A?
8. Suppose $A = QR$ is the QR factorization of a real, $m \times n$ matrix. From (2.51) we see that if $\mathbf{x}$ satisfies $R\mathbf{x} = Q^T\mathbf{b}$ it might still be true that $A\mathbf{x} \neq \mathbf{b}$. (Why? Because $QQ^T\mathbf{b}$ may not equal $\mathbf{b}$.) Find an example where $\mathbf{x}$ satisfies $R\mathbf{x} = Q^T\mathbf{b}$ and $A\mathbf{x} \neq \mathbf{b}$.

9. Suppose $A = QR$ is the QR factorization of a real, $m \times n$ matrix. We saw that if $\mathbf{x}$ satisfies $R\mathbf{x} = Q^T\mathbf{b}$ it might still be true that $A\mathbf{x} \neq \mathbf{b}$. Now, suppose additionally that $\mathbf{b} = Q\mathbf{c}$ for some $n \times 1$ vector $\mathbf{c}$. Show that this is enough to guarantee that $A\mathbf{x} = \mathbf{b}$ if $R\mathbf{x} = Q^T\mathbf{b}$. In your explanation(s), do not use specific values for $m, n, A, Q, R, \mathbf{b}$, or $\mathbf{c}$.

In problems 10 and 11, note the QR factorization and use that to find all least squares solutions.

10. $$\begin{bmatrix} 2/3 & 1/\sqrt{2} \\ 2/3 & -1/\sqrt{2} \\ 1/3 & 0 \end{bmatrix} \begin{bmatrix} 6 & -3 \\ 0 & \sqrt{2} \end{bmatrix} \mathbf{x} = \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}$$

11. $$\begin{bmatrix} 2/3 & 1/\sqrt{5} \\ 2/3 & 0 \\ -1/3 & 2/\sqrt{5} \end{bmatrix} \begin{bmatrix} 3 & -3 \\ 0 & \sqrt{5} \end{bmatrix} \mathbf{x} = \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}$$

12. For the system $A\mathbf{x} = \mathbf{b}$ given by

$$\begin{bmatrix} 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 \\ -1/\sqrt{2} & 0 & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \mathbf{x}$$
$$= \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix},$$

(a) Find all least squares solutions, (b) find the l.s.s. of minimum norm, and (c) find the Moore–Penrose generalized inverse of A.

13. Why did we say, "If $m > n$" before Theorem 2.36; in particular, why did we not mention the case $m < n$?
14. Suppose A is a real, $m \times n$ matrix and $A = QR$ is its QR factorization. Explain why $||A\mathbf{x}||^2 = ||R\mathbf{x}||^2$ for all $\mathbf{x}$ in $\mathbb{R}^n$.
15. Suppose $A = [\mathbf{a}_1 \quad \mathbf{a}_2]$ is a real, $m \times 2$ matrix. Suppose we don't know the entries of A but we do know that $\langle \mathbf{a}_1, \mathbf{a}_1 \rangle = 2$, $\langle \mathbf{a}_1, \mathbf{a}_2 \rangle = -1$, and $\langle \mathbf{a}_2, \mathbf{a}_2 \rangle = 2$. Without specifying the entries of A, find the QR factorization of A. Your Q should be written in terms of $\mathbf{a}_1, \mathbf{a}_2$.
16. Suppose the columns of a real, $m \times n$ matrix A are linearly independent. Use the QR factorization to explain why there is only one l.s.s. of $A\mathbf{x} = \mathbf{b}$, no matter what $\mathbf{b}$ is.

In problems 17–20, find an SVD of the given matrix. Use exact values, that is, do not make decimal approximations. Check that $A = U\Sigma V^T$.

17. $\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$
18. $\begin{bmatrix} 1 & 0 & -5 \\ 0 & -2 & 0 \\ -5 & 0 & 1 \end{bmatrix}$
19. $\begin{bmatrix} 2 & 1 \\ 4/\sqrt{5} & -8/\sqrt{5} \\ 2 & 1 \end{bmatrix}$
20. $\frac{1}{3\sqrt{2}} \begin{bmatrix} 3\sqrt{5} & \sqrt{5} & 2\sqrt{10} \\ 0 & -16 & 4\sqrt{2} \\ 3\sqrt{5} & \sqrt{5} & 2\sqrt{10} \end{bmatrix}$
21. Recall from the explanation of the SVD that $\mathbf{x}^\star = V_1 S^{-1} U_1^T \mathbf{b}$ is a l.s.s. of $A\mathbf{x} = \mathbf{b}$. Now, use (2.60) to explain why we can rewrite $\mathbf{x}^\star = V_1 S^{-1} U_1^T \mathbf{b}$ as (2.61).
22. Verify that the Moore–Penrose inverse A^+ we found in Example 2.33 is a solution for X that satisfies properties (2.67)(a), (b), (c), and (d).
23. Verify that the Moore–Penrose inverse A^+ we found in Example 2.34 is a solution for X that satisfies properties (2.67)(a), (b), and (d).
24. Suppose $A^T A$ is invertible. Then the unique l.s.s. of $A\mathbf{x} = \mathbf{b}$ is given by $\mathbf{x} = (A^T A)^{-1} A^T \mathbf{b}$. Because the minimum norm l.s.s. has to be the only l.s.s. in this situation, find a formula for the Moore–Penrose inverse A^+. Also, verify that this A^+ is a solution for X that satisfies properties (2.67)(a) and (c).
25. If A is a real, symmetric, positive definite matrix, find its SVD and relate it to the results of the spectral theories of Section 2.6.
26. Suppose A is a real, symmetric, $n \times n$ matrix having eigenvalues $\lambda_1, \ldots, \lambda_n$ and corresponding set of eigenvectors $\{\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(n)}\}$ is an o.n. basis for $\mathbb{R}^n$. Assume also that $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$. Find an SVD of A and relate it to the results of the spectral theories of Section 2.6.
27. (a) Write down an example of a 3×3 matrix A whose nine entries are -1, 0, 0, 0, 1, 1, 1, 1, 1, and (b) find the *SVD* of the matrix A you wrote in part (a).
28. Suppose a square matrix A has QR factorization QR. Explain why A is orthogonally similar to the matrix RQ, the *reverse multiplication*. This result will be used by the QR—algorithm mentioned in Section 8.5. [Hint: Multiply $A = QR$ on the left by Q^T and, after that, on the right by Q.]
29. (a) Write down an example of a 3×3 matrix A that is in RREF, which has exactly one entry of 2 and exactly one entry of -1, and (b) find the *SVD* factorization of the matrix A you wrote in part (a).

2.8 Factorizations: LU and Cholesky

2.8.1 LU Factorizations

One drawback to using the SVD factorization is that it assumes we can find the eigenvalues and eigenvectors of the symmetric matrix B. As we will see in Chapter 8, for a "large"

matrix, finding these things accurately can be a difficult problem from the practical point of view; even finding eigenvectors by solving homogeneous linear systems can have practical difficulties. Here we will briefly mention two other types of factorizations that are very useful. In addition to studying these methods further in Chapter 8, we will see how the LU factorization is a building block for many methods, including for solving finite difference approximations of partial differential equations in Chapter 12.

Recall that in Section 1.2, we mentioned that all elementary row operations can be implemented by multiplication on the left by elementary matrices. For example, to get a row echelon form of the matrix A below, the operations $R_1 \leftrightarrow R_2$, followed by $-\frac{1}{5}R_1+R_3 \rightarrow R_3$, followed by $R_2 \leftrightarrow R_3$ on

$$A = \begin{bmatrix} 0 & 0 & 1 & 4 \\ 10 & 0 & 15 & 10 \\ 2 & 1 & 2 & 2 \end{bmatrix}$$

can be implemented, respectively, by

$$E_1 \triangleq \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad E_2 \triangleq \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\frac{1}{5} & 0 & 1 \end{bmatrix}, \quad \text{and} \quad E_3 \triangleq \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

You can check that

$$E_3E_2E_1A = \begin{bmatrix} 10 & 0 & 15 & 10 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 4 \end{bmatrix} \triangleq U, \tag{2.68}$$

where U is an upper triangular matrix and a row echelon form for A.

The matrix E_2 is lower triangular and the matrices E_1 and E_3 are neither upper nor lower triangular. E_1 and E_3 are "permutation matrices," so called because they change the sequence of rows. For example, E_1 changes the list of rows R_1, R_2, R_3 to R_2, R_1, R_3.

To reduce a matrix to row echelon form, we may need to do row interchanges, that is, we may need to do "pivoting." In fact, as we will see in Section 8.4, it often improves the accuracy of the Gauss–Jordan algorithm to do "implicit partial pivoting," that is, row interchanges that move numbers of larger relative magnitude into pivot positions.

Theorem 2.41

Every matrix A has a general LU **factorization**

$$EA = LU \tag{2.69}$$

where

E is a product of permutation matrices
L is a lower triangular matrix
U is an upper triangular matrix

We can rewrite (2.69) as

$$L^{-1}EA = U \tag{2.70}$$

In (2.70), all of the permutation matrices operate on A *first*, *after which* a lower triangular matrix L^{-1} operates on EA to get U.

Unfortunately, (2.68) does not fit the form (2.70), because the lower triangular matrix E_2 is stuck in between the permutation matrices E_1 and E_3. But, always we can operate on A with all of the permutations matrices first.

In the above example,

$$E_3E_1A = E_3\begin{bmatrix}10 & 0 & 15 & 10\\ 0 & 0 & 1 & 4\\ 2 & 1 & 2 & 2\end{bmatrix} = \begin{bmatrix}10 & 0 & 15 & 10\\ 2 & 1 & 2 & 2\\ 0 & 0 & 1 & 4\end{bmatrix}$$

At this point, the operation E_2 does no good. Instead, we need to operate using

$$E_4 \triangleq \begin{bmatrix}1 & 0 & 0\\ -\frac{1}{5} & 1 & 0\\ 0 & 0 & 1\end{bmatrix}.$$

In fact,

$$E_4(E_3E_1A) = E_4\begin{bmatrix}10 & 0 & 15 & 10\\ 2 & 1 & 2 & 2\\ 0 & 0 & 1 & 4\end{bmatrix} = \begin{bmatrix}10 & 0 & 15 & 10\\ 0 & 1 & -1 & 0\\ 0 & 0 & 1 & 4\end{bmatrix} \triangleq U.$$

Denote $E = E_3E_1$. If we define $L^{-1} \triangleq E_4$, then

$$L = E_4^{-1} = \begin{bmatrix}1 & 0 & 0\\ \frac{1}{5} & 1 & 0\\ 0 & 0 & 1\end{bmatrix}.$$

Putting everything together, we have in this example

$$EA = LU = \begin{bmatrix}1 & 0 & 0\\ \frac{1}{5} & 1 & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}10 & 0 & 15 & 10\\ 0 & 1 & -1 & 0\\ 0 & 0 & 1 & 4\end{bmatrix},$$

where

E is a product of permutation matrices
L is lower triangular
U is upper triangular

If no row interchanges were used, then we would have

$$A = LU,$$

which is a special case of the LU factorization but which is not always possible to achieve.

L is not only lower triangular but it is also invertible because all of its diagonal entries are nonzero.

To solve a system $A\mathbf{x} = \mathbf{b}$, we note that all permutation matrices are invertible, so we can use (2.68) to rewrite the system as

$$E^{-1}LU\mathbf{x} = \mathbf{b},$$

or

$$LU\mathbf{x} = E\mathbf{b}, \tag{2.71}$$

If we define $\mathbf{y} \triangleq U\mathbf{x}$, then solving $A\mathbf{x} = \mathbf{b}$ is equivalent to solving

$$\mathbf{y} = U\mathbf{x} \tag{2.72}$$

for $\mathbf{x}$ in terms of $\mathbf{y}$ and then solving

$$L\mathbf{y} = E\mathbf{b} \tag{2.73}$$

2.8.2 Cholesky Factorizations

In the special case that A is real, symmetric, and positive semi-definite, we can find a lower triangular matrix L so that $U = L^T$ gives an LU factorization.

Theorem 2.42

(**Cholesky factorization**) If A is real, symmetric, and positive semi-definite, we can find a lower triangular matrix L such that $A = LL^T$.

Why? The next, simple example will suggest how to explain the general result. For the general derivation, please see the book by John Rice mentioned in the "Read more about it" at the end of this section.

A matrix that can be written in the form LL^T must be symmetric and positive semi-definite.

Example 2.35

Find a Cholesky factorization of $A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$.

Method: We want to find a lower triangular matrix $L = \begin{bmatrix} \ell_{11} & 0 \\ \ell_{21} & \ell_{22} \end{bmatrix}$ satisfying $A = LL^T$, that is,

$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} = LL^T = \cdots = \begin{bmatrix} \ell_{11}^2 & \ell_{11}\ell_{21} \\ \ell_{11}\ell_{21} & \ell_{21}^2 + \ell_{22}^2 \end{bmatrix}.$$

The $(1,1)$ entry of A requires $\ell_{11}^2 = 2$, so one Cholesky factorization can use $\ell_{11} = \sqrt{2}$. After that, both the $(1,2)$ and $(2,1)$ entries of A require that $-1 = \ell_{11}\ell_{21}$, hence $\ell_{21} = -\frac{1}{\sqrt{2}}$. Finally, the $(2,2)$ entry of A requires $2 = \ell_{21}^2 + \ell_{22}^2 = \frac{1}{2} + \ell_{22}^2$. One Cholesky factorization of A is given by

$$A = \begin{bmatrix} \sqrt{2} & 0 \\ -\frac{1}{\sqrt{2}} & \frac{\sqrt{3}}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ -\frac{1}{\sqrt{2}} & \frac{\sqrt{3}}{\sqrt{2}} \end{bmatrix}^T . \bigcirc$$

Another Cholesky factorization instead would use $\ell_{11} = -\sqrt{2}$ and $\ell_{21} = \frac{1}{\sqrt{2}}$.

Theorem 2.43

(Positive definiteness) Suppose A is a real, symmetric matrix. Then A is positive definite if, and only if, A has a Cholesky factorization $A = LL^T$ where L is lower triangular and its diagonal entries are all positive.

Learn More About It

A good reference for this chapter is *Computations and Mathematical Software*, by John R. Rice, McGraw-Hall, Inc., 1981. In particular, that book has "pseudo-code" for algorithms to implement the LU factorization, for example, by the Crout algorithm, and the Cholesky factorization. Another good reference for much of the material in this chapter is *Matrix Analysis for Scientists and Engineers*, by Alan J. Laub, mentioned at the end of Section 2.7.

2.8.3 Problems

In problems 1 and 2, find an LU factorization of the given tri-diagonal matrix. Use exact values, that is, do not make decimal approximations.

1. $A_3 = \begin{bmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{bmatrix}$
2. $A_4 = \begin{bmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \end{bmatrix}$
3. [Small project] For all $k \geq 3$, find an LU factorization of the $k \times k$ tri-diagonal matrix

$$A_k = \begin{bmatrix} -2 & 1 & 0 & . & . & . & & & 0 \\ 1 & -2 & 1 & & & & & & . \\ 0 & 1 & -2 & . & & & & & . \\ . & & & . & . & . & & & . \\ . & & & & . & . & . & & \\ . & & & & & . & 1 & -2 & 1 \\ 0 & . & . & . & & & 0 & 1 & -2 \end{bmatrix}.$$

Your conclusions for L and U should be $k \times k$ matrices, with unspecified k.

In problems 4–7, find a Cholesky factorization of the given matrix.

4. $A = \begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix}$

5. $A = \begin{bmatrix} 3 & -2 \\ -2 & 2 \end{bmatrix}$

6. $A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix}$

7. $A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$

[Hint: If A is 3×3, partition

$$L = \left[\begin{array}{ccc|c} & L_{11} & & O \\ - & - & & - - \\ \ell_{31} & \ell_{32} & & \ell_{33} \end{array}\right], \quad \text{where} \quad L_{11} \text{ is } 2 \times 2,$$

and correspondingly

$$A = \left[\begin{array}{c|c} A_{11} & A_{12} \\ - - & - - \\ A_{21} & A_{22} \end{array}\right].$$

When you calculate LL^T, you will see that you need $A_{11} = L_{11}L_{11}^T$.]

2.9 Rayleigh Quotient

The Rayleigh quotient gives a method for approximating eigenvalues. This method has been used in vibration problems for systems with a finite number of masses, also known as discrete systems. The method is particularly useful for estimating the lowest and highest eigenvalues of a real, symmetric matrix. Also, we will see in Section 9.7 that there is a powerful generalization of the Rayleigh quotient to boundary value problems, that is, continuous systems.

Suppose λ is an eigenvalue of a real, symmetric matrix A with corresponding eigenvector $\mathbf{x}$. Then,

$$\mathbf{x}^T(A\mathbf{x}) = \mathbf{x}^T(\lambda\mathbf{x}) = \lambda(\mathbf{x}^T\mathbf{x}) = \lambda||\,\mathbf{x}\,||^2.$$

This motivates

Definition 2.23

The **Rayleigh quotient** of a real matrix A is defined by

$$\mathcal{R}_A(\mathbf{x}) \triangleq \frac{\langle \mathbf{x}, A\mathbf{x} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle} = \frac{\mathbf{x}^T A \mathbf{x}}{||\,\mathbf{x}\,||^2}, \text{ for } \mathbf{x} \neq \mathbf{0}. \tag{2.74}$$

From the discussion preceding the definition, we have

Remark

If a real, symmetric matrix A has eigenvalue λ and corresponding eigenvector $\mathbf{x}$, then

$$\mathcal{R}_A(\mathbf{x}) = \lambda.$$

This explains why the Rayleigh quotient can produce eigenvalues. Turning theory into practice requires more work, some of which we will do in this section, some of which will be in Chapter 8, and some of which will be beyond the scope of this book.

Example 2.36

It is known that $\mathbf{x} \triangleq [-1 \quad 0 \quad 1]^T$ is an eigenvector of the matrix

$$A = \begin{bmatrix} 2 & 2 & 4 \\ 2 & -1 & 2 \\ 4 & 2 & 2 \end{bmatrix}.$$

Verify that the Rayleigh quotient produces an eigenvalue of A.

Method: We calculate

$$\mathcal{R}_A(\mathbf{x}) = \frac{1}{||\mathbf{x}||^2}\,\mathbf{x}^T A \mathbf{x} = \frac{1}{2}\,[-1 \quad 0 \quad 1] \begin{bmatrix} 2 & 2 & 4 \\ 2 & -1 & 2 \\ 4 & 2 & 2 \end{bmatrix} \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

$$= \frac{1}{2}\,[-1 \quad 0 \quad 1] \begin{bmatrix} 2 \\ 0 \\ -2 \end{bmatrix} = \frac{1}{2} \cdot (-4).$$

Direct calculations show that $A\mathbf{x} = -2\mathbf{x}$, that is, $\mathcal{R}_A(\mathbf{x}) = -2$ is an eigenvalue of A for which this $\mathbf{x}$ is a corresponding eigenvector. ○

If A is $n \times n$, then

$$\mathcal{R}_A(\mathbf{x}) \triangleq \frac{\sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j}{\sum_{i=1}^n x_i^2}.$$

Theorem 2.44

The Rayleigh quotient has the property that for any $\mathbf{x} \neq \mathbf{0}$ and any scalar $\beta \neq 0$, $\mathcal{R}_A(\beta\,\mathbf{x}) = \mathcal{R}_A(\mathbf{x})$.

Corollary 2.12

$\{\mathcal{R}_A(\mathbf{x}) : \mathbf{x} \neq \mathbf{0}\} = \{\mathcal{R}_A(\mathbf{x}) : ||\,\mathbf{x}\,|| = 1\}$.

These are the same as Theorem 2.33 in Section 2.6 and Corollary 2.11 in Section 2.6 written in the notation of a Rayleigh quotient.

2.9.1 A Rayleigh Theorem

If A is a real, symmetric matrix, then it has eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and a corresponding set of eigenvectors $\{\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(n)}\}$ that is an o.n. basis for $\mathbb{R}^n$. So, the real, orthogonal matrix $Q \triangleq [\mathbf{q}^{(1)} \quad \ldots \quad \mathbf{q}^{(n)}]$ diagonalizes A, that is, $A = QDQ^T$, where $D = diag(\lambda_1, \ldots, \lambda_n)$. For any vector $\mathbf{x} \neq \mathbf{0}$, we define $\widehat{\mathbf{x}} = \dfrac{1}{||\,\mathbf{x}\,||}\,\mathbf{x}$ and calculate

$$\mathcal{R}_A(\widehat{\mathbf{x}}) = \frac{1}{||\,\widehat{\mathbf{x}}\,||^2}\,\widehat{\mathbf{x}}^T A\widehat{\mathbf{x}} = \frac{1}{1}\,\widehat{\mathbf{x}}^T QDQ^T\widehat{\mathbf{x}} = (Q^T\widehat{\mathbf{x}})^T D(Q^T\widehat{\mathbf{x}}).$$

Defining

$$\mathbf{y} = Q^T\widehat{\mathbf{x}} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix},$$

we have

$$\mathcal{R}_A(\widehat{\mathbf{x}}) = (\,Q^T\widehat{\mathbf{x}})^T D(\,Q^T\widehat{\mathbf{x}}) = \mathbf{y}^T D\mathbf{y} = \cdots = \sum_{i=1}^{n} \lambda_i y_i^2 \tag{2.75}$$

Because $\mathbf{y}$ is a unit vector, $-1 \leq y_i \leq 1$ for $i = 1, \ldots, n$. It follows that the largest value that $\mathcal{R}_A(\widehat{\mathbf{x}})$ can be is the largest of the numbers $\lambda_1, \ldots, \lambda_n$ and the smallest that $\mathcal{R}_A(\widehat{\mathbf{x}})$ can be is the smallest of the numbers $\lambda_1, \ldots, \lambda_n$. Equation (2.75) establishes

Theorem 2.45

Suppose A is a real, symmetric $n \times n$ matrix. Then the values

$$\lambda_1 \triangleq \max\{\mathcal{R}_A(\mathbf{x}) : \mathbf{x} \text{ satisfying } \mathbf{x} \neq \mathbf{0}\} \quad \text{and} \quad \lambda_n \triangleq \min\{\mathcal{R}_A(\mathbf{x}) : \mathbf{x} \text{ satisfying } \mathbf{x} \neq \mathbf{0}\}$$

exist, are eigenvalues of A, and there are eigenvectors $\mathbf{x}^{(1)}$, $\mathbf{x}^{(n)}$, respectively, that "achieve" the values λ_1 and λ_n, that is,

$$\lambda_1 = \mathcal{R}_A(\mathbf{x}^{(1)}) \quad \text{and} \quad \lambda_n = \mathcal{R}_A(\mathbf{x}^{(n)}).$$

Example 2.37

Use the Rayleigh quotient to estimate the minimum and maximum eigenvalues of the matrix

$$A = \begin{bmatrix} -4 & 2 & 3 \\ 2 & -5 & 1 \\ 3 & 1 & -8 \end{bmatrix}.$$

Method: It follows from Theorem 2.37 that

$$\lambda_1 = \max\{\mathcal{R}_A(\mathbf{x}) : \mathbf{x} \text{ satisfying } \mathbf{x} \neq \mathbf{0}\}.$$

We use the *Mathematica* command

FindMaximum[{f[x,y,z], 1 ≥ x ≥ −1&&1 ≥ y ≥ −1&&1 ≥ z ≥ −1}, {x,y,z}],

where we replace the vector $\mathbf{x}$ by $[x \quad y \quad z]^T$, and calculate that

$$f(x,y,z) \triangleq \mathcal{R}_A(\mathbf{x}) = \frac{-4x^2 + 4xy - 5y^2 + 6xz + 2yz - 8z^2}{x^2 + y^2 + z^2}.$$

Mathematica gives output

$$\{-1.10996, \{x \to 0.258129, y \to 0.167868, z \to 0.136756\},$$

so the maximum eigenvalue of A is $\lambda_1 \approx -1.10996$. Similarly, the `FindMinimum` command gives us that the minimum eigenvalue of A is $\lambda_n \approx -9.60653$. ◯

In Section 5.3, we will study vibrations of spring–mass systems modeled by systems of differential equations $\ddot{\mathbf{x}} = A\mathbf{x}$. In that context, the matrix A will be negative definite. The frequencies of vibration will turn out to be $\omega = \sqrt{-\lambda}$, where λ is an eigenvalue of A. So, the minimum frequency of vibration will be equal to $\sqrt{-\lambda_1}$, where λ_1 is the maximum value of the Rayleigh quotient of A. For the matrix A of Example 2.37, the lowest frequency of vibration is $\omega_1 \approx 1.05354$ and highest frequency of vibration $\omega_3 \approx \sqrt{9.60653\ldots} \approx 3.09944$.

Example 2.38

Use the Rayleigh quotient to estimate the minimum and maximum eigenvalues of the matrix

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}.$$

Method: The Rayleigh quotient is

$$\mathcal{R}_A(\mathbf{x}) = \frac{2(x_1x_2 + x_1x_3 + x_1x_4 + x_2x_3 + x_2x_4 + x_3x_4)}{x_1^2 + x_2^2 + x_3^2 + x_4^2} \triangleq f(x_1, x_2, x_3, x_4).$$

The function $f(x_1, x_2, x_3, x_4)$ is *symmetrical* in the sense that interchanging the roles of x_i and x_j would have no effect on $\mathcal{R}_A(\mathbf{x})$. For example, $f(x_1, x_2, x_3, x_4) = f(x_2, x_4, x_3, x_1)$.

This suggests that a vector with maximum symmetry, such as $\mathbf{x}_+ = [1 \quad 1 \quad 1 \quad 1]^T$, and a vector with maximum "disorder," such as $\mathbf{x}_- = [1 \quad -1 \quad 1 \quad -1]^T$, might give good estimates for eigenvalues. We calculate $\mathcal{R}_A(\mathbf{x}_+) = \frac{12}{4} = 3$ and $\mathcal{R}_A(\mathbf{x}_-) = \frac{-4}{4} = -1$. This suggests guessing that the eigenvalues of A are $\lambda_1 = 3 \geq \lambda_2 \geq \lambda_3 \geq \lambda_4 = -1$. ◯

By the way, for the matrix A of Example 2.38, *Mathematica* calculates that the eigenvalues are $3, -1, -1, -1$, so the Rayleigh quotient and intuition did really well at guessing the eigenvalues.

Learn More About It

Below, Problems 2.9.2.7 and 2.9.2.8 are adapted from *Theory of Matrices*, by Peter Lancaster, Academic Press, 1969, specifically its Exercises 3.2.2 and 3.2.3. Lancaster's book is a very useful reference for the whole subject of matrices. In particular, the book discusses the "mini-max" technique and its relationship to the eigenvalues between the minimum and maximum eigenvalues of a real, symmetric matrix.

2.9.2 Problems

In problems 1 and 2, use the Rayleigh quotient to find the *exact* minimum and maximum eigenvalues of the 2×2 matrix. If possible, reduce the problem to a Calculus problem of finding the minimum and maximum values of a function of a single variable.

1. $A = \begin{bmatrix} 2 & 1 \\ 1 & -1 \end{bmatrix}$
2. $A = \begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix}$

In problems 3 and 4, use the Rayleigh quotient to find the approximations of the minimum and maximum eigenvalues of the 3×3 matrix. You may use *Mathematica* or MATLAB as in Example 2.37.

3. $A = \begin{bmatrix} 2 & \sqrt{3} & 0 \\ \sqrt{3} & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix}$
4. $A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 2 \end{bmatrix}$

In problems 5 and 6, use the Rayleigh quotient and intuition, as in Example 2.38, to find estimates of the minimum and maximum eigenvalues of the matrix.

5. $A = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}$

6. $A = \begin{bmatrix} 0 & 1 & -1 & 1 & -1 \\ 1 & 0 & 1 & -1 & 1 \\ -1 & 1 & 0 & 1 & -1 \\ 1 & -1 & 1 & 0 & 1 \\ -1 & 1 & -1 & 1 & 0 \end{bmatrix}$

7. Use the vector $\mathbf{x} = [0 \ \ldots \ 0 \ 1 \ 0 \ \ldots \ 0]^T = e^{(i)}$, the ith column of the identity matrix I_n, to explain why $\lambda_n \leq a_{ii} \leq \lambda_1$ for each of the diagonal elements of a real, symmetric matrix A.
8. Use the vector $\mathbf{x} \triangleq [1 \ \ 1 \ \ \ldots \ \ 1]^T$, the "*ones*" vector, to explain why $\lambda_n \leq \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} \leq \lambda_1$ for the elements of a real, symmetric matrix A.
9. Suppose A is a real, symmetric matrix. Let $\mathbf{q}_1$ and $\mathbf{q}_n$ be unit eigenvectors of A corresponding to λ_1 and λ_n, respectively, the maximum and minimum eigenvalues of A. Define $\mathbf{x}(t) = t\mathbf{q}_1 + (1-t)\mathbf{q}_n$ for $0 \leq t \leq 1$.
   a. Use the Pythagorean theorem to explain why $||\mathbf{x}(t)||^2 = t^2 + (1-t)^2$. Explain why that guarantees that $\mathbf{x}(t) \neq \mathbf{0}$ for $0 \leq t \leq 1$.
   b. Define a function of a single variable by $f(t) \triangleq \mathcal{R}_A\big(\mathbf{x}(t)\big)$. Explain why $f(t)$ is a continuous function.
   c. For a real, symmetric matrix such as A, the interval $W = [\lambda_n, \lambda_1]$ is called the **numerical range**. Use the Intermediate Value Theorem of Calculus I to explain why for every number w in the interval W there is a nonzero vector $\mathbf{x}$ for which $w = \mathcal{R}_A(\mathbf{x})$. This explains why we refer to it as the numerical "range."
10. Suppose A is a real, symmetric $n \times n$ matrix, C is a real, $m \times n$ matrix, and $B = A + C^T C$. Explain why
    a. The maximum eigenvalue of B is greater than or equal to the maximum eigenvalue of A
    b. The minimum eigenvalue of B is greater than or equal to the minimum eigenvalue of A

2.10 Short Take: Inner Product and Hilbert Spaces

Suppose $\mathcal{V}$ is a vector space and we have an operation that produces a scalar, $\langle \mathbf{x}, \mathbf{y} \rangle$, from every choice of vectors $\mathbf{x}, \mathbf{y}$ in $\mathcal{V}$.

Definition 2.24

(a) We say $\langle \cdot, \cdot \rangle$ is an **inner product** if it satisfies the properties

$$\left\{ \begin{array}{l} \text{(i)} \ \ \langle \mathbf{x}, \mathbf{y} \rangle = \overline{\langle \mathbf{y}, \mathbf{x} \rangle} \\ \text{(ii)} \ \ \langle \alpha \mathbf{x}, \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle \\ \text{(iii)} \ \ \langle \mathbf{x}_1 + \mathbf{x}_2, \mathbf{y} \rangle = \langle \mathbf{x}_1, \mathbf{y} \rangle + \langle \mathbf{x}_2, \mathbf{y} \rangle \\ \text{(iv)} \ \ \langle \mathbf{x}, \mathbf{x} \rangle \geq 0, \text{with equality only if } \mathbf{x} = \mathbf{0} \end{array} \right\}, \tag{2.76}$$

for all $\mathbf{x}, \mathbf{y}, \mathbf{x}_1, \mathbf{x}_2$ in $\mathcal{V}$ and all scalars α. In (i), $\bar{}$ denotes complex conjugation, as in Section 2.1.

(b) If $\langle \cdot, \cdot \rangle$ is an inner product on $\mathcal{V}$ then we say $(\mathcal{V}, \langle \cdot, \cdot \rangle)$ is an **inner product space**, or "i.p. space" for short. If the scalars are the real numbers, then we call it a **real i.p. space**; if the scalars are the complex numbers, then we call it a **complex i.p. space**. There are other choices for the set of scalars besides the real or complex numbers, but we won't need those other choices.

We have already seen examples of an i.p. space in Section 2.3. Because of Theorem 2.12 in Section 2.3, we have:

Example 2.39

$\left(\mathbb{R}^n, \langle \cdot, \cdot \rangle\right)$ is a real i.p. space when given the usual inner product $\langle \mathbf{x}, \mathbf{y} \rangle \triangleq \mathbf{x}^T \mathbf{y}$. ◯

We have seen in Example 2.9 in Section 2.1 that eigenvalues and eigenvectors of a real matrix may be complex. So, we also need the next example.

Example 2.40

$\left(\mathbb{C}^n, \langle \cdot, \cdot \rangle\right)$ is a complex i.p. space when given the inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle \triangleq \mathbf{x}^T \overline{\mathbf{y}} = \sum_{j=1}^{n} x_j \overline{y_j},$$

where $\bar{}$ denotes the complex conjugate. ◯

Because of Examples 2.39 and 2.40, it is common to define a more universal notation:

$$\mathbf{x}^H \mathbf{y} \triangleq \begin{Bmatrix} \mathbf{x}^T \mathbf{y}, & \text{if } \mathcal{V} = \mathbb{R}^n \\ \mathbf{x}^T \overline{\mathbf{y}}, & \text{if } \mathcal{V} = \mathbb{C}^n \end{Bmatrix}. \tag{2.77}$$

Recall that in Theorem 2.29 in Section 2.6 and Problems 2.6.3.17 and 2.6.3.18, we defined an alternative inner product on $\mathbb{R}^n$ by

$$\langle \mathbf{x}, \mathbf{y} \rangle_W \triangleq \langle W\mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T W^T \mathbf{y}.$$

Example 2.41

If W is a real, positive definite, symmetric matrix, then $\left(\mathbb{R}^n, \langle \cdot, \cdot \rangle_W\right)$ is a real i.p. space.

It is because a vector space may have more than one inner product that Definition 2.24(b) adopted the notation $(\mathcal{V}, \langle \cdot, \cdot \rangle)$ that pairs a vector space with an inner product.

Example 2.42

Let $\mathcal{P}_n$ be the set of all polynomials of degree less than or equal to n with real coefficients, that is, functions of the form

$$p(x) = a_0 + a_1 x + \cdots + a_n x^n,$$

where $a_0, a_1, \ldots, a_n$ are real constants. If $p(x), q(x)$ are in $\mathcal{P}_n$, define

$$\langle p, q \rangle \triangleq \int_{-1}^{1} p(x)q(x)\,dx. \tag{2.78}$$

Then $(\mathcal{P}_n, \langle \cdot, \cdot \rangle)$ is a real i.p. space.

To be brief and to avoid mathematical reasoning where possible, we will omit the explanation for property (d) in Example 2.42, specifically, why the only real polynomial of degree less than or equal to n for which $0 = \langle p, p \rangle = \int_{-1}^{1} p(x)p(x)dx = \int_{-1}^{1} |p(x)|^2 dx$ is the zero polynomial, that is, the polynomial whose coefficients are $a_0 = a_1 = \cdots = a_n = 0$.

Properties (a), (b), and (c) follow from linearity of the operation of integration.

2.10.1 Linear Functionals and Operators

Definition 2.25

Suppose $\mathcal{V}$ is an i.p. space.

(a) A **functional** on $\mathcal{V}$ is a function f whose inputs are vectors in $\mathcal{V}$ and whose outputs are scalars. The scalars are real (or complex numbers) if the i.p. space $\mathcal{V}$ is real (or complex, respectively). The "machine picture" is in Figure 2.12.

(b) A functional f is **linear** if

$$f(\alpha \mathbf{x} + \beta \mathbf{y}) = \alpha f(\mathbf{x}) + \beta f(\mathbf{y}),$$

for all vectors $\mathbf{x}, \mathbf{y}$ and scalars α, β.

Example 2.43

Explain why a linear functional f must have $f(\mathbf{0}) = 0$.

Why? For $\alpha = 0, \beta = 0$ and any vectors $\mathbf{x}, \mathbf{y}$, linearity of f implies that

$$f(\mathbf{0}) = f(0 \cdot \mathbf{x} + 0 \cdot \mathbf{y}) = 0 \cdot f(\mathbf{x}) + 0 \cdot f(\mathbf{y}) = 0 + 0 = 0. \bigcirc$$

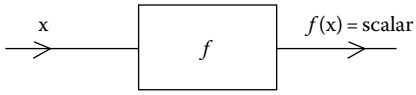


FIGURE 2.12
Functional as a machine.

Example 2.44

Suppose $\mathcal{V}$ is an i.p. space and $\mathbf{z}$ is a fixed vector in $\mathcal{V}$. Then,

$$f(\mathbf{x}) \triangleq \langle \mathbf{x}, \mathbf{z} \rangle$$

defines a linear functional on $\mathcal{V}$.

Why? For all $\mathbf{x}, \mathbf{y}$ in $\mathcal{V}$ and scalars α, β,

$$f(\alpha\mathbf{x} + \beta\mathbf{y}) = \langle \alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z} \rangle = \langle \alpha\mathbf{x}, \mathbf{z} \rangle + \langle \beta\mathbf{y}, \mathbf{z} \rangle = \alpha\langle \mathbf{x}, \mathbf{z} \rangle + \beta\langle \mathbf{y}, \mathbf{z} \rangle = \alpha f(\mathbf{x}) + \beta f(\mathbf{y}). \bigcirc$$

Definition 2.26

Suppose $\mathcal{V}$ is an i.p. space. A **linear operator** on $\mathcal{V}$ is a function A whose inputs are vectors in $\mathcal{V}$ and whose outputs are also vectors in $\mathcal{V}$ and satisfies

$$A(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha A(\mathbf{x}) + \beta A(\mathbf{y})$$

for all vectors $\mathbf{x}, \mathbf{y}$ and scalars α, β.

Example 2.45

Suppose $\mathcal{V}$ is $\mathbb{R}^n$ (or $\mathbb{C}^n$), and is given the usual inner product defined in (2.77). If A is a real (or complex), $n \times n$ matrix, then

$$A(\mathbf{x}) \triangleq A\mathbf{x}$$

defines a linear operator on $\mathbb{R}^n$ (respectively, $\mathbb{C}^n$.)

Because of this example, usually we will write the output of a linear operator on a vector $\mathbf{x}$ as $A\mathbf{x}$ rather than as $A(\mathbf{x})$.

Remark

In Example 2.43 we saw that every linear functional f has $f(\mathbf{0}) = 0$. A similar thing is true concerning linear operators: if A is any linear operator, then $A\mathbf{0} = \mathbf{0}$.

Example 2.46

Suppose $\mathbf{q}$ is a unit vector in $\mathbb{R}^n$. Then

$$A(\mathbf{x}) \triangleq \left(\mathbf{q}^T\mathbf{x}\right)\mathbf{q}$$

defines a linear operator on $\mathbb{R}^n$.

Why? For all $\mathbf{x}, \mathbf{y}$ in $\mathbb{R}^n$ and scalars α, β,

$$A\left(\alpha\mathbf{x} + \beta\mathbf{y}\right) = \left(\mathbf{q}^T\left(\alpha\mathbf{x} + \beta\mathbf{y}\right)\right)\mathbf{q} = \left(\alpha\left(\mathbf{q}^T\mathbf{x}\right) + \beta\left(\mathbf{q}^T\mathbf{y}\right)\right)\mathbf{q} = \alpha\left(\mathbf{q}^T\mathbf{x}\right)\mathbf{q} + \beta\left(\mathbf{q}^T\mathbf{y}\right)\mathbf{q}$$
$$= \alpha A\mathbf{x} + \beta A\mathbf{y}. \bigcirc$$

Orthogonal projection matrices, P, that we studied in Section 2.4 and used throughout Chapter 2, are examples of linear operators on $\mathbb{R}^n$.

2.10.2 Norm and Bounded Linear Operators

Suppose $(\mathcal{V}, \langle\cdot,\cdot\rangle)$ is an i.p. space.

Definition 2.27

The corresponding **norm** on $\mathcal{V}$ is given by $||\,\mathbf{x}\,|| \triangleq \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle}$.

Theorem 2.46

Suppose $(\mathcal{V}, \langle\cdot,\cdot\rangle)$ is an i.p. space. Then the corresponding norm $||\cdot||$ satisfies the properties in (2.9) in Section 2.3, that is, that for all $\mathbf{x}, \mathbf{y}$ in $\mathcal{V}$ and scalars α,

$$\left\{\begin{array}{l} \text{(a)}\ \ ||\alpha\mathbf{x}|| = |\alpha|\ ||\,\mathbf{x}\,|| \\ \text{(b)}\ \ ||\mathbf{x}+\mathbf{y}|| \le ||\,\mathbf{x}\,|| + ||\,\mathbf{y}\,|| \\ \text{(c)}\ \ ||\,\mathbf{x}\,|| \ge 0, \text{with equality only if } \mathbf{x} = \mathbf{0} \\ \text{(d)}\ \ ||\,\mathbf{x}\,||^2 = \langle \mathbf{x}, \mathbf{x}\rangle \\ \text{(e)}\ \ ||\mathbf{x}+\mathbf{y}||^2 = ||\,\mathbf{x}\,||^2 + ||\,\mathbf{y}\,||^2 + 2\langle \mathbf{x}, \mathbf{y}\rangle \\ \text{(f)}\ \ |\langle \mathbf{x}, \mathbf{y}\rangle| \le ||\,\mathbf{x}\,||\ ||\,\mathbf{y}\,|| \end{array}\right\}. \tag{2.79}$$

Note that Theorem 2.46(f) is the "Cauchy–Schwarz" inequality stated previously as Theorem 2.13(e) in Section 2.3, that is, (2.9)(e) in Section 2.3.

Example 2.47

If $\mathcal{V}$ is $\mathbb{R}^n$ (or $\mathbb{C}^n$) and is given the usual inner product $\langle \mathbf{x}, \mathbf{y}\rangle \triangleq \mathbf{x}^H\mathbf{y}$, then the corresponding norm satisfies

$$||\,\mathbf{x}\,||^2 = \mathbf{x}^H\mathbf{x} = \sum_{j=1}^{n} |x_j|^2.$$

Example 2.48

For all $\mathbf{x}, \mathbf{y}$ in an i.p. space,

$$\Big|\,||\,\mathbf{x}\,|| - ||\,\mathbf{y}\,||\,\Big| \le ||\,\mathbf{x}-\mathbf{y}\,||\,.$$

Why? First, note that $a \triangleq ||\,\mathbf{x}\,|| - ||\,\mathbf{y}\,||$ is a real number. If b is a nonnegative real number, an inequality of the form $|a| \le b$ is equivalent to $a \le b$ and $-b \le a$. So, we want to explain both $||\,\mathbf{x}\,|| - ||\,\mathbf{y}\,|| \le ||\,\mathbf{x}-\mathbf{y}\,||$ and $-||\,\mathbf{x}-\mathbf{y}\,|| \le ||\,\mathbf{x}\,|| - ||\,\mathbf{y}\,||$.

First, by the triangle inequality, that is, (2.9)(b) in Section 2.3, with $\mathbf{x}$ replaced by $\mathbf{x}-\mathbf{y}$, we have

$$||\mathbf{x}|| = ||(\mathbf{x}-\mathbf{y})+\mathbf{y}|| \leq ||\mathbf{x}-\mathbf{y}|| + ||\,\mathbf{y}\,||.$$

Subtract $||\mathbf{y}|$ from the left- and right-hand sides to get

$$||\mathbf{x}|| - ||\,\mathbf{y}\,|| \leq ||\mathbf{x}-\mathbf{y}||. \tag{2.80}$$

Second, in that inequality, swap the roles of $\mathbf{x}$ and $\mathbf{y}$ to get

$$||\mathbf{y}|| - ||\,\mathbf{x}\,|| \leq ||\mathbf{y}-\mathbf{x}||.$$

But, using (2.9)(a) in Section 2.3, we have

$$||\mathbf{y}-\mathbf{x}|| = ||(-1)(\mathbf{x}-\mathbf{y})|| = |-1|\,||\mathbf{x}-\mathbf{y}|| = ||\mathbf{x}-\mathbf{y}||,$$

so

$$||\mathbf{y}|| - ||\,\mathbf{x}\,|| \leq ||\mathbf{x}-\mathbf{y}||.$$

Multiply through by (-1), which changes the direction of the inequality, to get

$$||\mathbf{x}|| - ||\,\mathbf{y}\,|| = (-1)\left(||\mathbf{y}|| - ||\,\mathbf{x}\,||\right) \geq (-1)||\mathbf{x}-\mathbf{y}||,$$

that is,

$$-||\mathbf{x}-\mathbf{y}|| \leq ||\mathbf{x}|| - ||\,\mathbf{y}\,||.$$

This and (2.79) explain the result. ◯

Definition 2.28

Suppose A is a linear operator on an i.p. space $(\mathcal{V}, \langle\cdot,\cdot\rangle)$ and $||\cdot||$ denotes the corresponding norm. A is **bounded** if there is a constant $M \geq 0$ satisfying

$$||\,A\mathbf{x}\,|| \leq M||\,\mathbf{x}\,||, \text{ for } \textit{all} \text{ nonzero } \mathbf{x} \text{ in } \mathcal{V}. \tag{2.81}$$

If A is bounded, then the *smallest* such number M is denoted $||\,A\,||$, called the ("induced") norm of A.

Remark

1. $||\,A\mathbf{x}\,|| \leq ||\,A\,||\,||\,\mathbf{x}\,||$, for *all* $\mathbf{x}$ in $\mathcal{V}$.
2. $||\,A\,|| = 0$ only if $A = O$, the zero operator that satisfies $A\mathbf{x} = \mathbf{0}$ for all $\mathbf{x}$ in $\mathcal{V}$.
3. $||\,A\,||$ is the smallest number M for which it is true that

$$||\,A\mathbf{x}\,|| \leq M||\,\mathbf{x}\,||, \text{ for } \textit{all} \text{ unit vectors } \mathbf{x}.$$

Theorem 2.47

If A is a real (or complex), $n \times n$ matrix, then it is automatically a bounded linear operator on $\mathbb{R}^n$ (or $\mathbb{C}^n$, respectively) and

$$|| A || \leq || A ||_F \triangleq \sqrt{\sum_{j=1}^{n} \sum_{k=1}^{n} |a_{jk}|^2}. \tag{2.82}$$

$|| \, ||_F$ is called the **Frobenius norm**.

Theorem 2.48

Give $\mathbb{R}^n$ its usual norm $|| \mathbf{x} || = \sqrt{x_1^2 + \cdots + x_n^2}$.

(a) If Q is an $n \times n$ real, orthogonal matrix then considered as a linear operator on $\mathbb{R}^n$ it has $|| Q || = 1$.
(b) If P is a nonzero, orthogonal projection on $\mathbb{R}^n$ then $|| P || = 1$.

Why? For (a), do Problem 2.4.4.12. For (b), suppose P is any nonzero, orthogonal projection. First, let us explain why there must exist a nonzero $\mathbf{x}$ for which $P\mathbf{x} = \mathbf{x}$: because P is nonzero, there must be at least one $\mathbf{y}$ for which $P\mathbf{y} \neq \mathbf{0}$. After that, define $\mathbf{x} \triangleq P\mathbf{y}$. Because P is a projection, $P^2 = P$. It follows that

$$P\mathbf{x} = P(P\mathbf{y}) = P^2\mathbf{y} = P\mathbf{y} = \mathbf{x}.$$

From this we can conclude $M \geq 1$ in order for (2.81) to be true. Why? Because $\mathbf{x} = P\mathbf{x}$ implies that $|| \mathbf{x} || = || P\mathbf{x} || \leq M || \mathbf{x} ||$, hence $M \geq 1$. It follows that $|| P ||$, the least such M, must have $|| P || \geq 1$.

Second, let us explain why $|| P || \leq 1$: for all vectors $\mathbf{x}$ in $\mathcal{V}$, $\mathbf{x} = P\mathbf{x} + (I - P)\mathbf{x}$ is a decomposition into a sum or orthogonal vectors, by Theorem 2.17 in Section 2.3. By the Pythagorean theorem, that is, Theorem 2.15 in Section 2.3,

$$|| \mathbf{x} ||^2 = || P\mathbf{x} ||^2 + || (I - P)\mathbf{x} ||^2 \geq || P\mathbf{x} ||^2 + 0 = || P\mathbf{x} ||^2 \tag{2.83}$$

because Theorem 2.46(c) implies that $|| (I - P)\mathbf{x} ||^2 \geq 0$. Take the square root of both sides of (2.83) to get $|| \mathbf{x} || \geq || P\mathbf{x} ||$. Because this is true for all $\mathbf{x}$ in $\mathcal{V}$, the number $M = 1$ makes (2.81) true. So, P is bounded and has $|| P || \leq 1$.

So, $|| P || \geq 1$ and $|| P || \leq 1$, so we must have $|| P || = 1$. □

Example 2.49

If A is a bounded linear operator on an i.p. space and has an eigenvalue λ, explain why $|\lambda| \leq || A ||$.

Method: Let $\mathbf{q}$ be a unit eigenvector of A corresponding to eigenvalue λ, that is, $A\mathbf{q} = \lambda\mathbf{q}$ and $||\,\mathbf{q}\,|| = 1$. We calculate, using (2.79)(a),

$$|\,\lambda\,| = |\,\lambda\,| \cdot 1 = |\,\lambda\,|\,||\,\mathbf{q}\,|| = ||\,\lambda\mathbf{q}\,|| = ||\,A\mathbf{q}\,|| \leq ||\,A\,||\,||\,\mathbf{q}\,|| = ||\,A\,|| \cdot 1 = ||\,A\,||. \bigcirc$$

Example 2.49 is an example of "soft analysis." It explains why we can still say a lot in general even if we don't specify much about the particular space and linear operator. "Hard analysis" gets into the nitty-gritty details of a particular situation and can often produce more useful results.

Corollary 2.13

If A is a real, symmetric $n \times n$ matrix whose eigenvalues, including multiplicities, are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, then $||\,A\,|| = \max_{1 \leq j \leq n} |\lambda_j| = \max\{|\lambda_1|, |\lambda_n|\}$. Note that λ_1 and/or λ_n may be negative.

Why? The fact that $\max_{1 \leq j \leq n} |\lambda_j| = \max\{|\lambda_1|, |\lambda_n|\}$ follows from the assumption that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Note that λ_1 and/or λ_n may be negative.

By the result of Example 2.49, $|\lambda_j| \leq ||\,A\,||$ for $j = 1, \ldots, n$. It follows that

$$||\,A\,|| \geq \max_{1 \leq j \leq n} |\lambda_j|.$$

Here's how to explain why there is equality instead of inequality in the previous line: Using a corresponding set of eigenvectors $\{\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(n)}\}$ that is an o.n. basis for $\mathbb{R}^n$, we can expand any $\mathbf{x}$ in $\mathbb{R}^n$ as $\mathbf{x} = c_1\mathbf{q}^{(1)} + \cdots + c_n\mathbf{q}^{(n)}$. We calculate, using Parseval's identity Corollary 2.6(b) in Section 2.4, that

$$||\,A\mathbf{x}\,||^2 = ||A(c_1\mathbf{q}^{(1)} + \cdots + c_n\mathbf{q}^{(n)})||^2 = ||\lambda_1 c_1\mathbf{q}^{(1)} + \cdots + \lambda_n c_n\mathbf{q}^{(n)}||^2 = \sum_{i=1}^{n} |\lambda_i c_i|^2.$$

Use Parseval's identity a second time to get $||\,\mathbf{x}\,||^2 = \sum_{i=1}^{n} |c_i|^2$, so

$$||\,A\mathbf{x}\,||^2 = \sum_{i=1}^{n} |\lambda_i|^2 |c_i|^2 \leq \sum_{i=1}^{n} \left(\max_{1 \leq i \leq n} \{|\lambda_i|\}\right)^2 |c_i|^2 = \left(\max_{1 \leq i \leq n} \{|\lambda_i|\}\right)^2 \sum_{i=1}^{n} |c_i|^2, \tag{2.84}$$

hence

$$||\,A\mathbf{x}\,||^2 \leq \left(\max_{1 \leq i \leq n} \{|\lambda_i|\}\right)^2 ||\,\mathbf{x}\,||^2.$$

Take the square root of both sides of (2.84) to conclude that $||\,A\,|| \leq \max\{|\lambda_1|, |\lambda_n|\}$.

Because $||\,A\,|| \geq \max\{|\lambda_1|, |\lambda_n|\}$ and $||\,A\,|| \leq \max\{|\lambda_1|, |\lambda_n|\}$, the equality holds and the desired result follows. □

Definition 2.29

Suppose $\mathcal{V}$ is an i.p. space and A is a linear operator.

a. We say A is **one-to-one** if whenever $x_1 \neq x_2$, it follows that $Ax_1 \neq Ax_2$. [Note: because A is linear, this is the same as saying that the equation $Ax = 0$ has only the trivial solution $x = 0$.]
b. Suppose A is one-to-one. The **algebraic inverse** of A is the linear operator B defined implicitly by $By =$(the unique x such that $Ax = y$).
c. If the algebraic inverse is also bounded, we write $B = A^{-1}$ and we say that A is **invertible**.

In (2.23) in Section 2.5, we defined the concept of $^\perp$ for a vector subspace of $\mathbb{R}^n$. This generalizes to any i.p. space.

Definition 2.30

Suppose $\mathcal{V}$ is an i.p. space and $\mathcal{W}$ is a vector subspace of $\mathcal{V}$. The **orthogonal complement** of $\mathcal{W}$ is defined by

$$\mathcal{W}^\perp \triangleq \{\mathbf{x} \text{ in } \mathcal{V} : \ \langle \mathbf{w}, \mathbf{x} \rangle = 0 \text{ for } \textit{all } \mathbf{w} \text{ in } \mathcal{W}\}.$$

2.10.3 Convergence, Cauchy Completeness, and Hilbert Spaces

Recall that in Calculus we studied sequences $\{x_k\}_{k=1}^{\infty}$ of real numbers and we were particularly interested in convergence of sequences. For example, if $x_k = 1 - 10^{-k}$ for $k = 1, 2, 3, \ldots$, then we see that the sequence, that is,

$$.9, \ .99, \ .999, \ .9999, \ldots$$

converges to 1. Sometimes we write this as $x_k \to 1$, or, more explicitly, as

$$x_k \to 1 \text{ as } k \to \infty.$$

We are *not* saying that the sequence ever *arrives* at 1 but rather that the terms in the sequence get arbitrarily closer and closer to 1 the further we go along the sequence.

Suppose x_k is the diameter, in cm, of ball bearings produced by a machine and that ideally we want our bearings to have diameter 1 cm. If the machine's production run makes bearings whose diameters in cm. are successively,

$$.9, \ .99, \ .999, \ .9999, \ldots$$

then we would say that after a few bad bearings the machine settles down to produce better and better bearings. Define a positive number ε to be an "error tolerance" we demand

of our product. We want to find a K_ε, possibly dependent on ε, so that

$$|x_k - 1| < \varepsilon, \quad \text{for } \textit{all } k \geq K_\varepsilon.$$

For example, we may choose $K_{.01} = 3$ because $.999, .9999, \ldots$ are all within .01 of 1. Similarly, $K_{.05} = 3$ will work, but $K_{.001}$ has to be chosen to be at least 4. Note that $(K_\varepsilon - 1)$ may be the number of bearings rejected by an error tolerance of ε.

Definition 2.31

A sequence $\{\mathbf{x}_k\}_{k=1}^{\infty}$ in an i.p. space is

a. **Bounded** if there is an $M \geq 0$ such that $||\mathbf{x}_k|| \leq M$ for all k
b. **Converges** to $\mathbf{x}_\infty$ if for *all* values of $\varepsilon > 0$, we can choose K_ε such that

$$|| \mathbf{x}_k - \mathbf{x}_\infty || < \varepsilon, \text{ for } \textit{all } k \geq K_\varepsilon.$$

If $\{\mathbf{x}_k\}_{k=1}^{\infty}$, converges to $\mathbf{x}_\infty$, then we write $\mathbf{x}_k \to \mathbf{x}_\infty$. Also, we say $\{\mathbf{x}_k\}_{k=1}^{\infty}$ is **convergent** if there exists some $\mathbf{x}_\infty$ to which the sequence converges.

So, convergence of a sequence is the same as our being able to impose an *arbitrarily small* error tolerance. Of course, the number of objects rejected may become arbitrarily large as we impose an arbitrarily fine error tolerance.

In the real world, we don't demand mathematical convergence of the sequence of bearings produced by an infinite production run, for two reasons: First, we are not interested in producing an infinite number of bearings, and second we trade off the economic costs of "achieving" perfection, or getting arbitrarily close to it, versus the goal of producing reliable, high-quality bearings.

Nevertheless, the mathematical concept is relevant because people design and improve machines that are capable of greater and greater production and of making finer and finer quality product. The mathematical, ideal concept orients us to do better and better.

Theorem 2.49

$\mathbf{x}_k \to \mathbf{x}_\infty$ in an i.p. space $\mathcal{V}$ if, and only if, the sequence of real numbers $\{||\mathbf{x}_k - \mathbf{x}_\infty||\}_{k=1}^{\infty}$ converges to 0 in $\mathbb{R}^1$.

Example 2.50

If $\mathbf{x}_k \to \mathbf{x}_\infty$ and A is a bounded linear operator, then $A\mathbf{x}_k \to A\mathbf{x}_\infty$.

Why? $|| A\mathbf{x}_k - A\mathbf{x}_\infty || = || A(\mathbf{x}_k - \mathbf{x}_\infty) || \leq || A || \, | \mathbf{x}_k - \mathbf{x}_\infty || \to 0.$ ◯

Definition 2.32

Suppose $\mathcal{V}$ is an i.p. space.

a. If $\mathcal{S} \subseteq \mathcal{V}$, that is, $\mathcal{S}$ is a subset of $\mathcal{V}$, and every sequence $\{\mathbf{x}_k\}_{k=1}^{\infty}$ in $\mathcal{S}$ that is convergent has its limit in $\mathcal{S}$, then we say $\mathcal{S}$ is **closed**.
b. A vector subspace $\mathcal{W}$ in $\mathcal{V}$ is a **closed subspace** if $\mathcal{W}$ is a closed subset of $\mathcal{V}$.

Related to the definition of convergence is another concept.

Definition 2.33

A sequence $\{\mathbf{x}_k\}_{k=1}^{\infty}$ satisfies the **Cauchy criterion**, or is **Cauchy**, for short, if for *all* values of $\varepsilon > 0$ we can choose K_ε such that

$$|| \mathbf{x}_k - \mathbf{x}_\ell || < \varepsilon, \text{ for } \textit{all } k \text{ and } \ell \geq K_\varepsilon.$$

Theorem 2.50

A sequence of real numbers is convergent if, and only if, it is Cauchy.

Now we have enough background knowledge to see a crucial definition.

Definition 2.34

An i.p. space is **Cauchy complete** if every Cauchy sequence in the space is convergent. A **Hilbert space** is an i.p. space that is Cauchy complete.

Often we use the symbol $\mathcal{H}$ to denote a Hilbert space.

Theorem 2.51

$\mathbb{R}^n$ and $\mathbb{C}^n$, when given the usual inner product $\langle \mathbf{x}, \mathbf{y} \rangle \triangleq \mathbf{x}^H \mathbf{y}$, are Hilbert spaces.

Theorem 2.51 is a generalization of Theorem 2.50.

Unfortunately, not all i.p. spaces are Hilbert spaces. For example, $\mathcal{P}$, the space of all polynomials when given the inner product in (2.78), is not Cauchy complete because we can explain why the sequence of polynomials $p_k(t) \triangleq \sum_{j=0}^{k} \frac{1}{j!} t^j$, the partial sums of the Maclaurin series for e^t, is Cauchy but does not converge to a *polynomial*. Intuitively, an

i.p. space that is not Cauchy complete has "holes," for example, we're missing the place in $\mathcal{P}$ where e^t would be if it were a polynomial.

2.10.4 Bounded Linear Functionals and Operator Adjoint

Definition 2.35

Suppose f is a linear functional on an i.p. space $(\mathcal{V}, \langle\cdot,\cdot\rangle)$ and $||\cdot||$ denotes the corresponding norm. f is **bounded** if there is a constant $M \geq 0$ satisfying

$$|f(\mathbf{x})| \leq M||\,\mathbf{x}\,||, \text{ for } \textit{all} \text{ nonzero } \mathbf{x} \text{ in } \mathcal{V}.$$

If f is bounded, then the *smallest* such number M is called $||f||$.

In a real i.p. space, $|\ |$ denotes absolute value and in a complex i.p. space $|\ |$ denotes modulus of a complex number.

Theorem 2.52

(Riesz representation theorem) If f is a bounded linear functional on a Hilbert space $\mathcal{H}$, then there exists a vector $\mathbf{z}$ such that

$$f(\mathbf{x}) = \langle \mathbf{x}, \mathbf{z}\rangle, \text{ for } \textit{all } \mathbf{z} \text{ in } \mathcal{H}.$$

Definition 2.36

Suppose A is a bounded linear operator on a Hilbert space, $(\mathcal{H}, \langle\cdot,\cdot\rangle)$.

a. If there is a linear operator A^* that satisfies

$$\langle A\mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{x}, A^*\mathbf{y}\rangle, \text{ for } \textit{all } \mathbf{x}, \mathbf{y} \text{ in } \mathcal{V},$$

then we call A^* the **adjoint** of A.

b. If $A = A^*$ then we say A is **self-adjoint** or **Hermitian**.

Although we won't really give the explanation here, it is the Riesz representation theorem, that is, Theorem 2.52, that explains why the next result is true for *any* Hilbert space.

Theorem 2.53

If A is a bounded linear operator on a Hilbert space, then A^* exists.

Example 2.51

Suppose $\mathcal{H} = \mathbb{R}^n$ is given the usual inner product $\mathbf{x}^T\mathbf{y}$. If A is a real, $n \times n$ matrix then, considered as a bounded linear operator on $\mathcal{H}$, $A^* = A^T$.

Why? For *all* $\mathbf{x}, \mathbf{y}$ in $\mathbb{R}^n$,

$$\mathbf{x}^T A^T \mathbf{y} = (A\mathbf{x})^T \mathbf{y} = \langle A\mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{x}, A^*\mathbf{y}\rangle = \mathbf{x}^T A^* \mathbf{y}.$$

So, we need $A^* = A^T$. ◯

Example 2.52

Suppose $\mathcal{H} = \mathbb{C}^n$ is given the usual inner product $\mathbf{x}^H\mathbf{y}$. If A is a (possibly) complex $n \times n$ matrix then, considered as a bounded linear operator on $\mathcal{H}$, $A^* = \overline{A^T}$, where $^-$ denotes the complex conjugate.

Why? You will explain this result in Problem 2.10.8.5. ◯

2.10.5 Application to Signal Restoration

Example 2.53

On the Hilbert space $\mathcal{H} = \mathbb{C}^n$ with the usual inner product $\langle \mathbf{x}, \mathbf{y}\rangle = \mathbf{x}^H\mathbf{y} = \mathbf{x}^T\overline{\mathbf{y}}$, suppose W is a Hermitian matrix, that is, satisfies $W = W^*$. If W is positive definite, that is, $\langle W\mathbf{x}, \mathbf{x}\rangle > 0$ for all $\mathbf{x} \neq \mathbf{0}$, then

$$\langle \mathbf{x}, \mathbf{y}\rangle_W \triangleq \langle W\mathbf{x}, \mathbf{y}\rangle = \mathbf{x}^T W^T \overline{\mathbf{y}}$$

defines an inner product on $\mathcal{H}$.

This is a generalization of Example 2.41.
One example of a "signal restoration" problem (Cadzow 1997) deals with the equation

$$\mathbf{x} = A\mathbf{a} + \mathbf{w},$$

where
x is a **measurement signal**
A is a known $m \times n$ complex matrix
w is a Gaussian (random) noise vector. The mathematical problem is to find a vector of the form $A\mathbf{a}$, where **a** is in $\mathbb{C}^n$, so that the squared error functional

$$(\mathbf{x} - A\mathbf{a})^H W^T \overline{(\mathbf{x} - A\mathbf{a})}$$

is minimized. Here, W^T is the matrix inverse of the covariance matrix of **w**.

This is a generalized least squares problem similar to the one you studied in Problem 2.6.3.18, the only difference being that the vectors and matrices may be complex. Nevertheless, a similar conclusion holds, specifically that **a** should satisfy the normal equations

$$A^*WA\mathbf{a} = A^*W\mathbf{x}.$$

If it happens that A^*WA is invertible, then the unique l.s.s. would be

$$\mathbf{a} = (A^*WA)^{-1}A^*W\mathbf{x}.$$

This and much further applications of Hilbert space to signal processing are in the article by Cadzow.

2.10.6 Projection and Minimization

Definition 2.37

A bounded linear operator P on a Hilbert space $\mathcal{H}$ is a **projection** if it satisfies $P^2 = P = P^*$.

Theorem 2.54

Suppose $\mathbf{q}_1, \ldots, \mathbf{q}_n$ is an o.n. set in a Hilbert space $\mathcal{H}$. For all $\mathbf{x}$ in $\mathcal{H}$, define

$$P\mathbf{x} \triangleq \sum_{j=1}^{n} \langle \mathbf{x}, \mathbf{q}_j \rangle \mathbf{q}_j.$$

Then

a. P is a bounded linear operator and a projection on $\mathcal{H}$
b. For all $\mathbf{x}, \mathbf{y}$, $\langle P\mathbf{x}, (I - P)\mathbf{y} \rangle = 0$
c. For all $\mathbf{x}$, $P\mathbf{x}$ is the vector in $\mathcal{W} \triangleq Span\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$ that is closest to $\mathbf{x}$, that is,

$$\min\{||\mathbf{x} - \mathbf{v}|| \ : \ \mathbf{v} \text{ in } \mathcal{W}\} = ||\mathbf{x} - P\mathbf{x}||.$$

Why? (a) Linearity of P follows from properties (ii) and (iii) of Definition 2.24. To explain why $P^2 = P$, that is, $P^2\mathbf{x} = P\mathbf{x}$ for all $\mathbf{x}$, using linearity we calculate

$$P^2\mathbf{x} = P(P\mathbf{x}) = P\left(\sum_{j=1}^{n} \langle \mathbf{x}, \mathbf{q}_j \rangle \mathbf{q}_j\right) = \sum_{j=1}^{n} \langle \mathbf{x}, \mathbf{q}_j \rangle P\mathbf{q}_j = \sum_{j=1}^{n} \langle \mathbf{x}, \mathbf{q}_j \rangle \mathbf{q}_j = P\mathbf{x},$$

because for $\ell = 1, 2, \ldots$

$$P\mathbf{q}_\ell = \sum_{j=1}^{n} \langle \mathbf{q}_\ell, \mathbf{q}_j \rangle \mathbf{q}_j = \sum_{j=1}^{n} \delta_{\ell j}\mathbf{q}_j = \mathbf{q}_\ell.$$

As for explaining why $P = P^*$, for all $\mathbf{x}.\mathbf{y}$, we have

$$\langle \mathbf{x}, P^*\mathbf{y} \rangle \triangleq \langle P\mathbf{x}, \mathbf{y} \rangle = \left\langle \sum_{j=1}^{n} \langle \mathbf{x}, \mathbf{q}_j \rangle \mathbf{q}_j \, , \mathbf{y} \right\rangle = \sum_{j=1}^{n} \langle \mathbf{x}, \mathbf{q}_j \rangle \left\langle \mathbf{q}_j, \mathbf{y} \right\rangle, \tag{2.85}$$

by definition of P^*.

Using properties (i) and (ii) of Definition 2.24, we have $\langle \mathbf{v}, \alpha \mathbf{w} \rangle = \overline{\alpha} \langle \mathbf{v}, \mathbf{w} \rangle$, so

$$\langle \mathbf{x}, P\mathbf{y} \rangle = \left\langle \mathbf{x}, \sum_{j=1}^{n} \langle \mathbf{y}, \mathbf{q}_j \rangle \mathbf{q}_j \right\rangle = \sum_{j=1}^{n} \overline{\langle \mathbf{y}, \mathbf{q}_j \rangle} \langle \mathbf{x}, \mathbf{q}_j \rangle = \sum_{j=1}^{n} \langle \mathbf{x}, \mathbf{q}_j \rangle \langle \mathbf{q}_j, \mathbf{y} \rangle = \langle P\mathbf{x}, \mathbf{y} \rangle. \tag{2.86}$$

Using (2.85) and (2.86), $\langle \mathbf{x}, P^* \mathbf{y} \rangle = \langle \mathbf{x}, P\mathbf{y} \rangle$ for all $\mathbf{x}, \mathbf{y}$. So, $P = P^*$.

The boundedness of P follows from the explanation of Theorem 2.48.

(b) The explanation is the same as for Theorem 2.17 in Section 2.3.

(c) The explanation is the same as for the best approximation property of P_A explained in Section 2.5, particularly using (2.24) in Section 2.5. □

2.10.7 Weak Convergence and Compactness

Definition 2.38

A sequence $\{\mathbf{x}_k\}_{k=1}^{\infty}$ in a Hilbert space $\mathcal{H}$ is **weakly convergent** if there is some $\mathbf{x}_\infty$ such that

$$\text{for } \textit{all} \text{ fixed } \mathbf{y} \text{ in } \mathcal{H}, \ \langle \mathbf{x}_k, \mathbf{y} \rangle \to \langle \mathbf{x}_\infty, \mathbf{y} \rangle,$$

in which case we write $\mathbf{x}_k \rightharpoonup \mathbf{x}_\infty$.

In conversational English, we can think of weak convergence as "convergence in every direction $\mathbf{y}$."

Theorem 2.55

If $\mathbf{x}_k \to \mathbf{x}_\infty$ then $\mathbf{x}_k \rightharpoonup \mathbf{x}_\infty$.

Weak convergence really is different from convergence. For example, in Problem 2.10.8.6, you will explain why $\mathbf{x}_k \to \mathbf{x}_\infty$ implies that $||\mathbf{x}_k|| \to ||\mathbf{x}_\infty||$. Because some Hilbert spaces have *infinite dimension*, convergence in every direction does *not* imply convergence. A weakly convergent sequence* could have $\mathbf{x}_k \rightharpoonup \mathbf{0}$, but $||\mathbf{x}_k|| = 1$ for all $k = 1, 2, \ldots$ implies $\{\mathbf{x}_k\}_{k=1}^{\infty}$ cannot converge to $\mathbf{0}$.

Definition 2.39

If $\{\mathbf{x}_k\}_{k=1}^{\infty}$ is a sequence, then a **subsequence** is a partial, ordered list $\{\mathbf{x}_{k_\ell}\}_{\ell=1}^{\infty}$ of items chosen from the original sequence.

* For example, if $\{\mathbf{e}^{(k)}\}_{k=1}^{\infty}$ is an o.n. set, let $\mathbf{x}_k = \mathbf{e}^{(k)}$.

For example, the sequence

$$1, \frac{1}{2}, 2, \frac{2}{3}, 3, \frac{3}{4}, 4, \frac{4}{5}, \ldots$$

has a subsequence

$$\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \ldots$$

In this example, the subsequence is convergent even though the original sequence is not.

Definition 2.40

(a) A set $\mathcal{S}$ in a Hilbert space is **compact** if for every sequence in $\mathcal{S}$ there is a convergent subsequence whose limit is in S.
(b) A set $\mathcal{S}$ in a Hilbert space is **weakly compact** if every sequence in $\mathcal{S}$ has a weakly convergent subsequence whose weak limit is in S.

Theorem 2.56

(Banach-Alaoglu)

(a) Suppose $\mathcal{H}$ is a Hilbert space and $r > 0$ is a constant. Then the sets

$$\{\mathbf{x} : ||\,\mathbf{x}\,|| \leq r\} \text{ and } \{\mathbf{x} : ||\,\mathbf{x}\,|| = r\}$$

are weakly compact.
(b) Every bounded sequence in a Hilbert space has a weakly convergent subsequence.

Theorem 2.56 is powerful, particularly because of the concept defined next and the theorem that follows.

Definition 2.41

An operator A on a Hilbert space $\mathcal{H}$ is **compact** if for every weakly convergent sequence $\{\mathbf{x}_k\}_{k=1}^{\infty}$ the sequence $\{A\mathbf{x}_k\}_{k=1}^{\infty}$ is convergent.

Theorem 2.57

(Spectral theory) Suppose A is a bounded, linear, compact, self-adjoint operator on a Hilbert space $\mathcal{H}$. Then A has real eigenvalues λ_k and corresponding eigenvectors $\mathbf{q}_k$ such

that $\{\mathbf{q}_k\}$ is an o.n. basis for $\mathcal{H}$ and for all $\mathbf{x}$ in $\mathcal{H}$

$$A\mathbf{x} = \sum_j \lambda_j \langle \mathbf{x}, \mathbf{q}_j \rangle \mathbf{q}_j.$$

In the summation notation, we just wrote $\sum_j$ rather than $\sum_{j=1}^{\infty}$. This is because a Hilbert space may be finite dimensional or infinite dimensional. Theorem 2.57 can be considered as a generalization of Theorem 2.24 in Section 2.6 (Spectral decomposition of a real, symmetric matrix) to (possibly) infinite dimensional space. Because of Example 2.50, a linear operator on $\mathbb{R}^n$ is self-adjoint if it corresponds to an $n \times n$ real, symmetric matrix.

Learn More About It

Very good expositions of higher mathematics useful for science and engineering, including Hilbert spaces and the generalization known as Banach spaces, are found in (1) *Elements of Applicable Functional Analysis*, by Charles W. Groetsch, Marcel Dekker, Inc., 1980, and (2) *Linear Operator Theory in Engineering and Science*, (v. 40 of Applied Mathematical Sciences), by Arch W. Naylor and George R. Sell, Springer-Verlag, 1982.

2.10.8 Problems

1. Suppose p is a polynomial and define the *average value* of p by $\overline{p} = \frac{1}{2}\int_{-1}^{1} p(x)\,dx$ and the *root mean square of p* by $p_{rms} = \left(\frac{1}{2}\int_{-1}^{1} |p(x)|^2\,dx\right)^{1/2}$. Use the Cauchy–Schwarz inequality for the inner product defined in (2.78) and the function $q(x) \equiv 1$ to explain why $|\overline{p}| \leq p_{rms}$.
2. Suppose A and B are bounded linear operators on an i.p. space $\mathcal{V}$. Define BA by $(BA)(x) \triangleq B(Ax)$. You may assume that BA is a linear operator on $\mathcal{V}$. Explain why BA is also a bounded linear operator and why $||\,BA\,|| \leq ||\,B\,||\,||\,A\,||$.
3. Suppose $\mathcal{V}$ is an i.p. space and A is a one-to-one, bounded, linear operator. If B is the algebraic inverse of A and B is also a bounded linear operator, then $||\,B\,|| \geq (||\,A\,||)^{-1}$. [Hints: Use the fact that $BA = I$, along with $||\,I\,|| = 1$ and the result of Problem 2.10.8.2.]
4. Suppose $\mathcal{V}$ is an i.p. space and A is a one-to-one, bounded linear operator. Define $|||\,\mathbf{x}\,||| \triangleq ||\,A\mathbf{x}\,||$. Explain why $|||\,\mathbf{x}\,|||$ also defines a norm on $\mathcal{V}$.
5. Suppose $\mathcal{H} = \mathbb{C}^n$ is given the usual inner product $\mathbf{x}^H\mathbf{y}$. If A is a (possibly) complex $n \times n$ matrix then, considered as a bounded linear operator on $\mathcal{H}$, explain why $A^* = \overline{A^T}$, where $^-$ denotes the complex conjugate.
6. Explain why $\mathbf{x}_k \to \mathbf{x}_\infty$ implies that $||\mathbf{x}_k|| \to ||\mathbf{x}_\infty||$. [Hint: Use Example 2.48.]
7. Assume $\mathbf{x}_k \rightharpoonup \mathbf{x}_\infty$ *and* $\mathbf{y}_k \to \mathbf{y}_\infty$. Explain why the sequence of scalars $\langle \mathbf{x}_k, \mathbf{y}_k \rangle \to \langle \mathbf{x}_\infty, \mathbf{y}_\infty \rangle$.
8. Find the values of the constants c_1 and c_2 so that $\int_{-1}^{1} |x^2 - c_1 - c_2 x|^2 dx$ is minimized.
9. Suppose $\{\mathbf{u}^1, \ldots, \mathbf{u}^n\}$ is an orthonormal set of vectors in an inner product space $\mathcal{V}$ and $\lambda_1, \ldots, \lambda_n$ are complex numbers with $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$. Define

an operator A by $Ax \triangleq \sum_{j=1}^{n} \lambda_j \langle \mathbf{x}, \mathbf{u}^{(j)} \rangle \mathbf{u}^{(j)}$. You may assume that A is a linear operator. Explain why A is bounded and find $||A||$.

10. Suppose $\mathcal{V}$ is an i.p. space and $\mathcal{W}_1$ and $\mathcal{W}_2$ are vector subspaces of $\mathcal{V}$ with $\mathcal{W}_1 \subseteq \mathcal{W}_2$, that is, every vector $\mathbf{w}_1$ in $\mathcal{W}_1$ is also in $\mathcal{W}_2$. Explain why $\mathcal{W}_2^{\perp} \subseteq \mathcal{W}_1^{\perp}$.
11. Suppose A is a bounded linear operator on a Hilbert space $\mathcal{H}$. Suppose A is invertible. Explain why $\left(A^{-1}\right)^* = (A^*)^{-1}$.
12. Suppose A is a bounded linear operator on a Hilbert space $\mathcal{H}$. Suppose there is a positive number α such that for *all* $\mathbf{x}$ in $\mathcal{H}$ we have $||A\mathbf{x}|| \geq \alpha ||\mathbf{x}||$. Explain why A is invertible and $||A^{-1}|| \leq \alpha^{-1}$.
13. Use the Cauchy–Schwarz inequality to derive Theorem 2.47, that is, $||A|| \leq ||A||_F$.
14. Consider the functions $f_0(x) \equiv 1$, $f_1(x) = x$, and $f_2(x) = x^2$ defined on the interval $[0, 1]$. Define $\mathcal{V} = Span\{f_0(x), f_1(x), f_2(x)\}$. Find an orthonormal basis for $\mathcal{V}$ considered as a vector subspace of $L^2(0, 1)$ with the scalars being $\mathbb{R}$, with its usual inner product $\langle f(x), g(x) \rangle \triangleq \int_0^1 f(x)\, g(x)\, dx$.

Key Terms

adjoint: Definition 2.36 in Section 2.10
algebraic inverse: Definition 2.29 in Section 2.10
algebraic multiplicity: Definition 2.3 in Section 2.1
algorithm: during explanation of Theorem 2.16 in Section 2.3
best approximation: (2.24) in Section 2.5
bounded linear functional: Definition 2.35 in Section 2.10
bounded linear operator: Definition 2.28 in Section 2.10
bounded sequence: Definition 2.31 in Section 2.10
Cauchy complete: Definition 2.34 in Section 2.10
Cauchy criterion, **Cauchy**: Definition 2.33 in Section 2.10
Cauchy–Schwarz: Theorem 2.13(e) in Section 2.3
characteristic equation: Definition 2.2 in Section 2.1
characteristic polynomial: Definition 2.2 in Section 2.1
Cholesky factorization: Theorem 2.42 in Section 2.8
closed set, **closed subspace**: Definition 2.31 in Section 2.10
column space: after Corollary 2.3 in Section 2.4
compact: Definition 2.40 in Section 2.10
compact operator: Definition 2.41 in Section 2.10
complex conjugate: after Example 2.7 in Section 2.1; Figure 2.2
complex i.p. space: Definition 2.24 in Section 2.10
complex number: after Example 2.7 in Section 2.1
sequence converges, **convergent**: Definition 2.31 in Section 2.10
correlation coefficient: after Theorem 2.14 in Section 2.3
deficient: Definition 2.8 in Section 2.2
degenerate eigenvalue: Definition 2.3 in Section 2.1
diagonalizable, **diagonalizes**: Definition 2.6 in Section 2.2
eigenspace: Definition 2.5 in Section 2.2
eigenvalue, **eigenvector**: Definition 2.1 in Section 2.1

Frobenius norm: Theorem 2.47 in Section 2.10
functional, **linear**: Definition 2.25 in Section 2.10
generalized eigenvalue problem: Problem 2.1.6.29, (2.43) in Section 2.6
geometric *LU* **factorization**: Theorem 2.41 in Section 2.8
geometric multiplicity: Definition 2.4 in Section 2.1
generalized normal equations: Problem 2.6.3.18
generalized, **weighted**, **least squares** problem: Problem 2.6.3.18
Gram matrix, **Grammian**: before Example 2.21 in Section 2.5
Gram–Schmidt process: Theorem 2.16 in Section 2.3
Hermitian: Definition 2.36 in Section 2.10
Hilbert space: Definition 2.34 in Section 2.10
Householder matrix: Problem 2.4.4.17
imaginary part: after Example 2.7 in Section 2.1; Figure 2.2
inertia matrix: after (2.42) in Section 2.6
inner product: Definition 2.9 in Section 2.3, Definition 2.24 in Section 2.10
inner product space: Definition 2.24 in Section 2.10
invertible: Definition 2.29 in Section 2.10
Kronecker delta: after Definition 2.14 in Section 2.3
least squares solution (l.s.s.): after (2.22) in Section 2.5
length: Definition 2.10 in Section 2.3
linear operator: Definition 2.26 in Section 2.10
LU **factorization**: (2.69) in Section 2.8
magnitude: Definition 2.10 in Section 2.3
mass matrix: after (2.42) in Section 2.6
MATLAB and *Mathematica*: after Theorem 2.5 in Section 2.1
measurement signal: Example 2.53 in Section 2.10
Moore–Penrose generalized inverse: (2.66) in Section 2.7
norm: Definitions 2.10 in Section 2.3, (2.27) in Section 2.10
normal equations: Theorem 2.22 in Section 2.5
normalized: after Theorem 2.5 in Section 2.1, after Example 2.17 in Section 2.3
numerical range: Problem 2.9.2.9
one-to-one: Definition 2.29 in Section 2.10
orthogonal: Definition 2.12 in Section 2.3
orthogonal complement: Definition 2.30 in Section 2.10
orthogonal matrix: Definition 2.16 in Section 2.4
orthogonal projection: Definition 2.15 in Section 2.3
orthogonal set of vectors: Definition 2.13 in Section 2.3
orthogonally similar: Definition 2.18 in Section 2.6
orthonormal set of vectors: Definition 2.14 in Section 2.3
Parseval identities: Corollary 2.6 in Section 2.4
positive definite: Definition 2.19 in Section 2.6
positive semi-definite: Definition 2.19 in Section 2.6
principal minors: Definition 2.21 in Section 2.6
projection: Definition 2.36 in Section 2.10
pseudo-diagonal: Definition 2.22 in Section 2.7
real: after Example 2.7 in Section 2.1; Figure 2.2
range: after Corollary 2.3 in Section 2.4
Rayleigh quotient: Definition 2.23 in Section 2.9
real i.p. space: Definition 2.24 in Section 2.10

real part: after Example 2.7 in Section 2.1; Figure 2.2
reduced SVD: (2.56) in Section 2.7
regression line: Example 2.21 in Section 2.5
residual: after (2.22) in Section 2.5
self-adjoint: Definition 2.37 in Section 2.10
similar: Definition 2.7 in Section 2.2
simple eigenvalue: Definition 2.3 in Section 2.1
singular value decomposition (SVD): Theorem 2.37 in Section 2.7
spectral formula, (Spectral decomposition): Theorem 2.24 in Section 2.6
Spectral Theorem, **spectra**: Theorem 2.23 in Section 2.6
stiffness matrix: after (2.42) in Section 2.6
subsequence: Definition 2.40 in Section 2.10
symmetric: Definition 2.17 in Section 2.6
thin SVD: (2.56) in Section 2.7
triangle inequality: (2.9)(f) in Section 2.3
unit vector: after Example 2.17 in Section 2.3
weakly compact: Definition 2.41 in Section 2.10
weakly convergent: Definition 2.39 in Section 2.10

MATLAB® commands

`colspace(A)`: after Corollary 2.3 in Section 2.4

`eig`: after Theorem 2.5 in Section 2.1

`>[V,D]=eig(A)`: after Theorem 2.5 in Section 2.1

`>format long`: after Theorem 2.5 in Section 2.1

Mathematica Commands

`Eigenvalues[A],Eigenvectors[A]`: after Theorem 2.5 in Section 2.1

`Eigenvalues[A,k],Eigenvectors[A,k]`: after Theorem 2.5 in Section 2.1

`Eigensystem[A],Eigensystem[A,k]`: after Theorem 2.5 in section 2.1

RegionPlot3D[ac > 1 && 4a − b^2 > 0 && 4ac − 4a −4b − b^{2}c > 4, {a, 0, 6}, {b, −4, 3}, {c, 1, 6}]: after Example 2.28 in Section 2.6

FindMaximum[{f[x,y,z], 1 ≥ x ≥ −1&&1 ≥ y ≥ −1 &&1 ≥ z ≥ −1}, {x,y,z}]: Example 2.37 in Section 2.9

References

Bryan, K. and Leise, T. The $25,000,000,000* eigenvector: The linear algebra behing Google. *SIAM Review*, 48, 569–581, 2006.

Cadzow, J.A. Signal restoration, in R.C. Dorf (editor-in-chief), *The Electrical Engineering Handbook*, 2nd edn. CRC Press/IEEE Press, Boca Raton, FL, 1997, Section 14.4.

Naylor, A.W. and Sell, G.R. *Linear Operator Theory in Engineering and Science*, Vol. 40, Applied Mathematical Sciences, Springer-Verlag, New York, 1982.

3

Scalar ODEs I: Homogeneous Problems

3.1 Linear First-Order ODEs

3.1.1 Scalar ODEs

A **scalar ordinary differential equation** (ODE) is an equation that relates the derivative(s) of a *single* function of one variable to possibly the independent variable and the function itself. For example,

$$\frac{dy}{dt}(t) = ky(t)$$

is the ODE for exponential growth or decay that you saw in calculus.

In this section and chapter, we will be most concerned with *methods* for solving ODEs. At the end of Section 3.2, we will also present three basic "existential results" that give a firm foundation for all of the techniques we will learn.

The **order** of the ODE is the highest derivative of that function in the equation. We will study first-order ODEs that can be written in the form

$$\frac{dy}{dt}(t) = f\big(t, y(t)\big),$$

for some function f. Often, our notation will suppress the dependence on the independent variable and write the ODEs as

$$\frac{dy}{dt} = f(t, y). \tag{3.1}$$

Having the derivative with respect to t tells us that y is a function of t, so $y = y(t)$ and $\dot{y} = \dot{y}(t)$. More generally, ODEs could also take the form $F(t, y, \dot{y}) = 0$, that is, involve $\dot{y}$ implicitly. For example, we might try to solve $(\dot{y})^2 = 2(1 - \cos y)$ after rewriting it in the forms of $\dot{y} = \pm\sqrt{2(1 - \cos y)}$.

Definition 3.1

A **solution** of an ODE (3.1) is a function $y(t)$ defined on an open interval I for which the derivative also exists on I and satisfies (3.1) on I.

Open intervals can be of the form $(a,b) = \{t : a < t < b\}$, where $a < b$, $(-\infty,\infty) = \{t : -\infty < t < \infty\}$, $(-\infty,b) = \{t : -\infty < t < a\}$, or $(a,\infty) = \{t : a < t < \infty\}$.

A solution has to satisfy the ODE on an open interval. A practical way to find the interval is to substitute the supposed solution into the ODE and during, or after, that, choose the open interval I. It is usually safe to treat the issue of the open interval as an afterthought as long as you don't ignore it completely.

Along with a single first-order differential equation, we may also have an **initial condition**, for example, $y(0) = 1$. The combination of a differential equation and an initial condition is called an **initial value problem** (IVP). Usually, when solving an IVP, first we should find all of the solutions of the differential equation and then satisfy the initial condition by solving for an arbitrary constant.

Example 3.1

Find the solution of the ODE $\dot{y} + 10y = -2\sin 2t$ that also satisfies the IC $y(0) = 1$. You may assume that the ODE has solutions

$$y(t) = Ce^{-10t} + \frac{1}{52}(2\cos 2t - 10\sin 2t),$$

for any value of the constant C, as we will derive in Example 3.6.

Method: Substitute the solutions into the IC to get

$$1 = y(0) = C \cdot 1 + \frac{1}{52}(2 \cdot 1 - 0) = C + \frac{1}{26},$$

so we should choose $C = \frac{25}{26}$. The solution of the IVP is

$$y(t) = \frac{25}{26}e^{-10t} - \frac{1}{26}(\cos 2t - 5\sin 2t). \bigcirc$$

By the way, the ODE in Example 3.1 can come from a series RC circuit with an alternating current (AC) source, where $y(t)$ is the loop current. You will see an example of this in Problem 4.1.5.16.

There is an additional notion of a solution of an ODE: An **implicit solution** is a curve, for example, $x^2 + y^2 = 4$, on which $\frac{dy}{dx} = f(x,y)$ except possibly at finitely many points. If such a curve passes through a given point (x_0, y_0), then we say that the implicit solution also satisfies the initial condition $y(x_0) = y_0$. If asked for a solution, we should find a solution rather than an implicit solution, if at all possible.

3.1.2 Linear First-Order ODEs

The **standard form** of linear first-order ODEs is

$$\dot{y} + p(t)y = f(t). \tag{3.2}$$

ODE (3.2) is called **homogeneous** if $f(t) \equiv 0$. Physically, this can occur in the simplest model for friction acting to slow an object, $\dot{v} + pv = 0$.

A little algebra may be needed to put a linear first-order ODE into standard form, for example, the ODE

$$\frac{1}{2}\,t^2\dot{y} + ty = e^{-2t}$$

can be multiplied through by $\frac{2}{t^2}$ to rewrite it as

$$\dot{y} + \frac{2}{t}\,y = \frac{2}{t^2}\,e^{-2t}.$$

The *integrating factor method* allows us to find a formula for solutions of ODE (3.2). First, let's see how it works in an example.

Example 3.2

Find all solutions of

$$\dot{y} + 2y = 5e^{-3t}. \tag{3.3}$$

Method: If we could somehow combine the two terms on the left-hand side (LHS) of the ODE, then hopefully we could solve for y. The product rule says that the derivative of a product yields two terms. As they are now, the two terms on the LHS of (3.3) cannot be combined to be the derivative of one term. But, if we multiply through (3.3) by e^{2t}, we have

$$e^{2t}\,\dot{y} + 2e^{2t}y = 5e^{2t}e^{-3t}. \tag{3.4}$$

The two terms on the LHS add up to $\frac{d}{dt}\big[\,e^{2t}y\,\big]$, so (3.4) can be rewritten as

$$\frac{d}{dt}\big[\,e^{2t}y\,\big] = 5e^{2t}e^{-3t} = 5e^{-t}.$$

Indefinite integration with respect to t of both sides yields

$$e^{2t}y = -5e^{-t} + c_1,$$

and then multiply through by e^{-2t} to have

$$y = -5e^{-3t} + c_1e^{-2t},$$

where c_1 is an arbitrary constant. This gives all* of the solutions. ○

Now let's return to the standard linear first-order differential (3.2), $\dot{y} + p(t)y = f(t)$. In Example 3.2, multiplication of both sides of the ODE by e^{2t} appeared "out of the blue." While the terms $\dot{y} + 2y$ cannot be combined using the product rule, $e^{2t}\,\dot{y} + 2e^{2t}y$ can be. How did we know that multiplying through by e^{2t} would be so useful?

* The algebraic steps and the step of integrating neither created spurious solutions nor lost valid solutions, at least on an interval of t on which $e^{2t}y$ and $5e^{-t}$ are differentiable and nonzero. This is because of Rolle's theorem, which states that $\frac{dw}{dt} = 0$ is equivalent to $w = c_1$, where c_1 is an arbitrary constant.
More generally, in ODE (3.2), $p(t)$ and $f(t)$ being continuous will be useful technical assumptions for finding *all* of the solutions.

The factor e^{2t} is called an integrating factor. In general, if multiplication by $\mu(t)$ of the two terms on the LHS of (3.2) leads to the two terms being a single derivative, then we say $\mu(t)$ is an **integrating factor**:

$$\mu(t)\left(\dot{y} + p(t)y\right) = \mu(t)\,\dot{y} + \mu(t)p(t)y = \frac{d}{dt}\big[\,\mu(t)y\,\big].$$

But, according to the product rule,

$$\frac{d}{dt}\big[\,\mu(t)y\,\big] = \mu(t)\,\dot{y} + \frac{d\mu(t)}{dt}\,y,$$

so we need $\dfrac{d\mu}{dt} = \mu p(t)$. This is a differential equation whose solutions are $\mu(t) = Ce^{\int p(t)dt}$, where C is any constant. We could find all such solutions for $\mu(t)$, but we need *only one* integrating factor,

$$\mu(t) = e^{\int p(t)\,dt}, \tag{3.5}$$

to multiply through (3.2) so as to be able to use the product rule. So,

Example 3.3

Find an integrating factor for $\dot{y} + 2y = 5e^{-3t}$.

Method: The ODE is in standard form: $\mu(t) = e^{\int p(t)dt} = e^{\int 2dt} = e^{2t}$ is an integrating factor. ◯

Example 3.4

Find all solutions of ODE $t\dot{y} + 2y = 5e^{-3t}$.

Method: To find an integrating factor, first put the PDE into standard form (3.2):

$$\dot{y} + \frac{2}{t}\,y = 5\,\frac{1}{t}\,e^{-3t}. \tag{3.6}$$

Now that the equation is in *standard form*, we can use (3.5):

$$\mu(t) = e^{\int p(t)dt} = e^{\int \frac{2}{t}dt} = e^{2\ln(t)} = e^{\ln(t^2)} = t^2.$$

Multiply through (3.6) by $\mu(t)$ to get

$$t^2\dot{y} + 2ty = 5te^{-3t},$$

use the product rule to rewrite it as

$$\frac{d}{dt}\big[\,t^2y\,\big] = 5te^{-3t},$$

and then integrate both sides with respect to t to get

$$t^2y = 5\int te^{-3t}\,dt.$$

The latter can be done using the method of integration by parts: Let $u = t$ and $dv = e^{-3t}dt$, so $du \triangleq \frac{du}{dt}dt = 1dt = dt$ and $v = \int dv = \int e^{-3t}dt = -\frac{1}{3}e^{-3t}$. So,

$$\int te^{-3t}dt = \int udv = uv - \int vdu = -\frac{1}{3}te^{-3t} - \int \left(-\frac{1}{3}e^{-3t}\right)dt.$$

So,

$$t^2y = 5\left(-\frac{1}{3}te^{-3t} - \frac{1}{9}e^{-3t} + c_1\right),$$

where c_1 is an arbitrary constant. So, all solutions of $t\dot{y} + 2y = 5e^{-3t}$ are given by

$$y = -\frac{5}{9}\left(3t^{-1} + t^{-2}\right)e^{-3t} + c_1t^{-2}, \text{ where } c_1 \text{ is an arbitrary constant. } \bigcirc$$

Analogous to Definitions 1.17 and 1.19 in Section 1.3 for *homogeneous* systems of linear algebraic equations, we have

Definition 3.2

The **general solution** of a first-order linear *homogeneous* ODE

$$\dot{y} + p(t)y = 0 \tag{3.7}$$

has the form

$$y_h = c_1y_1(t),$$

if for every solution $y^*(t)$ of (3.7) there is a value of the scalar c_1 giving $y^*(t) = c_1y_1(t)$. In this case, we call the set of one function $\{y_1(t)\}$ a **complete set of basic solutions**, and we call the function $y_1(t)$ a **basic solution** of (3.7).

Unlike for systems of linear algebraic equations, for first-order linear homogeneous ODEs, we don't have a concept of "free variables" or "basic variables." But also unlike for systems of linear algebraic equations, we have a nice formula for a basic solution: denote

$$y_1(t) = e^{-\int p(t)dt}.$$

Because we have such a formula for a basic solution, first-order linear homogeneous ODEs are simpler than systems of linear algebraic equations, at least in this way.

Again, analogous to Definitions 1.21 and 1.22 in Section 1.4 for nonhomogeneous systems of linear algebraic equations, we have

Definition 3.3

(a) A **particular solution** of a first-order linear nonhomogeneous ODE (3.2) is any function $y_p(t)$ that satisfies (3.2).

(b) $y(t) = y_p(t) + c_1y_1(t)$ is the **general solution** of ODE (3.2) if for every solution $y^*(t)$ of (3.2), there is a value of the scalar c_1 giving $y^*(t) = y_p(t) + c_1y_1(t)$.

Analogous to Theorem 1.18 in Section 1.4, we have

Theorem 3.1

If $y_p(t)$ is any particular solution of first-order linear nonhomogeneous ODE (3.2) and $y_1(t)$ is any basic solution of the corresponding first-order linear homogeneous ODE (3.7), then $y(t) = y_p(t) + c_1 y_1(t)$ is the general solution of ODE (3.2).

This result will follow from the theory found at the end of Section 3.2.

The integrating factor method produces in one process all of the solutions, including both a particular solution and the solutions of the corresponding homogeneous problem. Additionally, there is an unusually helpful aspect of the method: After multiplying the standard form of the ODE through by the integrating factor, the product rule acts as an *automatic check* that most of the preceding work has been done correctly. I advise you to always check that the product rule works out in the problems you do. If the product rule doesn't seem to be working, then it should act as an *alarm bell* to alert you to an error having happened. Unfortunately, like some error messages a computer might give you, this alarm bell doesn't tell you specifically *what* the error was.

The likely sources of error are (1) an algebra error when putting the ODE into standard form, (2) misidentifying the function $p(t)$, for example, by ignoring a minus sign, (3) an error in doing the integral $\int p(t)\,dt$, (4) an error in using the exponential function and possibly a logarithm, (5) an error in multiplying both sides of the ODE in standard form by $\mu(t)$, (6) an error in doing the integral $\int \mu(t)p(t)dt$, and (7) an algebra error in dividing both sides of $\mu(t)y = \cdots$ by $\mu(t)$.

As we've seen, along with a single first-order differential equation, we may also have an initial condition. When solving an IVP involving a linear differential equation, usually we should first find all solutions of the differential equation and, only then, satisfy the initial condition by solving for the arbitrary constant.

Example 3.5

Solve the IVP $\dot{y} + \dfrac{5}{10-2t}y = 4, y(0) = 23$.

Method: The ODE is in standard form and has an integrating factor

$$e^{\int \frac{5}{10-2t}dt} = e^{-\frac{5}{2}\ln|10-2t|} = e^{\ln\left(|10-2t|^{-5/2}\right)} = |10-2t|^{-5/2}.$$

But, because

$$|a| = \left\{ \begin{array}{ll} a, & \text{if } a > 0 \\ -a, & \text{if } a < 0 \end{array} \right\},$$

we don't need the absolute value in the integrating factor: $\mu(t) = (10-2t)^{-3/2}$ will do.* (Remember that we only need to find one integrating factor.) We multiply $\mu(t)$ through

* Because we're solving on an interval containing $t = 0$ and $(10-2t)^{-5/2}$ doesn't involve the square root of a negative number near $t = 0$.

the ODE and use the product rule to get

$$\frac{d}{dt}\left[(10-2t)^{-5/2}y\right] = (10-2t)^{-5/2}\dot{y} + (10-2t)^{-5/2}\,\frac{5}{10-2t}\,y = 4(10-2t)^{-5/2}.$$

Integrate both sides with respect to t to get

$$(10-2t)^{-5/2}y = \int 4(10-2t)^{-5/2}dt = \frac{4}{3}\,(10-2t)^{-3/2} + c_1;$$

hence, all solutions of the ODE are given by

$$y(t) = \frac{4}{3}\,(10-2t) + c_1(10-2t)^{5/2}.$$

Substitute in the initial condition to get

$$23 = y(0) = \frac{40}{3} + c_1 \cdot 10^{5/2},$$

so $c_1 = \frac{29}{3}\left(10^{-5/2}\right)$. The solution of the IVP is

$$y(t) = \frac{4}{3}\,(10-2t) + \frac{29}{3}\left(\frac{10-2t}{10}\right)^{5/2},$$

that is,

$$y(t) = \frac{4}{3}\,(10-2t) + \frac{29}{3}\left(\frac{5-t}{5}\right)^{5/2}. \bigcirc$$

Note that this solution only exists on the time interval $-\infty < t \leq 5$.

Example 3.6

For ODE $\dot{y} + 10y = -2\sin 2t$, find the general solution and a particular solution.

Method: It's easy to see that e^{10t} is an integrating factor. Multiply through by it to get

$$\frac{d}{dt}\left[e^{10t}y\right] = e^{10t}\,\dot{y} + 10e^{10t}y = -2e^{10t}\sin 2t,$$

so the solutions are given by

$$e^{10t}y = -\int 2e^{10t}\sin 2t dt. \tag{3.8}$$

To do this kind of integral, use integration by parts twice followed by the "recursion" technique, or consult a book of formulas that includes

$$\int e^{at}\cos bt\,dt = \frac{e^{at}}{a^2+b^2}\cdot\left(b\sin bt + a\cos bt\right) + c \tag{3.9}$$

and

$$\int e^{at}\sin bt\,dt = \frac{e^{at}}{a^2+b^2}\cdot\left(a\sin bt - b\cos bt\right) + c, \tag{3.10}$$

where c is an arbitrary constant, or use a symbolic manipulation capable calculator or software package. In our problem here, we have

$$e^{10t}y = -2\int e^{10t}\sin 2t dt = \frac{-2e^{10t}}{104}\,(10\sin 2t - 2\cos 2t) + c_1.$$

Using this in (3.8), the general solution of the ODE is

$$y = c_1 e^{-10t} - \frac{1}{52}(10\sin 2t - 2\cos 2t).$$

Since $y = c_1 e^{-10t}$ is a solution of the corresponding homogeneous ODE,

$$y_p(t) = -\frac{1}{52}(10\sin 2t - 2\cos 2t)$$

is a particular solution of the original, nonhomogeneous ODE. ○

3.1.3 Steady-State and Transient Solutions

We saw in Example 3.6 that the solutions of $\dot{y} + 10y = -2\sin 2t$ are

$$y(t) = c_1 e^{-10t} - \frac{1}{52}(10\sin 2t - 2\cos 2t).$$

For any constant c_1, we know that $\lim_{t\to\infty} c_1 e^{-10t} = 0$, that is, $c_1 e^{-10t} \to 0$ as $t \to \infty$. This inspires a definition that is fundamental to engineering:

Definition 3.4

A **transient solution** of a linear ODE is a function $z(t)$ that (a) is either a particular solution or a solution of the corresponding homogeneous ODE and (b) has $\lim_{t\to\infty} z(t) = 0$.

Let $y_T(t)$ denote a transient solution. Intuitively, a transient solution is that part of a solution that becomes insignificant *eventually*.

The definition does *not* say that a transient solution must be a solution, by itself, of the ODE. As we will see, often, but not always, a transient solution turns out to be a solution of the corresponding linear homogeneous ODE. But in Example 4.7 in Section 4.1, we will see that the transient solution may not be a homogeneous solution. On the other hand, in Example 3.6, $y_T(t) = c_1 e^{-10t}$ is both a homogeneous solution and a transient solution.

There is another part to the solutions of $\dot{y} + 10y = -2\sin 2t$:

$$y_S(t) \triangleq y(t) - y_T(t) = -\frac{1}{52}(10\sin 2t - 2\cos 2t).$$

Note that $\lim_{t\to\infty} y_S(t)$ does not exist, due to oscillation.

Definition 3.5

A function $f(t)$ is **bounded as** $t \to \infty$ if there is *no* sequence of times t_n for which $\lim_{n\to\infty} |f(t_n)| = \infty$.

For example, $f(t) \triangleq e^t \cos 2t$ is *not* bounded as $t \to \infty$ because $f(n\pi) = |e^{n\pi}\cos 2n\pi| = e^{n\pi} \to \infty$ as $n \to \infty$.

Definition 3.6

(a) If a solution of a linear ODE can be written as $y(t) = y_T(t) + y_S(t)$, where

- $y_T(t)$ is a transient solution, and
- $y_S(t)$ is bounded as $t \to \infty$ and does *not* have $\lim_{t\to\infty} y_S(t) = 0$,

then we say $y_S(t)$ is a **steady-state solution**.

(b) Alternatively, if a solution $y(t)$ of a linear ODE is bounded as $t \to \infty$ and does *not* have $\lim_{t\to\infty} y(t) = 0$, then we say that $y(t) = y_S(t)$ is **itself a steady-state solution**.

So, in Example 3.6,

$$y_S(t) = -\frac{1}{52}(10\sin 2t - 2\cos 2t)$$

is a steady-state solution.

The definition does *not* say that a steady-state solution must be a solution, by itself, of the ODE. As we will see, often, but not always, a steady-state solution turns out to be a particular solution of the linear nonhomogeneous ODE. But in Example 4.7 in Section 4.1, we will see that the steady-state solution is not a particular solution.

Figure 3.1 illustrates a steady-state solution for the solution of Example 3.6 with $c_1 = 1.5$, that is, the solution of $\dot{y} + 10y = -2\sin 2t$, $y(0) = \frac{20}{13}$. Notice that as t increases, the transient solution, $1.5e^{-10t}$, becomes insignificant, leaving the solution, $y(t)$, looking like the steady-state oscillation $y_S(t)$.

Example 3.7

Consider the IVP $\dot{y} = -by + t$, $y(0) = 3$. Let b be an unspecified positive constant. (a) Solve the IVP, and (b) find the transient solution and the steady-state solution, if they exist.

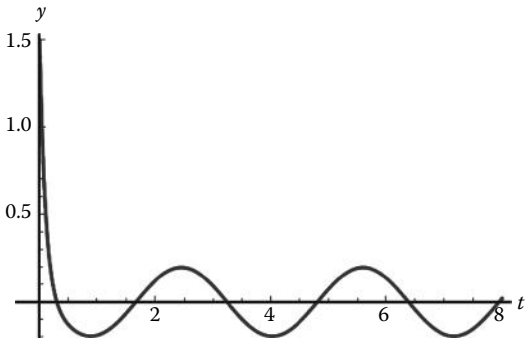


FIGURE 3.1
Example 3.6.

Method: (a) The only thing unusual about this example is that we don't know the value of the positive constant b. But our usual method works well: first, put the ODE into the standard form

$$\dot{y} + by = t.$$

Using integrating factor $\mu(t) = e^{\int b dt} = e^{bt}$, we have

$$\frac{d}{dt}\left[e^{bt}y\right] = e^{bt}\left(\dot{y} + by\right) = te^{bt},$$

so the solutions are given implicitly by

$$e^{bt}y = \int te^{bt}dt = t \cdot \frac{e^{bt}}{b} - \int \frac{e^{bt}}{b}dt = \frac{e^{bt}}{b^2}(bt - 1) + c_1.$$

Multiplying through by e^{-bt}, we find that the solutions are

$$y = c_1 e^{-bt} + \frac{bt - 1}{b^2},$$

where c_1 is an arbitrary constant.
(b) By definition, $y_T \triangleq c_1 e^{-bt}$ is a transient solution because (a) it is a solution of the corresponding homogeneous ODE and (b) $\lim_{t\to\infty} c_1 e^{-bt} = 0$ because the problem assumed $b > 0$. This agrees with our intuition that a transient solution should be that part of the solution that becomes insignificant eventually. But there is no steady-state solution in this example, because $w(t) \triangleq y(t) - y_T = \frac{bt-1}{b^2}$ is not bounded as $t \to \infty$, because we assumed that b is nonzero. ◯

Example 3.8

Find a formula for the solutions of $\dot{y} + p(t)y = f(t)$. (Such a formula may also be called a "closed form" solution.)

Method: Because we were not given formulas for $p(t)$ and $f(t)$, the best that we can hope for is a formula in terms of an integrating factor $\mu(t)$. Multiplying through $\dot{y}+p(t)y = f(t)$ by $\mu(t)$, we have

$$\mu(t)\dot{y} + p(t)\mu(t)y = \mu(t)f(t).$$

Because $\mu(t)$ satisfies $\dot{\mu}(t) = p(t)\mu(t)$, the product rule explains why this ODE is equivalent to

$$\frac{d}{dt}[\mu(t)y] = \mu(t)f(t).$$

So, either

$$y(t) = \frac{1}{\mu(t)}\left(c_1 + \int \mu(t)f(t)\,dt\right)$$

or

$$y(t) = \frac{1}{\mu(t)}\left(c_1 + \int_{t_0}^{t} \mu(s)f(s)\,ds\right),$$

where t_0 is a constant, is a formula for all solutions of $\dot{y} + p(t)y = f(t)$, where c_1 is an arbitrary constant. ◯

3.1.4 Problems

In problems 1–4, find all solutions of the ODE.

1. $\dot{y} + y = e^{-2t}$
2. $t\dot{y} + y = e^{-2t}$
3. $t\dot{y} - 3y = t^4$
4. $\dot{y} - \frac{1}{t+1}y = (t^2 - t - 2)$
5. Suppose $y(t) = t^2 \ln(t)$ solves $t\dot{y} - 2y = t^2$. Solve the IVP $t\dot{y} - 2y = t^2$, $y(1) = 3$.

In problems 6–17, solve the IVP.

6. $\dot{y} + 3y = e^{-t}$, $y(0) = -1$
7. $t\dot{y} + 3y = \frac{e^{-2t}}{t^2}$, $y(1) = -1$
8. $x\frac{dy}{dx} - 3y = x^4$, $y(1) = -1$
9. $t\dot{y} + (t-1)y + t^2 = 0$, $y(1) = -2$
10. $t\dot{y} = -y - 3t$, $y(1) = 1$
11. $\dot{A} = 4 - \frac{6A}{100-2t}$, $A(0) = 10$
12. $\dot{y} + 3y = 5te^{-t}$, $y(0) = 1$
13. $\dot{y} = -3y + 2t$, $y(1) = 0$
14. $\dot{y} = 5t - \frac{3}{t}y$, $y(2) = -4$
15. $\dot{y} + \frac{2t}{1+t^2}y = \frac{1}{t}$, $y(2) = -1$
16. $t\dot{y} + y = 4$, $y(1) = 3$
17. $\dot{y} + \frac{t+1}{t}y = \frac{1}{t}e^{-2t}$, $y(1) = 0$

In problems 18–21, find the steady-state solution.

18. $\dot{y} + \frac{y}{t} = \cos t$, $y(\frac{\pi}{2}) = 0$
19. $\dot{y} = -y + \sin t$, $y(0) = 1$
20. $\dot{y} = -\frac{1}{2}y - 1 + 2\sin t$, $y(0) = 0$
21. $\dot{y} + y = \cos 2t$, $y(0) = 0$
22. Let α be an unspecified parameter. Find the solution of the IVP $\dot{y} - \alpha y = t$, $y(0) = 3$. Your final conclusion(s) should be in terms of α. [Caution: $\alpha = 0$ may be a special case.]
23. Let α be an unspecified positive parameter. For the IVP $\dot{y} + \alpha y = 2$, $y(0) = 1$,
    (a) Find the solution, in terms of α.
    (b) Find an approximate value(s) of α for which $y(1) = 2$. Because $y(1)$ is a function of α, setting 2 equal to $y(1)$ could give a difficult equation to be solved for α. Technology, for example, a graphing calculator, may be useful for that.
24. The loop current I in a series RL circuit with constant voltage source E_0 satisfies $L\dot{I} + RI = E_0$, by Kirchhoff's voltage law. Assume that R and L are, as usual, constants. Assume that initially there is no current in the circuit. Find the current

as a function of time, and find the steady-state solution. When will the current be $(1 - e^{-1})$ times the steady-state current? [This time is called the "rise time."]

Newton's law of cooling says that a hot object in cool surroundings will lose heat at a rate proportional to the difference between the temperature of the object and the medium that surrounds it. Assuming the specific heat is roughly constant during this process, the temperature of the object, T, satisfies an ODE of the form

$$\dot{T} = -\alpha(T - M),$$

where
M is the constant temperature of the medium
α is a positive constant

---

25. At 1:00 pm. a hot object was brought into a room whose temperature is kept at a constant 20°C. Suppose that the object's temperature was 250°C at 1:03 pm. and 200°C at 1:04 pm. What was the object's temperature at 1:00 pm.?
26. When a cake was removed from an oven, the cake's temperature was 360°F. Fifteen minutes later, its temperature was 80°F. The kitchen's temperature is a constant 65°F. When was the cake's temperature 100°F?
27. [CSI-ODE] Suppose a coroner found the temperature of a dead person was 34.8°C at 11 am. and was 34.3°C a half an hour later. The police noted the room's temperature seemed to be a constant 21.1°C. Give an estimate for the person's time of death, assuming that living people usually have a temperature of about 36.95°C.
28. The situation is basically the same as in Problem 3.1.4.27 except that we note that living people usually have a temperature that lies in the interval of about 36.6°C–37.2°C, that is, 36.95°C, is not completely reliable. Give an *interval* estimate for the person's time of death, for example, something such as "The person died sometime between about 9:05 am and about 9:55 am."
29. Suppose that a 0.5 kg falling object experiences an air resistance force whose magnitude in newtons is 4 times its speed in m/s. The object is released from rest, and assume $g \approx 9.81\,\text{m/s}^2$. Be careful to state a physical coordinate system for this situation and then write down an IVP for the velocity and solve it. What is the steady-state velocity?
30. [Redheffer and Port] Suppose that you're in a canoe that passed the finish line in a race at time $t = 0$. Although you stopped paddling, your canoe continues drifting forward, going a distance y_T by time $t = T$ and a total distance y_{2T} by time $t = 2T$. Assuming that the water exerts a resistive force proportional to the speed of the canoe but that there is no wind or water current, explain why if you wait long enough the position of the canoe will be arbitrarily close to the number

$$y_\infty = \frac{y_T^2}{2y_T - y_{2T}}.$$

31. Suppose that the number of acres occupied by a certain plant satisfies two assumptions: (i) Goats are consuming the plant at a rate of 10 acres per year, and (ii) in

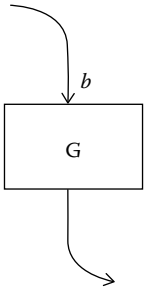


FIGURE 3.2
Problem 3.1.4.32.

the absence of goats, the acreage occupied by the plant would increase at a rate proportional to the current acreage.

(a) Write down an ODE to model this situation. Carefully define your variables.

(b) Find all solutions of the ODE that you wrote in part (a).

32. Glucose is dripping into the bloodstream of a patient at a constant rate of bg/m. At the same time, the patient's body uses up its glucose at a rate proportional to the amount of glucose in the bloodstream, with constant of proportionality k. An illustration of this situation is shown in Figure 3.2. We call this situation a **one-compartment model**.

Define your variables and set up an ODE for the amount of glucose in the patient's bloodstream.

Assuming that initially there is no glucose in the patient, find the steady-state amount of glucose in the patient's bloodstream, in terms of b and k. How long will it take for the amount of glucose in the bloodstream to reach $(1 - e^{-5})$ times the steady-state amount? Again, your conclusion should be in terms of k and b.

33. A rocket burns fuel to generate thrust and this also causes the rocket to continually decrease in mass. Newton's second law of motion says that $\frac{d}{dt}[mv] = \Sigma Forces$, where m is the mass of the rocket and v is the velocity of the rocket, assuming its motion is along a straight line. From this we can explain why the ODE describing the motion of the rocket motion is $-mg = m\dot{v} + u\dot{m}$, where u is a constant, and we assume the gravitational force is constant and there are no resistive forces. Assume $v(0) = v_0$, $g = 32$ ft/s^2, $m = m_0\left(1 - \frac{t}{200}\right)$, $0 \le t \le 190$, and m_0 is a constant. Find the velocity as a function of time. What is the velocity when the rocket stops burning, that is, when $t = 190$ s, assuming the mass of the payload is negligible?

34. Table 3.1 has hypothetical data for the temperature, T, of an object.

(a) Which of the following models is better for this data? Why?

Model #1: $\dot{T} = -\alpha T$

Model #2: $\dot{T} = -\alpha(T - M)$

TABLE 3.1

Choosing a Model

Time (s)	0	20	40	60	80	100	120	180	300
Temperature (°C)	140	80	50	45	35	33	32	30	28

(b) For the model you chose in part (a), estimate the numerical values of the physical parameters. Include the units in your conclusions.

Suppose we can find a formula for the indefinite integral for $\mu(t)$, but we can't find a formula for the indefinite integral $\int \mu(t)p(t)\,dt$. Then, in the method of integrating factor, instead of doing an indefinite integral of both sides of $\frac{d}{dt}\big[\,\mu\, y\,\big] = \mu(t)f(t)$, we could do a definite integral of both sides with respect to t. Note that $\frac{d}{dt}[z] = f(t)$ is equivalent to $z(t) = z(0) + \int_0^t f(s)\,ds$.

35. Solve the IVP $\dot{y} + 2ty = 1,\ y(0) = 3$.
36. Solve the IVP $y' = -3x^2y - 2x,\ y(1) = 0$, where $' = \frac{d}{dx}$.

Suppose there is no explicit formula for $\int p(t)dt$. That would seem to doom finding an integrating factor, but this is not true. We want to find an integrating factor, $\mu(t)$, that is, a function that satisfies the differential equation $\frac{d\mu}{dt} = \mu p(t)$. Separation of variables gives us

$$\frac{d\mu}{\mu} = p(t)\,dt.$$

Instead of doing an indefinite integration of both sides of this, do a definite integration: For example,

$$\int_1^\mu \frac{d\nu}{\nu} = \int_0^t p(s)\,ds,$$

that is,

$$\ln|\mu| - \ln|1| = \int_0^t p(s)\,ds.$$

Raise e to both sides of

$$\ln|\mu| - 0 = \int_0^t p(s)\,ds$$

to get $|\mu| = e^{\int_0^t p(s)ds}$; hence, $\pm\mu = e^{\int_0^t p(s)ds}$. Since we need only one integrating factor, let

$$\mu = e^{\int_0^t p(s)ds}.$$

37. Find all solutions of the ODE $\dot{y} + e^{t^2}y = 1$.

3.2 Separable and Exact ODEs

3.2.1 Separable ODEs

An ODE that can be written in the form

$$\frac{dy}{dt} = f(t)g(y), \tag{3.11}$$

is called **separable**. Usually, we can use integrals to find most of the solutions of an ODE (3.11).

Example 3.9

Find all solutions of $\dfrac{dy}{dt} = \dfrac{2ty^2}{t^2+1}$.

Method: Recall from Calculus I that the differential dy is defined by $dy \triangleq \dfrac{dy}{dt}\,dt$. It follows that $\dfrac{dy}{dt}$ can be written as the ratio of differentials:

$$\frac{dy}{dt} = \frac{dy}{dt}.$$

So, we can rewrite the ODE as

$$\frac{dy}{dt} = y^2\,\frac{2t}{t^2+1}.$$

If $y \neq 0$, divide both sides by y^2 and multiply both sides by dt to get

$$\frac{dy}{y^2} = \frac{2t}{t^2+1}dt.$$

Now we see why it is called the ODE "separable": All of the y dependence, including the dy, is on one side, and all of the t dependence, including the dt, is on the other side of the equation. [By the way, the use of differentials is a historical reason why we call this subject "differential" equations rather than "derivative" equations.]

We will see later that it can be justified to integrate both sides of this differential equation:

$$\int \frac{dy}{y^2} = \int \frac{2t}{t^2+1}\,dt.$$

This gives

$$-y^{-1} + c = \ln|t^2+1| + \tilde{c},$$

that is,

$$y = \frac{-1}{\ln(t^2+1) + c_1}, \text{where} -\infty < c_1 < \infty.$$

Even though there appeared to be two arbitrary constants c and $\tilde{c}$, they can be combined into a single arbitrary constant, $c_1 = \tilde{c} - c$. In future problems, we will use a single

arbitrary constant rather than having to "reinvent the wheel" in the future by combining two arbitrary constants each time.

But did we find *all* of the solutions? No. Recall that one of the algebraic steps was enabled by "If $y \neq 0$..." *What if* $y = 0$? Then we have to start from scratch and find another solution technique. But this turns out to be even easier than the many steps of algebra and calculus we had to do to get the aforementioned solutions involving c_1.

If $y = 0$, the right-hand side (RHS) of the differential equation becomes $\dot{y} = 0^2 \frac{2t}{t^2+1} \equiv 0$, and that is solvable by letting $y(t) \equiv 0$, that is, $y(t)$ is *identically equal* to 0, that is, $y(t) = 0$ for all t. In fact, this completes our work, that is, we have found *all* of the solutions:

$$\left\{ \begin{array}{c} y(t) \equiv 0 \\ \text{and} \\ y(t) = -\left(\ln|t^2+1| + c_1\right)^{-1}, \text{where } c_1 \text{ is an arbitrary constant} \end{array} \right\}. \bigcirc \qquad (3.12)$$

We will see why it makes sense to integrate both sides of $\dfrac{dy}{g(y)} = f(t)\,dt$ to get solutions satisfying

$$\int \frac{dy}{g(y)} = \int f(t)\,dt.$$

It seems very strange to equate two integrals, one integrating with respect to y and the other integrating with respect to t, but we will explain why it makes sense:

Theorem 3.2

Suppose $y(t)$ is a solution of (3.11), that is, $\dot{y} = f(t)g(y)$, on some open interval $I = (a, b)$ and $g(y(t)) \neq 0$ for *all* t in I. It follows that $\displaystyle\int \frac{dy}{g(y)} = \int f(t)\,dt$.

Why? Because $g(y(t)) \neq 0$ for all t in I, divide $\dfrac{dy}{dt}(t) = f(t)g(y(t))$ through by $g(y(t))$ to get $\dfrac{1}{g(y(t))}\dfrac{dy}{dt}(t) = f(t)$. Next, the indefinite integral with respect to t of both sides gives

$$\int \frac{1}{g(y(t))}\frac{dy}{dt}(t)dt = \int f(t)\,dt.$$

According to the method of substitution, the LHS of the preceding equation is

$$\int \frac{1}{g(y(t))}\frac{dy}{dt}(t)\,dt = \int \frac{1}{g(y)}dy,$$

so the result follows. □

Theorem 3.2 does not address the case where a solution $y(t)$ has $g(y(t)) = 0$ for some t in I. As we saw in Example 3.9, we can have *constant solution(s)* that is not studied in

Theorem 3.2. In general, if for some constant c, $y(t) \equiv c$ solves ODE (3.11), then the ODE says that $0 \equiv \dfrac{dy}{dt} = f(t)g(y(t)) \equiv f(t)g(c)$, for all t in I. Unless $f(t) \equiv 0$ on interval I, we need to have $g(c) = 0$. [And if $f(t) \equiv 0$ on I, then our differential equation is $\dfrac{dy}{dt} = f(t)g(y(t)) \equiv 0 \cdot g(y(t)) \equiv 0$ on I, which again says that $y(t) \equiv$ *constant* on I.]

Example 3.10

Find all constant solutions of $\dot{y} = 2t(y^2 - 3y)$.

Method: If $y(t) \equiv c$ is a constant solution, then $0 \equiv \dot{y} = 2t(c^2 - 3c)$. It follows that $0 = c^2 - 3c$, that is, $0 = c(c-3)$. The constant solutions are $y(t) \equiv 0$ and $y(t) \equiv 3$. ◯

3.2.2 Exact ODEs

Definition 3.7

An ODE in the form

$$M(t,y) + N(t,y)\frac{dy}{dt} = 0 \tag{3.13}$$

is called **exact** if there is a continuously differentiable function $\phi(t,y)$ satisfying

$$M(t,y) = \frac{\partial \phi}{\partial t} \quad \text{and} \quad N(t,y) = \frac{\partial \phi}{\partial y}. \tag{3.14}$$

Example 3.11

Verify that the ODE

$$2t\cos y - t^2 \sin y \frac{dy}{dt} = 0 \tag{3.15}$$

is exact.

Method: The ODE is exact because $\phi(t,y) \triangleq t^2 \cos y$ satisfies

$$2t\cos y = \frac{\partial}{\partial t}\big[\, t^2 \cos y \,\big] \text{ and } -t^2 \sin y = \frac{\partial}{\partial y}\big[\, t^2 \cos y \,\big]. \text{ ◯}$$

Recall that when taking the partial derivative with respect to t, we treat y as if it were a constant, and similarly, when taking the partial derivative with respect to y, we treat t as if it were a constant.

Is having such a function ϕ a curiosity, or is it useful in a general way?

Theorem 3.3

If ODE (3.13) is exact, then the curves $\phi(t,y) = C$ satisfy the ODE.

Why? First, here's a fact about partial derivatives: If $y = y(t)$ is differentiable, then

$$\frac{d}{dt}[\,\phi(t, y(t))\,] = \frac{\partial \phi}{\partial t}(t, y(t)) + \frac{\partial \phi}{\partial y}(t, y(t)) \cdot \frac{dy}{dt}(t). \tag{3.16}$$

This chain rule fact makes sense because, if we define $\triangle y \triangleq y(t+\triangle t) - y(t)$,

$$\begin{aligned}
&\lim_{\triangle t \to 0} \frac{\phi(t+\triangle t, y(t+\triangle t)) - \phi(t, y(t))}{\triangle t} \\
&\quad = \lim_{\triangle t \to 0} \frac{\phi(t+\triangle t, y(t+\triangle t)) - \phi(t, y(t+\triangle t))}{\triangle t} + \lim_{\triangle t \to 0} \frac{\phi(t, y(t+\triangle t)) - \phi(t, y(t))}{\triangle t} \\
&\quad = \lim_{\triangle t \to 0} \frac{\phi(t+\triangle t, y(t+\triangle t)) - \phi(t, y(t+\triangle t))}{\triangle t} \\
&\qquad + \left(\lim_{\triangle y \to 0} \frac{\phi(t, y(t)+\triangle y) - \phi(t, y(t))}{\triangle y} \right) \left(\lim_{\triangle t \to 0} \frac{\triangle y}{\triangle t} \right) \\
&\quad = \frac{\partial \phi}{\partial t}(t, y(t)) + \frac{\partial \phi}{\partial y}(t, y(t)) \cdot \frac{dy}{dt}(t).
\end{aligned}$$

So, using that fact and the exactness, we can rewrite the ODE as

$$0 = M(t, y(t)) + N(t, y(t)) \frac{dy}{dt}(t) = \frac{\partial \phi}{\partial t}(t, y(t)) + \frac{\partial \phi}{\partial y}(t, y(t)) \frac{dy}{dt}(t) = \frac{d}{dt}[\,\phi(t, y(t))].$$

Taking the indefinite integral of both sides with respect to t gives $C = \phi(t, y(t))$, where C is an arbitrary constant. □

For example, in Example 3.11, the curves

$$t^2 \cos y = C$$

satisfy ODE (3.15). Indeed, ODE (3.15) is also separable, and the method of separation of variables gives the same solution curves $t^2 \cos y = C$. In fact, as you will explain in Problem 3.2.4.22, all separable ODEs can be rewritten in a form that is exact.

If an ODE $M(t, y) + N(t, y)\frac{dy}{dt} = 0$ is exact, then Clairaut's theorem says we must have

$$\frac{\partial}{\partial y}\big[\,M(t, y)\,\big] = \frac{\partial}{\partial y}\left[\frac{\partial \phi}{\partial t}\right] = \frac{\partial^2 \phi}{\partial y \partial t} = \frac{\partial^2 \phi}{\partial t \partial y} = \frac{\partial}{\partial t}\left[\frac{\partial \phi}{\partial y}\right] = \frac{\partial}{\partial t}\big[\,N(t, y)\big].$$

Theorem 3.4

If ODE (3.13) is exact, then it must be true that

$$\frac{\partial}{\partial y}\big[\,M(t, y)\big] = \frac{\partial}{\partial t}\big[\,N(t, y)\big], \tag{3.17}$$

which is called the **exactness criterion**.

The theorem follows as earlier from Clairaut's theorem. The exactness criterion is a *necessary condition* because if ODE (3.13) is exact, then necessarily the ODE satisfies the exactness criterion. On the other hand, if M and N satisfy the exactness criterion in a nice region in the ty-plane, why is there a function ϕ as desired by the definition of "exact ODE?" An explanation for that would involve mathematics that is more advanced than we care to give here. But this result is related to "potential flow" for ideal fluids, existence of potentials in multivariable calculus, and existence of a "harmonic conjugate" in complex variable theory. So, even though an explanation involves more advanced mathematics, the *result* being explained may be of great significance to engineers.

Example 3.12

Decide whether the ODE as follows is exact, and if so,

(a) Find all solutions

(b) Find the solution passing through the point $(1, 2)$ in the ty-plane:

$$y^3 - 2t^2 + \left(3ty^2 + y\right)\frac{dy}{dt} = 0. \tag{3.18}$$

Method: (a) First, we check the exactness criterion. [If this fails, then the ODE is not exact and there is nothing left to do in the problem because of the "if so." If the instructions said to solve under all circumstances, then we would have to find some other method!]

$$3y^2 = \frac{\partial}{\partial y}\left[\, y^3 - 2t^2 \,\right] = \frac{\partial}{\partial y}\left[\, M(t,y) \,\right] \stackrel{?}{=} \frac{\partial}{\partial t}\left[\, N(t,y) \,\right] = \frac{\partial}{\partial t}\left[\, 3ty^2 + y \,\right] = 3y^2,$$

so, yes, ODE (3.18) is exact.

We want to find a "potential function" $\phi(t,y)$ satisfying (3.14). So, we need to have

$$y^3 - 2t^2 = M(t,y) = \frac{\partial}{\partial t}\left[\, \phi(t,y) \,\right];$$

hence,

$$\phi(t,y) = \int \left(y^3 - 2t^2\right) \partial t = ty^3 - \frac{2}{3}t^3 + f(y),$$

where $f(y)$ is an arbitrary function of *only* y. Our symbol $\int ... \partial t$ is shorthand for the operation of anti-partial differentiation with respect to t.

The reason we have an arbitrary function $f(y)$ instead of an arbitrary constant is because $\frac{\partial}{\partial t}[f(y)] \equiv 0$. Note also that because $f(y)$ is a function of y alone, $\frac{\partial}{\partial y}[f(y)] = \frac{df}{dy}$.

From (3.14), $\phi(t,y)$ must also satisfy

$$\left(3ty^2 + y\right) = N(t,y) = \frac{\partial}{\partial y}[\phi(t,y)] = \frac{\partial}{\partial y}\left[ty^3 - \frac{2}{3}t^3 + f(y)\right] = 3ty^2 + \frac{df}{dy},$$

so

$$y = \frac{df}{dy}.$$

We have $f(y) = \frac{1}{2}y^2$; we could add an arbitrary constant, but it would turn out to be redundant because our solutions are the curves $\phi(t,y) = C$.

Putting everything together, we have that the solutions of ODE (3.18) are the curves

$$C = \phi(t,y) = ty^3 - \frac{2}{3}t^3 + \frac{1}{2}y^2,$$

where C is an arbitrary constant.

(b) Saying the curve passes through the point $(1,2)$ in the ty-plane means that at $t = 1$, $y = 2$. We can satisfy this initial condition by solving for C:

$$C = \left(ty^3 - \frac{2}{3}t^3 + \frac{1}{2}y^2 \right) \Big|_{(t,y)=(1,2)} = 8 - \frac{2}{3} + 2.$$

The solution of the IVP is the curve

$$\frac{28}{3} = ty^3 - \frac{2}{3}t^3 + \frac{1}{2}y^2. \bigcirc$$

While it would be nice to find an **explicit solution**, that is, y given as an explicit function of t, and while it would be possible by using formulas for the roots of a cubic equation, we will skip that here because the formula is a little complicated and is not very informative. Instead, we used *Mathematica*^{TM} to get a graph, by using the command

$$\text{ContourPlot}\left[\text{ty}^3 - \frac{2}{3}\text{t}^3 + \frac{1}{2}\text{y}^2, \{\text{t}, -4, 2\}, \{\text{y}, -10, 10\}, \text{Contours} \to \left\{\frac{28}{3}\right\}, \right.$$
$$\left. \text{PerformanceGoal} \to \text{"Quality"}, \ \text{ContourShading} \to \text{False} \right]$$

to get Figure 3.3. The lower branch of the curve is not part of the solution of the IVP because it doesn't pass through the point $(1,2)$. We also plotted a zoomed-out view, in Figure 3.3b, to feel more confident that the two branches don't connect in the ty-plane.

Caution: The solutions of an exact ODE are the curves $\phi(t,y) = C$, where C is an (arbitrary) constant, not the function $\phi(t,y)$.

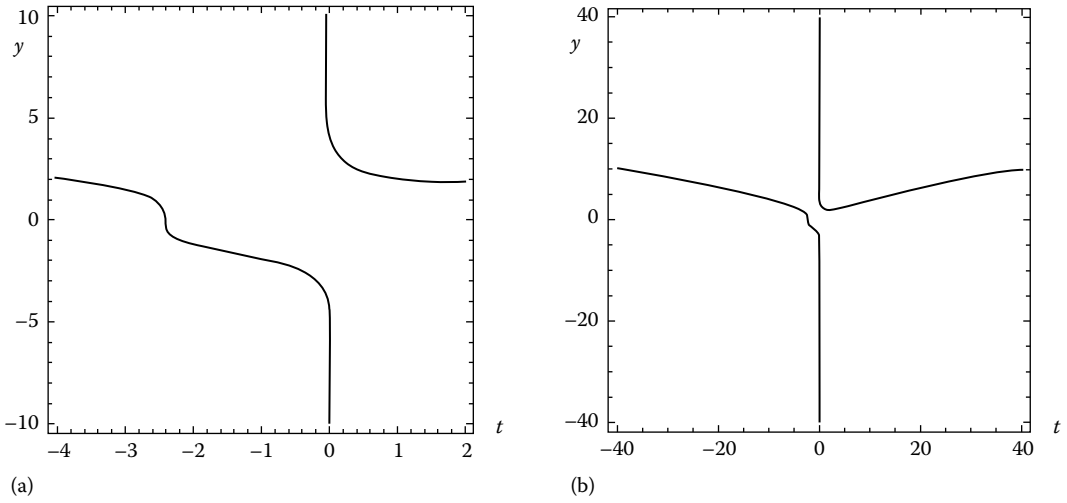


FIGURE 3.3
Example 3.12: (a) solutions curves and (b) zoomed out view of solution curves.

3.2.3 Existence of Solution(s) of an IVP

Unlike a system of linear algebraic equations, it is not obvious how to decide whether a given IVP has a solution or how to find the solution(s). We will explain some of the basic results on existence of solution(s) and present some examples. Unfortunately, the explanations for the theorems are more mathematically advanced than we choose to give, but we will refer the reader to other books or articles. Once again, even though an *explanation* involves a lot of pure mathematics, the *results* being explained are of great significance to engineers. While an engineer might use a numerical method (see Chapter 8) to approximate the solution of an IVP, it would be good to know that there *is* a solution to be approximated!

Theorem 3.5

(Peano's existence theorem) The IVP

$$\left\{\begin{array}{l} \dfrac{dy}{dt} = f(t,y) \\ \\ y(t_0) = y_0 \end{array}\right\} \tag{3.19}$$

has at least one solution as long as $f(t,y)$ is continuous on a closed rectangle

$$\mathcal{R}_{\alpha,\beta} \triangleq \{(t,y)\colon t_0 - \alpha \le t \le t_0 + \alpha,\ y_0 - \beta \le y \le y_0 + \beta\},$$

for some positive scalars α, β.

As far as it goes, this is a good result. However, it doesn't say how many solutions there are for the IVP. In science and engineering, we prefer that a mathematical model of a physical system should give only one prediction about future behavior. Also, the theorem does not tell us for how long a time interval the solution exists. If it turns out that the solution only exists for 10^{-40} s in the future, that would probably not be very useful for making predictions in a physical problem.

Example 3.13

For the IVP

$$\left\{\begin{array}{l} \dfrac{dy}{dt} = -\dfrac{3}{2}\,y^{1/3} \\ \\ y(1) = 0 \end{array}\right\}, \tag{3.20}$$

(a) explain why the method of separation of variables fails because, by itself, it produces no solution, and (b) explain why there are infinitely many solutions.

Method: (a) Separation of variables gives

$$\int \frac{dy}{y^{1/3}} = -\int \frac{3}{2}\,dt,$$

so it seems that the solutions are given by

$$\frac{3}{2}y^{2/3} = -\frac{3}{2}t + c,$$

where c is an arbitrary constant. Multiplying through by $\frac{2}{3}$ and renaming $\bar{c} = \frac{2}{3}c$, we have

$$y^{2/3}(t) = -t + \bar{c}.$$

The initial condition gives $0 = y(1) = -1 + \bar{c}$, so we have $\bar{c} = 1$. We get

$$y^{2/3}(t) = 1 - t. \tag{3.21}$$

Now, because $\big(y(t)\big)^{2/3} = \Big(\big(y(t)\big)^{1/3}\Big)^2 \geq 0$, the implicit solution given by (3.21) exists only for $t \leq 1$. But according to our definition in Section 3.1, a *solution* is a function $y(t)$ defined on an open interval I that satisfies the ODE on I and the IC $y(1) = 0$. So, a solution has to be defined on an open interval I containing $t = 1$ inside. Possible intervals I are (a, b), where $a < 1 < b$, or $(-\infty, b)$, where $b > 1$, or (a, ∞), where $a < 1$, or $(-\infty, \infty)$.

But our alleged solution doesn't exist for $t > 1$. A solution of the IVP has to exist on an *open* interval containing $t = 1$, so we reach a contradiction. The method of separation of variables, by itself, doesn't produce a solution!

This problem seems to be a bit "unfair" or "tricky." And it seems at first as if the difficulty is just that mathematicians messed up the definition of "solution." But the subject would be meaningless without a serious definition of "solution," and there is a very good reason to require solutions to exist both forward and backward in time: Engineering applications of physics should be able to predict what will happen in the future. Even "predicting" what happened in the past can be difficult.

But once you see the solution for part (b), it will appear to be more reasonable. Mathematics is like that sometimes.*

(b) The only problem with what we got from separation of variables, $y(t) = (1-t)^{3/2}$, is that it doesn't exist for $t > 1$. The trick is to find some way to "extend" the solution to the right of $t = 1$. Here's the inspiration: The initial condition is $y(1) = 0$, and $y(t) \equiv 0$ is a constant solution of this ODE. Let's define $y(t) = 0$ for $t > 1$, that is,

$$y(t) = \begin{cases} (1-t)^{3/2}, & t < 1 \\ 0, & t \geq 1 \end{cases}. \tag{3.22}$$

This is, indeed, an honest to goodness solution. Not only that, but (3.22) can be generalized to give infinitely many solutions, one for every value of the constant $\gamma \leq 1$:

$$y(t) = \begin{cases} (\gamma - t)^{3/2}, & t < \gamma \\ 0, & t \geq \gamma \end{cases}.$$

* Here's an apocryphal story about the American eccentric mathematician Norbert Weiner, who invented cybernetics: Once, during a lecture on the Fourier transform, a subject for which he made fundamental contributions, he was asked a question. He silently pondered it for a few minutes, abruptly left the lecture hall, and reappeared about an hour later. He said, "It's obvious," without further comment on it, and then resumed his lecture at the point where he had been interrupted. Apparently, after one of the greatest minds of the twentieth century had thought about it for an hour, he arrived at a conclusion but decided it was so trivial as to not be worth mentioning.

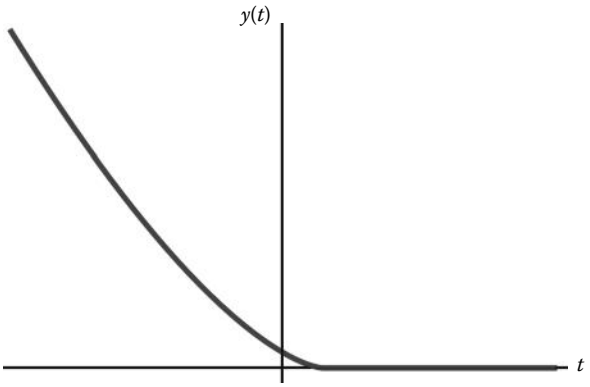


FIGURE 3.4
Example 3.13.

A typical solution graph is shown in Figure 3.4. This $y(t)$ is differentiable everywhere, with

$$\frac{dy}{dt}(t) = \begin{cases} -\frac{3}{2}(\gamma - t)^{1/2}, & t < \gamma \\ 0, & t \geq \gamma \end{cases},$$

which does equal the RHS of the ODE,

$$-\frac{3}{2}y(t)^{1/3} = -\frac{3}{2} \cdot \begin{cases} \left((t-\gamma)^{3/2}\right)^{1/3}, & t < \gamma \\ 0, & t \geq \gamma \end{cases} = \begin{cases} -\frac{3}{2} \cdot (t-\gamma)^{1/2}, & t < \gamma \\ 0, & t \geq \gamma \end{cases},$$

for all t.

In addition, there is a constant solution $y(t) \equiv 0$. ◯

Theorem 3.6

(Picard's existence and uniqueness theorem) Suppose that both $f(t, y)$ and $\frac{\partial f}{\partial y}(t, y)$ are continuous on a closed rectangle $\mathcal{R}_{\alpha,\beta}$, for some positive scalars α, β. Then there is an open time interval containing t_0 on which IVP (3.19) has exactly one solution.

This version of Picard's existence theorem addresses the "future predictions" issue but at the cost of demanding more "smoothness" of the function $f(t, y)$ on the RHS of the ODE, compared to Theorem 3.5.

Theorem 3.7

(Picard's theorem with interval of existence) Suppose that both $f(t, y)$ and $\frac{\partial f}{\partial y}(t, y)$ are continuous on a closed rectangle $\mathcal{R}_{\alpha,\beta} \triangleq \{(t, y) : t_0 - \alpha \leq t \leq t_0 + \alpha,\ y_0 - \beta \leq y \leq y_0 + \beta\}$, for some positive scalars α, β and where t_0 and y_0 are the same as in the initial condition

$y(t_0) = y_0$. Suppose that there are positive constants M and K such that for all (t, y) in $\mathcal{R}_{\alpha,\beta}$, we have

$$|f(t,y)| \le M \text{ and } \left|\frac{\partial f}{\partial y}(t,y)\right| \le K.$$

If we choose $\bar{\alpha}$ and $\bar{\beta}$ sufficiently small that

$$0 < \bar{\alpha} \le \alpha, 0 < \bar{\beta} \le \beta, M\bar{\alpha} \le \bar{\beta}, \text{ and } K\bar{\alpha} < 1,$$

then IVP (3.19) has exactly one solution on the time interval $I_{\bar{\alpha}} \triangleq [t_0 - \bar{\alpha}, t_0 + \bar{\alpha}]$ and the points $\big(t, y(t)\big)$ remain in the closed rectangle $\mathcal{R}_{\bar{\alpha},\bar{\beta}}$ for all t in the interval $I_{\bar{\alpha}}$.

Theorem 3.7 adds more information beyond the result of Theorem 3.6 by giving a specific time interval of existence. Most books don't explicitly state Theorem 3.7 but do include that information in their explanations for Theorem 3.6. Theorem 3.7 is illustrated in Figure 3.5.

The condition that $M\bar{\alpha} \le \bar{\beta}$ has a good physical interpretation: $\left|\frac{dy}{dt}\right| = |f\big(t, y(t)\big)|$ is the speed of an object whose position, $y(t)$, satisfies ODE (3.19), so M is an upper bound on that speed. So, $M\bar{\alpha} \le \bar{\beta}$ says, by distance = speed $\times$ time, that the object can't go further away from $y(t_0)$ than a distance of $\bar{\beta}$, that is, $|y(t) - y_0| \le \bar{\beta}$. So, $M\bar{\alpha} \le \bar{\beta}$ implies that the graph of $y(t)$ versus t stays inside the rectangle $\mathcal{R}_{\bar{\alpha},\bar{\beta}}$.

Both Peano's existence Theorem 3.5 and Picard's existence and uniqueness Theorem 3.6 have what are known as "sufficient conditions." What this means is that, for example, in Peano's theorem, "*if* $f(t, y)$ is continuous...*then* the IVP has a solution" says that *if* we can explain why $f(t, y)$ is continuous, *then* that will be *sufficient* to get the conclusion that the IVP has a solution.

But a sufficient condition may not be "necessary."

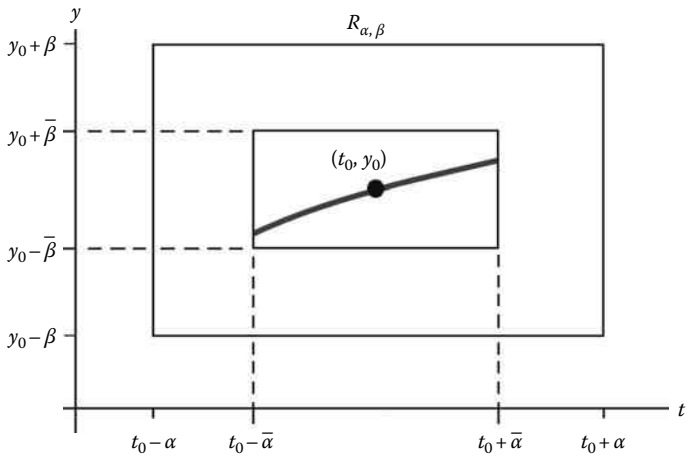


FIGURE 3.5
Picard's Theorem 3.7.

Example 3.14

Explain why the IVP $\dot{y} = \dfrac{y}{t}$, $y(0) = 0$ has a solution even though $f(t, y)$ is not continuous.

Method: In fact, separation of variables produces infinitely many solutions $y(t) = Kt$, for any constant K, even though $f(t, y) = \dfrac{y}{t}$ is continuous on no rectangle that contains $(t_0, y_0) = (0, 0)$ strictly inside. ○

Learn More About It

To see explanations of an improved version of Picard's existence Theorem 3.6, hence also an explanation for Theorem 3.7, see *An Introduction to Ordinary Differential Equations*, Earl A. Coddington, Dover publications, 1989, © 1961. For a more advanced point of view of Peano's existence theorem, see *Ordinary Differential Equations*, Jack K. Hale, Robert E. Krieger Publishing Company, © 1980, or *Metric Spaces*, E. T. Copson, Cambridge University Press, © 1988.

3.2.4 Problems

1. Solve the ODE $\left(t - \dfrac{t^2}{2y} + e^{-3y}\right)\dot{y} = e^{2t} - y + t\ln(y)$.
2. Find the solution of $\dfrac{dy}{dt} = \dfrac{1}{2(1+t^2)y}$ that passes through $(-1, 1)$ in the (t, y)-plane.
3. For the ODE $\dot{y} = \dfrac{t(y-1)}{t^2+1}$,
   (a) Find the solution, $y_2(t)$, that passes through the point $(t, y) = (1, 3)$.
   (b) Find the solution, $y_1(t)$, that passes through the point $(t, y) = (1, 1)$.
   (c) Graph $y_1(t)$ and $y_2(t)$ on the same set of axes and label which one is which.

The amount, A, of a radioactive element decays at a rate proportional to the amount remaining, so $\dot{A} = -\alpha A$ where α is a constant, called the **decay constant**. The **half-life**, denoted by $t^\star$, of a radioactive substance is the time it takes for its amount to be reduced to one-half of the initial amount.

4. Find a formula for $t^\star$ in terms of the constant α.
5. The wood of an Egyptian sarcophagus (burial case) is found to contain 63% of the carbon-14 that would be in a present-day sample. What is the approximate age of the sarcophagus, assuming that the half-life of carbon-14 is about 5730 years?
6. A certain isotope of radium decays radioactively. After 44.5 years, 2% of it has decayed. What is its half-life?

The effect of uncertainty in initial data:
In real life, we cannot measure quantities with absolute precision.

The goal of problems 8–10 is to explore how this fact of life can alter, sometimes significantly, conclusions drawn from ODE models.

Suppose that the earth's human population, $P(t)$, satisfies the differential equation

$$(\star)\ \dot{P} = kP,$$

where k is a constant that measures the rate of growth of the population and t is the time, in years. (This model can give accurate predictions over a short period of time but is probably not good over a "long" time interval; that leaves open the question just how long is a long time interval!)

---

7. Suppose that the earth's population on July 1, 1980 was $4,473,000,000$ and that on July 1, 1987, it was $5,055,000,000$. Assuming that P satisfies model $(\star)$, when would the earth's human population reach 10 billion?

Suppose that in real life, we can't really measure P with absolute precision. If T is the time when the earth's population reaches 10 billion, then we would have a range of values for T. For each of problems 8–10, before starting your symbol- and number-crunching, use your common sense, assisted by sketching solution(s) for P versus t, to predict how the value of T should change due to the $\pm$ uncertainty in the data.

---

8. Suppose that the earth's population on July 1, 1980, was $(4,473,000,000 \pm 50,000,000)$ and that on July 1, 1987, it was $5,055,000,000$. Assuming that P satisfies model $(\star)$, when would the earth's human population reach 10 billion? Your answer should be an *interval* of time predictions. Check your answer using common sense: which should be bigger, the T corresponding to $P(0) = 4.523$ billion or to $P(0) = 4.423$ billion?
9. Suppose that the earth's population on July 1, 1980, was $(4,473,000,000 \pm 50,000,000)$ and that on July 1, 1987, it was $(5,055,000,000 \pm 50,000,000)$. Assuming that P satisfies model $(\star)$, when would the earth's human population reach 10 billion?
10. Suppose that the earth's population on July 1, 1980, was $(4,473,000,000 \pm 10,000,000)$ and that on July 1, 1987, it was $(5,055,000,000 \pm 10,000,000)$. Assuming that P satisfies model $(\star)$, when would the earth's human population reach 10 billion?
11. A particle moves on the y-axis in such a way that its velocity is proportional to the square of its distance from the origin. At time $t = 0$, the particle is located at $y = 2$. At time $t = 3$, the particle is located at $y = 4$.
    (a) Find the position of the particle as a function of time t.
    (b) At what time does the particle reach position $y = 8$?
    (c) At what time does the particle reach $y = 1000$?
    (d) How long does the particle "live?"
12. For ODE $2xy\frac{dy}{dx} + y^2 - 1 = 0$, find the solutions passing through the points (a) $(1, 0)$ and (b) $(0, 1)$ in the xy-plane.
13. Solve the IVP

$$\frac{dy}{dx} = \frac{2x + y}{3 + 3y^2 - x},\ y(0) = 1.$$

[Hint: First rewrite the ODE in the form $M(x, y) + N(x, y)\frac{dy}{dx} = 0$.]

14. Solve the IVP

$$\frac{dy}{dx} = \frac{-x + \cos y}{2 + x \sin y},\ y\left(\frac{\pi}{2}\right) = -\frac{\pi}{2}.$$

[Hint: First rewrite the ODE in the form $M(x,y) + N(x,y)\frac{dy}{dx} = 0$.]

15. Solve the ODE

$$\dot{y} = -\frac{\sin y + y\cos t - 4}{\sin t + t\cos y + y}.$$

[Hint: First rewrite the ODE in the form $M(t,y) + N(t,y)\frac{dy}{dt} = 0$.]

16. Solve the ODE $(1 + x\cos(xy))\frac{dy}{dx} + y\cos(xy) = 0$.
17. Here's a model for the pharmacokinetics of alcohol: Let $x = x(t)$ be the concentration of alcohol in a person, so $0 \le x \le 1$, t be the time in hours, and A, k be constants. The model has $x(t)$ satisfying the ODE $(\star)\quad \dot{x} = -\frac{kx}{A+x}$ after a person has stopped absorbing alcohol. Assume that initially the person's alcohol concentration is 0.024, three times the legal limit in many states.
    (a) How long does it take for the person's concentration to fall within the legal limit of 0.008, assuming $A = 0.005$ and $k = 0.01$?
    (b) Choose new values for the constants A, k and discuss how the conclusion changes. Hypothesize some personal characteristics, for example, gender, body mass index (BMI), and age that could affect the conclusions and thus could be modeled by the values of A, k.
    (c) What term(s) would you add to ODE $(\star)$ to model the situation where a person is still absorbing alcohol at a constant rate?
18. Solve the ODE $(x\sin(xy) - 2y + \cos x)\frac{dy}{dx} = -y\sin(xy) + x + y\sin x$. [Hint: First rewrite the ODE in the form $M(x,y) + N(x,y)\frac{dy}{dx} = 0$.]
19. Find two distinct solutions of the IVP $\dot{y} = \sqrt{y-1},\ y(2) = 1$.
20. Find two distinct solutions of the IVP $\dot{y} = y^{1/5},\ y(2) = 0$.
21. Find two distinct solutions of the IVP $\dot{y} = y^{2/3},\ y(0) = 0$.
22. Explain why all separable ODEs can be rewritten in a form that is exact!
23. For the IVP

$$\left\{\begin{array}{l} \dfrac{dy}{dt} = y^2 \\ \\ y(0) = 3 \end{array}\right\},$$

    (a) Explain why the method of separation of variables produces a solution that exists only on a time interval of the form $-\infty < t < \delta$.
    (b) Explain why Picard's theorem guarantees the existence and uniqueness of a solution on an interval of the form $-\bar{\alpha} \le t \le \bar{\alpha}$.

(c) Compare the *largest* value of $\bar{\alpha}$ that Picard's theorem produces with the value of δ. Compare Picard's theorem theoretical guarantee of a time interval of existence with the actual time interval of existence.

24. Find an IVP of the form

$$\dot{y} = f(t, y),\ y(0) = 0,$$

other than that of Example 3.14, which has a solution even though $f(t, y)$ is not continuous on any rectangle that contains $(0, 0)$ strictly inside.

3.3 Second-Order Linear Homogeneous ODEs

Second-order, linear, constant coefficients homogeneous ordinary differential equations (LCCHODEs) have the standard form

$$\ddot{y} + p\dot{y} + qy = 0, \tag{3.23}$$

where $\dot{} = \dfrac{d}{dt}$ and "p" and "q" are constants. We will find the general solution by first guessing solutions and then turning that guessing process into a method.

3.3.1 Spring–Mass–Damper Systems

But first, a word from our "sponsors": Why should we care about (3.23)? For one thing, (3.23) is a model for vibrations basic to engineering. Figure 3.6 shows an unforced oscillator

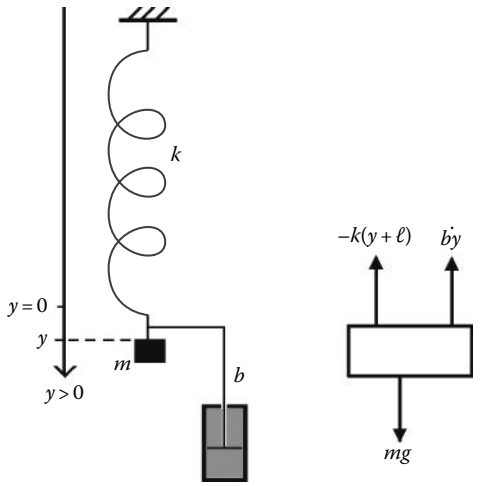


FIGURE 3.6
Vertical spring–mass–damper system.

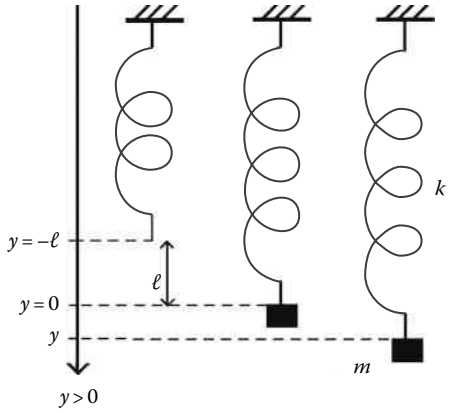


FIGURE 3.7
Vertical spring–mass system.

system with positive spring constant k, a positive constant mass m at the end of the spring, and a nonnegative constant damping coefficient b.

Let $y = y(t)$ be the displacement downward from the equilibrium position. Newton's second law of motion implies

$$m\ddot{y} + b\dot{y} + ky = 0. \tag{3.24}$$

Why? First, the spring–mass–damper system has vertical motion, so our model includes the force of gravity.

The basic differential equation is Newton's second law of motion

$$m\frac{dv}{dt} = \Sigma\, Forces.$$

As shown in Figure 3.7, there may be a damping device attached to the mass and spring.

There are three forces on the mass: (1) the downward force of gravity on the mass, (2) the spring's restoring force, which points toward the equilibrium position, and (3) (possibly) the damper's resisting force, which points opposite to the velocity vector. According to Hooke's law, the amount the spring is stretched by an object whose mass is m is $\ell = mg/k$.

We should explicitly state the coordinate system. We usually use the coordinate system where $y > 0$ is downward and $y = 0$ is the equilibrium position of the spring–mass–damper system. The free-body diagram has the spring force pointing up because the spring has been stretched beyond the equilibrium length. The damping force points up as long as the velocity points down.

From Newton's second law of motion,

$$m\frac{dv}{dt} = \Sigma\, Forces = F_{gravity} + F_{restoring} + F_{resisting} = mg - k(y + \ell) - bv.$$

Because $k\ell = mg$, that is, the force of gravity cancels the part of the spring force that opposes the stretch of the spring by a length ℓ, so we get

$$m\frac{dv}{dt} = -ky - b\dot{y},$$

hence (3.24).

If $y = 0$ is located where the end of the spring was *before* the damper and mass were attached, then $F_{restoring} = -ky$. In this coordinate system, the force of gravity is not canceled out, that is, the differential equation would be $m\ddot{y} + b\dot{y} + ky = mg$, which we won't know how to solve until Section 4.1. So it is convenient to use a coordinate system with $y = 0$ located at the equilibrium position of the spring–mass–damper system.

Note also that if $y > 0$ is upward, then $F_{gravity} = -mg$, not $+mg$, and $F_{restoring} = -k(y - \ell)$, so that again, when in equilibrium, the force of gravity cancels part of the spring force. So, for spring–mass–damper problems, the direction of $y > 0$ is not crucial, as long as we are consistent!

As to units of measurement, again, we should be consistent. For example, suppose that a problem says, "Assume a resistance whose magnitude in newtons is 20 times the magnitude of the instantaneous velocity v, in meters per second." It's good to "follow the lead" of the narrative of the problem. Here, the magnitude of the resistance force is in terms of velocity in meters per second, so it would be best to measure all lengths in meters. Since $|F_{resistive}| = b|v|$, where b is measured in units of newtons/(m/s), this says $b = 20$ if forces are measured in newtons, lengths are measured in meters, and time is measured in seconds.

Example 3.15

Assume that a mass of m kg is attached to the end of a vertical spring, and assume that a weight of (mg) N would stretch the spring by 3 m, where $g \approx 9.81$ m/s^2. Assume a damping force whose magnitude in newtons is 5 times the magnitude of the instantaneous velocity v, in meters per second. If the system is released from rest from a point $\frac{1}{3}$ m above the equilibrium position of the spring–mass–damper system, write down an IVP to model this physical problem.

Method: Assume the coordinate system is "$y > 0$ is downward and $y = 0$ is the equilibrium position of the object" and that y is measured in meters. At equilibrium, a weight of mg N stretches the spring $\ell = 3$ m beyond its natural length. So $mg = 3k$, where k is the spring constant; hence, $k = \frac{mg}{3}$. The third sentence tells us that $5|v| = |F_{resistive}| = b|v|$, so $b = 5$. It follows that the ODE is $m\ddot{y} + 5\dot{y} + \frac{mg}{3}y = 0$, that is, $\ddot{y} + \frac{5}{m}\dot{y} + \frac{g}{3}y = 0$.

As for the initial conditions, "...released from rest..." says that $\dot{y}(0) = 0$. Because "... the system is released from rest from a point $\frac{1}{3}$ m *above* the equilibrium position of the spring–mass–damper system..." and "$y > 0$ is downward," $y(0) = -\frac{1}{3}$.

To summarize, the IVP is

$$\left\{\begin{array}{l} \ddot{y} + \frac{5}{m}\dot{y} + \frac{g}{3}y = 0 \\ y(0) = -\frac{1}{3},\ \ \dot{y}(0) = 0 \end{array}\right\}. \bigcirc$$

From experience with the physical situation modeled by (3.24), we know that it is appropriate to specify also *two* pieces of initial data: the initial position, $y(t_0) = y_0$, and the initial velocity $v(t_0) = v_0$. Because $v(t) = \dot{y}(t)$, the second datum can be written as $\dot{y}(t_0) = \dot{y}_0$.

An **IVP** for a *second-order* ODE consists of the ODE along with the *two* initial conditions $y(t_0) = y_0, \dot{y}(t_0) = \dot{y}_0$.

3.3.2 Series *RLC* Circuit

A second example well known to engineers is a DC series "*RLC*" circuit with an inductor of strength L henrys, a capacitor of strength C farads, and a resistor of strength R ohms (Ω). Assume that R, L, and C are, as usual, constants. It has loop current I, as shown in Figure 3.8, satisfying

$$L\ddot{I} + R\dot{I} + \frac{1}{C}I = 0. \tag{3.25}$$

Why? The voltage drop across the resistor is RI, the voltage drop across the capacitor is $\frac{1}{C}q$, where q is the charge on the capacitor, and the voltage drop across the inductor is $L\dot{I}$. Kirchhoff's voltage law yields

$$L\frac{dI}{dt} + RI + \frac{1}{C}q = V_0.$$

Differentiate both sides of this equation with respect to t and use the fact that $I = \dot{q}$ to get (3.25).

This points to the power of mathematics: If we learn something about (3.23), then we will have learned something about both all unforced damped oscillators and all DC series electrical *RLC* circuits.

Figure 3.9 shows the shapes of some typical graphs of solutions of (3.24). The labels "overdamped," "critically damped," and "underdamped" will be explained shortly. By the way, some graphs are labeled by both overdamped and critically damped because those two cases can produce roughly the same graph, to within the discernment of the human visual system.

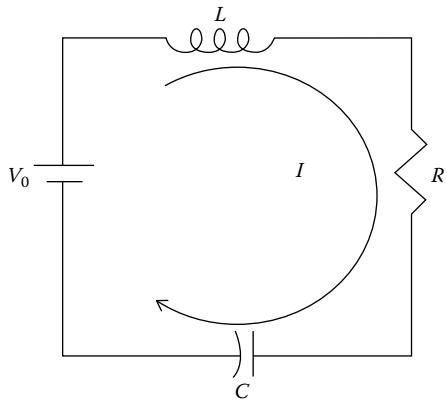


FIGURE 3.8
DC RLC series circuit.

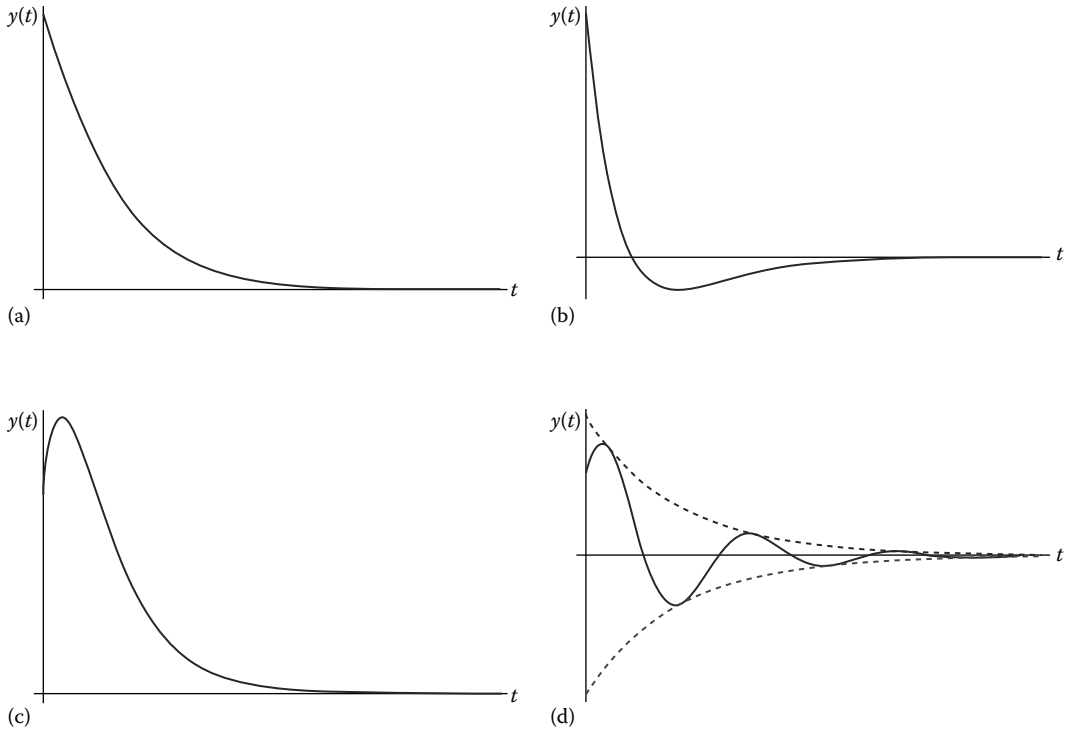


FIGURE 3.9
Shapes of solutions of mass–spring–damper system. (a) Critically damped or overdamped, (b) critically damped or overdamped, (c) critically damped or overdamped, and (d) underdamped.

To get started guessing solutions, recall that in Calculus I or II we studied the basic exponential growth or decay model

$$\dot{y} = -ay,$$

where "a" is a constant. Rewrite the ODE as

$$\dot{y} + ay = 0,$$

and guess solutions of the form

$$y = e^{st},$$

where later, the constant s will be chosen so as to be useful. Substitute $y = e^{st}$ and $\dot{y} = se^{st}$ into the ODE to get

$$0 = se^{st} + ae^{st} = (s + a)e^{st}.$$

So, if $s = -a$, we have a solution of the form $y = e^{-at}$. Moreover, we can check that for any constant C,

$$y(t) = Ce^{-at}$$

will also be a solution. You saw these solutions in a Calculus I or II course and in Problems 3.2.4.4 through 3.2.4.10.

For (3.23), let's try the same thing: Substitute $y = e^{st}$ into that ODE to get

$$0 = \dddot{y} + p\ddot{y} + qy = s^2e^{st} + pse^{st} + qe^{st} = (s^2 + ps + q)e^{st}.$$

The **characteristic equation**

$$s^2 + ps + q = 0$$

has two solutions, counting multiplicity, for the constant s:

$$s = -\frac{p}{2} \pm \frac{\sqrt{p^2 - 4q}}{2}.$$

Those solutions are also called the roots of the **characteristic polynomial**, $\mathcal{P}(s) \triangleq s^2+ps+q$.

For the vertical spring–mass–damper system modeled by (3.24), the characteristic equation is $ms^2+bs+k = 0$. This has two solutions for the constant s, counting multiplicity:

$$s = -\frac{b}{2m} \pm \frac{\sqrt{b^2 - 4mk}}{2m}.$$

Those solutions are also called the roots of the characteristic polynomial, $\mathcal{P}(s) \triangleq ms^2+bs+k$.

There are three cases for the roots of a quadratic polynomial in s:

- Distinct real roots, s_1, s_2, called the **overdamped** case for the mass–spring–damper ODE (3.24) [corresponding to $b^2 > 4mk$]
- Only one real distinct root, s_1, called the **critically damped** case for (3.24) [corresponding to $b^2 = 4mk$]
- A complex conjugate pair of roots, $\alpha \pm i\nu$, where $\nu \neq 0$, called the **underdamped** case for (3.24) [corresponding to $b^2 < 4mk$].

It is useful to define the **critical damping** by $b_{crit} = \sqrt{4mk}$, that is, the strength of damping that produces the critically damped case. Surprisingly, the critical damping definition is also useful in the context of forced vibrations of a spring–mass–damper system, as we will see in Section 4.2.

For an ODE, the time constant indicates how long it takes for a solution to decay to $\frac{1}{e}$ of its initial value. For the damped harmonic oscillator, all of the solutions are transient. The **time constant** τ for the ODE can be *defined* by

$$\tau \triangleq \frac{1}{r_{min}},$$

where r_{min} is the slowest "decay rate." For the overdamped case, $r_{min} = \frac{b-\sqrt{b^2-4mk}}{2m}$. For the underdamped and critically damped cases, $r_{min} = \frac{b}{2m}$.

Because each solution $y(t)$ may include many different decaying exponential functions, "weighted" by constants, we *can't* guarantee that $y(\tau) = \frac{1}{e}\, y(0)$. Nevertheless, for physical intuition, it is still useful to think of the time constant as being about how long it takes for the solution to decay in a standard way.

You may be wondering about the phrase "characteristic equation," which we first saw in Section 2.1, where it allowed us to find the eigenvalues of a matrix. In fact, as we will see in Section 5.2, the "characteristic equation" of a LCCHODE and the "characteristic equation" of a matrix are deeply related, so the solutions of those equations are related!

Example 3.16

Solve an ODE that models a spring–mass–damper system where the mass is 0.5 kg, that mass would stretch the spring by $\frac{9.81}{15}$ m, and the damping device exerts a force whose magnitude in newtons is 4 times the magnitude of the instantaneous velocity v, in meters per second. Assume the acceleration of gravity is 9.81 m/s^2.

Method: In the MKS system of units, we identify $m = 0.5$. Because of Hooke's Law, $mg = k\ell$, where ℓ is the stretch of the spring in equilibrium. So, $k = \frac{mg}{\ell} = \frac{(0.5)(9.81)}{9.81/15} = 7.5\,\text{N/m}$. The damping coefficient is $b = 4$, in N/(m/s). The ODE is $0.5\ddot{y} + 4\dot{y} + 7.5y = 0$, that is,

$$\ddot{y} + 8\dot{y} + 15y = 0. \tag{3.26}$$

The characteristic equation is $0 = s^2 + 8s + 15 = (s+3)(s+5)$, so $y = e^{-3t}$ and $y = e^{-5t}$ are both solutions of the ODE. Moreover, we can multiply each of those by arbitrary constants and then add to have what we hope are solutions:

$$y(t) = c_1 e^{-3t} + c_2 e^{-5t}. \tag{3.27}$$

Let's check that these are solutions: For any values of the constants c_1, c_2, substitute (3.27) into (3.26) to see that

$$\ddot{y} + 8\dot{y} + 15y = \frac{d^2}{dt^2}\left(c_1 e^{-3t} + c_2 e^{-5t}\right) + 8\frac{d}{dt}\left[c_1 e^{-3t} + c_2 e^{-5t}\right] + 15\left(c_1 e^{-3t} + c_2 e^{-5t}\right)$$

$$= (9 - 24 + 15)\, c_1 e^{-3t} + (25 - 40 + 15)\, c_2 e^{-5t} = 0 \cdot c_1 e^{-3t} + 0 \cdot c_2 e^{-5t} = 0. \bigcirc$$

Analogous to solving a homogeneous system of linear algebraic equations, we see that in Example 3.16, the solutions were in the form $y = c_1 y_1(t) + c_2 y_2(t)$, that is, the general linear combination of solutions $y_1(t), y_2(t)$.

Theorem 3.9 will tell us we have found all of the solutions in Example 3.16. Again, analogous to Definition 3.2 in Section 3.1 for first-order linear ODEs, we have

Definition 3.8

The **general solution** of a second-order linear homogeneous ODE

$$\ddot{y} + p(t)\dot{y} + q(t)y = 0 \tag{3.28}$$

has the form

$$y_h = c_1 y_1(t) + c_2 y_2(t)$$

if for every solution $y^*(t)$ of (3.28), there are values of constants c_1, c_2 giving $y^*(t) = c_1 y_1(t) + c_2 y_2(t)$. In this case, we call the set of functions $\{y_1(t), y_2(t)\}$ a **complete set of basic solutions**. Each of the functions $y_1(t), y_2(t)$ is called a **basic solution** of (3.28).

These definitions do not require that the coefficients p, q be constants.

Unlike for systems of linear algebraic equations, for second-order linear homogeneous ODEs we don't have a concept of "free variables" or "basic variables." But as we have begun to see using our guessing method, for *constant coefficients* linear homogeneous ODEs, there will be *formulas* for the complete set of basic solutions.

Theorem 3.8

(Existence and uniqueness) Suppose $p(t), q(t)$, and $f(t)$ are continuous on an open interval I. Then for all values of initial time t_0 inside I and initial data y_0 and $\dot{y}_0$, the IVP

$$\left\{ \begin{array}{l} \ddot{y} + p(t)\dot{y} + q(t)y = f(t) \\ \\ y(t_0) = y_0, \;\; \dot{y}(t_0) = \dot{y}_0 \end{array} \right\} \tag{3.29}$$

has exactly one solution $y(t)$ on the interval I.

Theorem 3.9

(Existence of a complete set of basic solutions of a linear homogeneous second-order ODE) Suppose $p(t)$ and $q(t)$ are continuous on an open interval I. Then the linear homogeneous second-order ODE

$$\ddot{y} + p(t)\dot{y} + q(t)y = 0 \tag{3.30}$$

has a complete set of basic solutions $\{y_1(t), y_2(t)\}$.

Why? Theorem 3.9 follows from Theorem 3.8: Pick *any* t_0 inside I. By Theorem 3.8, each of the two IVPs

$$\left\{ \begin{array}{l} \ddot{y}_1 + p(t)\dot{y}_1 + q(t)y_1 = 0 \\ y_1(t_0) = 1, \;\; \dot{y}_1(t_0) = 0 \end{array} \right\}, \qquad \left\{ \begin{array}{l} \ddot{y}_2 + p(t)\dot{y}_2 + q(t)y_2 = 0 \\ y_2(t_0) = 0, \;\; \dot{y}_2(t_0) = 1 \end{array} \right\}$$

has a solution. Suppose $y^*(t)$ is any solution of $\ddot{y} + p(t)\dot{y} + q(t)y = 0$ on I and define constants $c_1 = y^*(t_0)$ and $c_2 = \dot{y}^*(t_0)$. Let

$$z(t) = y^*(t) - \big(c_1 y_1(t) + c_2 y_2(t)\big).$$

It's easy to see that (1) $z(t)$ satisfies ODE $\ddot{y} + p(t)\dot{y} + q(t)y = 0$ on I, (2) $z(t_0) = 0$, and (3) $\dot{z}(t_0) = 0$. For example, to see that (2) is true, recall that $y_1(t_0) = 1$ and $y_2(t_0) = 0$, so $z(t_0) = y^*(t_0) - \big(c_1 y_1(t_0) + c_2 y_2(t_0)\big) = y^*(t_0) - \big(y^*(t_0) \cdot 1 + \dot{y^*}(t_0) \cdot 0\big) = y^*(t_0) - y^*(t_0) = 0$. But, the unique solution of the IVP

$$\begin{Bmatrix} \ddot{z} + p(t)\dot{z} + q(t)z = 0 \\ z(t_0) = 0, \;\; \dot{z}(t_0) = 0 \end{Bmatrix}$$

is easy to find: $z(t) \equiv 0$! So, $y^*(t) - \big(c_1 y_1(t) + c_2 y_2(t)\big) \equiv 0$, that is, $y^*(t) = c_1 y_1(t) + c_2 y_2(t)$. □

3.3.3 The Underdamped Case

Example 3.17

For the ODE

$$\ddot{y} + 2\dot{y} + 10y = 0. \tag{3.31}$$

find as many solutions as possible.

Method: The characteristic equation is $0 = s^2 + 2s + 10 = (s+1)^2 + 9$, so the roots of the characteristic polynomial are

$$s = -1 \pm i3.$$

This would seem to say that the solutions of the ODE are

$$\tilde{y}_1(t) = e^{(-1+i3)t}, \; \tilde{y}_2(t) = e^{(-1-i3)t}.$$

The presence of the symbol $i \triangleq \sqrt{-1}$ in a complex conjugate pair is reminiscent of the complex conjugate pair of eigenvalues and eigenvectors that we saw in Section 2.1. [If you're an electrical engineer, use j instead of i.]

In a sense, we have gotten ahead of ourselves, because the ˜'s over the y's will later signify that we are dealing with complex-valued quantities. So far, we don't know how these functions $\tilde{y}_1(t), \tilde{y}_2(t)$ behave. Also, since our original ODE had only *real* coefficients, 2 and 10, it is probably not appropriate to have solutions that involve the symbol i.

Euler's formula is

$$e^{i\theta} \triangleq \cos\theta + i\sin\theta.$$

We can take this to be a definition of a shorthand notation, or instead, we could define* the exponential function in terms of a power series and then derive† Euler's formula using the power series. We'll assume that the laws of exponents are true even for complex exponents

---

* $e^z \triangleq 1 + z + \frac{1}{2!}z^2 + \frac{1}{3!}z^3 + \cdots$

† $e^{i\theta} \triangleq 1 + i\theta + \frac{1}{2!}(i\theta)^2 + \frac{1}{3!}(i\theta)^3 + \cdots = \Big(1 - \frac{1}{2!}(\theta)^2 + \frac{1}{4!}(\theta)^4 \pm \cdots\Big) + i\Big(\theta - \frac{1}{3!}(\theta)^3 + \frac{1}{5!}(\theta)^5 \pm \ldots\Big) = \cos\theta + i\sin\theta,$ using the MacLaurin series for the cosine and sine functions. These series converge for all real θ.

Using Euler's formula and the law for addition of exponents, we can rewrite the solutions of Example 3.17:

$$\begin{aligned} y(t) &= \widetilde{c}_1\widetilde{y}_1(t) + \widetilde{c}_2\widetilde{y}_2(t) = \widetilde{c}_1 e^{(-1+i3)t} + \widetilde{c}_2 e^{(-1-i3)t} = \widetilde{c}_1 e^{-t+i3t} + \widetilde{c}_2 e^{-t-i3t} \\ &= \widetilde{c}_1 e^{-t}e^{i3t} + \widetilde{c}_2 e^{-t}e^{-i3t} = e^{-t}\big(\widetilde{c}_1\,(\cos 3t + i\sin 3t) + \widetilde{c}_2\,(\cos(-3t) + i\sin(-3t))\big) \\ &= e^{-t}\big(\widetilde{c}_1\,(\cos 3t + i\sin 3t) + \widetilde{c}_2\,(\cos 3t - i\sin 3t)\big) \\ &= e^{-t}\big((\widetilde{c}_1 + \widetilde{c}_2)\cos 3t + i\,(\widetilde{c}_1 - \widetilde{c}_2)\sin 3t\big) \\ &\triangleq e^{-t}\,(c_1\cos 3t + c_2\sin 3t)\,, \end{aligned}$$

where $c_1 \triangleq \widetilde{c}_1 + \widetilde{c}_2$ and $c_2 \triangleq i\,(\widetilde{c}_1 - \widetilde{c}_2)$. In the future, we will not have to go through the aforementioned process again.

So the solutions of Example 3.17 can be written in the form

$$y(t) = c_1 y_1(t) + c_2 y_2(t),$$

where

$$y_1(t) \triangleq e^{-t}\cos 3t = \mathcal{R}e(e^{(-1+i3)t}),\ y_2(t) \triangleq e^{-t}\sin 3t = \mathcal{I}m(e^{(-1+i3)t}).$$

Because $y_1(t)$ and $y_2(t)$ are real-valued functions, they don't have $\widetilde{}$'s over them. ◯

So that we don't have to "reinvent the wheel" in the future, we state

Theorem 3.10

Consider ODE $\ddot{y} + p\dot{y} + q = 0$. Suppose p, q are real constants and the characteristic polynomial has a complex conjugate pair of roots $s = \alpha \pm i\nu$, where α and ν are real and $\nu \neq 0$. Then all solutions of the ODE are given by

$$y(t) = c_1 e^{\alpha t}\cos \nu t + c_2 e^{\alpha t}\sin \nu t = e^{\alpha t}\big(c_1\cos \nu t + c_2 \sin \nu t\big),$$

where c_1, c_2 are arbitrary constants. So, $\{e^{\alpha t}\cos \nu t, e^{\alpha t}\sin \nu t\}$ is a complete set of basic solutions.

Why? The result follows from substituting each of the proposed basic solutions into the ODE to explain why they are indeed solutions. Theorem 3.13 will explain why $y_1(t) \triangleq e^{\alpha t}\cos \nu t$ and $y_2(t) \triangleq e^{\alpha t}\sin \nu t$ give a complete set of basic solutions. □

Example 3.18

For ODE (3.24), that is, $m\ddot{y} + b\dot{y} + ky = 0$, assume we have the underdamped case, that is, the characteristic polynomial has a complex conjugate pair of roots $s = \alpha \pm i\nu$ where α, ν are real and $\nu \neq 0$. Find the solutions of the ODE, in terms of the physical parameters m, b, k.

Method: The characteristic polynomial is $\mathcal{P}(s) = ms^2 + bs + k$, whose roots are

$$s = -\frac{b}{2m} \pm \frac{\sqrt{b^2 - 4mk}}{2m}\,.$$

But, to be in the underdamped case, we need to have $b^2 - 4mk < 0$, so we can rewrite

$$\sqrt{b^2 - 4mk} = \sqrt{-(4mk - b^2)} = i\sqrt{4mk - b^2}\,.$$

The roots of our characteristic polynomial are

$$s = \alpha \pm i\nu \triangleq -\frac{b}{2m} \pm i\,\frac{\sqrt{4mk - b^2}}{2m}\,.$$

By Theorem 3.10, the solutions in the underdamped case are given by

$$y(t) = e^{-bt/(2m)}\big(c_1 \cos \nu t + c_2 \sin \nu t\big),$$

where we define the **quasi-frequency** by

$$\nu \triangleq \frac{\sqrt{4mk - b^2}}{2m} = \sqrt{\frac{k}{m} - \left(\frac{b}{2m}\right)^2} = \sqrt{\omega_0^2 - \left(\frac{b}{2m}\right)^2},$$

and we define the "undamped natural frequency" to be $\omega_0 \triangleq \sqrt{\frac{k}{m}}$. ○

Surprisingly, the quantity defined to be the quasi-frequency reappears when studying the "frequency response" of forced vibrations of a spring–mass–damper system, as we will see in Section 4.2.

3.3.4 The Amplitude and Phase Form

Example 3.19

For the **undamped** oscillator ODE

$$m\ddot{y} + ky = 0, \tag{3.32}$$

(a) Find all solutions, and

(b) Rewrite them in the **amplitude and phase form**

$$y(t) = A\cos(\omega_0 t - \delta), \tag{3.33}$$

where the **amplitude** is A, a nonnegative constant; the **natural frequency** is ω_0, a positive constant; and the **phase** is δ, a constant with $-\pi < \delta \leq \pi$.

Method: (a) The characteristic polynomial is $\mathcal{P}(s) = ms^2 + k$, so its roots are $s = \pm i\omega_0$, where

$$\omega_0 = \frac{\sqrt{4mk}}{2m} = \sqrt{\frac{k}{m}}\,. \tag{3.34}$$

According to Theorem 3.10, the solutions of the ODE are

$$y(t) = e^{0\cdot t}\big(c_1 \cos \omega_0 t + c_2 \sin \omega_0 t\big) = c_1 \cos \omega_0 t + c_2 \sin \omega_0 t,$$

where c_1, c_2 are arbitrary constants.

(b) We equate the form of the solutions we have with the desired, new form (3.33) and use the trigonometric identity for the cosine of a difference:

$$y(t) = \underline{c_1} \cos \omega_0 t + \underset{\sim}{c_2} \sin \omega_0 t = A\cos(\omega_0 t - \delta) = A\big(\cos \omega_0 t \cos \delta + \sin \omega_0 t \sin \delta\big)$$

$$= \underline{A\cos\delta} \cos \omega_0 t + \underset{\sim}{\underline{A\sin\delta}} \sin \omega_0 t.$$

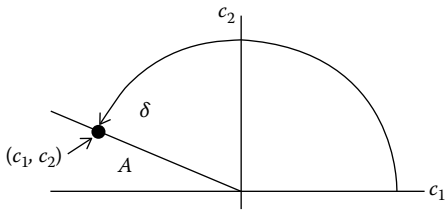


FIGURE 3.10
Amplitude and phase picture.

So, we need

$$c_1 = A\cos(\delta) \text{ and } c_2 = A\sin(\delta). \tag{3.35}$$

We recognize this as both how the trigonometric functions are defined and how polar coordinates are defined. A picture, such as Figure 3.10, "says a thousand words."*

We have

$$A = \sqrt{c_1^2 + c_2^2} \quad \text{and} \quad \tan(\delta) = \frac{c_2}{c_1}, \tag{3.36}$$

where the quadrant in which δ is found is the quadrant in which the point (c_1, c_2) is found. For future reference, we note that

$$\delta = \left\{ \begin{array}{rl} \arctan(c_2/c_1), & \text{if } c_1 > 0 \\ \pi + \arctan(c_2/c_1), & \text{if } c_1 < 0 \\ \frac{\pi}{2}, & \text{if } c_1 = 0 \text{ and } c_2 > 0 \\ -\frac{\pi}{2}, & \text{if } c_1 = 0 \text{ and } c_2 < 0 \\ D.N.E., & \text{if } (c_1, c_2) = (0,0) \end{array} \right\}. \bigcirc \tag{3.37}$$

Example 3.20

Find the exact solution of the IVP

$$\left\{ \begin{array}{c} \ddot{y} + \dot{y} + y = 0 \\ y(0) = -5,\ \dot{y}(0) = 3 \end{array} \right\}. \tag{3.38}$$

and express the solution in the exact form of

$$y(t) = Ae^{\alpha t}\cos(\nu t - \delta), \tag{3.39}$$

where $A, \nu > 0$ and $-\pi < \delta \leq \pi$.

Method: The characteristic polynomial $s^2 + s + 1$ has roots

$$s = \frac{-1 \pm \sqrt{1^2 - 4 \cdot 1 \cdot 1}}{2 \cdot 1} = -\frac{1}{2} \pm i\frac{\sqrt{3}}{2},$$

* "The mathematician Fred Almgren's son Robert is also a mathematician. He says, 'A lot of fathers give life advice to their sons. My father used to tell me, 'Son, when you're in doubt, when you don't know which way to turn or what to do, I want you to remember two things. First, draw a picture. Second, integrate by parts," from p. 182 of *Mathematical Aprocrypha*, by Steven George Krantz.

so the solutions of the ODE are

$$y(t) = c_1 e^{-t/2} \cos\left(\frac{\sqrt{3}\,t}{2}\right) + c_2 e^{-t/2} \sin\left(\frac{\sqrt{3}\,t}{2}\right).$$

The first IC requires $-5 = y(0) = c_1 \cdot 1 + c_2 \cdot 0 = c_1$, so

$$y(t) = -5e^{-t/2} \cos\left(\frac{\sqrt{3}\,t}{2}\right) + c_2 e^{-t/2} \sin\left(\frac{\sqrt{3}\,t}{2}\right).$$

Before applying the second IC, we calculate

$$\begin{aligned}\dot{y}(t) = \frac{5}{2}\, e^{-t/2}\cos\left(\frac{\sqrt{3}\,t}{2}\right) + \frac{5\sqrt{3}}{2}\, e^{-t/2} \sin\left(\frac{\sqrt{3}\,t}{2}\right) - \frac{1}{2}\, c_2 e^{-t/2}\sin\left(\frac{\sqrt{3}\,t}{2}\right) \\ + \frac{\sqrt{3}}{2}\, c_2 e^{-t/2}\cos\left(\frac{\sqrt{3}\,t}{2}\right).\end{aligned}$$

The second IC requires $3 = \dot{y}(0) = \frac{5}{2} + \frac{\sqrt{3}}{2}\, c_2$, so $c_2 = \frac{1}{\sqrt{3}}$. The solution to the IVP is

$$\begin{aligned}y(t) &= -5e^{-t/2} \cos\left(\frac{\sqrt{3}\,t}{2}\right) + \frac{1}{\sqrt{3}} e^{-t/2} \sin\left(\frac{\sqrt{3}\,t}{2}\right) \\ &= e^{-t/2}\left(5\cos\left(\frac{\sqrt{3}\,t}{2}\right) + \frac{1}{\sqrt{3}} \sin\left(\frac{\sqrt{3}\,t}{2}\right)\right).\end{aligned}$$

The two terms in the parentheses can be combined into *amplitude and phase form*, with the decay factor, $e^{-t/2}$ coming along for the ride:

$$-5\cos\left(\frac{\sqrt{3}\,t}{2}\right) + \frac{1}{\sqrt{3}} \sin\left(\frac{\sqrt{3}\,t}{2}\right) = A\cos(\nu t - \delta),$$

so we must choose $\nu = \frac{\sqrt{3}}{2}$. A and δ are found by drawing the picture in Figure 3.11 and using the formulas for amplitude and phase: $A = \sqrt{(-5)^2 + \left(\frac{1}{\sqrt{3}}\right)^2} = \sqrt{\frac{76}{3}}$ and $\tan(\delta) = \frac{1/\sqrt{3}}{-5}$, with $c_1 < 0$, so $\delta = \pi + \arctan\left(\frac{-1}{5\sqrt{3}}\right)$. Using $\arctan(-x) = -\arctan(x)$, the exact solution, in the desired final form, is

$$y(t) = \sqrt{\frac{76}{3}}\, e^{-t/2} \cos\left(\frac{\sqrt{3}}{2}\, t - \pi + \arctan\left(\frac{1}{5\sqrt{3}}\right)\right). \bigcirc$$

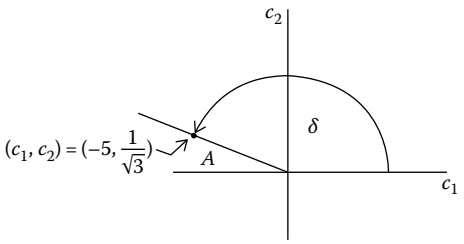


FIGURE 3.11
Amplitude and phase in Example 3.20.

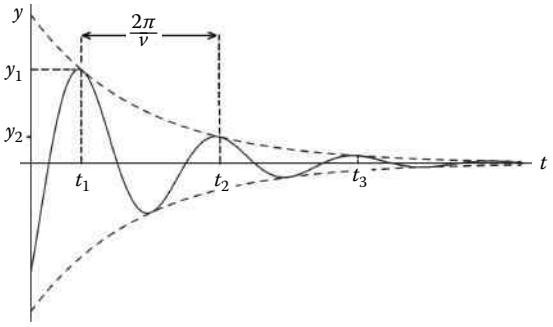


FIGURE 3.12
Typical underdamped solution.

3.3.5 Figures of Merit in Graphs of Underdamped Solutions

In the underdamped case, solutions have the form $y(t) = Ae^{\alpha t}\cos(\nu t - \delta)$. Think of this as a decaying, time-dependent amplitude, $Ae^{\alpha t}$, multiplying a sinusoidal function, $\cos(\nu t - \delta)$. A typical solution graph is shown in Figure 3.12, where the **envelope** consists of the two dashed curves $y = \pm e^{\alpha t}$: We can explain [see Problem 3.3.8.47] why the relative (also known as "local") maxima occur at times $t = t_1, t_2, t_3, \ldots$, which occur every $\frac{2\pi}{\nu}$ units of time; this explains why we refer to ν as the quasi-*frequency*. Even though the relative maxima do *not* occur exactly at the times when $\cos(\nu t - \delta) = 1$, we have $\cos(\nu t_1 - \delta) = \cos(\nu t_2 - \delta) = \cos(\nu t_3 - \delta) = \cdots$. Because of this, we have

$$\frac{y(t_2)}{y(t_1)} = \frac{y(t_3)}{y(t_2)} \cdots = \frac{Ae^{\alpha t_2}\cos(\nu t_2 - \delta)}{Ae^{\alpha t_1}\cos(\nu t_1 - \delta)} = e^{\alpha(t_2 - t_1)} = e^{2\pi\alpha/\nu}.$$

The **logarithmic decrement** is defined to be $D \triangleq \ln\big(y(t_2)/y(t_1)\big) = \dfrac{2\pi\alpha}{\nu}$, so

$$\nu = \frac{2\pi\alpha}{D}.$$

Example 3.21

Assuming mass $m = 1$, find the damping coefficient and the spring constant that can produce the solution graphed in Figure 3.13.

Method: First, for an ODE $\ddot{y} + b\dot{y} + ky = 0$, the roots of the characteristic polynomial $s^2 + bs + k$ are

$$s = -\frac{b}{2} \pm \frac{\sqrt{b^2 - 4k}}{2}.$$

In order to have infinitely many oscillations, albeit of decaying amplitude, our ODE must be in the underdamped case. So, as in Example 3.18, the roots of the characteristic polynomial have to be

$$s = \alpha \pm i\nu.$$

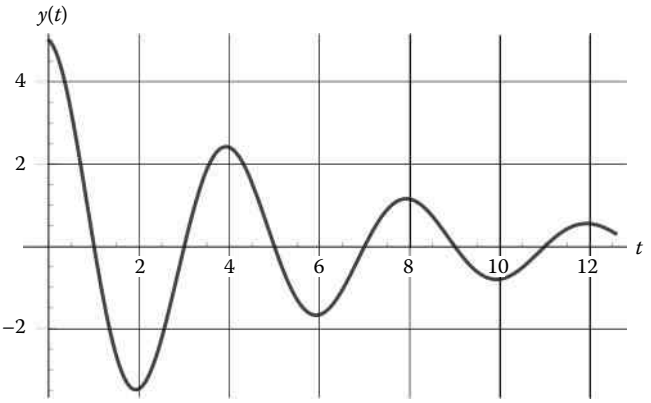


FIGURE 3.13
Example 3.21.

Here, $m = 1$ yields

$$\alpha = -\frac{b}{2}, \quad \nu = \sqrt{k - \left(\frac{b}{2}\right)^2}. \tag{3.40}$$

On the graph given in Figure 3.13, we cannot tell for sure that the first relative maximum occurs at $t = 0$, but we can estimate that there are relative maxima $y_1 \approx 2.5$, at $t_1 \approx 4.0$, and $y_2 \approx 1.2$, at $t_2 \approx 8.0$, to two significant digits. The **quasi-period** is

$$\mathsf{T} \triangleq \frac{2\pi}{\nu} \approx 8.0 - 4.0 = 4.0;$$

hence,

$$\nu \approx \frac{2\pi}{4.0} \approx 1.570796327,$$

and the logarithmic decrement is

$$\frac{2\pi\alpha}{\nu} = D = \ln\left(\frac{y_2}{y_1}\right) \approx \ln\left(\frac{1.2}{2.5}\right) \approx -0.7339691751.$$

So,

$$-\frac{b}{2} = \alpha = \frac{\nu D}{2\pi} = \frac{D}{\mathsf{T}} \approx -0.1834922938;$$

thus, the damping coefficient is $b \approx 0.3669845875$. It follows from (3.40) that the spring constant is

$$k = \nu^2 + \left(\frac{b}{2}\right)^2 \approx 2.501070523.$$

To summarize,

$$b \approx 0.3669845875, \quad k \approx 2.501070523.$$

Rounded off to the two significant digits of our graphical data, $b \approx 0.37, k \approx 2.5$. ◯

3.3.6 The Critically Damped Case

Example 3.22

For the IVP

$$\ddot{y} + 8\dot{y} + 16y = 0, \tag{3.41}$$

find as many solutions as possible.

Method: The characteristic polynomial $s^2 + 8s + 16 = (s+4)^2$ has roots $s = -4, -4$, so it seems that the solutions of the ODE are

$$y(t) = c_1 e^{-4t} + c_2 e^{-4t} = ce^{-4t},$$

where $c = c_1 + c_2$ is only one arbitrary constant. But our experience with the overdamped and underdamped cases says that we need two arbitrary constants if we are to solve for two initial conditions, and the general Theorem 3.9 says that we have two different solutions $y_1(t)$ and $y_2(t)$, where, for example, $y_2(0) = 0, \dot{y}_2(0) = 1$. No such function $y_2(t)$ appears from using the roots of the characteristic polynomial. What can we do?

Here's some general advice: if we don't know what to do, try using what we do know! That's vague. In this situation, we know we have a solution $y(t) = e^{-4t}$. What can we do with it? Multiplying it by a constant doesn't help, so try multiplying it by a function to be determined:

$$y(t) = v(t)e^{-4t}.$$

Substitute this into the ODE, after finding

$$\dot{y} = \dot{v}e^{-4t} - 4ve^{-4t}, \quad \ddot{y} = \ddot{v}e^{-4t} - 8\dot{v}e^{-4t} + 16ve^{-4t},$$

to get

$$0 = \ddot{y} + 8\dot{y} + 16y = (\ddot{v}e^{-4t} - 8\dot{v}e^{-4t} + \cancel{16ve^{-4t}}) + 8(\dot{v}e^{-4t} - \cancel{4ve^{-4t}}) + \cancel{16ve^{-4t}} = \ddot{v}.$$

The solutions of $0 = \ddot{v}$ are $v(t) = c_1 t + c_2$, so we get solutions

$$y = v(t)e^{-4t} = (c_1 t + c_2)e^{-4t} = c_1 e^{-4t} + c_2 te^{-4t},$$

where c_1, c_2 are arbitrary constants. By "accident" we got the solutions, ce^{-4t}, that we started with. ◯

As an aside, when we tried $y(t) = v(t)e^{-4t}$, we were using the method of "reduction of order."

Rather than "reinventing the wheel" in the future, it's useful to state

Theorem 3.11

(Solutions in the critically damped case) Consider the ODE $\ddot{y} + p\dot{y} + q = 0$. Suppose p, q are constants and the characteristic polynomial has a repeated real root $s = \alpha$. Then all solutions of the ODE are given by

$$y(t) = c_1 e^{\alpha t} + c_2 te^{\alpha t}, \quad \text{where } c_1, c_2 \text{ are arbitrary constants.}$$

So, $\{e^{\alpha t}, te^{\alpha t}\}$ is a complete set of basic solutions.

3.3.7 The Wronskian Determinant

Definition 3.9

Suppose $y_1(t), y_2(t)$ are any two functions. Their **Wronskian determinant** is defined to be

$$W(y_1, y_2)(t) \triangleq \begin{vmatrix} y_1(t) & y_2(t) \\ \dot{y}_1(t) & \dot{y}_2(t) \end{vmatrix}.$$

Example 3.23

$$W(e^{\alpha t}\cos \nu t, e^{\alpha t}\sin \nu t) = \begin{vmatrix} e^{\alpha t}\cos \nu t & e^{\alpha t}\sin \nu t \\ \alpha e^{\alpha t}\cos \nu t - \nu e^{\alpha t}\sin \nu t & \alpha e^{\alpha t}\sin \nu t + \nu e^{\alpha t}\cos \nu t \end{vmatrix}$$

$$= e^{\alpha t}\cos \nu t \left(\alpha e^{\alpha t}\sin \nu t + \nu e^{\alpha t}\cos \nu t\right) - e^{\alpha t}\sin \nu t \left(\alpha e^{\alpha t}\cos \nu t - \nu e^{\alpha t}\sin \nu t\right)$$

$$= \cdots = \nu e^{2\alpha t}. \bigcirc$$

Theorem 3.12

(Abel's theorem) Suppose y_1, y_2 are two solutions of the same ODE (3.28), that is, $\ddot{y}+p(t)\dot{y}+q(t)y = 0$, on an open interval I. Then

$$W(y_1, y_2)(t) = \exp\Big(-\int_{t_0}^{t} p(\tau)d\tau\Big)\, W(y_1, y_2)(t_0) \tag{3.42}$$

for any t_0, t inside I.

Why? We will see why $W(t) \triangleq W(y_1, y_2)(t)$ satisfies the first-order homogeneous ODE

$$\frac{dW}{dt} = -p(t)W(t),$$

and then the rest will be in Problem 3.3.8.46. We have

$$\begin{aligned} \dot{W} = \frac{d}{dt}\left[y_1(t)\dot{y}_2(t) - y_2(t)\dot{y}_1(t)\right] &= \left(\dot{y}_1(t)\dot{y}_2(t) + y_1(t)\ddot{y}_2(t)\right) - \left(\dot{y}_2(t)\dot{y}_1(t) + y_2(t)\ddot{y}_1(t)\right) \\ &= \left(\dot{y}_1(t)\dot{y}_2(t) - \dot{y}_2(t)\dot{y}_1(t)\right) + \left(y_1(t)\ddot{y}_2(t) - y_2(t)\ddot{y}_1(t)\right) \\ &= \begin{vmatrix} \dot{y}_1(t) & \dot{y}_2(t) \\ \dot{y}_1(t) & \dot{y}_2(t) \end{vmatrix} + \begin{vmatrix} y_1(t) & y_2(t) \\ \ddot{y}_1(t) & \ddot{y}_2(t) \end{vmatrix}. \end{aligned}$$

The first determinant is zero because it has two equal rows.

Both of y_1, y_2 solve the same ODE (3.28), that is, $\ddot{y}_1 + p(t)\dot{y}_1 + q(t)y_1 = 0, \ddot{y}_2 + p(t)\dot{y}_2 + q(t)y_2 = 0$ on I. So, using the effects of elementary row operations on determinants,

$$\dot{W} = 0 + \begin{vmatrix} y_1(t) & y_2(t) \\ -p(t)\dot{y}_1 - q(t)y_1 & -p(t)\dot{y}_2 - q(t)y_2 \end{vmatrix} \underset{-q(t)R_1+R_2\to R_2}{=} \begin{vmatrix} y_1(t) & y_2(t) \\ -p(t)\dot{y}_1(t) & -p(t)\dot{y}_2(t) \end{vmatrix}$$

$$\underset{R_2\leftarrow -p(t)R_2}{=} -p(t)\begin{vmatrix} y_1(t) & y_2(t) \\ \dot{y}_1(t) & \dot{y}_2(t) \end{vmatrix} = -p(t)W(t). \square$$

Theorem 3.13

Suppose $y_1(t), y_2(t)$ are solutions of the same linear homogeneous ODE (3.28), that is, $\ddot{y} + p(t)\dot{y} + q(t)y = 0$, on an open interval I, and $p(t), q(t)$ are continuous on I. Then $W(y_1, y_2)(t) \neq 0$ for all t in I if, and only if, $\{y_1(t), y_2(t)\}$ is a complete set of basic solutions of ODE (3.28) on I.

Why? The explanation is similar to that of Theorem 3.9: Given any $y^*(t)$ that solves ODE (3.28) on I, we can choose constants c_1, c_2 such that

$$\begin{bmatrix} y^*(t_0) \\ \dot{y}^*(t_0) \end{bmatrix} = \begin{bmatrix} y_1(t_0) & y_2(t_0) \\ \dot{y}_1(t_0) & \dot{y}_2(t_0) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$$

after using Theorem 3.12 to see that the invertibility of the 2×2 matrix follows from the fact that its Wronskian determinant $W(y_1, y_2)(t_0) \neq 0$ for all t_0 in I. After that, Lemma 1.3 in Section 1.7 and the uniqueness in Theorem 3.8 explain why $y^*(t) = c_1 y_1(t) + c_2 y_2(t)$. $\square$

3.3.8 Problems

In problems 1–6, solve the ODE. If all solutions have $\lim_{t\to\infty} y(t) = 0$, determine the time constant.

1. $\ddot{y} + 8\dot{y} + 15y = 0$
2. $2\ddot{y} + 12\dot{y} + 18y = 0$
3. $\ddot{y} + \dot{y} - \frac{15}{4}y = 0$
4. $\ddot{y} + 2\dot{y} + 65y = 0$
5. $\ddot{y} + 8\dot{y} + 18y = 0$
6. $\ddot{y} + 4\dot{y} + 7y = 0$

In problems 7–9, solve the IVP.

7. $\ddot{y} - 2\dot{y} - 3y = 0,\ y(0) = 5,\ \dot{y}(0) = 7$
8. $\ddot{y} + \frac{3}{2}\dot{y} + \frac{1}{2}y = 0,\ y(0) = 0,\ \dot{y}(0) = 1$
9. $\ddot{y} + \dot{y} + \frac{9}{4}y = 0,\ y(0) = 7,\ \dot{y}(0) = 0$

In problems 10–13, solve the IVP.

10. $\ddot{y} + 8\dot{y} - 20y = 0,\ y(0) = 1,\ \dot{y}(0) = 2$
11. $\ddot{y} + 3\dot{y} - 10y = 0,\ y(0) = 1,\ \dot{y}(0) = -3$
12. $\ddot{y} + 8\dot{y} + 16y = 0,\ y(0) = -1,\ \dot{y}(0) = 2$
13. $\ddot{y} + \dot{y} + \frac{1}{4}y = 0,\ y(0) = -1,\ \dot{y}(0) = 2$

In problems 14–21, solve the IVPs and express the solution in the amplitude-phase form (3.39). Give conclusions using exact values, as simply as possible.

14. $\ddot{y} + \dot{y} + y = 0,\ y(0) = -2,\ \dot{y}(0) = 4$
15. $\ddot{y} + \dot{y} + y = 0,\ y(0) = -2,\ \dot{y}(0) = -2$
16. $\ddot{y} + \dot{y} + \frac{17}{4}y = 0,\ y(0) = -1,\ \dot{y}(0) = 2$
17. $\ddot{y} + 2\dot{y} + 5y = 0,\ y(0) = -2,\ \dot{y}(0) = 6$
18. $\ddot{y} + 4\dot{y} + 5y = 0,\ y(0) = 2,\ \dot{y}(0) = 0$
19. $\ddot{y} + 4\dot{y} + 5y = 0,\ y(0) = -1,\ \dot{y}(0) = 2 + \sqrt{3}$
20. $\ddot{y} + 4\dot{y} + 10y = 0,\ y(0) = 2,\ \dot{y}(0) = 0$
21. $5\ddot{y} + 20\dot{y} + 60y = 0,\ y(0) = -2,\ \dot{y}(0) = 0$
22. For each of the ODEs following, indicate which case it's in with abbreviations "O" for overdamped, "C" for critically damped, or "U" for underdamped.
    (a) $\ddot{y} + 4\dot{y} + 2y = 0$
    (b) $\ddot{y} + 4\dot{y} + 4y = 0$
    (c) $4\ddot{y} + 4\dot{y} + 2y = 0$
    (d) $4\ddot{y} + 4\dot{y} + \frac{1}{2}y = 0$
    (e) $4\ddot{y} + 8\dot{y} + 8y = 0$
23. Suppose $y(t)$ satisfies a spring–mass–damper ODE $\ddot{y} + b\dot{y} + ky = 0$. For each of parts (a), (b), and (c), the behavior of $y(t)$ is described and you are asked to give a specific *numerical* value for each of the physical parameters b, k that can produce such a solution $y(t)$.
    (a) $y(t)$ is in the overdamped case
    (b) $y(t) = Ae^{\alpha t}\cos(\nu t - \delta)$, where $\alpha > 0$
    (c) $y(t)$ is in the critically damped case
24. Shown in Figure 3.14 are graphs of solutions of two different ODEs of the form $\ddot{y} + b\dot{y} + y = 0$. Decide which of the pairs of parameter values could *conceivably* give the graphs, and explain why. More than one pair may be correct.
    (a) $b_1 = 1$ and $b_2 = 2$
    (b) $b_1 = 0.5$ and $b_2 = 2$
    (c) $b_1 = 1$ and $b_2 = 0.8$
    (d) $b_1 = 1$ and $b_2 = 1.5$
    (e) $b_1 = 1.2$ and $b_2 = 1.5$
25. Shown in Figure 3.15 are graphs of solutions of two different ODEs of the form $\ddot{y} + 2\dot{y} + k_1 y = 0$ and $\ddot{y} + 2\dot{y} + k_2 y = 0$. Decide which of the pairs of parameter

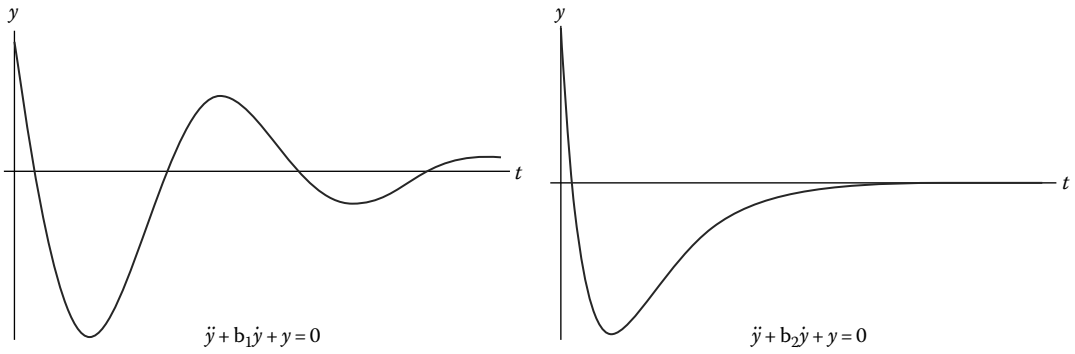


FIGURE 3.14
Problem 3.3.8.24.

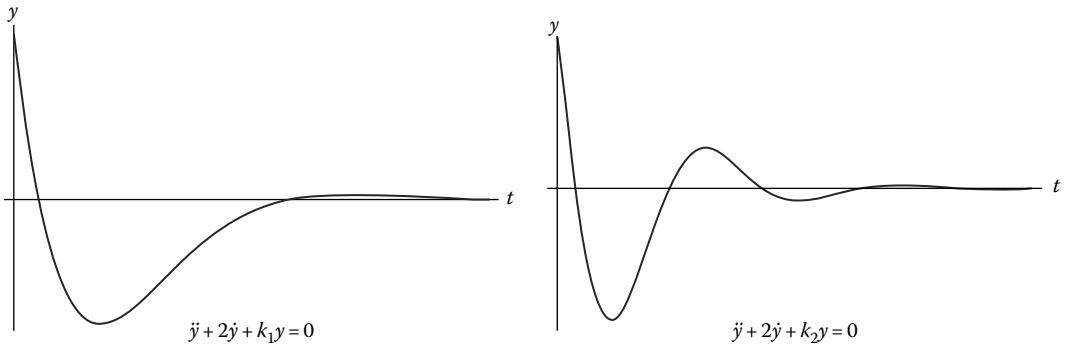


FIGURE 3.15
Problem 3.3.8.25.

values could *conceivably* give the graphs, and explain why. More than one pair may be correct.

(a) $k_1 = 1$ and $k_2 = 2$

(b) $k_1 = 0.5$ and $k_2 = 2$

(c) $k_1 = 1$ and $k_2 = 0.8$

(d) $k_1 = 1$ and $k_2 = 1.5$

(e) $k_1 = 1.2$ and $k_2 = 1.5$

26. An object, whose mass is 4 kg, stretches a spring 0.784 m. The mass is released from a point 0.3 m above the equilibrium position with a downward velocity of 10 m/s. Assuming there is no damping or other frictional forces, find the position of the mass as a function of time and find all times when the mass reaches the maximum displacement below the equilibrium position.

27. An object, whose mass is 2 kg, stretches a spring 0.392 m. At all times, the damping device gives a resisting force whose magnitude in newtons is 20 times the magnitude of the velocity in m/s. The mass is released from the equilibrium position with an upward velocity of 5 m/s. How far from the equilibrium position does the object get?

28. A mass on a vertical spring has position $y(t) = e^{-t}\left(\cos(\frac{t}{2}) + \sin(\frac{t}{2})\right)$.
    (a) When is the first time $t \geq 0$ when the mass passes through the equilibrium position?
    (b) Find an ODE for which $y(t)$ is a solution.
29. Suppose $y(t) = 2e^{-3t}\cos(t - \frac{\pi}{4})$ solves an ODE of the form $\ddot{y} + b\dot{y} + ky = 0$. Find the exact values of b, k. Of course, show all your work that explains how you found b, k.
30. Assume that the vertical position of a wheel in an automobile suspension system is given by $y(t) = -e^{-t}\cos 3t + \sqrt{3}e^{-t}\sin 3t$.
    (a) When is the first time $t \geq 0$ when the wheel passes through the equilibrium position?
    (b) What is the maximum *absolute* deviation of the wheel from the equilibrium position?
    (c) Find an ODE for which $y(t)$ is a solution.
31. Suppose that a series *RLC* circuit has current I satisfying $\frac{1}{2}\ddot{I} + 10\dot{I} + \frac{1}{C}I = 0$. Assume that C is, as usual, a constant. Give a criterion on C that guarantees that the current I does *not* oscillate infinitely often.
32. Suppose that $I(t) = 5e^{-2t}\cos(\sqrt{3}\,t - \frac{\pi}{3})$ is a current that solves an ODE of the form $5\ddot{I} + R\dot{I} + \frac{1}{C}I = 0$ for a series *RLC* circuit. Assume that R and C are, as usual, constants. Find the exact values of the parameters R, in ohms, and C, in farads. Assume the "5" in the ODE is short for 5 H.
33. Suppose that $y(t)$ satisfies an ODE of the form $\ddot{y} + p\dot{y} + qy = 0$, where p, q are constants. Suppose the graph of $y(t)$ has successive local maxima at the points $(t, y) = (2.0000, 5.1234)$ and $(4.0000, 3.9876)$. Find the values of p, q, correct to four decimal please.
34. Assuming mass $m = 1$, find the damping coefficient and the spring constant that can produce the solution graphed in Figure 3.16.
35. Given that $y_1(t)$ and $y_2(t)$ solve the same ODE $(\star)$ $\ddot{y} + p(t)\dot{y} + q(t)y = 0$ on an interval I and that $p(t)$ and $q(t)$ are continuous on that interval.

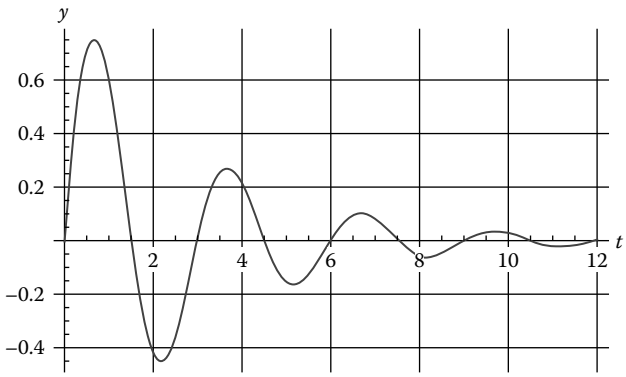


FIGURE 3.16
Problem 3.3.8.34.

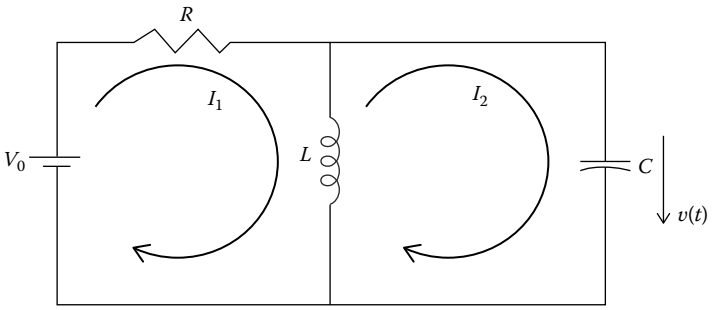


FIGURE 3.17
Problem 3.3.8.37.

(a) Which of the following is true?
(a1) $2y_1(t) - y_2(t)$ is a solution of $(\star)$ on I
(a2) $2y_1(t) - y_2(t)$ is not a solution of $(\star)$ on I
(a3) Without more information, we cannot tell whether (a1) or (a2) is true.
(b) Which of the following is true?
(b1) $c_1y_1(t) + c_2y_2(t)$ is a general solution of $(\star)$ on I
(b2) $c_1y_1(t) + c_2y_2(t)$ is not a general solution of $(\star)$ on I
(b3) Without more information, we cannot tell whether (b1) or (b2) is true.

36. For the DC series *RLC* circuit shown in Figure 3.8 and modeled by ODE (3.25), define the natural frequency by $\omega_0 = \sqrt{\frac{1}{LC}}$ and the **neper** "frequency" by $\alpha = \frac{R}{2L}$. Explain why the roots of the characteristic polynomial are $-\alpha \pm \sqrt{\alpha^2 - \omega_0^2}$.

37. (Small project) For the DC parallel *RLC* circuit shown in Figure 3.17, assume that L, R, and C are, as usual, constants.
(a) Explain why the voltage across the capacitor, $v(t)$, satisfies the ODE $C\ddot{v} + \frac{1}{R}\dot{v} + \frac{1}{L}v = 0$.
(b) Define the natural frequency by $\omega_0 = \sqrt{\frac{1}{LC}}$ and the **neper** "frequency" by $\alpha = \frac{1}{2RC}$. Explain why the roots of the characteristic polynomial are $-\alpha \pm \sqrt{\alpha^2 - \omega_0^2}$.
(c) Conclude that the natural frequency is the same whether the *RLC* circuit is series or parallel but that the formula for the neper "frequency" varies according to whether it is a series or parallel circuit.

38. Suppose $y(t)$ solves the IVP $\ddot{y} + \dot{y} - 6y = 0$, $y(0) = a$, $\dot{y}(0) = 0$, where "a" is an unspecified positive number. (Do not substitute in a specific value for "a.") Explain why $\lim_{t\to-\infty} y(t) = \infty = \lim_{t\to\infty} y(t)$.

39. Suppose that a critically damped spring–mass–damper system is released from rest. Explain why the maximum deviation from equilibrium occurs at the initial time. This is a completely general result!

40. Suppose that an overdamped spring–mass–damper system is released from rest. Explain why the maximum deviation from equilibrium occurs at the initial time. This is a completely general result!

41. For ODE

$$(1-2t)\ddot{y} + (1+4t^2)\dot{y} + (-2+2t-4t^2)y = 0,$$

(a) Explain why e^t and e^{t^2} are solutions.

(b) Find a complete set of basic solutions on some interval.

(c) Try to solve the IVP consisting of that ODE and the ICs $y(\frac{1}{2}) = 5,\ \dot{y}\left(\frac{1}{2}\right) = -3$.

(d) Does your difficulty in part (c) contradict the existence and uniqueness conclusion of Theorem 3.8?

42. Suppose ω and L are positive constants. Find all solutions of $\ddot{y} - \omega^2 y = 0$ that also satisfy the "boundary condition" $y(L) = 0$.

43. Suppose ω and L are positive constants. Find all solutions of $\ddot{y} - \omega^2 y = 0$ that also satisfy the "boundary condition" $\dot{y}(L) = 0$.

44. Suppose a second-order linear homogeneous ODE has coefficients that are continuous on an interval I and that $\{y_1(t), y_2(t)\}$ is a compete set of basic solutions on I. Use Theorem 3.13 to decide for which values of α the set of $\{\alpha y_1(t) + y_2(t), y_1(t) + 2\alpha y_2(t)\}$ is also a compete set of basic solutions on I.

45. Suppose a second-order linear homogeneous ODE has coefficients that are continuous on an interval I and that $\{y_1(t), y_2(t)\}$ is a compete set of basic solutions on I. Use Theorem 3.13 to decide for which values of α the set of $\{\alpha y_1(t) - 3y_2(t), y_1(t) - \alpha y_2(t)\}$ is also a compete set of basic solutions on I.

46. Using the idea and explanation that preceded Problem 3.1.4.37, explain why the solution of $\frac{dW}{dt} = -p(t)W(t)$ can be written in the form $W(t) = \exp\left(-\int_{t_0}^{t} p(\tau)d\tau\right) W(t_0)$.

47. Explain why the relative maxima of $y(t) = Ae^{\alpha t}\cos(\nu t - \delta)$ occur at times $t = t_1, t_2, t_3, \ldots$, which occur every $\frac{2\pi}{\nu}$ units of time. [This is why we call $T = \frac{2\pi}{\nu}$ the quasi-period.]

3.4 Higher-Order Linear ODEs

An IVP for an n-th order linear ODE has the standard form

$$\left\{\begin{array}{l} y^{(n)} + p_1(t)y^{(n-1)} + \cdots + p_{n-1}(t)\dot{y} + p_n(t)y = f(t), \\ \\ y(t_0) = y_0,\ \dot{y}(t_0) = y_0^{(1)}, \ldots, y^{(n-1)}(t_0) = y_0^{(n-1)} \end{array}\right\}. \tag{3.43}$$

Note that n initial conditions are imposed in (3.43).

Theorem 3.14

Suppose the functions $p_1(t), ..., p_n(t), f(t)$ are continuous on an open interval I. Then IVP (3.43) has exactly one solution on interval I.

Example 3.24

Find as many solutions of $\dddot{y} - 9\ddot{y} + 23\dot{y} - 15y = 0$ as possible.

Method: As in Section 3.3, we try solutions in the form $y = e^{st}$. Substitute that into the homogeneous ODE to get

$$0 = s^3e^{st} - 9s^2e^{st} + 23se^{st} - 15e^{st} = (s^3 - 9s^2 + 23s - 15)e^{st}.$$

So, we get a **characteristic equation** in s:

$$s^3 - 9s^2 + 23s - 15 = 0.$$

It's possible to factor the third-degree polynomial: $s^3 - 9s^2 + 23 - 15 = (s-1)(s-3)(s-5)$. So, we know the ODE has solutions e^t, e^{3t}, e^{5t}. Because the equation is linear and homogeneous, we can multiply each of these functions by an arbitrary constant and add to get solutions

$$y(t) = c_1e^t + c_2e^{3t} + c_3e^{5t}, \text{where } c_1, c_2, c_3 \text{ are arbitrary constants.} \bigcirc$$

By the way, here's some information about factoring: Denote the polynomial by $\mathcal{P}(s) = \underline{1}s^3 - 9s^2 + 23 - \underline{15}$. Standard advice when finding roots of a polynomial says that here we should *try*

$$s = \pm\frac{\text{All possible factors of 15}}{\text{all possible factors of 1}} = \pm1, \pm3, \pm5, \pm15.$$

Since $\mathcal{P}(1) = 0$, $(s-1)$ must be a factor of $\mathcal{P}(s)$. Long division then gives $\mathcal{P}(s) = (s-1)$ (*a second-degree polynomial*), and the latter can be factored using the quadratic formula.

Did we find *all* of the solutions in Example 3.24? To answer that, we need a little more theory.

Definition 3.10

The **general solution** of an n-th order linear homogeneous ODE

$$y^{(n)} + p_1(t)y^{(n-1)} + \cdots + p_{n-1}(t)\dot{y} + p_n(t)y = 0 \tag{3.44}$$

has the form

$$y_h(t) = c_1y_1(t) + c_2y_2(t) + \cdots + c_ny_n(t)$$

if for every solution $y^*(t)$ of (3.44), there are values of constants $c_1, c_2, \ldots, c_n$ giving $y^*(t) = c_1y_1(t) + c_2y_2(t) + \cdots + c_ny_n(t)$. In this case, we call the set of functions $\{y_1(t), y_2(t), \ldots, y_n(t)\}$ a **complete set of basic solutions**. Each of functions $y_1(t), \ldots, y_n(t)$ is called a **basic solution** of (3.44).

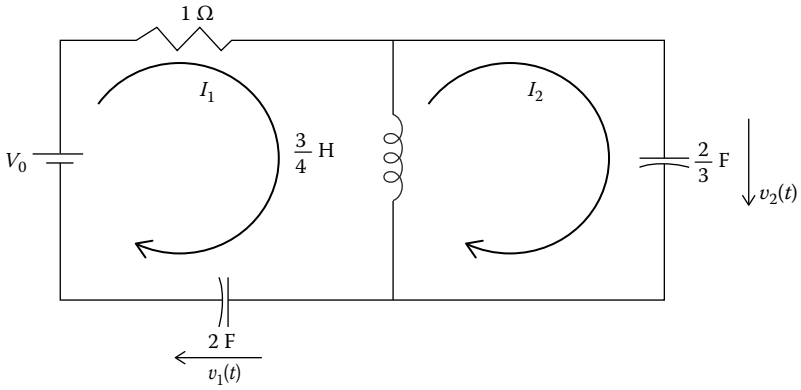


FIGURE 3.18
Example 3.25.

Example 3.25

Find the solution for the voltage $v_2(t)$ across the second capacitor in the DC circuit shown in Figure 3.18.

Method: The voltage, v, across a capacitor and current, I, through it satisfy $\dot{v} = \mathcal{C}^{-1}I$. Here,

$$\dot{v}_1 = \frac{1}{2}\, I_1, \text{ and } \dot{v}_2 = \frac{3}{2}\, I_2.$$

It follows that

$$\frac{2}{3}\ddot{v}_2 = \dot{I}_2. \tag{3.45}$$

Kirchhoff's voltage law in the first loop states that $V_0 = I_1 + \frac{3}{4}(\dot{I}_1 - \dot{I}_2) + v_1$. Take the derivative with respect to t of both sides to get

$$0 = \dot{I}_1 + \frac{3}{4}(\ddot{I}_1 - \ddot{I}_2) + \dot{v}_1\,. \tag{3.46}$$

Kirchhoff's voltage law in the second loop states that

$$0 = v_2 + \frac{3}{4}(\dot{I}_2 - \dot{I}_1) \tag{3.47}$$

The latter yields $v_2 = \frac{3}{4}\,\dot{I}_1 - \frac{3}{4}\,\dot{I}_2 = \frac{3}{4}\,\dot{I}_1 - \frac{1}{2}\,\ddot{v}_2$; hence,

$$\dot{I}_1 = \frac{4}{3}\, v_2 + \frac{2}{3}\,\ddot{v}_2. \tag{3.48}$$

It follows that

$$\frac{2}{3}\,\ddot{v}_2 = \dot{I}_1 - \frac{4}{3}\, v_2. \tag{3.49}$$

Take the derivative with respect to t of both sides of (3.47) to get

$$0 = \dot{v}_2 + \frac{3}{4}(\ddot{I}_2 - \ddot{I}_1). \tag{3.50}$$

Add to that (3.46) to get

$$0 = \dot{v}_2 + \dot{I}_1 + \dot{v}_1,$$

so,

$$\dot{I}_1 = -\dot{v}_2 - \dot{v}_1.$$

Substitute that into (3.49) to get

$$\frac{2}{3}\ddot{v}_2 = -\dot{v}_2 - \dot{v}_1 - \frac{4}{3}v_2\,.$$

But

$$\dot{v}_1 = \frac{1}{2}\,I_1,$$

so,

$$\frac{2}{3}\,\ddot{v}_2 = -\dot{v}_2 - \frac{1}{2}\,I_1 - \frac{4}{3}\,v_2. \tag{3.51}$$

Take the derivative with respect to t of both sides of (3.51) to get

$$\frac{2}{3}\,\dddot{v}_2 = -\ddot{v}_2 - \frac{1}{2}\,\dot{I}_1 - \frac{4}{3}\,\dot{v}_2. \tag{3.52}$$

Use (3.48), that is, $\dot{I}_1 = \frac{2}{3}\,\ddot{v}_2 + \frac{4}{3}\,v_2$, in (3.52) to get

$$\frac{2}{3}\,\dddot{v}_2 = -\ddot{v}_2 - \left(\frac{1}{3}\,\ddot{v}_2 + \frac{2}{3}\,v_2\right) - \frac{4}{3}\,\dot{v}_2$$

So,

$$\frac{2}{3}\,\dddot{v}_2 + \frac{4}{3}\,\ddot{v}_2 + \frac{4}{3}\,\dot{v}_2 + \frac{2}{3}\,v_2 = 0.$$

After multiplying through by $\frac{3}{2}$, the characteristic equation is $s^3 + 2s^2 + 2s + 1 = 0$. The standard advice says to try $s = \pm 1$. Here the advice works and gives root $s = -1$, so we can factor to get $0 = (s+1)(s^2 + s + 1)$. The roots are $s = -1, -\frac{1}{2} \pm i\,\frac{\sqrt{3}}{2}$, so the general solution is

$$v_2(t) = c_1 e^{-t} + c_2 e^{-t/2}\cos\left(\frac{\sqrt{3}t}{2}\right) + c_3 e^{-t/2}\sin\left(\frac{\sqrt{3}t}{2}\right),$$

where c_1, c_2, c_3 are arbitrary constants. ◯

Usually a characteristic equation will not be so simple and easy to solve, but we cooked this one up to be convenient and demonstrate the concept of general solution.

Theorem 3.15

Suppose the functions $p_1(t), \ldots, p_n(t)$ are continuous on an open interval I. Then the linear homogeneous n-th order ODE (3.44) has a general solution, that is, a complete set of n basic solutions, on interval I.

Why? Pick *any* t_0 inside I. By Theorem 3.14, each of the n IVPs

$$\left\{\begin{array}{l} y_1^{(n)}+p_1(t)y_1^{(n-1)}+\cdots+p_n(t)y_1=0 \\ \\ y_1(t_0)=1,\dot{y}_1(t_0)=0,\cdots,y^{(n-1)}(t_0)=0 \end{array}\right\},\quad \left\{\begin{array}{l} y_2^{(n)}+p_1(t)y_2^{(n-1)}+\cdots+p_n(t)y_2=0 \\ \\ y_2(t_0)=0,\dot{y}_2(t_0)=1,\cdots,y_2^{(n-1)}(t_0)=0 \end{array}\right\},\cdots$$

$$\left\{\begin{array}{l} y_n^{(n)}+p_1(t)y_n^{(n-1)}+\cdots+p_n(t)y_n=0 \\ \\ y_n(t_0)=0,\dot{y}_n(t_0)=0,\cdots,y_n^{(n-1)}(t_0)=1 \end{array}\right\}$$

has a solution. The rest of the explanation is similar to what we gave for Theorem 3.9 in Section 3.3. □

Unfortunately, in Example 3.24, the solutions e^t, e^{3t}, e^{5t} that we found are not the basic solutions $y_1(t), y_2(t), y_3(t)$ because, for example, e^t doesn't satisfy the initial conditions that $y_1(t), y_2(t)$, or $y_3(t)$ satisfies in the explanation for Theorem 3.15. But fortunately, we will have another tool in our toolbox.

Definition 3.11

Suppose $y_1(t), y_2(t), \ldots, y_n(t)$ are any n functions. Their **Wronskian determinant** is the $n\times n$ determinant defined by

$$W\left(y_1(t), y_2(t), \ldots, y_n(t)\right) \triangleq \begin{vmatrix} y_1(t) & . & . & . & y_n(t) \\ \dot{y}_1(t) & . & . & . & \dot{y}_n(t) \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ y_1^{(n-1)}(t) & . & . & . & y_n^{(n-1)}(t) \end{vmatrix}.$$

Example 3.26

$$W(e^t, e^{3t}, e^{5t}) = \begin{vmatrix} e^t & e^{3t} & e^{5t} \\ e^t & 3e^{3t} & 5e^{5t} \\ e^t & 9e^{3t} & 25e^{5t} \end{vmatrix} = \cdots = 16e^{9t}.$$

Theorem 3.16

(Abel's theorem) Suppose $y_1(t), y_2(t), \ldots, y_n(t)$ are n solutions of the same n-th order linear homogeneous ODE (3.44) on an open interval I on which the functions $p_1(t), \ldots, p_n(t)$ are all continuous. Then,

$$W(y_1, y_2, \ldots, y_n)(t) = \exp\left(-\int_{t_0}^{t} p_1(\tau)d\tau\right) W(y_1, y_2, \ldots, y_n)(t_0) \tag{3.53}$$

for any t_0, t inside I.

Why? Similar to the explanation of Theorem 3.12 in Section 3.3, but with need for greater use of determinants, we can explain why $W(t) \triangleq W(y_1, y_2, \ldots, y_n)(t)$ satisfies the first-order homogeneous ODE

$$\frac{dW}{dt} = -p_1(t)W(t).$$

The rest of the explanation is the same as in Section 3.3, specifically as in Problem 3.3.8.46. □

Theorem 3.17

Suppose $y_1(t), y_2(t), \ldots, y_n(t)$ are solutions of the same n-th order linear homogeneous ODE (3.44) on an open interval I and $p_1(t), \ldots, p_n(t)$ are all continuous on I. Then $W(y_1, y_2, \ldots, y_n)(\bar{t}) \neq 0$ for all $\bar{t}$ in I if, and only if, $\{y_1(t), y_2(t), \ldots, y_n(t)\}$ is a complete set of basic solutions of ODE (3.44) on I.

Why? The explanation is similar to what we gave for Theorem 3.13 in Section 3.3. □

Example 3.27

Solve the IVP $\left\{ \begin{array}{c} \dddot{y} - 9\ddot{y} + 23\dot{y} - 15y = 0 \\ y(0) = -1, \ \dot{y}(0) = 2, \ \ddot{y}(0) = 5 \end{array} \right\}$.

Method: In Example 3.24, we used a characteristic equation to find three solutions, e^t, e^{3t}, e^{5t}, for the third-order LCCHODE. In Example 3.26, we calculated their Wronskian, so Theorem 3.15 guarantees that $\{e^t, e^{3t}, e^{5t}\}$ is a complete set of basic solutions for the ODE. By definition,

$$y(t) = c_1 e^t + c_2 e^{3t} + c_3 e^{5t}$$

gives all the solutions of the ODE. To find constants c_1, c_2, c_3 so that $y(t)$ satisfies the three ICs, we can write the ICs in vector form as

$$\begin{bmatrix} -1 \\ 2 \\ 5 \end{bmatrix} = \begin{bmatrix} c_1 + c_2 + c_3 \\ c_1 + 3c_2 + 5c_3 \\ c_1 + 9c_2 + 25c_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \\ 1 & 9 & 25 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}.$$

The 3×3 matrix is the matrix whose determinant gives a Wronskian value of 16, at least at $t = 0$. It follows that we can invert the matrix, so

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \\ 1 & 9 & 25 \end{bmatrix}^{-1} \begin{bmatrix} -1 \\ 2 \\ 5 \end{bmatrix} = \begin{bmatrix} -3.25 \\ 3 \\ -0.75 \end{bmatrix},$$

the last step using a calculator. While calculators do not necessarily give exact conclusions, the facts that (1) the determinant of the matrix is 16 = (a power of 2) and (2) the values for c_1, c_2, c_3 are fractions whose denominators are powers of 2 give us confidence that these are the exact values. So, the solution of the IVP is

$$y(t) = -\frac{13}{4}e^t + 3e^{3t} - \frac{3}{4}e^{5t}. \bigcirc$$

3.4.1 The Zoo of Solutions of LCCHODEs

Suppose an n-th order homogeneous linear ODE has constant coefficients, that is, is a LCCHODE, so it is in the form

$$y^{(n)} + p_1 y^{(n-1)} + \cdots + p_{n-1}\dot{y} + p_n y = 0, \tag{3.54}$$

where $p_1, \ldots, p_n$ are constants. Substituting in $y(t) = e^{st}$ yields the corresponding **characteristic polynomial**

$$\mathcal{P}(s) \triangleq s^n + p_1 s^{n-1} + \cdots + p_{n-1}s + p_n, \tag{3.55}$$

which has n roots, including multiplicities.

If among those roots there is a real root, s_1, of multiplicity m_1, that is, $P(s) = (s-s_1)^{m_1}Q(s)$ where $Q(s_1) \neq 0$, then correspondingly LCCHODE (3.54) has solutions

$$e^{s_1 t}, \ldots, t^{m_1-1}e^{s_1 t}.$$

Since the ODE is linear and homogeneous, the principle of linear superposition tells us that the ODE has solutions

$$y(t) = c_1 e^{s_1 t} + \cdots + c_{m_1} t^{m_1-1} e^{s_1 t}, \tag{3.56}$$

where $c_1, \ldots, c_{m_1}$ are arbitrary constants. So, any positive integer power of t times an exponential function $e^{\alpha t}$ is in the zoo of all possible solutions of LCCHODEs. Of course, a characteristic polynomial's n roots may include other roots, so (3.56) may be only part of the whole story for solutions of that n-th order LCCHODE.

For example, in Example 3.22 in Section 3.3, ODE $\ddot{y} + 8\dot{y} + 16y = 0$ had characteristic polynomial $\mathcal{P}(s) = s^2 + 8s + 16 = (s+4)^2$ having a double root $s = -4, -4$; hence, s_1 has $m_1 = 2$, and the ODE had solutions e^{-4t}, te^{-4t}.

A special case of the preceding text is when $s_1 = 0$ is a root of multiplicity m_1. Because $e^{s_1 t} = e^{0\cdot t} = e^0 \equiv 1$, the ODE has among its solutions

$$y(t) = c_1 + c_1 t + \cdots + c_{m_1} t^{m_1-1}. \tag{3.57}$$

So, any polynomial in t is in the zoo of all possible solutions of LCCHODEs.

For example, ODE $y^{(m_1)}(t) = 0$ has solutions (3.57).

Suppose that among the n roots of the characteristic polynomial $\mathcal{P}(s)$, there is a complex conjugate pair $\alpha \pm i\nu$, where α, ν are real and $\nu \neq 0$, of multiplicity m, that is, $\mathcal{P}(s) = \left((s-\alpha)^2 + \nu^2\right)^m Q(s)$, where $Q(\alpha \pm i\nu) \neq 0$. Then LCCHODE (3.54) has among its solutions

$$e^{\alpha t}\cos \nu t, e^{\alpha t}\sin \nu t, \ \ldots\ , t^{m-1}e^{\alpha t}\cos \nu t, t^{m-1}e^{\alpha t}\sin \nu t.$$

By the principle of linear superposition, among the solutions of the LCCHODE are

$$y(t) = (c_1 + c_2 t + \cdots + c_m t^{m-1})e^{\alpha t}\cos \nu t + (d_1 + d_2 t + \cdots + d_m t^{m-1})e^{\alpha t}\sin \nu t,$$

where $c_1, \ldots, c_m; d_1, \ldots, d_m$ are arbitrary constants. The special case of $\alpha = 0$ is included.

Example 3.28

Suppose the characteristic polynomial of a LCCHODE has list of roots $s = -1, -1, -5, -3 \pm i2, -3 \pm i2, \pm i4, \pm i4$. Find all solutions of the ODE.

Method: By the principle of linear superposition, the solutions are

$$y(t) = c_1 e^{-t} + c_2 t e^{-t} + c_3 e^{-5t} + c_4 e^{-3t} \cos 2t + c_5 t e^{-3t} \cos 2t + d_4 e^{-3t} \sin 2t + d_5 t e^{-3t} \sin 2t + c_6 \cos 4t + c_7 t \cos 4t + d_6 \sin 4t + d_7 t \sin 4t,$$

where $c_1, \ldots, c_7$ and $d_4, \ldots, d_7$ are arbitrary constants. ◯

For an ODE, the time constant indicates how long it takes for a solution to decay to $\frac{1}{e}$ of its initial value. Suppose all solutions of a LCCHODE are transient. The time constant τ for that ODE can be *defined* by

$$\tau = \frac{1}{r_{min}},$$

where r_{min} is the slowest decay rate. For physical intuition, it is useful to think of the time constant as being about how long it takes for the solution to decay in a standard way.

3.4.2 Differential Operator Notation

Define the symbol D by

$$D[\, y \,] = \dot{y},$$

that is, D is the operator of differentiation. We can think of D as a machine whose input is a differentiable function, $y(t)$, and whose output is $\dot{y}(t)$, its derivative with respect to t, as illustrated in Figure 3.19.

D is a linear operator, because we know from Calculus I that

$$D[c_1 y_1(t) + c_2 y_2(t)] = c_1 D[y_1(t)] + c_2 D[y_2(t)],$$

for all constants c_1, c_2 and all differentiable functions $y_1(t), y_2(t)$.

We can define higher-order derivatives by

$$D^2[y(t)] \triangleq D[D[y(t)]] = \ddot{y}(t), D^3[y(t)] \triangleq D[D^2[y(t)]] = \dddot{y}(t), \cdots$$

and we can define L, an n-th order linear differential operator (LD-Op), by

$$L[y(t)] \triangleq D^n[y(t)] + p_1(t) D^{n-1}[y(t)] + \cdots + p_{n-1}(t) D[y(t)] + p_n(t) y(t).$$

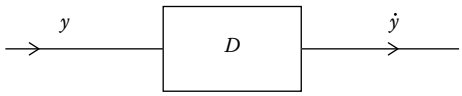


FIGURE 3.19
D Operator as a machine.

A special case is when the coefficients are constant so that we get a linear constant coefficients differential operator (LCCD-Op) L defined by

$$\begin{aligned} L[y(t)] &\triangleq D^n[y(t)] + p_1 D^{n-1}[y(t)] + \cdots + p_{n-1} D[y(t)] + p_n y(t) \\ &= (D^n + p_1 D^{n-1} + \cdots + p_{n-1} D + p_n)[y(t)]. \end{aligned}$$

What's nice about LCCD-Ops is that they can be factored just like polynomials, for example,

$$L = D^2 + 4D + 3 = (D+1)(D+3)$$

and the order of the factors does not matter:

$$(D+1)(D+3) = (D+3)(D+1).$$

Unfortunately, if an LD-Op L has at least one coefficient that is not constant, then we may not be able to factor L, and if an LD-Op can be factored, then the order of factors matters because of the product rule. For example,

$$\begin{aligned} (D(D+t))\,[y(t)] &\triangleq D[(D+t)[y(t)]] = D[\dot{y}(t) + ty(t)] = D[\dot{y}(t)] + D[ty(t)] \\ &= \ddot{y}(t) + \left(1 \cdot y(t) + t\dot{y}(t)\right) = \ddot{y}(t) + t\dot{y}(t) + y(t), \end{aligned}$$

versus

$$((D+t)D)\,[y(t)] \triangleq (D+t)[D[y(t)]] = (D+t)[\dot{y}(t)] = \ddot{y}(t) + t\dot{y}(t).$$

So, $D(D+t) \neq (D+t)D$. The order of operation matters just as the order of matrix multiplication matters. Fortunately, most of our study will be on LCCD-Ops!

Suppose $L = \mathcal{P}(D) \triangleq D^n + p_1 D^{n-1} + \cdots + p_{n-1} D + p_n$ is an LCCD-Op, where $\mathcal{P}$ is an n-th degree polynomial. When we substitute $y(t) = e^{st}$ into LCCHODE $0 = L[y(t)] = P(D)[y(t)]$, we get

$$\begin{aligned} 0 = \mathcal{P}(D)[e^{st}] &= (D^n + p_1 D^{n-1} + \cdots + p_{n-1} D + p_n)[e^{st}] \\ &= s^n e^{st} + p_1 s^{n-1} e^{st} + \cdots + p_{n-1} s e^{st} + p_n e^{st} = e^{st}\left(s^n + p_1 s^{n-1} + \cdots + p_{n-1} s + p_n\right). \end{aligned}$$

The characteristic polynomial is $\mathcal{P}(s) = s^n + p_1 s^{n-1} + \cdots + p_{n-1} s + p_n$, that is, the same polynomial in s rather than D. The characteristic polynomial, $\mathcal{P}(s)$, has exactly n roots, counting multiplicities.

Example 3.29

Find all solutions of the ODE

$$(D+1)^4 (D+3)\left((D-1)^2 + 3^2\right)^2 [y(t)] = 0. \tag{3.58}$$

Method: The characteristic polynomial is $(s+1)^4(s+3)\left((s-1)^2+3^2\right)^2$, which has roots $s=-1,-1,-1,-1,-3,1\pm i3,1\pm i3$. The solutions of LCCHODE (3.58) are

$$y(t)=(c_1+c_2t+c_3t^2+c_4t^3)e^{-t}+c_5e^{-3t}+(c_6+c_7t)e^t\cos 3t+(d_6+d_7t)e^t\sin 3t,$$

where $c_1,\ldots,c_7;d_6,d_7$ are arbitrary constants. ◯

3.4.3 Shift Theorem

We saw in Example 3.22 in Section 3.3 that ODE $(D^2+8D+16)[y(t)]=\ddot{y}+8\dot{y}+16y=0$ has a solution $y_1(t)=e^{-4t}$, and we saw that substituting in $y=e^{-4t}v(t)$ helped us to find a second solution, $y_2(t)=te^{-4t}$, after seeing that $\ddot{v}=0$. Let's rewrite what we saw as

$$(D+4)^2[e^{-4t}v(t)]=0 \text{ yielded } D^2[v]=0.$$

In fact, this is a special case of a general and useful result.

Theorem 3.18

(Shift theorem) For any positive integer k and constant α,

$$(D-\alpha)^k[e^{\alpha t}v(t)]=e^{\alpha t}D^k[v].$$

Why? We'll just explain why the result is true for $k=1$; for other positive integers k, we could use an inductive process. Notice that Theorem 3.18 is not about solving an equation but just states a property of differential operators and exponential functions.

We calculate, using the product and chain rules of Calculus I, that

$$\begin{aligned}(D-\alpha)[e^{\alpha t}v(t)]&=D[e^{\alpha t}v(t)]-\alpha\left(e^{\alpha t}v(t)\right)=\left(\cancel{\alpha e^{\alpha t}v(t)}+e^{\alpha t}\dot{v}(t)\right)-\cancel{\alpha e^{\alpha t}v(t)}\\&=e^{\alpha t}\dot{v}(t)=e^{\alpha t}D[v(t)]\,.\ \square\end{aligned}$$

Table 3.2 summarizes the zoo of solutions of LCCHODEs.

Note that the first line is a special case of the second line, because $e^{\alpha t}\equiv 1$ when $\alpha=0$. Likewise, the third line includes the functions $\cos \nu t$ and $\sin \nu t$ when $\alpha=0$.

TABLE 3.2

The Zoo of Solutions of Linear Homogeneous Constant Coefficients ODEs

Roots of Characteristic Polynomial	**Solutions *y(t)***
$0,0,\ldots,0$ [m times]	$1,t,\ldots,t^{m-1}$
$\alpha,\alpha,\ldots,\alpha$ [m times]	$e^{\alpha t},te^{\alpha t},\ldots,t^{m-1}e^{\alpha t}$
$\alpha\pm i\nu,\ldots,\alpha\pm i\nu$ [m times]	$e^{\alpha t}\cos \nu t,te^{\alpha t}\cos \nu t,\ldots,t^{m-1}e^{\alpha t}\cos \nu t$ and $e^{\alpha t}\sin \nu t,te^{\alpha t}\sin \nu t,\ldots,t^{m-1}e^{\alpha t}\sin \nu t$

3.4.4 Problems

In problems 1–3, solve the ODE.

1. $\dddot{y}+\ddot{y}-2y=0$. [Hint: Guess a simple-looking root of the characteristic polynomial.]
2. $y^{(6)}-y^{(4)}-2\ddot{y}=0$.
3. $(D+1)^3[\,y\,]=0$.

In problems 4 and 5, solve the IVP. Determine the time constant if all solutions have $\lim_{t\to\infty} y(t)=0$.

4. $(D^2-9)[\,y\,]=0,\ y(0)=3,\ \dot{y}(0)=-6$.
5. $\dddot{y}-2\ddot{y}-15\dot{y}=0,\ y(0)=0,\ \dot{y}(0)=0,\ \ddot{y}(0)=1$.

In problems 6 and 7, solve the IVP. [Hint: To factor the characteristic polynomial, substitute $s^2=r$ to get a quadratic equation in r.]

6. $(D^4+2D^2+1)[\,y\,]=0,\ y(0)=0,\ \dot{y}(0)=0,\ \ddot{y}(0)=0,\ \dddot{y}(0)=-2$.
7. $(D^4-2D^2-3)[\,y\,]=0,\ y(0)=0,\ \dot{y}(0)=0,\ \ddot{y}(0)=0,\ \dddot{y}(0)=-2$.

In problems 8–10, you are given a solution of the ODE. Use that information to help solve the ODE.

8. $y(t)=\cos t$ is a solution of ODE $\dddot{y}+3\ddot{y}+\dot{y}+3y=0$.
9. $y(t)=\sin t$ is a solution of ODE $y^{(4)}-8\dddot{y}+17\ddot{y}-8\dot{y}+16y=0$.
10. $y(t)=\sin t$ is a solution of ODE $y^{(4)}+2\dddot{y}+2\ddot{y}+2\dot{y}+y=0$.
11. Note that $(z\pm a)(z^2\mp az+a^2)=z^3\pm a^3$. Find the exact general solutions of
    (a) $\dddot{y}+8y=0$.
    (b) $\dddot{y}-2y=0$.
12. Find the solution for the voltage $v_2(t)$ across the second capacitor in the DC circuit shown in Figure 3.20.

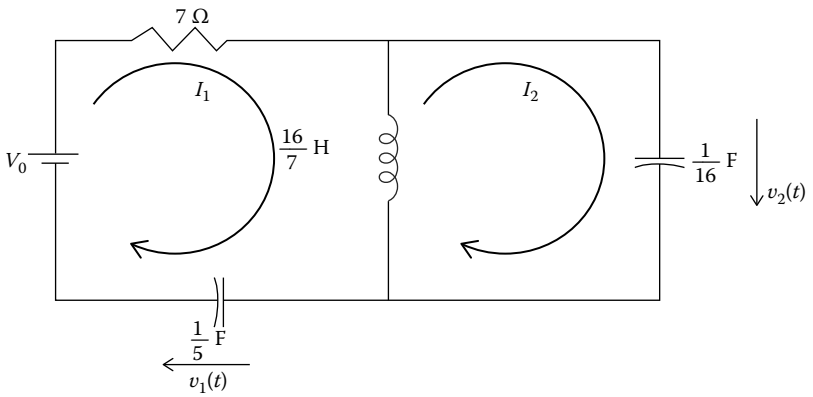


FIGURE 3.20
Problem 3.4.4.12.

13. Explain why Abel's Theorem 3.16 is true for $k = 3$. First, expand $W(t)$ along its first row, and then take its time derivative. After that, explain $\dot{W} = -p_1(t)W(t)$ by work similar to the explanation of Theorem 3.12 in Section 3.3.

3.5 Cauchy–Euler ODEs

Cauchy–Euler ODEs of the second-order have the standard form

$$r^2y'' + pry' + qy = 0, \tag{3.59}$$

where p, q are given real constants and, throughout this section, $'$ denotes $\frac{d}{dr}$. For this ODE, instead of guessing solutions of the form $y = e^{sr}$, it turns out that we will get solutions of the form

$$y(r) = r^n,$$

where n is to be chosen appropriately later.

When we substitute $y(r) = r^n$ into (3.59), we get

$$0 = r^2\left(n(n-1)r^{n-2}\right) + pr\left(nr^{n-1}\right) + qr^n = r^n\left(n(n-1) + pn + q\right).$$

So, we get a characteristic equation to be satisfied by n:

$$n(n-1) + pn + q = 0. \tag{3.60}$$

Example 3.30

Find as many solutions of

$$r^2y'' - 4ry' + 6y = 0 \tag{3.61}$$

as possible.

Method: The characteristic equation is $0 = n(n-1)-4n+6 = n^2-5n+6 = (n-2)(n-3)$, so we get solutions $y_1(r) = r^2$ and $y_2(r) = r^3$. Because the ODE is linear and homogeneous, we get solutions

$$y(r) = c_1r^2 + c_2r^3, \text{where } c_1, c_2 \text{ are arbitrary constants.}$$

In fact, this gives all of the solutions on any open interval I, as long as 0 is not in I, using Theorem 3.13 in Section 3.3: $W(y_1, y_2)(r) = \begin{vmatrix} r^2 & r^3 \\ 2r & 3r^2 \end{vmatrix} = r^4 \neq 0$ as long as $r \neq 0$. ◯

So, we see how to handle second-order Cauchy–Euler ODEs when the characteristic polynomial has two distinct real roots. The other two cases, when the characteristic polynomial has exactly one real root or a complex conjugate pair of roots, are trickier. To work on these cases, we need to study where the idea for trying $y = r^n$ came from.

Substitute $t = \ln(r)$, that is, $r = e^t$: if $y = y(r)$, denote $Y(t) \triangleq y(e^t)$. Using the chain rule for a function of a single variable, we have

$$\dot{Y} \triangleq \frac{d}{dt}\big[Y(t)\big] = \frac{d}{dr}\big[y(r)\big] \cdot \frac{dr}{dt} = \frac{dy}{dr} \cdot e^t = \frac{dy}{dr} \cdot r,$$

and, using the product rule and the chain rule, we have

$$\ddot{Y} \triangleq \frac{d^2}{dt^2}\big[Y(t)\big] = \frac{d}{dt}\big[\dot{Y}\big] = \frac{d}{dt}\left[\left(\frac{dy}{dr}\right) \cdot \left(e^t\right)\right] = \left(\frac{d}{dr}\left[\frac{dy}{dr}\right] \cdot \frac{dr}{dt}\right) \cdot \left(e^t\right) + \left(\frac{dy}{dr}\right) \cdot \left(\frac{d}{dt}\left[e^t\right]\right)$$

that is,

$$\ddot{Y} = \left(\frac{d^2y}{dr^2}\right) \cdot (e^t)^2 + \left(\frac{dy}{dr}\right) \cdot e^t = \left(\frac{d^2y}{dr^2}\right) \cdot r^2 + \left(\frac{dy}{dr}\right) \cdot r = r^2 \frac{d^2y}{dr^2} + \dot{Y}.$$

When we substitute

$$r^2 \frac{d^2y}{dr^2} = \ddot{Y} - \dot{Y} \text{ and } r\frac{dy}{dr} = \dot{Y} \tag{3.62}$$

into the original Cauchy–Euler ODE (3.59), we get

$$\left(\ddot{Y} - \dot{Y}\right) + p\dot{Y} + qY = 0,$$

that is,

$$\ddot{Y}(t) + (-1 + p)\dot{Y}(t) + qY(t) = 0.$$

When we substitute $Y(t) = e^{nt}$ into this LCCHODE, we get its characteristic equation

$$0 = n^2 + (-1 + p)n + q = n(n-1) + pn + q,$$

the same characteristic equation that we got from the original Cauchy–Euler ODE (3.59).

So, we can reuse everything about second-order LCCHODEs that we learned in Section 3.3. There, for the critically damped case, we got general solution $Y(t) = c_1 e^{\alpha t} + c_2 t e^{\alpha t}$. But, $t = \ln(r)$ and $r = e^t$, so

$$y(r) = Y(t) = c_1 e^{\alpha t} + c_2 t e^{\alpha t} = c_1 \left(e^t\right)^\alpha + c_2 \ln(r) \left(e^t\right)^\alpha = c_1 r^\alpha + c_2 \ln(r) r^\alpha.$$

So, our two solutions are $y_1(r) = r^\alpha$ and $y_2(r) = r^\alpha \ln(r)$, where α is real and α is the only root of the characteristic polynomial in (3.60).

Indeed, we calculate the Wronskian:

$$W(y_1, y_2)(r) = \begin{vmatrix} r^\alpha & r^\alpha \ln(r) \\ \alpha r^{\alpha-1} & \alpha r^{\alpha-1} \cdot \ln(r) + r^\alpha \cdot \frac{1}{r} \end{vmatrix} = \cdots = r^{2\alpha-1} \neq 0,$$

as long as $r \neq 0$. So, we have

Theorem 3.19

If the characteristic polynomial of the homogeneous second-order Cauchy–Euler ODE (3.59) has only one root, $n = \alpha$, then on any open interval I not containing $r = 0$, the general solution of the ODE is

$$y(r) = c_1 r^\alpha + c_2 \ln(r) r^\alpha,$$

that is, $\{r^\alpha, r^\alpha \ln(r)\}$ is a complete set of basic solutions on I.

Similarly, we can reuse what we learned about the complex solutions case for second-order LCCHODEs in Section 3.3. There, for the underdamped case, we got general solution $Y(t) = c_1 e^{\alpha t} \cos \nu t + c_2 t e^{\alpha t} \sin \nu t$. But $t = \ln(r)$ and $r = e^t$, so

$$y(r) = Y(t) = c_1 \left(e^t\right)^\alpha \cos(\nu \ln(r)) + c_2 \left(e^t\right)^\alpha \sin(\nu \ln(r)),$$

that is,

$$y(r) = c_1 r^\alpha \cos(\nu \ln(r)) + c_2 r^\alpha \sin(\nu \ln(r)).$$

So, our two solutions are $y_1(r) = r^\alpha \cos(\nu \ln(r))$ and $y_2(r) = r^\alpha \sin(\nu \ln(r))$, assuming $\alpha \pm i\nu$, where α, ν are real and $\nu \neq 0$, is a complex conjugate pair of roots of the characteristic polynomial in (3.60). In Problem 3.5.1.10, you will calculate the Wronskian, $W\left(y_1(r), y_2(r)\right)$, and explain why it is nonzero at all $r \neq 0$. So, we have

Theorem 3.20

If the characteristic polynomial of the homogeneous second-order Cauchy–Euler ODE (3.59) has a complex conjugate pair of roots $\alpha \pm i\nu$, where α, ν are real and $\nu \neq 0$, then on any open interval I not containing $r = 0$, the general solution is

$$y(r) = c_1 r^\alpha \cos(\nu \ln(r)) + c_2 r^\alpha \sin(\nu \ln(r)),$$

that is, $\{r^\alpha \cos(\nu \ln(r)), r^\alpha \sin(\nu \ln(r))\}$ is a complete set of basic solutions on I.

Example 3.31

Solve the IVP $\begin{Bmatrix} r^2 y'' + 3ry' + y = 0 \\ y(1) = -2,\ y'(1) = 5 \end{Bmatrix}$.

Method: The characteristic polynomial is $n(n-1) + 3n + 1 = n^2 + 2n + 1 = (n+1)^2$, so the characteristic polynomial has only one root, $n = -1$. By Theorem 3.19, the general solution of the ODE is

$$y(r) = c_1 r^{-1} + c_2 r^{-1} \ln(r),$$

where c_1, c_2 are arbitrary constants. Before substituting in the ICs, first calculate

$$y'(r) = \frac{d}{dr}\left(c_1 r^{-1} + c_2 r^{-1}\ln(r)\right) = -c_1 r^{-2} + c_2\left(-r^{-2}\ln(r) + r^{-1}\cdot\frac{1}{r}\right)$$
$$= r^{-2}\left(-c_1 + c_2(1-\ln(r))\right).$$

Substitute in the ICs to get

$$-2 = y(1) = c_1\cdot 1 + c_2\cdot 0 \quad \text{and} \quad 5 = y'(1) = -c_1 + c_2,$$

so $c_1 = -2$ and $c_2 = 3$. The solution of the IVP is

$$y(r) = -2r^{-1} + 3r^{-1}\ln(r) = r^{-1}\left(-2 + 3\ln(r)\right). \bigcirc$$

Example 3.32

Solve the IVP $\left\{ \begin{array}{c} r^2y'' + 3ry' + 10y = 0 \\ y(1) = -2,\ y'(1) = 5 \end{array} \right\}$.

Method: The characteristic equation is $0 = n(n-1)+3n+10 = n^2+2n+10 = (n+1)^2+9$, so the characteristic polynomial has roots, $n = -1\pm i3$. The general solution of the ODE is

$$y(r) = c_1 r^{-1}\cos(3\ln(r)) + c_2 r^{-1}\sin(3\ln(r)),$$

where c_1, c_2 are arbitrary constants. Before substituting in the ICs, use the product rule and the chain rule to calculate

$$y'(r) = \frac{d}{dr}\left[c_1 r^{-1}\cos(3\ln(r)) + c_2 r^{-1}\sin(3\ln(r))\right]$$
$$= c_1\left(-r^{-2}\cos(3\ln(r)) + r^{-1} - \sin(3\ln(r))\cdot\frac{3}{r}\right)$$
$$+ c_2\left(-r^{-2}\sin(3\ln(r)) + r^{-1}\cos(3\ln(r))\cdot\frac{3}{r}\right)$$
$$= r^{-2}\left(c_1\left(-\cos(3\ln(r)) - 3\sin(3\ln(r))\right) + c_2\left(-\sin(3\ln(r)) + 3\cos(3\ln(r))\right)\right).$$

Substitute in the ICs to get

$$-2 = y(1) = c_1\cdot 1 + c_2\cdot 0 \quad \text{and} \quad 5 = y'(1) = -c_1 + 3c_2,$$

so $c_1 = -2$ and $c_2 = 1$. The solution of the IVP is

$$y(r) = r^{-1}\left(-2\cos(3\ln(r)) + \sin(3\ln(r))\right). \bigcirc$$

Example 3.33

(Pressure in a thick-walled cylinder)* Let E be Young's modulus of elasticity, ν be Poisson's ratio, σ_r be stress in the radial direction, and p be pressure in a thick-walled cylinder. Then we can assume that

$$\sigma_r = \frac{E}{1-\nu^2}\left(\frac{du}{dr} + \nu\frac{u}{r}\right),$$

* (See Ugural and Fenster 2003, specifically Section 8.2).

where we have assumed there is no dependence on the angular variable, θ, and $u = u(r)$ is the deformation, also known as the displacement, from the reference state, for $a \le r \le b$.

Solve the boundary value problem (BVP)

$$\left\{ \begin{array}{l} r^2u''(r) + ru'(r) - u = 0 \\ \\ \sigma_r(a) = -p_i,\ \sigma_r(b) = -p_o \end{array} \right\},$$

where the subscripts i, o refer to the inner and outer walls of the cylinder.

Method: Each of the conditions $\sigma_r(a) = -p_i$ and $\sigma_r(b) = -p_o$ are called **boundary conditions**. The Cauchy–Euler ODE has characteristic equation $0 = n(n-1) + n - 1 = n^2 - 1$, so the characteristic polynomial has roots, $n = \pm 1$. The general solution of the ODE is

$$u(r) = c_1 r + c_2 r^{-1},$$

where c_1, c_2 are arbitrary constants. Before substituting in the boundary conditions, we calculate that

$$\sigma_r(r) = \frac{E}{1-\nu^2}\left(c_1 - c_2 r^{-2} + \nu c_1 + \nu c_2 r^{-2}\right) = \frac{E}{1-\nu^2}\left((1+\nu)c_1 + (\nu-1)c_2 r^{-2}\right).$$

Substitute in the boundary conditions to get

$$\left\{ \begin{array}{l} -\frac{1-\nu^2}{E}p_i = \frac{1-\nu^2}{E}\sigma_r(a) = (1+\nu)c_1 + a^{-2}(\nu-1)c_2 \\ \\ -\frac{1-\nu^2}{E}p_o = \frac{1-\nu^2}{E}\sigma_r(b) = (1+\nu)c_1 + b^{-2}(\nu-1)c_2 \end{array} \right\}.$$

Using the inverse of a 2×2 matrix, we get

$$\begin{aligned} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} &= -\frac{1-\nu^2}{E}\begin{bmatrix} \nu+1 & a^{-2}(\nu-1) \\ \nu+1 & b^{-2}(\nu-1) \end{bmatrix}^{-1}\begin{bmatrix} p_i \\ p_o \end{bmatrix} \\ &= -\frac{1-\nu^2}{E}\frac{1}{(\nu^2-1)(b^{-2}-a^{-2})}\begin{bmatrix} b^{-2}(\nu-1) & -a^{-2}(\nu-1) \\ -(\nu+1) & \nu+1 \end{bmatrix}\begin{bmatrix} p_i \\ p_o \end{bmatrix} \\ &= \frac{-a^2b^2}{E(b^2-a^2)}\begin{bmatrix} (\nu-1)(b^{-2}p_i - a^{-2}p_o) \\ \\ (\nu+1)(-p_i+p_o) \end{bmatrix} = \frac{1}{E(b^2-a^2)}\begin{bmatrix} (1-\nu)(a^2p_i - b^2p_o) \\ \\ (1+\nu)a^2b^2(p_i-p_o) \end{bmatrix}. \end{aligned}$$

It follows that the radial stress is

$$\begin{aligned} \sigma_r &= \frac{E}{1-\nu^2}\left((1+\nu)c_1 + (\nu-1)c_2 r^{-2}\right) \\ &= \frac{1}{(1-\nu^2)(b^2-a^2)}\left((1+\nu)(1-\nu)(a^2p_i - b^2p_o) + (\nu-1)(1+\nu)a^2b^2(p_i-p_o)r^{-2}\right) \\ &= \frac{1}{(b^2-a^2)}\left((a^2p_i - b^2p_o) - a^2b^2(p_i-p_o)r^{-2}\right) \end{aligned}$$

and

$$u(r) = c_1 r + c_2 r^{-1} = \frac{1}{E(b^2-a^2)}\left((1-\nu)(a^2p_i - b^2p_o)r + (1+\nu)a^2b^2(p_i-p_o)r^{-1}\right).\bigcirc$$

3.5.1 Problems

In problems 1–5, solve the ODE, where $'$ denotes $\frac{d}{dr}$.

1. $r^2y'' + 5ry' - 2y = 0$.
2. $r^2y'' + \frac{1}{4}y = 0$.
3. $r^2y'' + ry' + 4y = 0$.
4. $r^2y'' + 3ry' + 3y = 0$.
5. $r^2y'' + 5ry' + 4y = 0$.

In problems 6–9, solve the IVP. Let $'$ denote $\frac{d}{dr}$.

6. $r^2y'' + 6ry' + 6y = 0,\ y(2) = 0,\ y'(2) = 1$.
7. $r^2y'' - 2y = 0,\ y(e) = 0,\ y'(e) = 11$.
8. $r^2y'' - 3ry' + 4y = 0,\ y(e) = 2,\ y'(e) = -3$.
9. $r^2y'' - ry' + 5y = 0,\ y(1) = -2,\ y'(1) = 0$.
10. Explain why the Wronskian,

$$W(r^\alpha \cos(\nu \ln(r)), r^\alpha \sin(\nu \ln(r))$$

is nonzero at all $r \neq 0$, assuming $\nu \neq 0$.

11. In a quarter of a circular plate, a partial differential equation for steady-state heat conduction can lead to the ODE

$$r^2y'' + ry' - (2m)^2y = 0,$$

where m is a nonnegative integer. Find all of the solutions, in terms of m and r. [Hint: Consider separately the cases $m = 0$ and $m \geq 1$.]

12. In a circular plate, the "biharmonic" partial differential equation of solid mechanics can lead to the fourth order ODE

$$r^4y^{(4)} + 2r^3y^{(3)} - (2m^2 + 1)r^2y'' + (2m^2 + 1)ry' + m^2(m^2 - 4)y = 0,$$

where m is a nonnegative integer. This is a fourth-order ODE of Cauchy–Euler type.
   (a) Find the characteristic equation. [Hint: *Mathematica* may help when doing this.]
   (b) Find all of the solutions of the ODE, for $m = 0$, $m = 1$, and $m = 2$.

13. [Small Project] For the ODE of Problem 3.5.1.12, find all of the solutions in terms of m and r. [Hint: *Mathematica* may help when doing this.]

Key Terms

amplitude: (3.33) in Section 3.3
amplitude and phase form: Example 3.19 in Section 3.3
basic solution: Definition 3.2 in Section 3.1, Definition 3.8 in Section 3.3, Definition 3.10 in Section 3.4
bounded as $t \to \infty$: Definition 3.5 in Section 3.1
boundary conditions: Example 3.33 in Section 3.5
Cauchy–Euler ODE: (3.59) in Section 3.5
characteristic equation: before Example 3.16 in Section 3.3, Example 3.24 in Section 3.4
characteristic polynomial: (3.54) in Section 3.4
complete set of basic solutions: Definitions 3.2 in Section 3.1, (3.8) in Section 3.3, (3.10) in Section 3.4
critically damped: before Example 3.16 in Section 3.3
decay constant: Problems 3.2.4.4 through 3.2.4.6
differential operator: Section 3.4.2
envelope: before Example 3.21 in Section 3.3
Euler's Formula: Example 3.17 in Section 3.3
exact: Definition 3.7 in Section 3.2
exactness criterion: Theorem 3.4 in Section 3.2
explicit solution: after Example 3.12 in Section 3.2
general solution: Definitions 3.2 in Section 3.1, (3.3) in Section 3.1, (3.8) in Section 3.3, (3.10) in Section 3.4
half-life: Problems 3.2.4.4 through 3.2.4.6
homogeneous: after (3.2)
implicit solution: after Example 3.1 in Section 3.1
initial condition: before Example 3.1 in Section 3.1
initial value problem (**IVP**): before Example 3.1 in Section 3.1, before (3.25) in Section 3.3
integrating factor: before (3.5) in Section 3.1
logarithmic decrement: before Example 3.21 in Section 3.3
natural frequency: after (3.33) in Section 3.3
neper "frequency": Problems 3.3.8.36 and 3.3.8.37
one compartment model: Problem 3.1.4.32
order: beginning of Section 3.1
overdamped: before Example 3.16 in Section 3.3
particular solution: Definition 3.3 in Section 3.1
phase: (3.33) in Section 3.3
quasi-frequency: Example 3.18 in Section 3.3
quasi-period: after (3.40) in Section 3.3
scalar ordinary differential equation: beginning of Section 3.1
separable: after (3.11) in Section 3.2
series *RLC* circuit: after Theorem 3.13 in Section 3.3
Shift Theorem: Theorem 3.18 in Section 3.4
solution: Definition 3.1 in Section 3.1
spring–mass–damper systems: Section 3.3.1
standard form: (3.2) in Section 3.1
steady-state solution: Definition 3.6 in Section 3.1
time constant: before Example 3.16 in Section 3.16
transient solution: Definition 3.4 in Section 3.1

underdamped: before Example 3.16 in Section 3.3
undamped: Example 3.19 in Section 3.3
Wronskian determinant: Definition 3.9 in Section 3.3, Definition 3.11 in Section 3.4
Zoo of Solutions of LCCHODEs: before (3.54) in Section 3.4

Mathematica Command

$$\text{ContourPlot}[ty^3 - \frac{2}{3}t^3 + \frac{1}{2}y^2,\{t,-4,2\},\{y,-10,10\}, \text{Contours} \to \left\{\frac{28}{3}\right\},$$
$$\text{PerformanceGoal} \to \text{"Quality"}, \text{ContourShading} \to \text{False}]$$

Reference

Ugural, A.C. and Fenster, S.K. *Advanced Strength and Applied Elasticity*. Pearson, Upper Saddle River, NJ, 2003.

4

Scalar ODEs II: Nonhomogeneous Problems

4.1 Nonhomogeneous ODEs

The standard form of the n-th order linear ODE is, again,

$$y^{(n)} + p_1(t)y^{(n-1)} + \cdots + p_{n-1}(t)\dot{y} + p_n(t)y = f(t). \tag{4.1}$$

Definition 4.1

(a) A **particular solution** of an n-th order linear nonhomogeneous ODE (4.1) is any function $y_p(t)$ that satisfies (4.1).

(b) $y(t) = y_p(t)+c_1y_1(t)+c_2y_2(t)+\cdots+c_ny_n(t)$ is the **general solution** of ODE (4.1) if for every
solution $y^*(t)$ of (4.1), there are values for scalar constants $c_1, \ldots, c_n$ giving $y^*(t) = y_p(t) + c_1y_1(t) + c_2y_2(t) + \cdots + c_ny_n(t)$.

Analogous to Theorem 3.1 in Section 3.1, we have the following:

Theorem 4.1

If $y_p(t)$ is any particular solution of n-th order linear nonhomogeneous ODE (4.1) and $\{y_1(t), \ldots, y_n(t)\}$ is any complete set of basic solutions of the corresponding linear homogeneous ODE

$$y^{(n)} + p_1(t)y^{(n-1)} + \cdots + p_{n-1}(t)\dot{y} + p_n(t)y = 0, \tag{4.2}$$

then $y(t) = y_p(t) + c_1y_1(t) + c_2y_2(t) + \cdots + c_ny_n(t)$ is a general solution of ODE (4.1), where $c_1, \ldots, c_n$ are arbitrary constants.

Example 4.1

Suppose ODE $\ddot{y} + y = \cos t$ has a particular solution $y_p(t) = \frac{1}{2}\, t \sin t$. [We will see later in this section how to find this particular solution.] Solve the IVP

$$\left\{\begin{array}{c} \ddot{y} + y = \cos t \\ y(0) = 1,\ \ \dot{y}(0) = 5 \end{array}\right\}.$$

Method: Theorem 4.1 says that

$$y(t) = \frac{1}{2} t \sin t + c_1 y_1(t) + c_2 y_2(t)$$

is a general solution of the ODE, where $\{y_1(t), y_2(t)\}$ is a complete set of basic solutions of the corresponding linear homogeneous ODE $\ddot{y} + y = 0$ and c_1, c_2 are arbitrary constants. We saw how to solve this homogeneous ODE in Section 3.3: The characteristic polynomial, $s^2 + 1$, has a complex conjugate pair of roots, $s = \pm i$, so the general solution of $\ddot{y} + y = 0$ is $y(t) = c_1 \cos t + c_2 \sin t$: The general solution of the original, nonhomogeneous ODE $\ddot{y} + y = \cos t$ is

$$y(t) = \frac{1}{2} t \sin t + c_1 \cos t + c_2 \sin t,$$

where c_1, c_2 are arbitrary constants. Before using the ICs to solve for c_1 and c_2, calculate

$$\dot{y}(t) = \frac{1}{2} \sin t + \frac{1}{2} t \cos t - c_1 \sin t + c_2 \cos t.$$

Substituting $y(t)$ and $\dot{y}(t)$ into the ICs gives

$$\begin{Bmatrix} 1 & = y(0) & = 0 + c_1 \cdot 1 + c_2 \cdot 0, \\ 5 & = \dot{y}(0) & = 0 - c_1 \cdot 0 + c_2 \cdot 1 \end{Bmatrix},$$

so $c_1 = 1$ and $c_2 = 5$. The solution of the IVP is

$$y(t) = \frac{1}{2} t \sin t + \cos t + 5 \sin t. \bigcirc$$

4.1.1 Special Case: RHS from the Zoo and Constant Coefficients on LHS

Suppose a nonhomogeneous linear ODE has the special form

$$y^{(n)} + p_1 y^{(n-1)} + \cdots + p_{n-1} \dot{y} + p_n y = f(t), \tag{4.3}$$

where the coefficients $p_1, \ldots, p_n$ are constants and the right-hand side (RHS), $f(t)$, is a linear combination of functions that are in the zoo of solutions of linear, constant coefficients homogeneous ordinary differential equations (LCCHODEs). In this special case, we will see there is an efficient method for solving nonhomogeneous ODE (4.3).

For example,

$$f(t) = -\cos t + 17 e^{-2t} \sin 5t + 4t^3 - \frac{1}{3} e^{2t}$$

is such an RHS function, that is, is a linear combination of functions that are in the zoo.

Example 4.2

Find all solutions of ODE $\ddot{y} + 2\dot{y} + 5y = e^{-t}$.

Method: Define the differential operator L by $L[y] \triangleq \ddot{y} + 2\dot{y} + 5y$, that is, the left-hand side (LHS) of the nonhomogeneous ODE we're trying to solve. Let's guess a particular solution of the form $y_p(t) = Ae^{-t}$, where A is a *constant* to be determined later. Why? For this form of particular solution, no matter how many derivatives we take of Ae^{-t}, the result will be a multiple of Ae^{-t}, so $L[Ae^{-t}]$ will be a multiple of Ae^{-t} like the RHS, e^{-t}.

We substitute $y_p(t)$ into the ODE to get

$$\underline{e^{-t}} = L[y_p] = L[Ae^{-t}] = (D^2 + 2D + 5)[Ae^{-t}] = Ae^{-t} - 2Ae^{-t} + 5Ae^{-t} = \underline{4Ae^{-t}},$$

so $A = \frac{1}{4}$ and $y_p(t) = \frac{1}{4}e^{-t}$. By Theorem 4.1, the general solution of the nonhomogeneous ODE has the form $y = y_p + y_h = \frac{1}{4}e^{-t} + c_1y_1(t) + c_2y_2(t)$. The characteristic polynomial of the corresponding LCCHODE, $0 = L[y_h] = \ddot{y_h} + 2\dot{y_h} + 5y_h$, is $s^2 + 2s + 5$. Because the roots are $s = -1 \pm i2$, the general solution of the original nonhomogeneous ODE is

$$y = y_p + y_h = \frac{1}{4}e^{-t} + c_1e^{-t}\cos 2t + c_2e^{-t}\sin 2t,$$

where c_1, c_2 are arbitrary constants. ◯

4.1.2 The Method of Coefficients to Be Determined

This method is usually called the **method of undetermined coefficients**, but that makes it sound like we still have coefficients that were not determined after we finish. In fact, the whole point of the method is to figure out the correct form of a particular solution and then determine the values of the coefficients in that form, as we saw in Example 4.2.

Suppose $L = P(D)$ is an n-th order linear constant coefficients differential operator and $f(t)$ is a linear combination of functions in the "zoo." To solve $L[y] = f(t)$, here's the method:

Step 1. Find all n of the roots of the corresponding LCCHODE characteristic polynomial, $\mathcal{P}(s)$, and write them in a list, L_1. The list may include repetitions.

Step 2. Find all roots that correspond to the zoo functions that are in $f(t)$, and write them in a list L_2. The list may include repetitions. Suppose L_2 has m roots, including repetitions.

Step 3. Amalgamate the two lists to get a "superlist," L, which is organized so that all of its repetitions are next to each other. L has $m + n$ roots, including repetitions.

Step 4. Write down the general solution, $y(t)$, of an $(m + n)$-th order LCCHODE whose characteristic polynomial has roots identical to the superlist, L, including repetitions. (We don't need to actually write down that $(m + n)$-th order LCCHODE; we only need its solutions that we get from the superlist.)

Step 5. Let $y_p(t) = y(t) - y_h(t)$, where $y_h(t)$ is the general solution of the LCCHODE $L[y] = 0$. Note that $y_p(t)$ should have m coefficients that are constants to be determined in step 6.

Step 6. Substitute $y_p(t)$ into the original, nonhomogeneous ODE $L[y] = f(t)$, sort by common functions to get m equations for the m coefficients in $y_p(t)$, and solve to get $y_p(t)$.

Step 7. Write the general solution of the original, nonhomogeneous ODE in the form $y(t) = y_p(t) + y_h(t)$.

Step 8. *If* this is an IVP, that is, if there are initial conditions (ICs) to be satisfied, only *after* assembling the *general* solution $y = y_h + y_p$ do we substitute that into the ICs. Do NOT substitute a particular solution or a homogeneous solution into the ICs. □

Example 4.3

(Example 4.2 again) Find the form of a particular solution of ODE $\ddot{y} + 2\dot{y} + 5y = e^{-t}$.

Method: Characteristic polynomial $s^2 + 2s + 5$ gives list $\mathsf{L}_1 = -1 \pm i2$, and RHS $f(t) = e^{-t}$ gives list $\mathsf{L}_2 = -1$, so the superlist is $\mathsf{L} = -1 \pm i2, -1$. This gives $y(t) = c_1e^{-t}\cos 2t + c_2e^{-t}\sin 2t + c_3e^{-t}$, so $y_h(t) = c_1e^{-t}\cos 2t + c_2e^{-t}\sin 2t$ gives $y_p(t) = c_3e^{-t}$. The correct form of a particular solution is $y_p(t) = Ae^{-t}$, where A is a constant to be determined later after substituting y_p into $\ddot{y} + 2\dot{y} + 5y = e^{-t}$. ◯

Example 4.4

Find the form of a particular solution of ODE $\ddot{y} + 3\dot{y} + 2y = e^{-t}$.

Method: Characteristic polynomial $s^2 + 3s + 2$ gives list $\mathsf{L}_1 = -1, -2$, and RHS $f(t) = e^{-t}$ gives list $\mathsf{L}_2 = -1$, so the superlist is $\mathsf{L} = -1, -1, -2$. This gives $y(t) = c_1e^{-t} + c_2te^{-t} + c_3e^{-2t}$, so $y_h(t) = c_1e^{-t} + c_3e^{-2t}$ gives $y_p(t) = c_2te^{-t}$. The correct form of a particular solution is $y_p(t) = Ate^{-t}$, where A is a constant to be determined later after substituting y_p into $\ddot{y} + 4\dot{y} + 3y = e^{-t}$. ◯

There can be an interaction(s) between the differential operator, L, on the LHS of the nonhomogeneous ODE and the function $f(t)$ on the RHS. Some textbooks ask the reader to learn many special rules about how to choose the form of a particular solution. We prefer to learn one method that does not require remembering special cases

Example 4.5

Find the form of a particular solution of ODE $\ddot{y} + 4y = e^{-t} - \cos 2t$.

Method: Characteristic polynomial $s^2 + 4$ gives list $\mathsf{L}_1 = \pm i2$, and RHS $f(t) = e^{-t} - \cos 2t$ gives list $\mathsf{L}_2 = -1, \pm i2$, so the superlist is $\mathsf{L} = -1, \pm i2, \pm i2$. This gives $y(t) = c_1e^{-t} + c_2\cos 2t + c_3\sin 2t + c_4t\cos 2t + c_5t\sin 2t$, so $y_h(t) = c_2\cos 2t + c_3\sin 2t$ gives $y_p(t) = c_1e^{-t} + c_4t\cos 2t + c_5t\sin 2t$. The correct form of a particular solution is

$$y_p(t) = Ae^{-t} + Bt\cos 2t + Ct\sin 2t,$$

where A, B, C are constants to be determined. ◯

Example 4.6

Solve the IVP $\begin{Bmatrix} \ddot{y} + 3\dot{y} + 2y = e^{-t} \\ y(0) = -1,\ \dot{y}(0) = 2 \end{Bmatrix}$.

Method: From Example 4.4, the correct form of a particular solution is $y_p(t) = Ate^{-t}$, where A is a constant to be determined. First, we calculate derivatives

$$\dot{y_p} = A(1 \cdot e^{-t} - te^{-t}), \quad \ddot{y_p} = A(-2e^{-t} + te^{-t}).$$

Substituting these into the original nonhomogeneous ODE gives

$$\underline{1} \cdot e^{-t} = \ddot{y_p} + 3\dot{y_p} + 2y_p = A\left((-2e^{-t} + \cancel{te^{-t}}) + 3(e^{-t} - \cancel{te^{-t}}) + \cancel{2te^{-t}}\right) = \underline{A}e^{-t}. \qquad (4.4)$$

This determines that $A = 1$, so $y_p(t) = te^{-t}$. While solving Example 4.4, we found the solution of the corresponding LCCHODE, $y_h(t) = c_1e^{-t} + c_3e^{-2t}$. The general solution of the ODE is

$$y(t) = te^{-t} + c_1e^{-t} + c_3e^{-2t}.$$

While substituting y_p into the nonhomogeneous ODE, we calculated $\dot{y_p} = A(1-t)e^{-t}$, so $\dot{y} = (1-t)e^{-t} - c_1e^{-t} - 2c_2e^{-2t}$. We substitute the general solution into the ICs:

$$\left\{\begin{matrix} -1 & = y(0) & = 0 + c_1 + c_2 \\ 2 & = \dot{y}(0) & = 1 - c_1 - 2c_2 \end{matrix}\right\}.$$

This can be written as

$$\begin{bmatrix} 1 & 1 \\ -1 & -2 \end{bmatrix}\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

Using the inverse of the 2×2 matrix gives

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \frac{1}{-1}\begin{bmatrix} -2 & -1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \end{bmatrix},$$

so the solution of the IVP is

$$y(t) = te^{-t} - e^{-t} = (t-1)e^{-t}. \bigcirc$$

We notice that in (4.4), several terms canceled. This was not a coincidence, as we will see in Example 4.9.

Example 4.7

For the AC–LC circuit shown in Figure 4.1, find the current as a function of time, assuming the initial charge on the capacitor is 3 C and the initial current is 1 A. Also, find the steady-state current and its amplitude.

Method: Let $q = q(t)$ be the charge on the capacitor, in coulombs. Kirchhoff's voltage law yields ODE $L\ddot{q} + \frac{1}{C}q = V(t)$. The picture shows $L = 1$ H, $C = \frac{1}{9}$ F, and $V(t) = e^{-2t}$ volts. The IVP is

$$\left\{\begin{matrix} \ddot{q} + 9q = e^{-2t} \\ q(0) = 3,\ \dot{q}(0) = I(0) = 1 \end{matrix}\right\}.$$

The corresponding LCCHODE $\ddot{q} + 9q = 0$ has $s = \pm i3$ as the roots of its characteristic polynomial's, so $q_h(t) = c_1 \cos 3t + c_2 \sin 3t$, where c_1, c_2 are arbitrary constants.

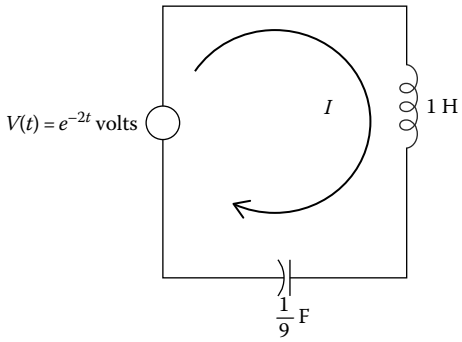


FIGURE 4.1
AC LC series circuit in Example 4.7.

The superlist is $\mathsf{L} = -2, \pm i3$, so $q(t) = c_1 \cos 3t + c_2 \sin 3t + c_3 e^{-2t}$, and thus, $q_p(t) = Ae^{-2t}$ is the correct form of a particular solution. Substituting that into the original, nonhomogeneous ODE gives

$$\underline{1} \cdot e^{-2t} = \ddot{q_p} + 9q_p = 4Ae^{-2t} + 9Ae^{-2t} = \underline{13A}e^{-2t};$$

hence, the general solution of the nonhomogeneous ODE is

$$q(t) = \frac{1}{13}e^{-2t} + c_1 \cos 3t + c_2 \sin 3t.$$

Substitute that into the ICs to get

$$\left\{\begin{array}{l} 3 = q(0) = \frac{1}{13} + c_1 \\ \\ 1 = \dot{q}(0) = -\frac{2}{13} - 0 \cdot c_1 + 3c_2 \end{array}\right\}.$$

This yields $c_1 = \frac{38}{13}$ and $c_2 = \frac{5}{13}$, and the solution of the IVP is

$$q(t) = \frac{1}{13}e^{-2t} + \frac{38}{13}\cos 3t + \frac{5}{13}\sin 3t.$$

Because $\lim_{t\to\infty} e^{-2t} = 0$, the transient solution is $q_T = \frac{1}{13}e^{-2t}$. Because, $q(t) - q_T = \frac{38}{13}\cos 3t + \frac{5}{13}\sin 3t$ is bounded as $t \to \infty$, the steady-state charge is

$$q_S(t) = \frac{38}{13}\cos 3t + \frac{5}{13}\sin 3t.$$

The steady-state current is

$$I_S(t) = \dot{q_S}(t) = \frac{3}{13}\left(-38\sin 3t + 5\cos 3t\right),$$

whose amplitude is

$$\frac{3}{13} \cdot \sqrt{(-38)^2 + 5^2} = \frac{3}{13}\sqrt{38^2 + 5^2} = \frac{3}{13}\sqrt{13 \cdot 113} = 3 \cdot \sqrt{\frac{113}{13}}. \bigcirc$$

In Example 4.7, the steady-state solution is not a particular solution, and the transient solution is not a homogeneous solution.

Example 4.8

For the IVP $\left\{\begin{array}{l} \ddot{y} + 2\dot{y} + 9y = \sin 2t \\ y(0) = 3,\ \dot{y}(0) = -2 \end{array}\right\}$, find the steady-state solution and its amplitude.

Method: The corresponding LCCHODE $\ddot{y} + 2\dot{y} + 9y = 0$ has its characteristic polynomial's roots being $-1 \pm i\sqrt{8}$, so $y_h(t) = c_1 e^{-t}\cos(\sqrt{8}\,t) + c_2 e^{-t}\cos(\sqrt{8}\,t)$, where c_1, c_2 are arbitrary constants. Because $\lim_{t\to\infty} y_h(t) = 0$, the steady-state solution can only possibly be $y_p(t)$.

The superlist is $\mathsf{L} = -1 \pm i\sqrt{8}, \pm i2$, so $y(t) = y_h(t) + c_3 \cos 2t + c_4 \sin 2t$, and thus, $y_p(t) = A\cos 2t + B\sin 2t$ is the correct form of a particular solution, where A, B are constants to be determined. Substitute that into the original, nonhomogeneous ODE to get

$$\begin{aligned} \sin 2t &= \ddot{y_p} + 2\dot{y_p} + 9y_p \\ &= (-4A\cos 2t - 4B\sin 2t) + 2(-2A\sin 2t + 2B\cos 2t) + 9(A\cos 2t + B\sin 2t) \\ &= (5A + 4B)\cos 2t + (-4A + 5B)\sin 2t. \end{aligned}$$

To match the LHS with the RHS, we sort by functions in common to get

$$\underline{0} \cdot \cos 2t + \underset{\sim}{1} \cdot \sin 2t = \underline{(5A + 4B)} \cos 2t + \underset{\sim\sim\sim\sim\sim}{(-4A + 5B)} \sin 2t;$$

hence,

$$\begin{Bmatrix} 0 = 5A + 4B \\ 1 = -4A + 5B \end{Bmatrix}.$$

Using the inverse of a 2×2 matrix, we get

$$\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} 5 & 4 \\ -4 & 5 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \frac{1}{41} \begin{bmatrix} 5 & -4 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \frac{1}{41} \begin{bmatrix} -4 \\ 5 \end{bmatrix}.$$

The steady-state solution of the IVP is

$$y_S(t) = y_p(t) = \frac{1}{41}(-4 \cos 2t + 5 \sin 2t),$$

and its amplitude is $\frac{1}{41}\sqrt{(-4)^2 + 5^2} = \frac{1}{\sqrt{41}}$. ○

We were only asked for the steady-state solution, so *in this problem*, we didn't need to use the ICs to solve for the constants c_1, c_2.

4.1.3 Justification for the Method

Suppose we want to solve a nonhomogeneous ODE $L[y] = f(t)$, where L is an n-th order linear constant coefficients ordinary differential operator and $f(t)$ is a linear combination of functions in the zoo of solutions of LCCHODEs.

Corresponding to $f(t)$ we have a list L_2 of m roots. From that list, we can construct an m-th order linear constant coefficients ordinary differential operator $\overline{L}$ that "annihilates" $f(t)$, in the sense that $\overline{L}[f(t)] = 0$. For example, in Example 4.5, for $f(t) = e^{-t} - \cos 2t$, we had list of roots $\mathsf{L}_2 = -1, \pm i2$, so $\overline{L} \triangleq (D + 1)(D - i2)(D + i2) = (D + 1)(D^2 + 4)$ annihilates $f(t)$. This is just a short way of saying that $(D+1)(D^2+4)[f(t)] = 0$, that is, $f(t)$ is a solution of LCCHODE $(D + 1)(D^2 + 4)[y] = 0$. [Note that $f(t)$, the RHS of the original, nonhomogeneous ODE $L[y] = f$, need not be a general solution, just some solution, of $\overline{L}[y] = 0$.]

We will explain why $y_p(t)$ is a solution of the $(n + m)$-order LCCHODE $\overline{L}L[y] = 0$ if $y_p(t)$ is a particular solution of the original nonhomogeneous $L[y] = f(t)$. This is because

$$L[y_p] = f \text{ implies } \overline{L}L[y_p] \triangleq \overline{L}[L[y_p]] = \overline{L}[f] = 0,$$

because $\overline{L}$ was chosen so as to annihilate f.

Moreover, we might as well let the form of a particular solution be $y_p(t) \triangleq y(t) - y_h(t)$, where $y(t)$ is the general solution of $\overline{L}L[y] = 0$ and $y_h(t)$ is the general solution of the corresponding n-th order LCCHODE $L[y] = 0$. Why? Adding $y_h(t)$ to $y_p(t)$ does not help to find $y_p(t)$, because if we did, we would wind up solving $f = L[y_p + y_h] = L[y_p] + L[y_h] = L[y_p] + 0$. □

4.1.4 Using a Shift Theorem

Recall from Theorem 3.18 in Section 3.4, a shift theorem, that for any positive integer n,

$$(D-\alpha)^n[e^{\alpha t}g(t)] = e^{\alpha t}D^n g(t) = e^{\alpha t}g^{(n)}(t).$$

For example, $(D-\alpha)[e^{\alpha t}g(t)] = e^{\alpha t}\dot{g}(t)$.

In Example 4.4, we found that $y_p(t) = Ate^{-t}$ is the correct form of a particular solution of ODE $\ddot{y}+3\dot{y}+2y = e^{-t}$. In Example 4.4, we substituted y_p into the nonhomogeneous ODE and then solved for the constant A. Here we will see an alternative way to substitute in y_p that takes advantage of a shift theorem.

Example 4.9

(Part of Example 4.6 again, done in a new way) Find a particular solution of $\ddot{y}+3\dot{y}+2y = e^{-t}$.

Method: Substitute $y_p(t) = Ate^{-t}$ into the ODE and use the (shift) Theorem 3.18 in Section 3.4 with $\alpha = -1$ and $f(t) = t$ to get

$$e^{-t} = L[\,y_p(t)] = (D^2+3D+2)[\,Ate^{-t}] = A(D+2)(D+1)[\,te^{-t}]$$
$$= A(D+2)[(D+1)[\,te^{-t}]] = A(D+2)[\,e^{-t}D[\,t]\,] = A(D+2)[\,e^{-t}] = Ae^{-t},$$

so $A = 1$, and $y_p(t) = te^{-t}$ is a particular solution. ◯

Each person can decide for himself or herself whether it is better/easier/more reliable to use a shift theorem or just take two derivatives of $y_p(t) = Ate^{-t}$ before substituting into the nonhomogeneous ODE. In more complicated examples, for example, higher-order ODEs, a shift theorem may become more advantageous in terms of accuracy and ease of use.

In any case, we can use a shift theorem as a way to check our work by using a different technique. In real life, any methods we have for checking our work are useful for reducing the number and severity of analytical errors.

Theorem 4.2

(Nonhomogeneous superposition principle) If $L[y_{1,p}] = f_1(t)$ and $L[y_{2,p}] = f_2(t)$, then $y(t) = y_{1,p} + y_{2,p}$ is a particular solution of the nonhomogeneous ODE $L[\,y\,] = f_1(t) + f_2(t)$.

Example 4.10

Solve $\ddot{y}+3\dot{y}+2y = 3e^{-t}+4e^{-4t}$.

Method: First, rewrite our ODE using differential operators:

$$(D+1)(D+2)[y_p] = 3e^{-t}+4e^{-4t}. \tag{4.5}$$

Using Theorem 4.2, we write a particular solution of the ODE in the form $y_p(t) = y_{1,p} + y_{2,p}$ where $(D+1)(D+2)[y_{1,p}] = 3e^{-t}$ and $(D+1)(D+2)[y_{2,p}] = 4e^{-4t}$.

For $(D+1)(D+2)[y_{1,p}] = 3e^{-t}$, our lists are $\mathsf{L}_1 = -1, -2$ and $\mathsf{L}_2 = -1$, so the superlist is $\mathsf{L} = -1, -1, -2$. We get $y(t) = c_1e^{-t} + c_2te^{-t} + c_3e^{-2t}$ and $y_h(t) = c_1e^{-t} + c_3e^{-2t}$, so

$$y_{1,p}(t) = Ate^{-t},$$

where A is a constant to be determined. Substitute that into the first nonhomogeneous problem and use a shift theorem to calculate

$$\underline{3}e^{-t} = (D+1)(D+2)[Ate^{-t}] = (D+2)[(D+1)[Ate^{-t}]] = (D+2)[Ae^{-t}D[t]]$$
$$= (D+2)[Ae^{-t}] = (-1+2)Ae^{-t} = \underline{A}e^{-t};$$

hence, $A = 3$.

For the second nonhomogeneous problem, $(D+1)(D+2)[y_{2,p}] = 4e^{-4t}$, our lists are $\mathsf{L}_1 = -1, -2$ and $\mathsf{L}_2 = -4$, so the superlist is $\mathsf{L} = -1, -2, -4$. We get $y(t) = c_1e^{-t} + c_2e^{-2t} + c_4e^{-4t}$ and $y_h(t) = c_1e^{-t} + c_2e^{-2t}$, so

$$y_{2,p}(t) = Be^{-4t},$$

Substitute that into the second nonhomogeneous problem and calculate

$$\underline{4}e^{-4t} = (D+1)(D+2)[Be^{-4t}] = (-4+1)(-4+2)Be^{-4t} = \underline{\underline{6B}}e^{-4t};$$

hence, $B = \frac{2}{3}$.

So, $y_p(t) = y_{1,p} + y_{2,p} = 3te^{-t} + \frac{2}{3}e^{-4t}$. The general solution of the original ODE is

$$y_p(t) = e^{-4t} + 3te^{-t} + c_1e^{-t} + c_2e^{-2t},$$

where c_1, c_2 are arbitrary constants. ◯

4.1.5 Problems

In problems 1–8, solve the ODE.

1. $\ddot{y} + 5\dot{y} + 6y = 3e^{-t}$
2. $\ddot{y} + 4\dot{y} + 6y = t + 3e^{-2t}$
3. $\ddot{y} + 5\dot{y} + 6y = 2e^{-3t}$
4. $y'' + 2y = xe^{-x}$, where $'$ denotes $\frac{d}{dx}$
5. $\ddot{y} - y = e^{-t} + 5e^{-2t}$
6. $\ddot{y} + 4\dot{y} + 5y = \sin 2t$
7. $\ddot{y} + \dot{y} - 12y = 5e^{-4t}$
8. $\ddot{y} + 4y = e^{-t}\cos 2t$
9. ODE ($\star$) $\ddot{y} + 3\dot{y} + 2y = \cos(e^t)$ has a particular solution $y(t) = -e^{-2t}\cos(e^t)$. Solve the IVP consisting of ODE ($\star$) and ICs $y(0) = 0$, $\dot{y}(0) = 0$.
10. Suppose that $y_1(x) = \sin x$ solves ODE ($\star$) $y'(x) + 2xy = \cos x + 2x\sin x$, where $' = \frac{d}{dx}$.
    (a) Find the general solution of ODE ($\star$).
    (b) Solve the IVP consisting of ODE ($\star$) and IC $y(0) = 5$.
11. Given that ODE $y(t) = -\frac{1}{9}te^t + \frac{1}{6}t^2e^t$ solves ($\star$) $\ddot{y} + \dot{y} - 2y = te^t$, find the solution of the IVP consisting of ODE ($\star$) and ICs $y(0) = 0, \dot{y}(0) = -2$.
12. Suppose α is a nonzero constant. Find all solutions of $\dot{y} + \alpha y = e^{-t}$ in as simple form as possible. [Caution: $\alpha = 1$ may be a special case.] Your final conclusion(s) should be in terms of α.

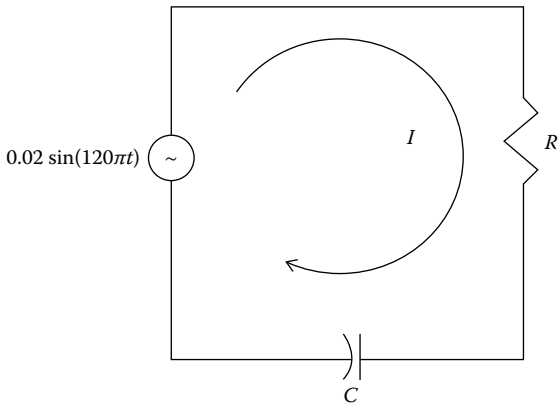


FIGURE 4.2
AC–RC series circuit in Problem 4.1.5.16.

In problems 13–15, solve the IVP.

13. $\ddot{y} - y = e^{-2t},\ y(0) = \dot{y}(0) = 0$
14. $\ddot{y} - y = e^{t},\ y(0) = \dot{y}(0) = 0$
15. $\ddot{y} + \dot{y} + 5y = 10,\ y(0) = \dot{y}(0) = 0$
16. For the AC–*RC*− series circuit shown in Figure 4.2, find the solution of the IVP $R\dot{q} + \frac{1}{C}\,q = 0.02\sin(120\pi\, t),\ q(0) = 0.001$. Let $q = q(t)$ be the charge on the capacitor, in units of coulombs, and measure R in ohms and C in farads. Note that $\omega = 120\pi$ is a frequency of 60 Hz, commonly used in electrical systems. As usual, assume that R and C are constants.
    (a) Using the method of undetermined coefficients
    (b) Using the integrating factor method of Section 3.1

In problems 17–24, find the steady-state solution and its amplitude.

17. $\ddot{y} + 2\dot{y} + 5y = \sin 2t$
18. $\ddot{y} + 4\dot{y} + 5y = \sin t$
19. $\ddot{y} + 4\dot{y} + 5y = \cos t$
20. $\ddot{y} + 3\dot{y} + 5y = 6\cos 4t$
21. $\ddot{y} + 2\dot{y} + 5y = 3,\ y(\pi) = 7,\ \dot{y}(\pi) = 0$
22. $\ddot{y} + 9y = 3e^{-2t},\ y(0) = 0,\ \dot{y}(0) = -4$
23. $2\ddot{y} + 4y = f_0 e^{-2t},\ y(0) = \dot{y}(0) = 0$
24. $m\ddot{y} + ky = f_0 e^{-2t},\ y(0) = 2,\ \dot{y}(0) = -1$
    Your final conclusion should be in terms of the unspecified positive constant parameters m, k, f_0.

In problems 25 and 26, solve the IVP.

25. $\ddot{y} - 2\dot{y} + 2y = t,\ y(\pi) = 0,\ \dot{y}(\pi) = 0$

26. $\ddot{y} + 2\dot{y} = e^{-2t},\ y(0) = 0,\ \dot{y}(0) = -5$
27. Given that $\cos t$ is one solution of $\dddot{y} + 2\ddot{y} + \dot{y} + 2y = 0$, find the general solution of $\dddot{y} + 2\ddot{y} + \dot{y} + 2y = 4 - 3e^{-2t}$.
28. Suppose that $y_3(t)$ solves ODE $m\ddot{y} + b\dot{y} + ky = g(t)$ and $y_0(t)$ solves ODE $m\ddot{y} + b\dot{y} + ky = 0$. Let $y(t) = 2y_0(t) - y_3(t)$. What ODE does $y(t)$ satisfy, and why?
29. Suppose that $y_1(t) = -2t + 3$ solves ODE $\ddot{y} + 2\dot{y} + 5y = -10t + 11$ and that $y_2(t) = -\cos 2t - 4\sin 2t$ solves ODE $\ddot{y} + 2\dot{y} + 5y = -17\cos 2t$. Find the general solution of $\ddot{y} + 2\dot{y} + 5y = t - \frac{11}{10} + 2\cos 2t$.
30. Given that $y_1(t)$ and $y_2(t)$ solve the same ODE $(\star)$ $\ddot{y} + p(t)\dot{y} + q(t)y = e^t$ on an interval I and that $p(t)$ and $q(t)$ are continuous on that interval.
    (a) Which of the following is true? (a1) $2y_1(t) - y_2(t)$ is a solution of $(\star)$ on I, (a2) $2y_1(t) - y_2(t)$ is not a solution of $(\star)$ on I, or (a3) without more information, we cannot tell whether (a1) or (a2) is true.
    (b) Which of the following is true? (b1) $c_1y_1(t) + c_2y_2(t)$ is a general solution of $(\star)$ on I, (b2) $c_1y_1(t) + c_2y_2(t)$ is not a general solution of $(\star)$ on I, or (b3) without more information, we cannot tell whether (b1) or (b2) is true.

In problems 31 and 32, a function $y(t)$ is given. Find at least two of the possible second-order ODEs for which $y(t)$ is one of the solutions.

31. $y(t) = e^{-t} + e^{-2t} - e^{-3t}$
32. $y(t) = e^{-t} + te^{-t} + 2e^{-2t}$

4.2 Forced Oscillations

A series *RLC* circuit with AC voltage source, $V(t) = \frac{V_0}{\omega}\sin\omega t$, where V_0 is a constant, has loop current $I(t)$ satisfying Kirchhoff's voltage equation,

$$L\frac{dI}{dt} + RI + \frac{1}{C}q = V(t), \tag{4.6}$$

where the charge on the capacitor, q, satisfies $\dot{q} = I$. Assume that L, R, and C are, as usual, constants. Differentiate (4.6) to get

$$L\ddot{I} + R\dot{I} + \frac{1}{C}I = V_0\cos\omega t. \tag{4.7}$$

As illustrated in Figure 4.3, essentially the same ODE can appear in vibrations where there is a sinusoidal external force:

$$m\ddot{y} + b\dot{y} + ky = f_0\cos\omega t, \tag{4.8}$$

where f_0 is a constant.

We will see that the graphs of solutions of (4.8) take one of the shapes shown in Figure 4.4. We will use the method of undetermined coefficients to solve (4.8) or (4.7).

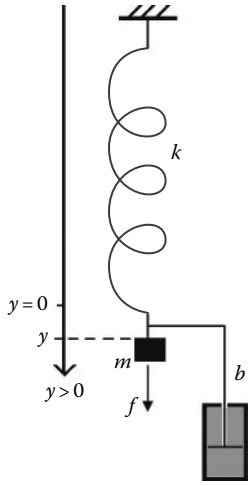


FIGURE 4.3
Forced spring–mass–damper system.

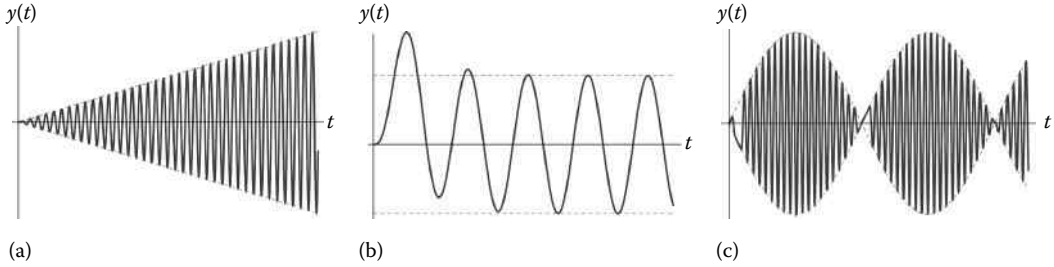


FIGURE 4.4
Forced oscillator examples: (a) pure resonance, (b) steady state, and (c) beats phenomenon.

4.2.1 The Resonance Case

Example 4.11

Solve the IVP

$$\left\{ \begin{matrix} \ddot{y} + 4y = 5\sin 2t \\ y(0) = 0,\ \dot{y}(0) = 0 \end{matrix} \right\}. \tag{4.9}$$

Method: The RHS function $5\sin 2t$ is in the zoo, corresponding to list of roots $\mathsf{L}_2 = \pm i2$. The corresponding LCCHODE, $\ddot{y} + 4y = 0$, gives list of roots $\mathsf{L}_1 = \pm i2$. The superlist is $\mathsf{L} = \pm i2, \pm i2$, so $y(t) = c_1 \cos 2t + c_2 \sin 2t + c_3 t\cos 2t + c_4 t \sin 2t$. Because $y_h(t) = c_1 \cos 2t + c_2 \sin 2t$, the correct form of a particular solution is

$$y_p(t) = At\cos 2t + Bt\sin 2t,$$

where A, B are constants to be determined. Using the product and the chain rules, we calculate

$$\dot{y}_p(t) = A(-2t\sin 2t + \cos 2t) + B(2t\cos 2t + \sin 2t)$$

and

$$\ddot{y}_p(t) = A(-4t\cos 2t - 4\sin 2t) + B(-4t\sin 2t + 4\cos 2t).$$

Substitute them into the original, nonhomogeneous ODE in (4.9) to get

$$\begin{aligned}\underline{0}\cdot\cos 2t + \underset{\sim}{5}\cdot\sin 2t &= \ddot{y_p} + 4y_p \\ &= A(\cancel{-4t\cos 2t} - 4\sin 2t) + B(\cancel{-4t\sin 2t} + 4\cos 2t) + 4(\cancel{At\cos 2t} + \cancel{Bt\sin 2t}) \\ &= \underline{4B}\cos 2t + \underset{\sim\sim}{(-4A)}\sin 2t,\end{aligned}$$

so $0 = 4B$ and $5 = -4A$. The general solution of ODE (4.9) is

$$y(t) = y_p(t) + y_h(t) = -\frac{5}{4}\,t\cos 2t + c_1\cos 2t + c_2\sin 2t,$$

where c_1, c_2 are arbitrary constants. It's useful to calculate that

$$\dot{y}(t) = \frac{5}{2}\,t\sin 2t - \frac{5}{4}\,\cos 2t - 2c_1\sin 2t + 2c_2\cos 2t.$$

Substituting into the ICs gives $0 = y(0) = c_1$, $0 = \dot{y}(0) = -\frac{5}{4} + 2c_2$. The solution of the IVP is

$$y(t) = -\frac{5}{4}\,t\cos 2t + \frac{5}{8}\sin 2t. \bigcirc$$

The graph of the solution is shown in Figure 4.5. The dashed lines $y = \pm\frac{5}{4}\,t$ give an approximate "envelope" for the graph of the solution; we say "approximate" because the graph escapes the envelope less and less as $t \to \infty$.

We can think of the particular solution $y_p(t) = -\frac{5}{4}\,t\cos 2t$ as being in the form

$$y_p(t) = A(t)\cos 2t,$$

where the time-dependent, growing amplitude is $A(t) = \frac{5}{4}\,t$.

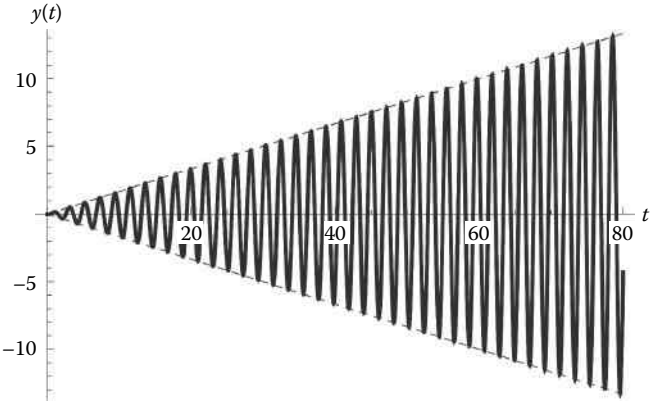


FIGURE 4.5
Example 4.11.

Note also that there is no transient solution, according to Definition 3.4 in Section 3.1, even though the term $\frac{5}{8}\sin 2t$ becomes insignificant compared to the term $y(t) = -\frac{5}{4}t\cos 2t$, for more and more values of t, as $t \to \infty$. This problem also does not have a steady-state solution, according to Definition 3.6 in Section 3.1, because (i) there is no transient solution and (ii) $y(t)$ is not bounded as $t \to \infty$, due to unbounded oscillation.

In Example 4.11, there was no damping, that is, $b = 0$, and there was a perfect match of the forcing frequency, $\omega = 2$, and the natural frequency, $\omega_0 = \sqrt{\frac{k}{m}} = 2$. We call this the **resonance** case. In general, this can be summarized as

$$\text{The resonance case : } b = 0 \text{ and } \omega = \omega_0.$$

In retrospect, when we did Example 4.1 in Section 4.1, we could have immediately known that this was in the resonance case by looking at the ODE and noticing that $b = 0$ and $\omega = \omega_0 = 1$.

Physically, in the resonance case, the mechanical system has *no* friction that dissipates energy, contrary to our experience in the real world, and the external force is continually pumping in energy at just the right frequency to match the unforced, natural frequency. An apocryphal story is that resonance was discovered when a Roman legion, marching in perfect formation, set a bridge into vibration of sufficient amplitude to send it crashing down.

In the real world, a mechanical system, and all but some superconducting electrical systems, always have some dissipation of energy due to friction or electrical resistance. Nevertheless, as we will see shortly, the system can have "practical resonance" solutions that are mathematically bounded as $t \to \infty$ *but* that get so large as to crash the system.

4.2.2 Steady-State Solution, Frequency Response, and Practical Resonance

As long as there is some damping, that is, $b > 0$, there cannot be pure resonance. Instead, there can be a steady-state oscillation.

Example 4.12

Let m, b, k, f_0 be unspecified constants, with m, b, k being positive. Find the steady-state solution and its amplitude for the IVP

$$\left\{\begin{array}{c} m\ddot{y} + b\dot{y} + ky = f_0\cos\omega t \\ y(0) = y_0,\ \dot{y}(0) = \dot{y}_0 \end{array}\right\}, \tag{4.10}$$

where ω is a positive constant.

Method: This example is general, that is, we don't substitute in specific values for the constant parameters m, b, k, f_0. Nevertheless, we can use the method of undetermined coefficients.

The RHS function $f_0\cos\omega t$ is in the zoo, corresponding to list of roots $\mathsf{L}_2 = \pm i\omega$. The corresponding LCCHODE, $m\ddot{y} + b\dot{y} + ky = 0$, gives list of roots $\mathsf{L}_1 = s_1, s_2$ where

$$s_1, s_2 = -\frac{b}{2m} \pm \frac{\sqrt{b^2 - 4mk}}{2m}.$$

The superlist is $\mathsf{L} = \pm i\omega, s_1, s_2$.

No matter whether the two roots s_1, s_2 are in the overdamped, critically damped, or underdamped case, the assumption that b is positive implies that the solutions of the corresponding homogeneous ODE go to zero as $t \to \infty$. Why? Specifically, in the overdamped case, both of s_1, s_2 are real and negative. In the underdamped case, the solutions are of the form $y(t) = Ae^{\alpha t}\cos(\nu t - \delta)$, where A is a constant and $\alpha = -\frac{b}{2m} < 0$. In the critically damped case, solutions are of the form $y(t) = (c_1 + c_2 t)e^{\alpha t}$, where $\alpha = -\frac{b}{2m} < 0$, so L'Hôpital's rule implies that solutions go to zero as $t \to \infty$.

It follows that $y_h(t) \to 0$ as $t \to \infty$, that is, $y_h(t)$ is a transient solution, according to Definition 3.4 in Section 3.1. It also follows that the correct form of the particular solution is

$$y_p(t) = A\cos\omega t + B\sin\omega t,$$

where A, B are constants to be determined. We calculate

$$\dot{y}_p(t) = \omega(-A\sin\omega t + B\cos\omega t) \quad \text{and} \quad \ddot{y}_p(t) = -\omega^2(A\cos\omega t + B\sin\omega t).$$

Substitute them into the original, nonhomogeneous ODE in (4.10) to get

$$\begin{aligned} &\underline{f_0}\cdot\cos\omega t + \underset{\sim}{0}\cdot\sin\omega t \\ &= -m\omega^2 A\cos\omega t - m\omega^2 B\sin\omega t - b\omega A\sin\omega t + b\omega B\cos\omega t + kA\cos\omega t + kB\sin\omega t \\ &= \underline{\left((k-m\omega^2)A + b\omega B\right)}\cos\omega t + \underbrace{\left((k-m\omega^2)B - b\omega A\right)}\sin\omega t, \end{aligned}$$

so $f_0 = (k-m\omega^2)A + b\omega B$ and $0 = (k-m\omega^2)B - b\omega A$. This can be written as a 2×2 system

$$\left\{\begin{array}{rcrl} (k-m\omega^2)A & + & b\omega B & = f_0 \\ -b\omega A & + & (k-m\omega^2)B & = 0 \end{array}\right\},$$

whose solution is

$$\begin{aligned} \begin{bmatrix} A \\ B \end{bmatrix} &= \begin{bmatrix} k-m\omega^2 & b\omega \\ -b\omega & k-m\omega^2 \end{bmatrix}^{-1}\begin{bmatrix} f_0 \\ 0 \end{bmatrix} \\ &= \frac{1}{(k-m\omega^2)^2 + (b\omega)^2}\begin{bmatrix} k-m\omega^2 & -b\omega \\ b\omega & k-m\omega^2 \end{bmatrix}\begin{bmatrix} f_0 \\ 0 \end{bmatrix}. \end{aligned}$$

So, a particular solution of the IVP is given by

$$y_p(t) = \frac{f_0}{(k-m\omega^2)^2 + (b\omega)^2}\left((k-m\omega^2)\cos\omega t + (b\omega)\sin\omega t\right).$$

Earlier discussion explained that this is also the steady-state solution because $y_h(t)$ is a transient solution, and the initial data $y_0, \dot{y}_0$ don't affect this! The steady-state solution is

$$y_s(t) = y_p(t),$$

which has

$$\text{Amplitude} = \sqrt{\left(\frac{(k-m\omega^2)f_0}{(k-m\omega^2)^2 + (b\omega)^2}\right)^2 + \left(\frac{b\omega f_0}{(k-m\omega^2)^2 + (b\omega)^2}\right)^2}$$

$$= \sqrt{\frac{\left((k-m\omega^2)^2+(b\omega)^2\right)}{\left((k-m\omega^2)^2+(b\omega)^2\right)^2} f_0^2},$$

that is,

$$\text{Amplitude} = \frac{|f_0|}{\sqrt{(k-m\omega^2)^2+(b\omega)^2}}. \bigcirc \tag{4.11}$$

Again, the amplitude of the steady-state solution does *not* depend on the IC.

In retrospect, when we did Example 4.8 in Section 4.1, just from looking at the ODE, we could have immediately known that this was in the steady-state solution case because $b = 2 > 0$. It's good to be able to know something about the solutions before finding them in full detail.

In fact, let's double-check our amplitude formula using that old example: $m = 1, b = 2$, $k = 9, f_0 = 1$, and $\omega = 2$ should give

$$\text{Amplitude} = \frac{1}{\sqrt{(9-4)^2+(2\cdot 2)^2}} = \frac{1}{\sqrt{41}}, \tag{4.12}$$

which agrees with the steady-state solution found in that problem.

We can think of the forcing function, $f_0 \cos \omega t$, as being the "input" for the mechanical system, and the steady-state solution, $y_S(t)$, as being the "output," depicted in Figure 4.6. The ratio

$$G \triangleq \frac{\text{Amplitude of the steady state}}{|f_0|} = \frac{1}{\sqrt{(k-m\omega^2)^2+(b\omega)^2}} \tag{4.13}$$

can be called the **frequency response**, which measures the ratio of the strength of the output to the strength of the input.

Shown in Figure 4.7a is a plot of the frequency response versus frequency for ODE

$$\ddot{y} + b\dot{y} + 4y = \cos \omega t \tag{4.14}$$

for two values of b, specifically $b = 0.1$ and 0.4. Shown in Figure 4.7b is another plot of the frequency response for ODE (4.14) for the values of $b = 1, 4, 4\sqrt{2}, 8$.

Example 4.13

Suppose that $y(t) = 2e^{-t} + 3\cos 2t - 4\sin 2t$ is a solution of an IVP of the form

$$\left\{ \begin{array}{l} m\ddot{y} + b\dot{y} + ky = f_0 \cos \omega t \\ \\ y(0) = y_0,\ \dot{y}(0) = \dot{y}_0 \end{array} \right\}. \tag{4.15}$$

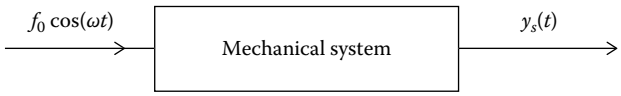


FIGURE 4.6
Forced vibrations input–output relationship.

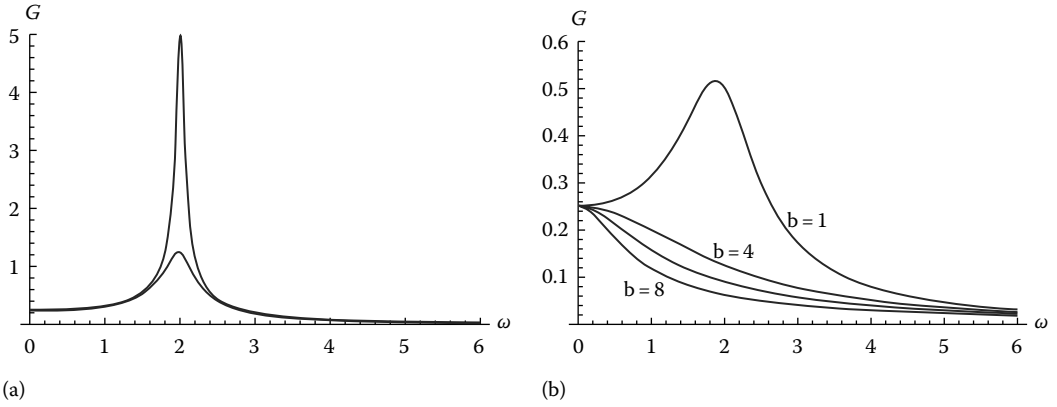


FIGURE 4.7
Frequency response (a) b = 0.1, 0.4 and (b) b = $1, 4, 4\sqrt{2}, 8$.

Assuming $b > 0$, find exact values of the constants $m, b, k, f_0, \omega, y_0, \dot{y}_0$.

Method: The initial data are easily found by substituting $t = 0$ into the given solution to get

$$y_0 = y(0) = 2 + 3 - 0 = 5.$$

Substitute $t = 0$ into the derivative of given solution, $\dot{y}(t) = -2e^{-t} - 6\sin 2t - 8\cos 2t$, to get

$$\dot{y}_0 = \dot{y}(0) = -10.$$

The assumption that b is a positive constant implies that the solutions of the corresponding homogeneous ODE are transient and that the parts of the given solution that aren't transient, that is, $3\cos 2t - 4\sin 2t$, must be a particular solution due to the RHS of ODE (4.15), that is, $f_0 \cos \omega t$. It follows that $\omega = 2$.

While we could find the amplitude of the steady-state solution, $y_S(t) = 3\cos 2t - 4\sin 2t$, and compare that with the formula for the amplitude given by (4.11), this would not be the best thing to try: There would be too many constant parameters to solve for, that is, m, b, k, f_0, at the same time, and the formula is nonlinear in m, b, k.

Instead, we can first substitute the transient solution, $y_T(t) = 2e^{-t}$, into the corresponding homogeneous ODE, $m\ddot{y} + b\dot{y} + ky = 0$, to get

$$0 = m\frac{d^2}{dt^2}\big[\,y_T\,\big] + b\frac{d}{dt}\big[\,y_T\,\big] + ky_T = m\frac{d^2}{dt^2}\big[\,2e^{-t}\,\big] + b\frac{d}{dt}\big[\,2e^{-t}\,\big] + k2e^{-t} = 2e^{-t}(m - b + k);$$

hence, we have a linear algebraic equation:

$$m - b + k = 0. \tag{4.16}$$

Also, use our first inference that $\omega = 2$ and substitute the steady-state solution, $y_S(t) = 3\cos 2t - 4\sin 2t$, into the original, nonhomogeneous ODE $m\ddot{y} + b\dot{y} + ky = f_0 \cos \omega t$ to get

$$\begin{aligned} \underline{f_0}\cos 2t + \underbrace{0}\cdot \sin 2t &= m\ddot{y_S} + b\dot{y_S} + ky_S \\ &= -4m(3\cos 2t - 4\sin 2t) + 2b(-3\sin 2t - 4\cos 2t) + k(3\cos 2t - 4\sin 2t) \\ &= \underline{(-12m - 8b + 3k)}\cos 2t + \underbrace{(16m - 6b - 4k)}\sin 2t; \end{aligned}$$

hence,

$$-12m - 8b + 3k = f_0 \tag{4.17}$$

and

$$16m - 6b - 4k = 0. \tag{4.18}$$

Equations (4.16) through (4.18) are three equations in four unknowns, m, b, k, f_0. On general principle, that would seem to be not enough equations, and that is true! If we look at the ODE in the IVP, we see that letting $m = 1$ *wouldn't* lose all of our ability to find *a* solution for the parameters. We don't need to find all the ODEs, just one.

We can rewrite the three (4.16) through (4.18), after substituting in $m = 1$, as

$$\left\{\begin{array}{rrrcr} -b & +k & & = & -1 \\ -8b & +3k & -f_0 & = & 12 \\ -6b & -4k & & = & -16 \end{array}\right\}. \tag{4.19}$$

We could use the inverse of a 3×3 matrix to solve for

$$\begin{bmatrix} b \\ k \\ f_0 \end{bmatrix},$$

but it is easier and more reliable to first find the solution for b, k to satisfy the first and third equations of (4.19), which don't involve f_0:

$$\begin{bmatrix} b \\ k \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ -6 & -4 \end{bmatrix}^{-1} \begin{bmatrix} -1 \\ -16 \end{bmatrix} = \frac{1}{10} \begin{bmatrix} -4 & -1 \\ 6 & -1 \end{bmatrix} \begin{bmatrix} -1 \\ -16 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix};$$

hence, $b = 2, k = 1$. After that, substitute into the second of the equations in (4.19), to get

$$f_0 = -12 - 8b + 3k = -12 - 8(2) + 3(1) = -25.$$

To summarize, the exact values are, assuming $m = 1$,

$$m = 1, b = 2, k = 1, f_0 = -25, \omega = 2, y_0 = 5, \text{ and } \dot{y}_0 = -10. \bigcirc$$

Define $\zeta \triangleq |\omega - \omega_0|/2$, a so-called **de-tuning parameter**. We see from formula (4.13) that we get a very large frequency response, that is, ratio of the steady-state output magnitude to the forcing function input magnitude, when $0 < b \ll 1$, that is, b is positive but very small, and ζ is also very small. This can be called **practical resonance**.

$$\textit{The practical resonance case}: \ 0 < b \ll 1 \textit{ and } 0 \leq \zeta \ll 1.$$

Example 4.14

If the damping is very small, the solution can look a lot like the pure resonance case for quite a long time! We illustrate this with Figure 4.8's graph of the solution of the IVP

$$\left\{\begin{array}{c} \ddot{y} + 0.04\,\dot{y} + 4y = \cos 2t \\ y(0) = 0, \ \dot{y}(0) = 0 \end{array}\right\}. \tag{4.20}$$

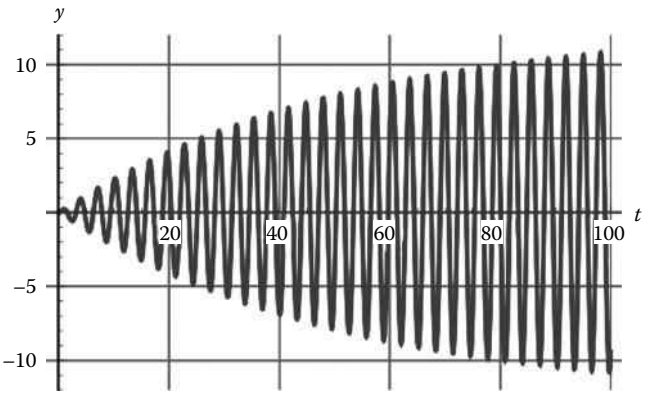


FIGURE 4.8
Example 4.14.

4.2.3 Maximum Frequency Response

In the apocryphal story, when the Roman legion marched in perfect formation over a bridge and the resulting vibrations sent the bridge crashing down, there was damping or friction in the system. The oscillations do not become unbounded, but when the magnitude of the solution becomes sufficiently large, then the system exceeds its design specifications enough to crash it.

Similarly, an airplane rudder can fail if its oscillations become sufficiently large, even though there is some damping in the system. But, to be honest, in the real world, the models are much more complicated than we have discussed so far, because there is nonlinearity in the ODEs. We will return to this issue in Chapter 18.

Also, the models at least start with nonlinear partial differential equations, although mathematical analysis may reduce the model to nonlinear ODEs.

Engineers interested in design problems care about optimization, that is, maximizing or minimizing some design objective. In this situation, one natural objective would be to enforce a limit on how large the steady-state oscillation can be, for example, in order to avoid having a bridge fall down, an electrical device "burn" out, or an airplane rudder jam in one position. While real-world problems often involve nonlinearity, sometimes we can get some insight into these phenomena by seeing the maximum magnitude of the frequency response of the steady-state oscillation.

Recall from (4.13) that the frequency response is

$$G = G(\omega) = \frac{1}{\sqrt{(k - m\omega^2)^2 + (b\omega)^2}}.$$

Some of the graphs of frequency response in Figure 4.7 showed $G(\omega)$ achieved a global maximum value.

In calculus I, we learned how to find the global maximum value of $G = G(\omega)$ as a function of ω in the interval $\mathcal{I} \triangleq (0, \infty)$, that is, where $0 < \omega < \infty$.

Theorem 4.3

Suppose a function $g = g(x)$ is defined and differentiable for x in the interval $(0, \infty)$. If $g'(x)$ is positive for $0 < x < x^\star$ and $g'(x)$ is negative for $x^\star < x < \infty$, then the global maximum value of g on $(0, \infty)$ is

$$g_{\max} = g(x^\star).$$

In our situation, it is easier to deal with the derivative of the square of $G(\omega)$ rather than $G(\omega)$ itself. We had another relevant result in calculus I:

Theorem 4.4

Suppose a function $g = g(x)$ is defined and differentiable for x in an interval $\mathcal{I}$. If $g(x) \geq 0$ for all x in the interval $\mathcal{I}$ and $f(x) \triangleq \big(g(x)\big)^2$ has global maximum value $f(x^*)$ on $\mathcal{I}$, then $g(x)$ has global maximum value $g(x^*)$ on $\mathcal{I}$.

Define

$$f(\omega) \triangleq (G(\omega))^2 = \frac{1}{(k - m\omega^2)^2 + (b\omega)^2}.$$

Using the chain rule, we calculate

$$\begin{aligned} f'(\omega) = \frac{df}{d\omega} &= \frac{-1}{\big((k - m\omega^2)^2 + (b\omega)^2\big)^2} \cdot \frac{d}{d\omega}\Big((k - m\omega^2)^2 + (b\omega)^2\Big) \\ &= \frac{-4m\omega(k - m\omega^2) + 2b^2\omega}{\big((k - m\omega^2)^2 + (b\omega)^2\big)^2}. \end{aligned}$$

In order to know where f' is positive or negative, it would help to know where $f' = 0$, that is, where the numerator of f' is zero:

$$0 = -4m\omega(k - m\omega^2) + 2b^2\omega = 2\omega\Big(-2m(k - m\omega^2) + b^2\Big) = 2\omega\Big(2m^2\omega^2 - 2mk + b^2\Big).$$

So, $f' = 0$ at

$$\omega = 0,\ \omega = \pm\sqrt{\frac{k}{m} - \frac{b^2}{2m^2}},$$

where we *assume* that $\dfrac{k}{m} - \dfrac{b^2}{2m^2} > 0$, that is, $b^2 < 2mk$. Recall from Section 3.3 that the critical damping is $b_{crit} \triangleq \sqrt{4mk}$ for the corresponding unforced system. Note that $b^2 < 2mk$ is equivalent to the damping being less than $\dfrac{1}{\sqrt{2}}$ times the critical damping.

We are only concerned with $\omega > 0$. Recall the definition $\omega_0 \triangleq \sqrt{\frac{k}{m}}$. Define

$$\omega^* \triangleq \sqrt{\omega_0^2 - \frac{b^2}{2m^2}}.$$

The numerator of f', that is,

$$2\omega\left(2m^2\omega^2 - 2mk + b^2\right) = 4m^2\omega\left(\omega^2 - \frac{k}{m} + \frac{b^2}{2m^2}\right) = 4m^2\omega\left(\omega^2 - (\omega^*)^2\right),$$

is positive for $0 < \omega < \omega^*$ and is negative for $\omega < \omega^* < \infty$, as long as $b^2 < 2mk$. It follows from Theorem 4.3 that $f_{\max} = f(\omega^*)$, so Theorem 4.4 implies that the maximum frequency response is

$$G_{\max} = G(\omega^*) = \frac{1}{\sqrt{(k - m(\omega^*)^2)^2 + (b\omega^*)^2}}.$$

In Problem 4.2.5.27, you will substitute $\omega^* \triangleq \sqrt{\omega_0^2 - \frac{b^2}{2m^2}}$ into $G_{\max}$ and get the next result.

Theorem 4.5

Suppose that the damping satisfies $b^2 < 2mk$, that is, $0 < b < \frac{1}{\sqrt{2}}\, b_{crit}$. Then the maximum frequency response is

$$G_{\max} = \frac{1}{\nu b},$$

where we recall that

$$\nu \triangleq \frac{\sqrt{4mk - b^2}}{2m} = \sqrt{\frac{k}{m} - \frac{b^2}{4m^2}}$$

is the quasi-frequency for the corresponding unforced system.

On the other hand, if b is greater than or equal to $\frac{1}{\sqrt{2}}$ times the critical damping for the corresponding unforced system, then you will explain in Problem 4.2.5.28 why $G'(\omega) < 0$ for all $\omega > 0$; hence, $\lim_{\omega \to 0^+} G(\omega) = \frac{1}{k}$ is the maximum frequency response.

4.2.4 Beats Phenomenon, Fast and Slow Frequencies, and Frequency Response

The third basic phenomenon occurs when there is zero damping, that is, $b = 0$, but the two frequencies do not match, that is, $\omega \neq \omega_0$.

Example 4.15

Solve the IVP

$$\left\{\begin{array}{c} \ddot{y} + 4y = 3\cos(\sqrt{5}\,t) \\ y(0) = 0,\ \dot{y}(0) = 0 \end{array}\right\} \tag{4.21}$$

and write the solution in the form of $y(t) = A(t)\cdot$ (a sinusoidal function).

Method: The RHS function $3\cos(\sqrt{5}\,t)$ is in the zoo, corresponding to list of roots $\mathsf{L}_2 = \pm i\sqrt{5}$. The corresponding homogeneous ODE, $\ddot{y} + 4y = 0$, gives list of roots $\mathsf{L}_1 = \pm i2$ and homogeneous solution $y(t) = c_1\cos 2t + c_2\sin 2t$. The superlist is $\mathsf{L} = \pm i\sqrt{5}, \pm i2$, so $y(t) = c_1\cos 2t + c_2\sin 2t + c_3\cos(\sqrt{5}\,t) + c_4\sin(\sqrt{5}\,t)$, and the correct form of a particular solution is

$$y_p(t) = A\cos(\sqrt{5}\,t) + B\sin(\sqrt{5}\,t),$$

where A, B are constants to be determined.

Substitute $\ddot{y}_p(t)$ into the original, nonhomogeneous ODE to get

$$\begin{aligned} \underline{3}\cdot\cos(\sqrt{5}\,t) + \underset{\sim}{0}\cdot\sin(\sqrt{5}\,t) &= \ddot{y_p} + 4y_p = \\ &= -5\left(A\cos(\sqrt{5}\,t) + B\sin(\sqrt{5}\,t)\right) + 4\left(A\cos(\sqrt{5}\,t) + B\sin(\sqrt{5}\,t)\right) \\ &= \underline{-A}\cos(\sqrt{5}\,t) + \underset{\sim}{(-B)}\sin(\sqrt{5}\,t); \end{aligned}$$

hence, $-A = 3, -B = 0$ and

$$y_p(t) = -3\cos(\sqrt{5}\,t).$$

The general solution of ODE (4.21) is

$$y(t) = y_p(t) + y_h(t) = -3\cos(\sqrt{5}\,t) + c_1\cos 2t + c_2\sin 2t,$$

where c_1, c_2 are arbitrary constants. Substitute this and its derivative into the ICs to get

$$0 = y(0) = -3 + c_1 \quad \text{and} \quad 0 = \dot{y}(0) = 0 + 2c_2.$$

The solution of the IVP is

$$y(t) = -3\cos(\sqrt{5}\,t) + 3\cos 2t.$$

Using one of the less well-known trigonometric identities, namely, the difference of cosine functions identity

$$-\cos\alpha + \cos\beta = 2\sin\left(\frac{\alpha-\beta}{2}\right)\sin\left(\frac{\alpha+\beta}{2}\right),$$

we can rewrite the solution of the IVP in the form

$$y(t) = 6\sin\left(\frac{\sqrt{5}\,t - 2t}{2}\right)\sin\left(\frac{2t + \sqrt{5}\,t}{2}\right) = A(t)\sin\left(\frac{(\sqrt{5}+2)t}{2}\right), \tag{4.22}$$

where the slowly varying amplitude is

$$A(t) \triangleq 6\sin\left(\frac{(\sqrt{5}-2)t}{2}\right). \bigcirc$$

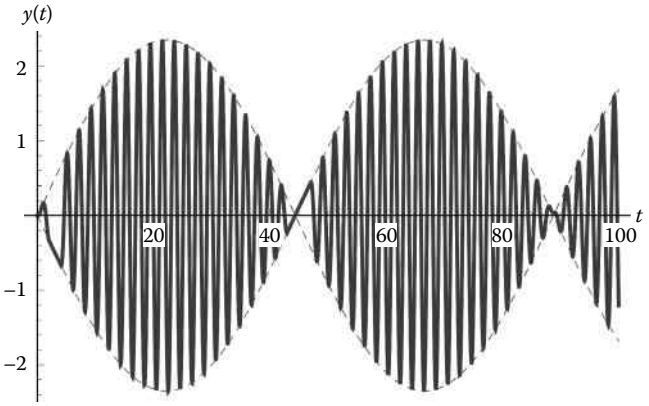


FIGURE 4.9
Example 4.15.

Another correct final conclusion for Example 4.15 would be

$$y(t) = \bar{A}(t) \sin\left(\frac{(\sqrt{5}-2)t}{2}\right), \text{ where } \bar{A}(t) = 6\sin\left(\frac{(\sqrt{5}+2)t}{2}\right). \tag{4.23}$$

We prefer (4.22) to (4.23) for the graphical reason that the former has a slowly varying amplitude rather than the quickly varying amplitude in the latter.

The dashed curves $y = \pm A(t) = \pm 6 \sin\left(\frac{(\sqrt{5}-2)t}{2}\right)$ in Figure 4.9 give an "envelope" for the graph of the solution.

The solution (4.22) has two frequencies, a slow-frequency

$$\zeta \triangleq \frac{\sqrt{5}-2}{2},$$

which is the same de-tuning parameter we mentioned in the steady-state oscillation case before Example 4.14, and a fast-frequency

$$\beta \triangleq \frac{\sqrt{5}+2}{2}.$$

The **beats phenomenon** consists of fast oscillations, $\sin \beta t$, within an envelope of slow oscillations, for example, $y = \pm 6 \sin \zeta t$. The slow frequency is sometimes called the "frequency of the beats." It is this slow frequency that a piano tuner uses to estimate how much to raise or lower the tension on a piano string in order to have it calibrated with a standard frequency, for example, that of a tuning fork.

More generally, the solution of an IVP

$$\left\{ \begin{array}{l} \ddot{y} + \omega_0^2 y = f_0 \cos \omega t \\ \\ y(0) = y_0,\ \dot{y}(0) = \dot{y}_0 \end{array} \right\}$$

can be written in the form

$$y(t) = a_0 \cos(\omega_0 t - \delta_0) + a_1 \cos(\omega t - \delta_1). \tag{4.24}$$

In general, it is *not* possible to use a trigonometric identity to rewrite $y(t)$ as a product of a fast oscillation with an envelope of slow oscillations. In general, the solution given by (4.24) is either a "quasiperiodic" function or a "periodic" function, depending upon the relationship between the forcing frequency, ω, and the natural frequency, ω_0.

Definition 4.2

(a) A function $y(t)$ is **periodic** if there is a positive constant T such that $y(t + \mathrm{T}) \equiv y(t)$ at all t in the domain of y. [Pictorially, a function is periodic if its graph is not changed when you shift it horizontally by T units.]

(b) A function $y(t)$ is **quasiperiodic** if it is *not* periodic, but it is the sum of two (or any finite number of) periodic functions.

A quasiperiodic function must involve more than one frequency; otherwise, it would automatically be periodic.

Unfortunately, there is no simple graphical interpretation of the concept of quasiperiodicity, although there are sophisticated interpretations involving all the horizontal shifts of a quasiperiodic function. There is an even more general concept of so-called "almost periodic" functions, and they show up in the general theory of ODEs but involve more advanced mathematics that we won't discuss.

A function of the form

$$y(t) = a_0 \cos(\omega_0 t - \delta_0) + a_1 \cos(\omega t - \delta_1) \tag{4.25}$$

is periodic if the ratio $\frac{\omega}{\omega_0}$ is a rational number $\frac{p}{q}$, that is, the ratio of two integers, in which case, $\mathrm{T} \triangleq \frac{2\pi p}{\omega}$ is a period. For example,

$$y(t) = \cos\left(3\sqrt{2}\, t - \frac{7}{4}\right) + 8\cos(5\sqrt{2}t)$$

is periodic with period $\mathrm{T} = \frac{2\pi 5}{5\sqrt{2}} = \pi\sqrt{2}$ because $\frac{\omega}{\omega_0} = \frac{5\sqrt{2}}{3\sqrt{2}} = \frac{5}{3} = \frac{p}{q}$ and

$$y(t + \pi\sqrt{2}) = \cos\left(3\sqrt{2}\,(t + \pi\sqrt{2}) - \frac{7}{4}\right) + 8\cos\left(5\sqrt{2}(t + \pi\sqrt{2})\right)$$

$$= \cos\left(3\sqrt{2}\, t + 6\pi - \frac{7}{4}\right) + 8\cos\left(5\sqrt{2}t + 10\pi\right) = \cos\left(3\sqrt{2}\, t - \frac{7}{4}\right) + 8\cos(5\sqrt{2}t) \equiv y(t).$$

While it is not possible to say much that is exactly correct about solutions with the beats phenomenon in the quasiperiodic case, a solution in the form (4.25) has crude bounds on its amplitude: Because $|\cos\theta| \leq 1$ for all θ and there is an inequality for real numbers called the triangle inequality, that is, $|a + b| \leq |a| + |b|$, we have

$$|y(t)| \leq |a_0| + |a_1|.$$

In addition, for some time interval, the solution will behave as if it has slow-frequency

$$\zeta \triangleq \frac{|\omega - \omega_0|}{2}$$

and fast-frequency

$$\beta \triangleq \frac{\omega + \omega_0}{2},$$

say for a time interval of length three of the slow periods.

4.2.5 Problems

1. Solve $\ddot{y} + 4y = -3\cos 2t$.

In problems 2–4, solve the IVP.

2. $\ddot{y} + 9y = 5\sin 3t,\ y(0) = \dot{y}(0) = 0$
3. $\ddot{y} + 16y = 4\cos 4t,\ y(0) = -1,\ \dot{y}(0) = 3$
4. $\ddot{y} + 8y = 5\sin 3t,\ y(0) = \dot{y}(0) = 0$

In problems 5–8, find the steady-state solution, express it in the amplitude-phase form, and state its amplitude.

5. $\ddot{y} + 2\dot{y} + 2y = \sin t$
6. $\ddot{y} + \dot{y} + 5y = \cos 2t$
7. $\ddot{y} + 2\dot{y} + 3y = \sin 2t$
8. $\ddot{y} + 2\dot{y} + y = 4\sin 3t$

In problems 9–12, find the solution and the steady-state solution, and express the latter in the amplitude-phase form and state its amplitude.

9. $\ddot{y} + 2\dot{y} + 10y = 74\cos 3t, y(0) = -1, \dot{y}(0) = 2$
10. $\ddot{y} + 4\dot{y} + 5y = \sin \pi t,\ y(0) = \dot{y}(0) = 0$
11. $\ddot{y} + 2\dot{y} + 6y = \sin 2t,\ y(0) = -3,\ \dot{y}(0) = 3 - 3\sqrt{5}$
12. $\ddot{y} + \dot{y} + 3y = \sin 2t,\ y(0) = 1,\ \dot{y}(0) = 0$

In problems 13 and 14, for the IVP, find (a) the frequency of the beats and (b) the maximum amplitude of the motion.

13. $\ddot{y} + 9y = 5\cos(\sqrt{8}\,t),\ y(0) = \dot{y}(0) = 0$
14. $\ddot{y} + 5y = 3\cos 2t,\ y(0) = \dot{y}(0) = 0$
15. Find the steady-state solution for $m\ddot{y} + b\dot{y} + ky = f_0 \sin \omega t$. Show all work, and leave your final conclusion in terms of the unspecified physical parameters m, b, k, f_0, ω, that is, do not substitute in specific values for them.

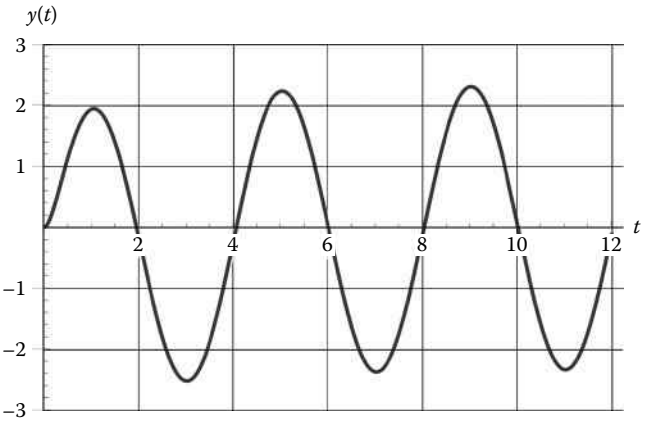


FIGURE 4.10
Problem 4.2.5.16

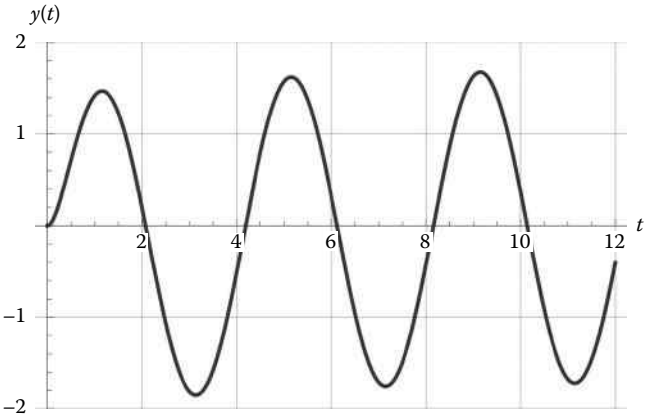


FIGURE 4.11
Problem 4.2.5.17.

16. Shown in Figure 4.10 is the graph of a solution of $\ddot{y} + 6\dot{y} + ky = 22\cos\omega t$. Determine (a) ω, (b) the amplitude of the steady-state solution, and (c) $\sqrt{k}$, the natural frequency of the system, using parts (a) and (b). For each part, cite specific information from the graph to help justify your conclusion(s).
17. Shown in Figure 4.11 is the graph of a solution of $\ddot{y} + 4\dot{y} + ky = 11\cos\omega t$. Determine (a) ω, (b) the amplitude of the steady-state solution, and (c) $\sqrt{k}$, the natural frequency of the system, using parts (a) and (b). For each part, cite specific information from the graph to help justify your conclusion(s).
18. Shown in Figure 4.12 are graphs of solutions of two different ODEs of the form $\ddot{y} + b\dot{y} + y = \cos(\omega_2 t)$. Decide which of the pairs of parameter values could *conceivably* give the graphs, and explain why. More than one pair may be correct. (a) $b_1 = 0.5$ and $\omega_1 = \sqrt{2}$, (b) $b_2 = 0.5$ and $\omega_2 = \sqrt{2}$, (c) $b_1 = 0.5$ and $\omega_1 = 1$, (d) $b_2 = 0$ and $\omega_2 = 1$, (e) $b_1 = 1$ and $\omega_1 = \sqrt{2}$, and (f) $b_2 = 0$ and $\omega_2 = \sqrt{2}$.

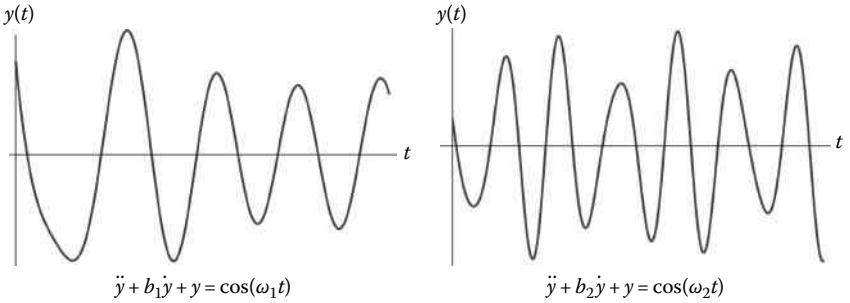


FIGURE 4.12
Problem 4.2.5.18.

19. For ODE $\ddot{y}+2d\dot{y}+y=\cos t$, let d be a positive, but adjustable, parameter. Describe the behavior of the steady-state solution as $d \to 0^+$. Interpret this in regard to the concepts of pure versus practical resonance.
20. The loop current in a series LC circuit with a voltage source satisfies $L\ddot{I} + \frac{1}{C} I = f_0 \sin(120\pi t)$. If $L = 10^{-3}$ H and the constant C is measured in farads, for which value(s) of C does the circuit exhibit
    (a) Pure resonance?
    (b) The beats phenomenon? In that case, find the slow, beats frequency in terms of C.
21. A front-loading washing machine is mounted on a rubber pad that acts like a spring; the weight of the machine depresses the pad by exactly 0.1 in. When the machine spins at $\frac{\omega}{2\pi}$ revolutions per second, the rotor exerts a vertical force of $f_0 \cos \omega t$ pounds on the machine. At what speed, in revolutions per second, will resonance vibrations occur, assuming we can neglect damping and $g \approx 32$ ft/s^2?
22. An oscillator problem $\ddot{y} + b\dot{y} + ky = f_0 \cos \omega t$ has a solution $y(t) = e^{-t} \cos 3t + 3 \cos 2t + 2 \sin 2t$. State the exact values of b, k, ω, and explain the reasoning process you used to find them.*
23. An oscillator problem $m\ddot{y} + b\dot{y} + ky = f_0 \cos \omega t$ has a solution $y(t) = 2t \sin 2t + 5 \cos(2t - \delta)$, for some constant δ. Assume the spring constant k is 6 N/m. State the exact values of m, b, f_0, ω, and explain the reasoning process you used to find them.
24. An oscillator problem $\ddot{y}+b\dot{y}+ky = f_0 \cos \omega t$ has a solution $y(t) = 2 \sin 3t - \cos 3t - \frac{1}{4} t \sin 3t$. State the exact values of b, k, f_0, ω, and explain the reasoning process you used to find them.
25. Suppose $y(t)$ satisfies a spring, mass, and possibly damped, ODE of the form $\ddot{y}+b\dot{y}+9y = f_0 \cos \omega t$. For each of parts (a), (b), (c), the behavior of $y(t)$ is described, and you are asked to give a specific *numerical* value for each of the physical parameters b, f_0, ω that can produce such a solution $y(t)$. For convenience, assume $f_0 \geq 0$ and $\omega > 0$.
    (a) $y(t) =$ (*steady-state solution*) + (*transient solution*).
    (b) $y(t)$ exhibits pure resonance.
    (c) $y(t)$ exhibits the beats phenomenon.

* Problems 22–24 were inspired by problems in Farlow et al. (2007).

26. The following is a model (Bearman, 1984) for oscillations of a rigid 2D bluff body placed normal to fluid flow and mounted on springs. The body is subject to a transverse fluid force due to the body's shedding of vortices and viscous-type damping force associated with the springs and their mounting:

$$m\ddot{y} + 4\pi m n_0 \delta_s \dot{y} + 4\pi^2 m n_0^2 y = \frac{1}{2}\overline{C}_y \varrho U^2 D \sin(2\pi n_v t + \varphi)$$

where ϱ is the fluid density, the fluid speed is U, and $\overline{C}_y$ is a coefficient measuring the transverse force of the vortices. D is a characteristic dimension of the body, usually the body width; k is the spring constant; m is the mass of the body; $\beta = 4\pi\delta_s m n_0$ is the damping coefficient; δ_s is the fraction of critical damping; and the natural (unforced, undamped) frequency in Hz is $n_0 = \frac{1}{2\pi}\sqrt{\frac{k}{m}}$.

Substitute into the ODE a steady-state solution in the form $y_p = \overline{y}\sin(2\pi n_v t)$ without the phase angle φ because it is assumed that the fluid force "leads" the oscillation. Use the trigonometric identity $\sin(\alpha+\varphi) = \sin\alpha\cos\varphi + \cos\alpha\sin\varphi$ on the RHS of the ODE to help explain why

$$\frac{n_0}{n_v} = \left(1 - \frac{\overline{C}_y}{4\pi^2}\cos\varphi\left(\frac{\varrho D^2}{2m}\right)\left(\frac{U}{n_0 D}\right)^2\left(\frac{\overline{y}}{D}\right)^{-1}\right)^{-1/2}$$

and

$$\frac{\overline{y}}{D} = \frac{\overline{C}_y}{8\pi^2}\sin\varphi\left(\frac{\varrho D^2}{2m\delta_s}\right)\left(\frac{U}{n_0 D}\right)^2\frac{n_0}{n_v}.$$

This explains why $\overline{y}$, the amplitude of the response, depends on the phase angle φ, which in turn depends on the forcing frequency, n_v, in Hz.

27. Suppose that the damping $b^2 < 2mk$. Substitute $\omega^* \triangleq \sqrt{\omega_0^2 - \frac{b^2}{2m^2}}$ into $G_{\max}$ and simplify it to find the nicer looking result that the maximum frequency response is $G_{\max} = \frac{1}{vb}$, the conclusion of Theorem 4.5.

28. Suppose that the damping $b^2 \geq 2mk$. (a) Explain why (a) $\frac{dG}{d\omega} < 0$ for all $\omega > 0$, (b) calculate $\lim_{\omega\to 0^+} G(\omega)$, and (c) explain why the maximum frequency response is $\frac{1}{k}$ but is not achieved for any $\omega > 0$.

29. Throughout this problem, suppose that $y(t) = 1.5 + 0.75\cos(t - \frac{\pi}{3}) - 0.5e^{-t/5}$ and ω_0, δ, γ, α, ε, and η are constants.
    (a) Find a nonhomogeneous ODE of the form $\ddot{y} + \omega_0^2 y = \delta + \gamma e^{-\alpha t}$ that has $y(t)$ as a solution.
    (b) Find a nonhomogeneous ODE of the form $(D^2 + \omega_0^2)(D + \varepsilon)[y] = \eta$ that has $y(t)$ as a solution.
    (c) Find a *homogeneous* ODE that has $y(t)$ as a solution.

30. Throughout this problem, suppose that the graph of $y(t)$ is drawn in Figure 4.13 and ω_0, δ, γ, ε, and η are constants. Assume $y(t)$ satisfies the ICs $y(0) = \dot{y}(0) = 1$.

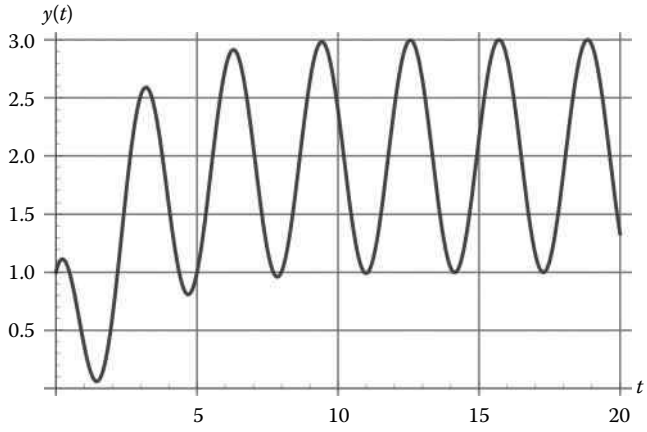


FIGURE 4.13
Problem 4.2.5.30.

(a) Find a nonhomogeneous ODE of the form $\ddot{y} + \omega_0^2 y = \delta + \gamma e^{-t/2}$ that has $y(t)$ as a solution.

(b) Using the $y(t)$ you found to help in solving part (a), find a nonhomogeneous ODE of the form $(D^2 + \omega_0^2)(D + \varepsilon)[\,y\,] = \eta$ that has $y(t)$ as a solution.

(c) Find a *homogeneous* ODE that has $y(t)$ as a solution.

31. Find the steady-state solutions for (a) $\dot{y} + \delta y = f_0 \cos \omega t$, (b) $\dot{y} + \delta y = f_0 \sin \omega t$, and (c) $\dot{y} + \delta y = f_0 \cdot (a \cos \omega t + b \sin \omega t)$.
For each of (a), (b), and (c), show all work, and leave your final conclusion in terms of the unspecified parameters δ, f_0, ω, and, for (c), a, b, that is, do not substitute in specific values for them.

32. [Small project] Find the frequency response, function for $\dot{y} + \delta y = f_0 \cos \omega t$ in terms of the unspecified parameters δ, f_0, ω. Get a result(s) concerning maximum frequency response such as we did in work preceding Theorem 4.5.

33. [Small project] Find the frequency response function for $\dot{y} + \delta y = f_0 \sin \omega t$ in terms of the unspecified parameters δ, f_0, ω. Get a result(s) concerning maximum frequency response such as we did in work preceding Theorem 4.5.

4.3 Variation of Parameters

The method of variation of parameters will be our second method for solving ODEs that are nonhomogeneous. Suppose our ODE can be written in the form

$$L[\,y\,] = f(t),$$

where L is a second-order linear differential operator (LD-op). It is legitimate to use Section 4.1's method of undetermined coefficients if L is a constant coefficients LD-op and f is a

linear combination of functions in the zoo of solutions of LCCHODEs. In this section, we will learn a more general method.

In this section, we will work only with second-order nonhomogeneous ODEs. If an ODE is of higher order, it is solvable using a generalization of the method to systems of ODEs that we will develop in Chapter 5.

By the way, in the original, 1951, version of the movie "The Day the Earth Stood Still," Klaatu, an extra terrestrial from a civilization far more advanced than ours, finds a scientist, Prof. Jacob Barnhardt, he can confide in. Klaatu helps the professor with a tough problem and finishes his comments with, "With variation of parameters, this is the answer."*

Let's start with an example we could have done in Section 4.1 but would have been tedious if solved by the method of undetermined coefficients. After we do this example, we will develop the new method in general.

Example 4.16

Solve the ODE by the method of variation of parameters developed as follows:

$$\ddot{y} + 5\dot{y} + 6y = t^2 e^{-t}. \tag{4.26}$$

Method: The corresponding homogeneous ODE, $\ddot{y} + 5\dot{y} + 6y = 0$, has characteristic polynomial $s^2 + 5s + 6 = (s + 2)(s + 3)$, so the general solution is

$$y(t) = c_1 e^{-2t} + c_2 e^{-3t}.$$

Let's try to find a solution of the nonhomogeneous ODE (4.26) in the form

$$y(t) = v_1(t)e^{-2t} + v_2(t)e^{-3t}, \tag{4.27}$$

where $v_1(t), v_2(t)$ are functions to be determined later. The idea of replacing constants c_1, c_2 by functions $v_1(t), v_2(t)$ is what we mean by **variation of parameters**, also known as **variation of constants**. The motivation for trying this comes from success we had in finding a second solution of the form $y_2(t) = v(t)e^{\alpha t}$ in the critically damped case in Section 3.3.

To substitute (4.27) into ODE (4.26), we will need to take two derivatives of $y(t)$. Using the product rule, we get

$$\dot{y}(t) = \dot{v}_1(t)e^{-2t} + \dot{v}_2(t)e^{-3t} - 2e^{-2t}v_1(t) - 3e^{-3t}v_2(t).$$

If we were to take another derivative, we would get both $\ddot{v}_1(t)$ and $\ddot{v}_2(t)$ appearing in what we would need to solve, and it sounds like a bad idea to replace one ODE involving one second derivative with a problem involving *two* second derivatives.

A couple of centuries ago, Joseph Louis Lagrange figured out how to make things work better. His idea was to make sure that $\ddot{v}_1(t), \ddot{v}_2(t)$ wouldn't be needed, by later enforcing the requirement that

$$\dot{v}_1(t)e^{-2t} + \dot{v}_2(t)e^{-3t} \equiv 0. \tag{4.28}$$

Assuming that $\dot{v}_1, \dot{v}_2(t)$ can be made to satisfy (4.28), we have

$$\dot{y}(t) = -2e^{-2t}v_1(t) - 3e^{-3t}v_2(t),$$

* Between the 41st and the 42nd minutes of the film.

and thus,

$$\ddot{y}(t) = -2e^{-2t}\,\dot{v}_1(t) - 3e^{-3t}\,\dot{v}_2(t) + 4e^{-2t}v_1(t) + 9e^{-3t}v_2(t).$$

Substituting all of that into the original, nonhomogeneous ODE (4.26), we get

$$\begin{aligned} t^2e^{-t} &= \ddot{y} + 5\dot{y} + 6y \\ &= -2e^{-2t}\,\dot{v}_1(t) - 3e^{-3t}\,\dot{v}_2(t) + \cancel{4e^{-2t}v_1(t)} + \cancel{9e^{-3t}v_2(t)} + 5(\cancel{-2e^{-2t}v_1(t)} - \cancel{3e^{-3t}v_2(t)}) \\ &\quad + 6(\cancel{v_1(t)e^{-2t}} + \cancel{v_2(t)e^{-3t}}), \end{aligned}$$

that is,

$$-2e^{-2t}\,\dot{v}_1(t) - 3e^{-3t}\,\dot{v}_2(t) = t^2e^{-t}. \tag{4.29}$$

It turns out that having so many terms cancel is no accident, as we will see when we discuss the method in general.

So, $\dot{v}_1(t), \dot{v}_2(t)$ should satisfy both (4.28) and (4.29), that is, should satisfy the system of linear equations

$$\left\{\begin{array}{rl} e^{-2t}\,\dot{v}_1(t) + e^{-3t}\,\dot{v}_2(t) & = 0 \\ -2e^{-2t}\,\dot{v}_1(t) - 3e^{-3t}\,\dot{v}_2(t) & = t^2e^{-t} \end{array}\right\}. \tag{4.30}$$

Using the inverse of a 2×2 matrix, we get

$$\begin{bmatrix} \dot{v}_1 \\ \dot{v}_2 \end{bmatrix} = \begin{bmatrix} e^{-2t} & e^{-3t} \\ -2e^{-2t} & -3e^{-3t} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ t^2e^{-t} \end{bmatrix} = \frac{1}{-e^{-5t}} \begin{bmatrix} -3e^{-3t} & -e^{-3t} \\ 2e^{-2t} & e^{-2t} \end{bmatrix} \begin{bmatrix} 0 \\ t^2e^{-t} \end{bmatrix} = \begin{bmatrix} t^2e^{t} \\ -t^2e^{2t} \end{bmatrix}.$$

We obtain, using integration by parts twice,

$$v_1(t) = \int \dot{v}_1(t)dt = \int t^2e^t dt = \cdots = t^2e^t - 2te^t + 2e^t + c_1,$$

and similarly,

$$v_2(t) = \int \dot{v}_2(t)dt = -\int t^2e^{2t}dt = \cdots = -\frac{1}{2}t^2e^{2t} + \frac{1}{2}te^{2t} - \frac{1}{4}e^{2t} + c_2,$$

where c_1, c_2 are arbitrary constants.

Putting everything together, we get solutions

$$\begin{aligned} y(t) = v_1(t)e^{-2t} + v_2(t)e^{-3t} &= \left((t^2 - 2t + 2)e^t + c_1\right)e^{-2t} + \left(\left(-\tfrac{1}{2}t^2 + \tfrac{1}{2}t - \tfrac{1}{4}\right)e^{2t} + c_2\right)e^{-3t} \\ &= \left(t^2 - 2t + 2 - \tfrac{1}{2}t^2 + \tfrac{1}{2}t - \tfrac{1}{4}\right)e^{-t} + c_1e^{-2t} + c_2e^{-3t}. \end{aligned}$$

By Theorem 4.1 in Section 4.1, the general solution of nonhomogeneous ODE (4.26) is

$$y(t) = \left(\frac{1}{2}\,t^2 - \frac{3}{2}\,t + \frac{7}{4}\right)e^{-t} + c_1e^{-2t} + c_2e^{-3t}, \tag{4.31}$$

where c_1, c_2 are arbitrary constants. ◯

In (4.31), we have both $y_p(t)$, a particular solution, as well as $y_h(t)$, a general solution of the corresponding homogeneous ODE. The method of variation of parameters automatically gives the general solution all at once, that is, we don't need to add $y_h(t)$ to a $y_p(t)$.

Also, if we are only interested in a particular solution, we can substitute $c_1 = c_2 = 0$ into (4.31) to find a $y_p(t)$.

In Problem 4.3.2.21, you will use the method of Section 4.1, that is, the method of undetermined coefficients, to solve (4.26). This will both check our conclusion in (4.31) and also give us an opportunity to decide whether, in this example, the method of undetermined coefficients would have been easier or more reliable than the method of variation of parameters.

In general, we should use the method of undetermined coefficients if it is legitimate to do so *and* the RHS of the nonhomogeneous ODE looks simple.

In general, to solve

$$L[y] \triangleq \ddot{y} + p(t)\dot{y} + q(t)y = f(t),$$

that is, where L is a second-order LD-op in *standard* form, here is the

4.3.1 Method of Variation of Parameters

Step 1. Find $\{y_1(t), y_2(t)\}$, a complete set of basic solutions of $\ddot{y} + p(t)\dot{y} + q(t)y = 0$, that is, the corresponding linear homogeneous ODE *in standard form*.

Step 2. Let

$$y(t) = y_1(t)v_1(t) + y_2(t)v_2(t), \tag{4.32}$$

where $v_1(t), v_2(t)$ are functions to be found later. Assume that

$$y_1(t)\dot{v}_1(t) + y_2(t)\dot{v}_2(t) \equiv 0, \tag{4.33}$$

and use that to calculate $\dot{y}(t)$ and then $\ddot{y}(t)$ from (4.32).

Step 3. Substitute $y, \dot{y}, \ddot{y}$ into the original nonhomogeneous ODE $\ddot{y} + p(t)\dot{y} + q(t)y = f(t)$, which is in standard form. Almost all of the terms cancel, leaving

$$\dot{y}_1(t)\dot{v}_1(t) + \dot{y}_2(t)\dot{v}_2(t) = f(t). \tag{4.34}$$

Step 4. Solve the system consisting of (4.33) and (4.34), that is,

$$\left\{\begin{array}{ll} y_1(t)\dot{v}_1(t) + y_2(t)\dot{v}_2(t) & = 0 \\ \dot{y}_1(t)\dot{v}_1(t) + \dot{y}_2(t)\dot{v}_2(t) & = f(t) \end{array}\right\}, \tag{4.35}$$

for $\dot{v}_1(t), \dot{v}_2(t)$.

Step 5. Integrate $\dot{v}_1(t)$ and $\dot{v}_2(t)$ with respect to t to get $v_1(t)$ and $v_2(t)$, including arbitrary constants c_1, c_2, and then substitute them into (4.32) to find the general solution of the nonhomogeneous ODE, $\ddot{y} + p(t)\dot{y} + q(t)y = f(t)$. ○

To solve system (4.35), we use the inverse of a 2×2 matrix in

$$\begin{bmatrix} \dot{v}_1 \\ \dot{v}_2 \end{bmatrix} = \begin{bmatrix} y_1(t) & y_2(t) \\ \dot{y}_1(t) & \dot{y}_2(t) \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ f(t) \end{bmatrix}, \tag{4.36}$$

so we need that the Wronskian determinant, $W(y_1, y_2)(t)$, never be zero. But Theorem 3.13 in Section 3.3 says that it would follow from the assumption we made that $\{y_1(t), y_2(t)\}$ is a complete set of basic solutions of the corresponding linear homogeneous ODE, $\ddot{y} + p(t)\dot{y} + q(t)y = 0$!

We claim that if we assume that both $y_1(t)$ and $y_2(t)$ are solutions of $L[y] = 0$ and we assume that $y_1(t)\dot{v}_1(t) + y_2(t)\dot{v}_2(t) \equiv 0$, then almost all of the terms cancel, leaving (4.34). Why? Because

$$\begin{aligned}
&L[v_1(t)y_1(t) + v_2(t)y_2(t)] \\
&\frac{d^2}{dt^2}\big[\, v_1(t)y_1(t) + v_2(t)y_2(t) \,\big] + p(t)\frac{d}{dt}\big[\, v_1(t)y_1(t) + v_2(t)y_2(t) \,\big] + q(t)\left(v_1(t)y_1(t) + v_2(t)y_2(t)\right) \\
&\quad = \dot{y}_1(t)\dot{v}_1(t) + \dot{y}_2(t)\dot{v}_2(t) + \ddot{y}_1(t)\dot{v}_1(t) + \ddot{y}_2(t)\dot{v}_2(t) + p(t)\left(\dot{y}_1(t)v_1(t) + \dot{y}_2(t)v_2(t)\right) \\
&\qquad + q(t)(v_1(t)y_1(t) + v_2(t)y_2(t)) \\
&\quad = \dot{y}_1(t)\dot{v}_1(t) + \dot{y}_2(t)\dot{v}_2(t) + v_1(t)\left(\ddot{y}_1(t) + p(t)\dot{y}_1(t) + q(t)y_1(t)\right) \\
&\qquad + v_2(t)\left(\ddot{y}_2(t) + p(t)\dot{y}_2(t) + q(t)y_2(t)\right) \\
&\quad = \dot{y}_1(t)\dot{v}_1(t) + \dot{y}_2(t)\dot{v}_2(t) + 0 + 0,
\end{aligned}$$

that is, (4.34).

Example 4.17

(Rotating disks of constant thickness) (Ugural and Fenster, 2003). Solve the ODE

$$r^2\frac{d^2u}{dr^2} + r\frac{du}{dr} - u = -\frac{(1-\nu^2)\varrho\,\omega^2 r^3}{E}, \tag{4.37}$$

where $u = u(r)$ is the deformation, also known as the displacement, from the reference state, and mass density ϱ, angular speed ω, Young's modulus of elasticity E, and Poisson's ratio ν are constants.

Method: First, we put the ODE into the standard form by dividing through by r^2:

$$L[u] \triangleq \frac{d^2u}{dr^2} + r^{-1}\frac{du}{dr} - r^{-2}u = -\eta\, r,$$

where the constant $\eta \triangleq (1-\nu^2)\varrho\omega^2/E$. The method of undetermined coefficients, in Section 4.1, doesn't apply to this ODE because the LD-op L does not have constant coefficients. The only method we can use is the method of variation of parameters.

The corresponding homogeneous ODE, $\frac{d^2u}{dr^2} + r^{-1}\frac{du}{dr} - r^{-2}u = 0$, is the same one we studied in Example 3.33 in Section 3.5: Substituting in $u(r) = r^n$ yields characteristic

polynomial $n(n-1)+n-1 = n^2-1 = (n+1)(n-1)$, so the functions $\{u_1(r) = r, u_2(r) = r^{-1}\}$ are a complete set of basic solutions. The method of variation of parameters says to try solutions in the form $u(r) = rv_1(r)+r^{-1}v_2(r)$, where $v_1(r), v_2(r)$ should satisfy the system (4.35). Denoting $\frac{d}{dr} = '$, this is

$$\left\{\begin{array}{ll} r\,v_1'(r) + r^{-1}\,v_2'(r) & = 0 \\ \\ v_1'(r) - r^{-2}\,v_2'(r) & = -\eta\,r \end{array}\right\}.$$

Using the inverse of a 2×2 matrix, we get

$$\begin{bmatrix} v_1' \\ v_2' \end{bmatrix} = \begin{bmatrix} r & r^{-1} \\ 1 & -r^{-2} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ -\eta\,r \end{bmatrix} = \frac{1}{-2r^{-1}} \begin{bmatrix} -r^{-2} & -r^{-1} \\ -1 & r \end{bmatrix} \begin{bmatrix} 0 \\ -\eta\,r \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}\eta\,r \\ \frac{1}{2}\eta\,r^3 \end{bmatrix}.$$

We get

$$v_1(r) = \int v_1'(r)dr = \int \left(-\frac{1}{2}\eta\,r\right) dr = -\frac{1}{4}\eta\,r^2 + c_1$$

and

$$v_2(r) = \int v_2'(r)dr = \int \frac{1}{2}\eta\,r^3 dr = \frac{1}{8}\eta\,r^4 + c_2,$$

where c_1, c_2 are arbitrary constants.

Putting everything together, we get solutions

$$u(r) = rv_1(r) + r^{-1}v_2(r) = \cdots = -\frac{\eta}{8}r^3 + c_1 r + c_2 r^{-1}.$$

Putting back in the definition of η, the general solution of the nonhomogeneous ODE (4.37) is

$$u(r) = -\frac{(1-\nu^2)\varrho\,\omega^2}{8E}r^3 + c_1 r + c_2 r^{-1},$$

where c_1, c_2 are arbitrary constants. ◯

Example 4.18

Solve the IVP

$$\left\{\begin{array}{l} \ddot{y} + 2\dot{y} + y = \sqrt{t}\,e^{-t} \\ y(0) = -4,\ \dot{y}(0) = 5 \end{array}\right\}. \tag{4.38}$$

Method: The corresponding homogeneous ODE, $\ddot{y} + 2\dot{y} + y = 0$, has characteristic polynomial $s^2 + 2s + 1 = (s+1)^2$, so $\{y_1(t) = e^{-t}, y_2(t) = te^{-t}\}$ is a complete set of basic solutions of that homogeneous ODE. Let's try to find a solution of the nonhomogeneous ODE (4.38) in the form

$$y(t) = v_1(t)e^{-t} + v_2(t)te^{-t}.$$

The functions $v_1(t), v_2(t)$ should satisfy system (4.35), which here is

$$\left\{\begin{array}{c} e^{-t}\,\dot{v}_1(t) + te^{-t}\,\dot{v}_2(t) = 0 \\ \\ -e^{-t}\,\dot{v}_1(t) + (1-t)e^{-t}\,\dot{v}_2(t) = \sqrt{t}\,e^{-t} \end{array}\right\}.$$

We get

$$\begin{bmatrix}\dot{v}_1\\ \dot{v}_2\end{bmatrix} = \begin{bmatrix} e^{-t} & te^{-t}\\ -e^{-t} & (1-t)e^{-t}\end{bmatrix}^{-1}\begin{bmatrix}0\\ \sqrt{t}\,e^{-t}\end{bmatrix} = \frac{1}{e^{-2t}}\begin{bmatrix}(1-t)e^{-t} & -te^{-t}\\ e^{-t} & e^{-t}\end{bmatrix}\begin{bmatrix}0\\ \sqrt{t}\,e^{-t}\end{bmatrix};$$

hence,

$$\begin{bmatrix}\dot{v}_1\\ \dot{v}_2\end{bmatrix} = \begin{bmatrix}-t^{3/2}\\ t^{1/2}\end{bmatrix}.$$

So,

$$v_1(t) = \int \dot{v}_1(t)\,dt = \int (-t^{3/2})\,dt = -\frac{2}{5}\,t^{5/2} + c_1,$$

and

$$v_2(t) = \int \dot{v}_2(t)\,dt = \int t^{1/2}\,dt = \frac{2}{3}\,t^{3/2} + c_2,$$

where c_1, c_2 are arbitrary constants.

Putting everything together, we get solutions

$$y(t) = v_1(t)e^{-t} + v_2(t)te^{-t} = \cdots = \frac{4}{15}\,t^{5/2}\,e^{-t} + c_1e^{-t} + c_2te^{-t}.$$

After calculating that

$$\dot{y} = \frac{2}{3}\,t^{3/2}e^{-t} - \frac{4}{15}\,t^{5/2}e^{-t} - c_1e^{-t} + c_2(1-t)e^{-t},$$

we address the ICs:

$$-4 = y(0) = 0 + c_1 + 0 \quad \text{and} \quad 5 = \dot{y}(0) = 0 - c_1 + c_2;$$

hence, $c_1 = -4, c_2 = 1$. The solution of the IVP is

$$y(t) = \frac{4}{15}\,t^{5/2}e^{-t} - 4e^{-t} + te^{-t} = \left(\frac{4}{15}\,t^{5/2} - 4 + t\right)e^{-t}. \bigcirc$$

Example 4.19

Find a formula for the solutions of $\ddot{y} + p(t)\dot{y} + q(t)y = f(t)$. (Such a formula may also be called a "closed form" solution.)

Method: The best that we can hope for is a formula in terms of $\{y_1(t), y_2(t)\}$, a complete set of basic solutions of the corresponding linear homogeneous ODE, $\ddot{y} + p(t)\dot{y} + q(t)y = 0$. Assume $p(t)$ and $q(t)$ are continuous on an open interval containing $t = 0$. Continuing from (4.36), we have

$$\begin{aligned}\begin{bmatrix}\dot{v}_1\\ \dot{v}_2\end{bmatrix} &= \begin{bmatrix}y_1(t) & y_2(t)\\ \dot{y}_1(t) & \dot{y}_2(t)\end{bmatrix}^{-1}\begin{bmatrix}0\\ f(t)\end{bmatrix} = \frac{1}{W(y_1,y_2)(t)}\begin{bmatrix}\dot{y}_2(t) & -y_2(t)\\ -\dot{y}_1(t) & y_1(t)\end{bmatrix}\begin{bmatrix}0\\ f(t)\end{bmatrix}\\ &= \frac{1}{W(y_1,y_2)(t)}\begin{bmatrix}-y_2(t)f(t)\\ y_1(t)f(t)\end{bmatrix}.\end{aligned}$$

Because, for example, $v_1(t) = c_1 + \int_0^t \dot{v}_1(s)\,ds$, where c_1 is an arbitrary constant,

$$y(t) = \left(c_1 + \int_0^t -\frac{y_2(s)f(s)}{W(y_1,y_2)(s)}\,ds\right)y_1(t) + \left(c_2 + \int_0^t \frac{y_1(s)f(s)}{W(y_1,y_2)(s)}\,ds\right)y_2(t), \tag{4.39}$$

where c_1, c_2 are arbitrary constants. Equation (4.39) gives a formula for all solutions of $\ddot{y} + p(t)\dot{y} + q(t)y = f(t)$. $\bigcirc$

4.3.2 Problems

In problems 1–12, solve the ODE.

1. $y'' + 4y = \frac{1}{\sin 2x}$, where $' = \frac{d}{dx}$
2. $\ddot{y} + y = \sec(t)\csc(t)$.
3. $\ddot{y} + y = \sec^2 t$.
4. $\dot{y} + 4\dot{y} + 4y = \frac{e^{-2t}}{t-1}$
5. $x^2y'' - 2xy' + 2y = x^3$, where $' = \frac{d}{dx}$
6. $x^2y'' - 2xy' + 2y = x^3e^{-3x}$, where $' = \frac{d}{dx}$
7. $x^2y'' - 5xy' + 8y = x^3e^{-x}$, where $' = \frac{d}{dx}$
8. $r^2y'' - 4ry' + 6y = r^4\cos(r)$, where $' = \frac{d}{dr}$
9. $r^2y'' - 6ry' + 12y = r^5\sin 2r$, where $' = \frac{d}{dr}$
10. $r^2y'' + 4ry' + 2y = e^{-r}$, where $' = \frac{d}{dr}$
11. $r^2y'' - 4ry' + 6y = r^2$, where $' = \frac{d}{dr}$
12. $\ddot{y} + 4\dot{y} + 5y = e^{-2t}\sec(t)$
13. Use the method of variation of parameters to solve $\ddot{y} + \dot{y} = e^{-t}$.
14. Use the method of variation of parameters to solve $\ddot{y} + 8\dot{y} + 16y = e^{-4t}$.

In problems 15–17, solve the IVP.

15. $\ddot{y} + 4\dot{y} + 4y = \sqrt{t}\,e^{-2t}$, $y(1) = -1$, $\dot{y}(1) = 0$
16. $y'' + y = \sec(x)$, $y(0) = -1$, $y'(0) = 0$, where $' = \frac{d}{dx}$
17. $r^2y'' + ry' + y = r^2$, $y(1) = 3$, $y'(1) = -1$, where $' = \frac{d}{dr}$
18. For ODE $x^2y'' - 4xy' + 6y = 0$, find the solutions satisfying the given ICs, if possible.
    (a) $y(1) = 0$, $y'(1) = -2$
    (b) $y(0) = 0$, $y'(0) = -2$
    (c) Why does your difficulty in part (b) not contradict the existence and uniqueness conclusion of Theorem 3.8 in Section 3.3?
19. Solve $xy'' - (x+2)y' + 2y = x^3$, where $' = \frac{d}{dx}$, given that the corresponding homogeneous ODE has among its solutions $y = e^x$ and $y = 1 + x + \frac{1}{2}x^2$.
20. Another way to find a particular solution of the ODE in Example 4.18 is to rewrite it as $(D+1)^2[y] = \sqrt{t}\,e^{-t}$, look for a solution in the form $y(t) = e^{-t}v(t)$, and use the shift theorem 3.18 in Section 3.4. [This method was shown to me by Daniel Schepler in the 1990s.]
21. Use the method of undetermined coefficients to solve (4.26), that is, $\ddot{y} + 5\dot{y} + 6y = t^2e^{-t}$. For this problem, give your opinion as to you would rather use undetermined coefficients or variation of parameters, as we did in Example 4.16.

22. Use the Wronskian and the inverse of a 3×3 matrix to find a formula for all solutions of the third-order linear ODE $\dddot{y} + p_1(t)\ddot{y} + p_2(t)\dot{y} + p_3(t)y = f(t)$ in terms of $f(t)$ and a complete set of basic solutions of the corresponding linear homogeneous ODE.

4.4 Laplace Transforms: Basic Techniques

We will see how "Laplace transforms" can turn many ODE problems into algebra problems. Engineers especially prize Laplace transforms because they give techniques to solve ODEs involving terms that are switched on or off. In addition, they give engineers convenient language and techniques for solving control problems "in the frequency domain."

Here, our emphasis is on *using* Laplace transforms rather than how they are defined or calculated using an improper integral, which will be discussed in Appendix B near the end of this book. For our purpose now, the Laplace transform of a function $f(t)$ is

$$\mathcal{L}[f(t)](s),$$

a function only of s. As depicted in Figure 4.14, $\mathcal{L}$ can be thought of as a machine, just as we thought of D, the operator of differentiation.

We need to know what $\mathcal{L}$ *does* so that we can use $\mathcal{L}$ to solve ODEs; justifications for the properties will be in Appendix B.

Theorem 4.6

Assuming the Laplace transforms exist, they satisfy the following properties:

(1) $\mathcal{L}[f(t) + g(t)](s) = \mathcal{L}[f(t)](s) + \mathcal{L}[g(t)](s)$.
(2) $\mathcal{L}[cf(t)](s) = c\,\mathcal{L}[f(t)](s)$, any constant c.
(3) $\mathcal{L}[e^{at}](s) = \frac{1}{s-a}$.
(4) $\mathcal{L}[\dot{y}](s) = s\mathcal{L}[y(t)](s) - y(0)$.

Usually, we denote by an uppercase letter, the Laplace transform of a lower case letter, for example, $F(s) \triangleq \mathcal{L}[f(t)](s)$.

Properties (1)−(4) suffice for solving some ODEs and IVPs.

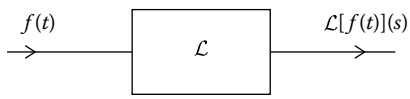


FIGURE 4.14
Laplace transform as a machine.

Example 4.20

Solve

$$\begin{Bmatrix} \dot{y} + 2y = 3e^{-4t} \\ y(0) = 5 \end{Bmatrix}. \tag{4.40}$$

Method: Take $\mathcal{L}$ of both sides of ODE (4.40) and use properties (1) and (2) of Theorem 4.6 to get

$$\mathcal{L}[\dot{y}] + 2\mathcal{L}[y] = 3\mathcal{L}[e^{-4t}].$$

Use properties (3) and (4) and the IC $y(0) = 5$ to get

$$\big(s\mathcal{L}[y] - 5\big) + 2\mathcal{L}[y] = 3 \cdot \frac{1}{s-(-4)}.$$

At this point, we have an algebra problem: solve

$$(s+2)Y(s) = 5 + \frac{3}{s+4}$$

for $Y(s) \triangleq \mathcal{L}[y(t)](s)$ in terms of s:

$$Y(s) = \frac{5}{s+2} + \frac{3}{(s+2)(s+4)}.$$

But, to solve the IVP means to find the solution $y(t)$, not $\mathcal{L}[y](s)$.

Using property (3), we could infer what $y(t)$ is if we could rewrite $Y(s)$ as a linear combination of terms of the form $\frac{1}{s-a}$. To do this, we use the techniques of partial fractions expansions from first-year calculus:

$$\frac{3}{(s+2)(s+4)} = \frac{A}{s+2} + \frac{B}{s+4}, \tag{4.41}$$

where *constants* A, B are to be determined.

To find A, B, multiply (4.41) through by the denominator of its LHS to get

$$3 = A(s+4) + B(s+2). \tag{4.42}$$

Substitute $s = -4$ into (4.42) to get

$$3 = A(-4+4) + B(-4+2) = -2B,$$

so $B = -\frac{3}{2}$. Similarly, substitute $s = -2$ into (4.42) to get

$$3 = A(-2+4) + B(-2+2) = 2A,$$

so $A = \frac{3}{2}$.

We have

$$Y(s) = \frac{5}{s+2} + \left(\frac{3/2}{s+2} - \frac{3/2}{s+4}\right) = \frac{13/2}{s+2} - \frac{3/2}{s+4}.$$

Thinking backward, we infer from $\mathcal{L}[y(t)](s)$ and property (3) that the solution is

$$y(t) = \frac{13}{2}e^{-2t} - \frac{3}{2}e^{-4t}. \bigcirc$$

By the way, the function $Y(s)$ is called the **solution in the frequency domain**.

Just from this example, we observe many things about the method of Laplace transforms for solving ODEs:

Remarks

1. *The ODE and IC were used together to find the solution.* In previous methods, we first found the general solution of the ODE *before* using the IC.
2. *After finding $Y(s)$, we had to "think backward" to find $y(t)$.* This will be formalized as follows by working with the "inverse Laplace transform."
3. *We used algebraic manipulations, including partial fractions, to get $Y(s)$ in a form suitable for finding $y(t)$.* For some simple IVPs, the method of Laplace transform may involve more work than techniques we learned before, such as the method of undetermined coefficients.

Recall that if an $n \times n$ matrix A is invertible, then

$$A\mathbf{x} = \mathbf{b} \iff \mathbf{x} = A^{-1}\mathbf{b}.$$

Similarly, we define the inverse Laplace transform, $\mathcal{L}^{-1}[\ ]$, by

$$\mathcal{L}[f(t)](s) = F(s) \iff f(t) = \mathcal{L}^{-1}[F(s)](t).$$

For example, property (3) and the corresponding inverse property say that

$$\mathcal{L}[e^{at}](s) = \frac{1}{s-a} \iff \mathcal{L}^{-1}\left[\frac{1}{s-a}\right] = e^{at}.$$

Now using this notation, in Example 4.20, we could have written the solution as

$$y(t) = \mathcal{L}^{-1}[Y(s)] = \mathcal{L}^{-1}\left[\frac{13/2}{s+2} - \frac{3/2}{s+4}\right] = \frac{13}{2}e^{-2t} - \frac{3}{2}e^{-4t}.$$

For every property of Laplace transforms, there is a corresponding property for inverse Laplace transforms.

In order to solve a wider variety of ODEs and IVPs, we need more facts about Laplace transforms.

Theorem 4.7

Laplace transforms exist with the following properties:

(5) $\mathcal{L}[t^n](s) = \frac{n!}{s^n}$, $n = 0, 1, 2, \ldots$.

(6) $\mathcal{L}[\cos \omega t](s) = \dfrac{s}{s^2 + \omega^2}$.

(7) $\mathcal{L}[\sin \omega t](s) = \dfrac{\omega}{s^2 + \omega^2}$.

(8) $\mathcal{L}[e^{at}f(t)](s) = F(s-a)$, where $F(s) = \mathcal{L}[f(t)](s)$.

(9) $\mathcal{L}[\ddot{y}](s) = s^2\mathcal{L}[y](s) - sy(0) - \dot{y}(0)$.

Property (8) can also be stated as

$$\mathcal{L}[\,e^{at}f(t)](s) = \mathcal{L}[f(t)]\Big|_{s\mapsto(s-a)},$$

which is equivalent to the property

$$\mathcal{L}^{-1}[\,F(s-a)] = e^{at}\cdot\mathcal{L}^{-1}[\,F(s)].$$

Properties (4) and (9) can be generalized to

$$\mathcal{L}[\,y^{(n)}](s) = s^n\mathcal{L}[\,y(s)] - s^{n-1}y(0) - s^{n-2}\dot{y}(0) - \cdots - sy^{(n-2)}(0) - y^{(n-1)}(0).$$

Example 4.21

Find the Laplace transforms of the following given functions: (a) $2t^3 - 5t + 7$, (b) $5\cos 2t - \sin 3t$, and (c) $e^{-2t}\cos t + e^{-t}\sin 2t + te^{-3t}$.

Method:

(a)

$$\mathcal{L}[\,2t^3 - 5t + 7] = 2\mathcal{L}[\,t^3] - 5\mathcal{L}[\,t] + 7\mathcal{L}[\,t^0] = 2\,\frac{3!}{s^4} - 5\,\frac{1!}{s^2} + 7\,\frac{0!}{s^1} = \frac{12}{s^4} - \frac{5}{s^2} + \frac{7}{s}.$$

(b)

$$\mathcal{L}[\,5\cos 2t - \sin 3t] = 5\,\frac{s}{s^2+2^2} - \frac{3}{s^2+3^2} = \frac{5s}{s^2+4} - \frac{3}{s^2+9}.$$

(c)

$$\begin{aligned}
&\mathcal{L}[\,e^{-2t}\cos t + e^{-t}\sin 2t + te^{-3t}]\\
&\quad= \mathcal{L}[\,\cos t]\Big|_{s\mapsto(s-(-2))} + \mathcal{L}[\,\sin 2t]\Big|_{s\mapsto(s-(-1))} + \mathcal{L}[\,t]\Big|_{s\mapsto(s-(-3))}\\
&\quad= \frac{s}{s^2+1}\Big|_{s\mapsto(s+2))} + \frac{2}{s^2+4}\Big|_{s\mapsto(s+1))} + \frac{1}{s^2}\Big|_{s\mapsto(s+3))}\\
&\quad= \frac{s+2}{(s+2)^2+1} + \frac{2}{(s+1)^2+4} + \frac{1}{(s+3)^2}.\ \bigcirc
\end{aligned}$$

Example 4.22

Find the inverse Laplace transforms of the following:

$$\text{(a) } \mathcal{L}^{-1}\left[\frac{3s+4}{s^2+2}\right], \quad \text{(b) } \mathcal{L}^{-1}\left[\frac{(s+3)+5}{(s+3)^2+4}\right], \quad \text{and} \quad \text{(c) } \mathcal{L}^{-1}\left[\frac{3s+8}{s^2+4s+5}\right].$$

Method:

(a)

$$\mathcal{L}^{-1}\left[\frac{3s+4}{s^2+2}\right] = 3\mathcal{L}^{-1}\left[\frac{s}{s^2+2}\right] + 4\mathcal{L}^{-1}\left[\frac{1}{s^2+2}\right] = 3\cos(\sqrt{2}\,t) + 4\frac{1}{\sqrt{2}}\,\mathcal{L}^{-1}\left[\frac{\sqrt{2}}{s^2+2}\right]$$

$$= 3\cos(\sqrt{2}\,t) + 2\sqrt{2}\sin(\sqrt{2}\,t).$$

(b)

$$\mathcal{L}^{-1}\left[\frac{(s+3)+5}{(s+3)^2+4}\right] = \mathcal{L}^{-1}\left[F(s-(-3))\right] = e^{-3t}\mathcal{L}^{-1}\left[F(s)\right] = e^{-3t}\mathcal{L}^{-1}\left[\frac{s+5}{s^2+4}\right]$$
$$= e^{-3t}\left(\cos 2t + \frac{5}{2}\sin 2t\right).$$

(c) The denominator is not of the form $s^2+\omega^2$ but can be rewritten in the form $(s-a)^2+\omega^2$ after completing the square:

$$s^2+4s+5 = \left(s^2+4s+4-4\right)+5 = \left(s^2+4s+4\right)-4+5 = (s+2)^2+1.$$

After that, we need to rewrite the numerator to be in terms of $(s+2)$ so that we can use property (8) backward:

$$\mathcal{L}^{-1}\left[\frac{3s+8}{s^2+4s+5}\right] = \mathcal{L}^{-1}\left[\frac{3s+8}{(s+2)^2+1}\right] = \mathcal{L}^{-1}\left[\frac{3\left((s+2)-2\right)+8}{(s+2)^2+1}\right]$$

$$= \mathcal{L}^{-1}\left[\frac{3(s+2)+2}{(s+2)^2+1}\right] = e^{-2t}\mathcal{L}^{-1}\left[\frac{3s+2}{s^2+1}\right] = e^{-2t}\left(3\cos t + 2\sin t\right). \bigcirc$$

Example 4.23

Solve

$$\left\{\begin{matrix} \ddot{y}+4\dot{y}+5y = e^{-t} \\ y(0)=-3, \dot{y}(0)=7 \end{matrix}\right\}. \tag{4.43}$$

Method: Using property (9) of Theorem 4.7, take $\mathcal{L}$ of both sides of ODE (4.43) to get

$$s^2\mathcal{L}[y](s) - s(-3) - 7 + 4\left(s\mathcal{L}[y](s) - (-3)\right) + 5\mathcal{L}[y](s) = \frac{1}{s+1}.$$

For convenience, denote $Y(s) = \mathcal{L}[y](s)$. Combining all terms that involve $Y(s)$ and moving all other terms to the RHS, we have

$$(s^2+4s+5)\mathcal{L}[y](s) = -3s-5+\frac{1}{s+1}.$$

So,

$$y(t) = \mathcal{L}^{-1}\left[\frac{-3s-5}{s^2+4s+5} + \frac{1}{(s+1)(s^2+4s+5)}\right].$$

To handle the second term, use the partial fractions expansion

$$\frac{1}{(s+1)(s^2+4s+5)} = \frac{A}{s+1} + \frac{Bs+C}{s^2+4s+5},$$

where A, B, C are constants to be determined. Multiply through by the denominator of the LHS:

$$1 = A(s^2+4s+5) + (Bs+C)(s+1). \tag{4.44}$$

Substitute $s=-1$ into (4.44) to get

$$1 = A\left((-1)^2+4(-1)+5\right) + (B(-1)+C)\cdot 0 = 2A.$$

Substitute $A = \frac{1}{2}$ into (4.44) and move terms to the LHS to get

$$1 - \frac{1}{2}(s^2 + 4s + 5) = (Bs + C)(s + 1).$$

After dividing through by $(s + 1)$, we have

$$Bs + C = \frac{1 - \frac{1}{2}s^2 - 2s - \frac{5}{2}}{s+1} = \frac{-\frac{1}{2}s^2 - 2s - \frac{3}{2}}{s+1} = -\frac{1}{2} \cdot \frac{s^2 + 4s + 3}{s+1} = -\frac{1}{2}\frac{\cancel{(s+1)}(s+3)}{\cancel{(s+1)}}$$
$$= -\frac{1}{2}(s + 3).$$

So,

$$y(t) = \mathcal{L}^{-1}\left[\frac{-3s-5}{s^2+4s+5} + \frac{\frac{1}{2}}{s+1} + \frac{-\frac{1}{2}(s+3)}{s^2+4s+5}\right]$$
$$= \frac{1}{2}\left(\mathcal{L}^{-1}\left[\frac{1}{s+1}\right] + \mathcal{L}^{-1}\left[\frac{-7s-13}{s^2+4s+5}\right]\right)$$
$$= \frac{1}{2}\left(e^{-t} + \mathcal{L}^{-1}\left[\frac{-7\left((s+2)-2\right)-13}{(s+2)^2+1}\right]\right) = \frac{1}{2}\left(e^{-t} + \mathcal{L}^{-1}\left[\frac{-7(s+2)+1}{(s+2)^2+1}\right]\right)$$
$$= \frac{1}{2}\left(e^{-t} + e^{-2t}\mathcal{L}^{-1}\left[\frac{-7s+1}{s^2+1}\right]\right) = \frac{1}{2}\left(e^{-t} + (-7\cos t + \sin t)e^{-2t}\right). \bigcirc$$

Remark

In Example 4.23, after we divided through by $(s+1)$, it was good that the polynomial $Bs+C$ turned out to be equal to a polynomial such as $-\frac{1}{2}(s+3)$. If we had *not* been able to cancel out factors of $(s+1)$, hence we had seemed to get $Bs+C$ *not* equal to a polynomial, then an alarm bell should go off. Like some computer error messages, an alarm bell may not tell us exactly where we made an error but merely that we've made an error somewhere.

Example 4.24

Find the steady-state solution of

$$\ddot{y} + 4\dot{y} + 5y = 6\cos 2t. \tag{4.45}$$

Method: Note that ICs are not given in this problem. Take $\mathcal{L}$ of both sides of ODE (4.45) to get

$$s^2\mathcal{L}[y](s) - sy(0) - \dot{y}(0) + 4\left(s\mathcal{L}[y](s) - y(0)\right) + 5\mathcal{L}[y](s) = \frac{6s}{s^2+4}.$$

For convenience, denote $Y(s) = \mathcal{L}[y](s)$. Solve for $Y(s)$ in terms of s:

$$Y(s) = \frac{sy(0) + \dot{y}(0) + 4y(0)}{s^2+4s+5} + \frac{6s}{(s^2+4)(s^2+4s+5)}.$$

Because $s^2+4s+5 = (s+2)^2+1$, the first terms in $y(t) = \mathcal{L}^{-1}[Y(s)]$ will be of the form e^{-2t} times cosine and sine functions and hence will be part of the transient solution. So, the ICs play no role in the steady-state solution *in this example*.

We use the partial fraction expansion

$$\frac{6s}{(s^2+4)(s^2+4s+5)} = \frac{As+B}{s^2+4} + \frac{Cs+E}{s^2+4s+5}, \tag{4.46}$$

where A, B, C, E are constants to be determined. Multiply through by the denominator of the LHS of (4.46) to get

$$6s = (As + B)(s^2 + 4s + 5) + (Cs + E)(s^2 + 4). \tag{4.47}$$

Unfortunately, substituting in real values of s will not wipe out any of the coefficients A, B, C, E. We could get four equations in A, B, C, E by sorting (4.47) by powers of s. Alternatively, we will substitute four convenient values of s into (4.47), specifically $s = 0, 1, -1, -2$ to get, respectively,

$$\begin{Bmatrix} @s=0: & 0= & 5B & +4E \\ @s=1: & 6= & 10(A+B) & +5(C+E) \\ @s=-1: & -6= & 2(-A+B) & +5(-C+E) \\ @s=-2: & -12= & (-2A+B) & +8(-2C+E). \end{Bmatrix}.$$

The solution of this system of four equations in four unknowns is

$$\begin{bmatrix} A \\ B \\ C \\ E \end{bmatrix} = \begin{bmatrix} 0 & 5 & 0 & 4 \\ 10 & 10 & 5 & 5 \\ -2 & 2 & -5 & 5 \\ -2 & 1 & -16 & 8 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 6 \\ -6 \\ -12 \end{bmatrix}.$$

Only $A = \frac{6}{65}$, $B = \frac{96}{65}$ figure into the steady-state solution, which is

$$y_S(t) = \frac{6}{65}\mathcal{L}^{-1}\left[\frac{s+16}{s^2+4}\right] = \frac{6}{65}\left(\cos 2t + 8\sin 2t\right). \bigcirc$$

Table 4.1 has a short summary of facts concerning Laplace transforms.

4.4.1 Problems

In problems 1–5, find the Laplace transform of the given function.

1. $-5e^{3t} + \sin 2t$
2. $\cos 3t + \sin(\frac{t}{2})$
3. $1 + at + \frac{1}{2!}(at)^2 + \frac{1}{3!}(at)^3$, where a is an unspecified constant
4. t^3e^{-2t}
5. $e^{t/2}\cos\left(\frac{\sqrt{3}}{2}t\right)$

In problems 6–11, find the inverse Laplace transform of the given function.

6. $\frac{5}{s^3} - \frac{3s-1}{s^2+4}$
7. $\frac{s-2}{(s+2)(s^2+1)}$
8. $\frac{s+1}{s^2-4s+5}$
9. $\frac{-s+4}{s^2+4s+7}$
10. $\frac{2s^2+4s+8}{s^3-4s}$
11. $\frac{s+6}{(s+3)^2(s^2+2s+2)}$

TABLE 4.1

First Table of Laplace Transforms, Where c, ω Are Positive Constants

Formula	$f(t)$	$F(s) = \mathcal{L}[f(t)]$
L1.1	$f(t) + g(t)$	$F(s) + G(s)$
L1.2	$cf(t)$	$cF(s)$
L1.3	e^{at}	$\frac{1}{s-a}$
L1.4	$\dot{y}$	$sY(s) - y(0)$
L1.5	t^n	$\frac{n!}{s^{n+1}}$, integer $n \geq 0$
L1.6	$\sin \omega t$	$\frac{\omega}{s^2+\omega^2}$
L1.7	$\cos \omega t$	$\frac{s}{s^2+\omega^2}$
L1.8	$e^{at}f(t)$	$F(s-a)$
L1.9	$\ddot{y}$	$s^2Y(s) - sy(0) - \dot{y}(0)$
L1.10	$step(t-c)$	$\frac{1}{s}\ e^{-cs}$
L1.11	$g(t)step(t-c)$	$e^{-cs}\mathcal{L}[g(t+c)](s)$
L1.12	$(f * g)(t)$	$F(s)G(s)$
L1.13	$t \sin \omega t$	$\frac{2\omega s}{(s^2+\omega^2)^2}$
L1.14	$t \cos \omega t$	$\frac{s^2-\omega^2}{(s^2+\omega^2)^2}$
L1.15	$\delta(t-c)$	e^{-cs}

12. Find $y(t) = \mathcal{L}^{-1}\left[\frac{s}{s^2+4s+13}\right]$ and express it in an amplitude-phase form $Ae^{\alpha t}\cos(\nu t - \delta)$.

In problems 13–16, use Laplace transforms to solve the given IVP. If the problem has a steady-state solution, then note what it is.

13. $\dot{y} - 2y = 3e^{4t},\ y(0) = -1$
14. $\dot{y} + 2y = \cos 4t,\ y(0) = -1$
15. $\ddot{y} + 3\dot{y} - 10y = 0,\ y(0) = 1,\ \dot{y}(0) = -3$
16. $\ddot{y} + 9y = 5\sin 2t,\ y(0) = 1,\ \dot{y}(0) = -3$
17. For the IVP $\ddot{y} + 9y = 10te^{-t},\ y(0) = 0,\ \dot{y}(0) = 0$, find the solution using (a) the method of undetermined coefficients, and (b) the method of Laplace transforms. Which method do you prefer for this problem?
18. For the IVP $\ddot{y} + 4y = e^{-t}\cos 2t,\ y(0) = 0,\ \dot{y}(0) = 0$, find the solution using (a) the method of undetermined coefficients and (b) the method of Laplace transforms. Which method do you prefer for this problem?
19. For ODE $\ddot{y} + 2\dot{y} + 2y = \sin t$, use Laplace transforms to find the steady-state solution.
20. Suppose $Y(s) = \frac{s-2}{(s^2+1)(s^2+2s+5)}$ is the Laplace transform of a solution of an IVP.
    (a) Find the steady-state solution of the IVP.
    (b) Find an IVP for which the given function $Y(s)$ is the Laplace transform of a solution.

In each of problems 21 and 22, suppose a forced oscillator problem has a solution, $y(t)$. *Without* finding the inverse Laplace transform, decide whether the ODE fits the case of

resonance, beats phenomenon, steady-state oscillation, or none of the above. Of course, give at least a brief explanation for how you reached your conclusion. In addition, if the solution is in the steady-state oscillation case, find the steady-state solution.

21. $\mathcal{L}[\,y(t)] = \frac{s+1}{s^2+3} - \frac{s+2}{s^2+1}$
22. $\mathcal{L}[\,y(t)] = \frac{s-1}{s^2+4s+10} - \frac{s+2}{s^2+9}$
23. The Taylor polynomial
$$p_n(t) \triangleq 1 + t + \frac{1}{2!}(t)^2 + \cdots + \frac{1}{n!}(t)^n$$
is an approximation of e^t. Find $\mathcal{L}\left[\,p_n(t)\right]$ and explain, using partial sums of a geometric series, why
$$\mathcal{L}\left[\,p_n(t)\,\right] \to \mathcal{L}\left[\,e^t\right], \quad \text{as} \quad n \to \infty, \text{ for } s > 1.$$

4.5 Laplace Transforms: Unit Step and Other Techniques

As mentioned earlier, Laplace transforms are particularly useful for solving ODEs involving terms that are switched on or off, for example, in electrical circuits. In this section, we will study and apply "unit step functions," as well as "convolution" and its relationship to the "transfer function" and "delta functions." In addition, we will work with periodic functions often used by engineers to analyze systems.

Definition 4.3

The **unit step function** switched on at $t = c$ is

$$step(t-c) \triangleq \begin{Bmatrix} 0, & t < c \\ 1, & t \geq c \end{Bmatrix}, \tag{4.48}$$

where c is a nonnegative constant.

As in the previous section, our emphasis is on *using* Laplace transforms rather than how they are defined or calculated using an improper integral. But we definitely need to know what the Laplace transform does.

Theorem 4.8

(Properties of unit step functions): if c is a nonnegative constant, then

(10) $\mathcal{L}[\,step(t-c)](s) = \frac{1}{s}\,e^{-cs}$.
(11) $\mathcal{L}[\,g(t)step(t-c)](s) = e^{-cs} \cdot \mathcal{L}[\,g(t+c)](s)$.

We refer to property (11) as the horizontal shift theorem because of its shift of the function g's input on the RHS.

Example 4.25

Find the Laplace transforms of the given functions:

(a) $\mathcal{L}[\,step(t-3)](s)$, (b) $\mathcal{L}[\,(t-3)step(t-3)](s)$, and (c) $\mathcal{L}[\,t\,step(t-3)](s)$.

Method:

(a) $\mathcal{L}[step(t-3)](s) = \frac{1}{s}e^{-3s}$.

(b) $\mathcal{L}[(t-3)step(t-3)](s) = e^{-3s}\mathcal{L}[\,g(t+3)]$, where $g(t) = (t-3)$. So,

$$\mathcal{L}[(t-3)step(t-3)](s) = e^{-3s}\mathcal{L}\Big[(t-3)\Big|_{t\mapsto(t+3)}\Big] = e^{-3s}\mathcal{L}[(t+3)-3] = e^{-3s}\mathcal{L}[t] = \frac{1}{s^2}e^{-3s}.$$

(c) $\mathcal{L}[t\,step(t-3)](s) = e^{-3s}\mathcal{L}[h(t+3)$, where $h(t) = t$. So,

$$\mathcal{L}[t\,step(t-3)](s) = e^{-3s}\mathcal{L}\Big[\,t\Big|_{t\mapsto(t+3)}\Big] = e^{-3s}\mathcal{L}[\,t+3] = \left(\frac{1}{s^2}+\frac{3}{s}\right)e^{-3s}. \bigcirc$$

As earlier, corresponding to every property of Laplace transforms is a property of the inverse Laplace transforms. Corresponding to property (11) is the next result.

Corollary 4.1

If $g(t) = \mathcal{L}^{-1}[\,G(s)]$, then

$$\mathcal{L}^{-1}\left[\,e^{-cs}G(s)\right] = g(t-c)step(t-c) = g(t)\Big|_{t\mapsto(t-c)}step(t-c).$$

Example 4.26

Find the inverse Laplace transforms of the given functions:
(a) $\mathcal{L}^{-1}\left[\frac{s}{s^2+4}\,e^{-s}\right]$ and (b) $\mathcal{L}^{-1}\left[\frac{1}{s^3}\,e^{-5s}\right]$.

Method:

(a) $\mathcal{L}^{-1}\left[\frac{s}{s^2+4}\,e^{-s}\right] = \mathcal{L}^{-1}\left[\frac{s}{s^2+4}\right]\Big|_{t\mapsto(t-1)}step(t-1) = \cos 2t\Big|_{t\mapsto(t-1)}step(t-1)$
$= \cos(2(t-1))\,step(t-1)$.

(b) $\mathcal{L}^{-1}\left[\,\frac{1}{s^3}\,e^{-5s}\right] = \mathcal{L}^{-1}\left[\,\frac{1}{2!}\cdot\frac{2!}{s^3}\,\right]\Big|_{t\mapsto(t-5)}step(t-5) = \frac{1}{2}t^2|_{t\mapsto(t-5)}step(t-5)$
$= \frac{1}{2}(t-5)^2 step(t-5). \bigcirc$

4.5.1 Writing a Function in Terms of Step Function(s)

Example 4.27

Write $f(t) = \begin{Bmatrix} 0, & 0 \le t < 2 \\ 1, & 2 \le t < 5 \\ 0, & 5 \le t < \infty \end{Bmatrix}$ in terms of step functions.

Method: Concerning only the time interval $0 \le t < 5, f(t)$ has a "1" switched on at $t = 2$. So, for $0 \le t < 5$, $f(t) = step(t-2)$. But then at $t = 5$, the "1" is switched off, or, equivalently, a "-1" is switched on. So,

$$f(t) = 1 \cdot step(t-2) + (-1) \cdot step(t-5) = step(t-2) - step(t-5). \bigcirc$$

Example 4.28

Write $f(t) = \begin{Bmatrix} 0, & 0 \le t < 2 \\ 1, & 2 \le t < 5 \\ t-4, & 5 \le t < \infty \end{Bmatrix}$ in terms of step functions.

Method 1: Just as for Example 4.27, on the interval $0 \le t < 5, f(t) = step(t-2)$. At $t = 5$, the "1" is switched off, or, equivalently, a "-1" is switched on, *and* also "$(t-4)$" is switched on. So,

$$f(t) = 1 \cdot step(t-2) + ((-1) + (t-4)) \cdot step(t-5) = step(t-2) + (t-5) \cdot step(t-5). \bigcirc$$

Method 2: We can rewrite $f(t)$ as

$$f(t) = \begin{Bmatrix} 0, & 0 \le t < 2 \\ 1, & 2 \le t < 5 \\ 1 - 1 + (t-4), & 5 \le t < \infty \end{Bmatrix} = \begin{Bmatrix} 0, & 0 \le t < 2 \\ 1, & 2 \le t < 5 \\ 1, & 5 \le t < \infty \end{Bmatrix} + \begin{Bmatrix} 0, & 0 \le t < 2 \\ 0, & 2 \le t < 5 \\ t-5, & 5 \le t < \infty \end{Bmatrix}$$

$$= \begin{Bmatrix} 0, & 0 \le t < 2 \\ 1, & 2 \le t < \infty \end{Bmatrix} + (t-5) \begin{Bmatrix} 0, & 0 \le t < 5 \\ 1, & 5 \le t < \infty \end{Bmatrix} = step(t-2) + (t-5)step(t-5). \bigcirc$$

We presented two slightly different methods, one that is essentially graphical and the other algebraic, so that you may choose which method *you* find to be the more comfortable and accurate. Of course, you have the option to use both methods to check your conclusions.

By the way, other authors use other notations for the unit step function such as $H(t-c), u_c(t)$, or $u(t-c)$. The H stands for Oliver Heaviside, a great electrical engineer, and the u stands for "unit step."

Example 4.29

Solve the IVP

$$\begin{Bmatrix} \dot{y} + \frac{1}{2}y = 3step(t-4) \\ y(0) = 5 \end{Bmatrix}. \tag{4.49}$$

The ODE models the RC series circuit with $R = 2\ \Omega$, $C = 1\ F$, and voltage source $V(t) = 6step(t-4)$ volts, a DC source switched on at $t = 4$, shown in Figure 4.15.

Method: Denote $Y(s) = \mathcal{L}[y(t)](s)$. Take the Laplace transform of both sides and use properties (4) and (10):

$$\left(s + \frac{1}{2}\right) Y(s) = 5 + \frac{3}{s} e^{-4s}.$$

To find

$$y(t) = \mathcal{L}^{-1}[Y(s)] = \mathcal{L}^{-1}\left[\frac{5}{s+\frac{1}{2}} + \frac{3}{s\left(s+\frac{1}{2}\right)} e^{-4s}\right],$$

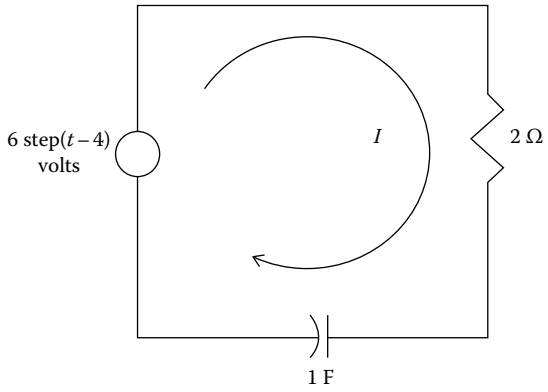


FIGURE 4.15
RC circuit with step-function voltage source.

use the partial fractions expansion

$$\frac{3}{s(s+\frac{1}{2})} = \frac{A}{s} + \frac{B}{s+\frac{1}{2}},$$

where A, B are constants to be determined.

Caution: The e^{-4s} factor is *not* part of the partial fractions work.

Multiply through by the denominator of the LHS to get

$$3 = A\left(s+\frac{1}{2}\right) + Bs.$$

Substitute $s = 0$ in to get $A = 6$; substitute $s = -\frac{1}{2}$ in to get $B = -6$.

So,

$$y(t) = \mathcal{L}^{-1}\left[\frac{5}{s+\frac{1}{2}} + \left(\frac{6}{s} - \frac{6}{s+\frac{1}{2}}\right)e^{-4s}\right] = 5e^{-t/2} + \left(6 - 6e^{-t/2}\right)\Big|_{t\mapsto(t-4)} step(t-4)$$
$$= 5e^{-t/2} + 6\left(1 - e^{-(t-4)/2}\right) step(t-4). \bigcirc$$

4.5.2 Graph of a Solution of an ODE Involving a Step Function

Recall that the graph of $y = g(t-c)$ is the same as the graph of $y = g(t)$ but shifted c units to the right. For example, the graph of Figure 4.16b, $6\left(1 - e^{-(t-4)/2}\right) step(t-4)$, is the same as the graph of Figure 4.16a, $6\left(1 - e^{-t/2}\right) step(t-0)$ shifted 4 units to the right.

To understand the graph of the solution of Example 4.29, that is,

$$y(t) = 5e^{-t/2} + 6\left(1 - e^{-(t-4)/2}\right) step(t-4),$$

it helps to rewrite it as

$$y(t) = 5e^{-t/2} + 6\left(1 - e^{-(t-4)/2}\right) \cdot \begin{Bmatrix} 0, & 0 \le t < 4 \\ 1, & t \ge 4 \end{Bmatrix} = 5e^{-t/2} + \begin{Bmatrix} 0, & t < 4 \\ 6\left(1 - e^{-(t-4)/2}\right), & t \ge 4 \end{Bmatrix}$$

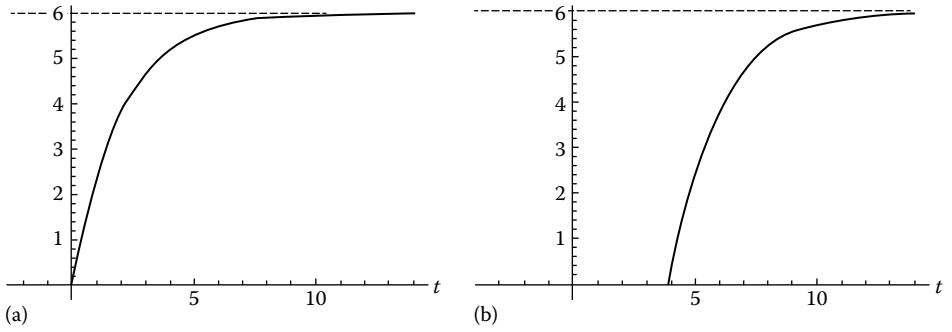


FIGURE 4.16
Horizontal shift: (a) $6\,(1-e^{-t/2})\;step(t-0)$ and (b) $6\,(1-e^{-(t-4)/2})\;step(t-4)$.

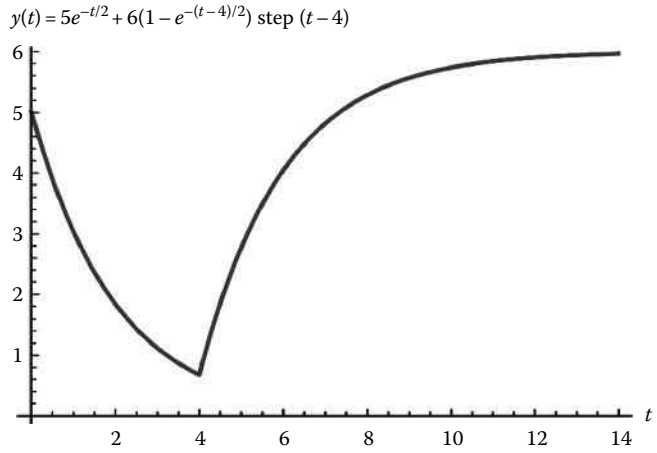


FIGURE 4.17
Solution of Example 4.29.

$$= \begin{Bmatrix} 5e^{-t/2}, & t < 4 \\ 5e^{-t/2} + 6\left(1 - e^{-(t-4)/2}\right), & t \geq 4 \end{Bmatrix}.$$

The graph of that solution is in Figure 4.17. We see clearly that the effect of the ODEs forcing term $3step(t-4)$ on the solution of the ODE is to switch on at $t = 4$ a horizontal shift of the function $6\left(1 - e^{-t/2}\right)$.

Notice that the graph of $y(t)$ is continuous at $t = 4$, but its derivative has a finite jump there. This makes sense because the ODE tells us that

$$\lim_{t \to 4^{\pm}} \dot{y}(t) = \lim_{t \to 4^{\pm}} -\frac{1}{2} y(t) + 3\,step(t-4) = \begin{Bmatrix} -\frac{5}{2}e^{-2}, & \text{as } t \to 4^{-} \\ -\frac{5}{2}e^{-2} + 3, & \text{as } t \to 4^{+} \end{Bmatrix}. \bigcirc$$

Besides switching on or off a source in an electrical circuit, another physical example of a step function occurs if we are driving, and for a certain time interval $[t_0, t_1]$, we drive

over "rumble strips" warning us to slow down or to stay off the shoulder of the road. The vibrations of the suspension system could be modeled by the ODE

$$m\ddot{y} + b\dot{y} + ky = f(t),$$

where $f(t)$ is a sum of **pulse** terms $f_0 \cdot \big(step(t - t_0) - step\ \ (t - t_1)\big)$.

4.5.3 Convolution

Definition 4.4

The "half-line" **convolution**, or "convolution," for short, of functions f and g is defined to be

$$(f * g)(t) \triangleq \int_0^t f(t-u)g(u)\,du. \tag{4.50}$$

Note that $(f * g)(t)$ is a function only of the variable t, because the u variable gets "integrated out" of the definite integral. Note also that t appears twice in the formula: as the upper limit of integration and inside the f function evaluation.

Note for Chapter 9 that in the subject of Fourier transforms, convolutions are integrals over the whole real line, that is, $\int_{-\infty}^{\infty}$.

Theorem 4.9

Properties of convolution

(a) $(f * g)(t) = (g * f)(t)$.
(b) $\big((f_1 + f_2) * g\big)\,(t) = (f_1 * g)(t) + (f_2 * g)(t)$.
(c) $\big((cf) * g\big)\,(t) = c(f * g)(t)$, if c is a constant.

Theorem 4.10

Property of Laplace transform of a convolution
(12) $\mathcal{L}[(f * g)(t)](s) = \mathcal{L}[f(t)](s) \cdot \mathcal{L}[g(t)](s)$.

Corresponding to property (12) is a property of inverse Laplace transforms:

$$\mathcal{L}^{-1}[F(s)G(s)] = \Big(\mathcal{L}^{-1}[F(s)] * \mathcal{L}^{-1}[G(s)]\Big)\,(t).$$

Example 4.30

Find the convolution of $e^t * t$ by two methods: (a) by using the definition of convolution and (b) using Theorem 4.10.

Method 1:

(a) Using a law of exponents and then integration by parts,

$$e^t * t \triangleq \int_0^t e^{t-u}\, u\, du = \int_0^t e^t e^{-u}\, u\, du = e^t \int_0^t e^{-u}\, u\, du = e^t \left[-ue^{-u} - e^{-u} \right]_0^t$$

$$= e^t \left(-te^{-t} - e^{-t} - (0-1) \right) = -t - 1 + e^t.$$

(b)

$$e^t * t = \mathcal{L}^{-1}\Big[\mathcal{L}[e^t * t]\Big] = \mathcal{L}^{-1}\Big[\mathcal{L}[\,e^t] \cdot \mathcal{L}[t]\Big] = \mathcal{L}^{-1}\left[\frac{1}{s-1} \cdot \frac{1}{s^2}\right].$$

Partial fractions gives

$$\frac{1}{(s-1)s^2} = \frac{A}{s-1} + \frac{B}{s} + \frac{C}{s^2},$$

where A, B, C are constants to be determined. Multiply through by the denominator of the LHS:

$$1 = As^2 + Bs(s-1) + C(s-1). \tag{4.51}$$

Substitute in $s = 0$ to get $1 = -C$, so $C = -1$, and substitute in $s = 1$ to get $1 = A$. Substitute the values $A = 1, C = -1$ into (4.51) to get

$$1 = s^2 + Bs(s-1) - (s-1).$$

Solve for B by putting all other terms on the LHS and then dividing through by $s(s-1)$:

$$B = \frac{\cancel{1} - s^2 + (s - \cancel{1})}{s(s-1)} = \frac{-s^2 + s}{s(s-1)} = \frac{\cancel{-s(s-1)}}{\cancel{s(s-1)}} = -1.$$

So, again,

$$e^t * t = \mathcal{L}^{-1}\left[\frac{1}{s-1} \cdot \frac{1}{s^2}\right] = \mathcal{L}^{-1}\left[\frac{1}{s-1} - \frac{1}{s} - \frac{1}{s^2}\right] = e^t - 1 - t. \bigcirc$$

Remark

One very quick "reality check" you can make on a calculation of a convolution is that $(f * g)(0)$ should be zero because $(f * g)(0) = \int_0^0 f(0-u)g(u)\, du = 0$ if the functions f and g are reasonable.

Example 4.31

Find $\mathcal{L}^{-1}\left[\dfrac{1}{(s+1)(s^2+4)}\right]$ using a convolution.

Method: $\mathcal{L}^{-1}\left[\dfrac{1}{(s+1)(s^2+4)}\right] = e^{-t} * \frac{1}{2}\sin 2t = \int_0^t e^{-(t-u)}\frac{1}{2}\sin 2u\, du$

$$= \int_0^t e^{-t}e^u \frac{1}{2}\sin 2u\, du = \frac{1}{2}e^{-t}\int_0^t e^u \sin 2u\, du = \frac{1}{2}e^{-t} \cdot \left[\frac{e^u}{1^2+2^2}(\sin 2u - 2\cos 2u)\right]_0^t$$

$$= \frac{1}{10}e^{-t}\left(e^{t}\left(\sin 2t - 2\cos 2t\right) + 2\right) = \frac{1}{5}e^{-t} + \frac{1}{10}\left(\sin 2t - 2\cos 2t\right). \bigcirc$$

Using convolution, we can get the next results.

Theorem 4.11

Properties of Laplace transforms of solutions of problems with resonance:

(13) $\mathcal{L}[\,t\sin\omega t](s) = \frac{2\omega s}{(s^2+\omega^2)^2}$.

(14) $\mathcal{L}[\,t\cos\omega t](s) = \frac{s^2-\omega^2}{(s^2+\omega^2)^2}$.

Example 4.32

Find (a) $\mathcal{L}^{-1}\left[\frac{9}{(s^2+9)^2}\right]$, and (b) $\mathcal{L}^{-1}\left[\frac{s^2}{(s^2+9)^2}\right]$.

Method: Using algebra work on numerators, we have

(a)

$$\mathcal{L}^{-1}\left[\frac{9}{(s^2+9)^2}\right] = \mathcal{L}^{-1}\left[\frac{\frac{1}{2}(s^2+9) - \frac{1}{2}(s^2-9)}{(s^2+9)^2}\right]$$

$$= \frac{1}{2}\mathcal{L}^{-1}\left[\frac{1}{s^2+9}\right] - \frac{1}{2}\mathcal{L}^{-1}\left[\frac{s^2-9}{(s^2+9)^2}\right] = \frac{1}{6}\sin 3t - \frac{1}{2}t\cos 3t.$$

(b)

$$\mathcal{L}^{-1}\left[\frac{s^2}{(s^2+9)^2}\right] = \mathcal{L}^{-1}\left[\frac{\frac{1}{2}(s^2+9) + \frac{1}{2}(s^2-9)}{(s^2+9)^2}\right] = \frac{1}{2}\mathcal{L}^{-1}\left[\frac{1}{s^2+9}\right] + \frac{1}{2}\mathcal{L}^{-1}\left[\frac{s^2-9}{(s^2+9)^2}\right]$$

$$= \frac{1}{6}\sin 3t + \frac{1}{2}t\cos 3t. \bigcirc$$

4.5.4 Convolution and Particular Solutions

Suppose p, q are constants. We want to find a formula for a particular solution, $y_p(t)$, of

$$\ddot{y} + p\dot{y} + qy = f(t). \tag{4.52}$$

Take the Laplace transform of the ODE:

$$(s^2 + ps + q)Y_p(s) - sy_p(0) - \dot{y}_p(0) - py_p(0) = F(s),$$

where $Y_p(s) = \mathcal{L}[\,y_p(t)](s)$ and $F(s) = \mathcal{L}[f(t)](s)$. Because we are interested only in a particular solution, we may ignore the $-sy(0) - \dot{y}(0) - py(0)$ terms. So, let

$$Y_p(s) = \frac{1}{s^2+ps+q}\cdot F(s).$$

So,

$$y_p(t) = \mathcal{L}^{-1}[Y_p(s)] = \mathcal{L}^{-1}\left[\frac{1}{s^2+ps+q}\cdot F(s)\right] = \mathcal{L}^{-1}\left[\frac{1}{s^2+ps+q}\right] * \mathcal{L}^{-1}[F(s)]$$
$$= \mathcal{L}^{-1}\left[\frac{1}{s^2+ps+q}\right] * f(t).$$

So,

$$y_p(t) = \int_0^t y_2(t-u)f(u)\,du,$$

where $y_2(t)$ is the solution of the IVP $\left\{\begin{array}{c}\ddot{y}+p\dot{y}+qy=0\\ y(0)=0\\ \dot{y}(0)=1\end{array}\right\}$. Why? Because $Y_2(s) = \mathcal{L}[\,y_2(t)]$ satisfies

$$(s^2+ps+q)Y_2(s) - s\cdot 0 - 1 - p\cdot 0 = 0,$$

that is, $Y_2(s) = \dfrac{1}{s^2+ps+q}$.

By the way, in Theorem 3.9 in Section 3.3, the function $y_2(t)$ was one of the two solutions in a complete set of basic solutions of the corresponding homogeneous ODE, $\ddot{y}+p\dot{y}+qy=0$.

Definition 4.5

If p, q are constants,

$$T(s) \triangleq \frac{1}{s^2+ps+q}$$

is called the **transfer function** for the problem $\ddot{y}+p\dot{y}+qy=f(t)$.

The zeros of the transfer function, that is, s satisfying $s^2+ps+q=0$, are called the **poles** of the system. The poles are, not coincidentally, the zeros of the characteristic polynomial of the corresponding LCCHODE.

Example 4.33

Express a particular solution of $\ddot{y}+\omega^2 y = f(t)$ using a convolution. Assume $\omega \neq 0$.

Method: Take the Laplace transform of the ODE to get

$$(s^2+\omega^2)Y_p(s) - sy(0) - \dot{y}(0) = F(s),$$

where $Y_p(s) = \mathcal{L}[y_p(t)](s)$ and $F(s) = \mathcal{L}[f(t)](s)$. Because we are interested only in a particular solution, we may ignore the $-sy(0) - \dot{y}(0)$ terms. We have

$$y_p(t) = \mathcal{L}^{-1}[Y_p(s)] = \mathcal{L}^{-1}\left[\frac{1}{s^2+\omega^2}F(s)\right] = \mathcal{L}^{-1}\left[\frac{1}{s^2+\omega^2}\right] * f(t).$$

So, a particular solution of the ODE is given by

$$y_p(t) = \int_0^t \frac{1}{\omega}\sin(\omega(t-u))f(u)\,du. \bigcirc$$

Recall that a function $f(t)$ is **periodic** if there is a positive constant T such that $f(t+\mathsf{T}) \equiv f(t)$ at all t in the domain of f. Pictorially, a function is periodic if its graph is not changed when you shift it horizontally by T units. Periodic functions are commonly found in engineering as inputs to, and/or outputs from, machinery.

Define a **square wave** over one period by

$$f(t) = \begin{cases} 1, & 0 < t < \frac{\mathsf{T}}{2} \\ -1, & \frac{\mathsf{T}}{2} < t < \mathsf{T} \end{cases}. \tag{4.53}$$

For $\mathsf{T} = 2\pi$, the graph of the periodic function f is shown in Figure 4.18.

Example 4.34

Solve the IVP $\begin{Bmatrix} \dot{y} + 2y = f(t) \\ y(0) = 0 \end{Bmatrix}$, where $f(t)$ is the square wave defined in (4.53).

Method: Take the Laplace transform to get

$$Y(s) = \frac{1}{s+2} \cdot F(s),$$

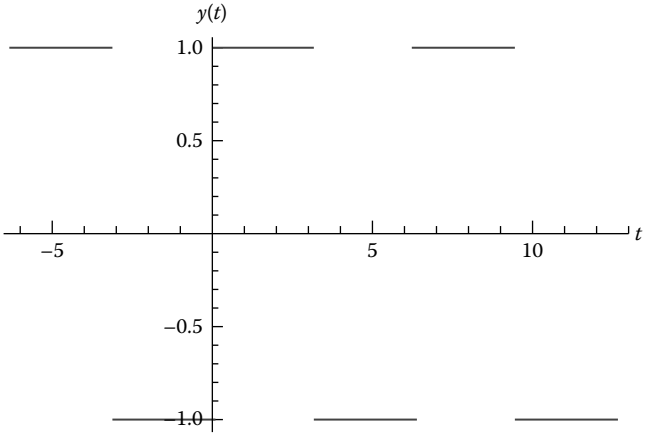


FIGURE 4.18
Square wave function.

where $Y(s) = \mathcal{L}[\,y(t)](s)$ and $F(s) = \mathcal{L}[f(t)](s)$. The solution can be expressed in terms of a convolution:

$$y(t) = \mathcal{L}^{-1}[Y(s)] = \mathcal{L}^{-1}\left[\frac{1}{s+2}\right] * F(s) = \int_0^t e^{-2(t-u)} f(u)du = e^{-2t} \cdot \int_0^t e^{2u} f(u)\, du. \bigcirc$$

This gives a form of the solution, but we can get even more information about the solution by evaluating the definite integral. Unfortunately, to do so requires a lot of tedious work by examining different intervals of t values. First, if $0 \le t \le \frac{\mathsf{T}}{2}$, then

$$I(t) \triangleq \int_0^t e^{2u} f(u)\, du = \int_0^t e^{2u} \cdot 1\, du = \left[\frac{e^{2u}}{2}\right]_0^t = \frac{1}{2}\left(e^{2t} - 1\right).$$

Next, if $\frac{\mathsf{T}}{2} \le t \le \mathsf{T}$, then a property of definite integration yields

$$\begin{aligned}\int_0^t e^{2u} f(u)\, du &= \int_0^{\frac{\mathsf{T}}{2}} e^{2u} \cdot 1\, du + \int_{\frac{\mathsf{T}}{2}}^t e^{2u} \cdot (-1)\, du = I\left(\frac{\mathsf{T}}{2}\right) + \left[-\frac{e^{2u}}{2}\right]_{\frac{\mathsf{T}}{2}}^t \\ &= \frac{1}{2}\left((e^{\mathsf{T}} - 1) - (e^{2t} - e^{\mathsf{T}})\right) = \frac{1}{2}\left(-e^{2t} - 1 + 2e^{\mathsf{T}}\right).\end{aligned}$$

If $\mathsf{T} \le t \le \frac{3\mathsf{T}}{2}$, then

$$\begin{aligned}\int_0^t e^{2u} f(u)\, du &= \int_0^{\mathsf{T}} e^{2u} f(u) du + \int_{\mathsf{T}}^t e^{2u} \cdot 1\, du = I(\mathsf{T}) + \left[-\frac{e^{2u}}{2}\right]_{\mathsf{T}}^t \\ &= \frac{1}{2}\left(-e^{2\mathsf{T}} - 1 + 2e^{\mathsf{T}} + (e^{2t} - e^{2\mathsf{T}})\right) = \frac{1}{2}\left(e^{2t} - 1 + 2e^{\mathsf{T}} - 2e^{2\mathsf{T}}\right).\end{aligned}$$

If $\frac{3\mathsf{T}}{2} \le t \le 2\mathsf{T}$, then

$$\begin{aligned}\int_0^t e^{2u} f(u) du &= \int_0^{\frac{3\mathsf{T}}{2}} e^{2u} f(u) du + \int_{\frac{3\mathsf{T}}{2}}^t e^{2u} \cdot (-1) du = I\left(\frac{3\mathsf{T}}{2}\right) + \left[-\frac{e^{2u}}{2}\right]_{\frac{3\mathsf{T}}{2}}^t \\ &= \frac{1}{2}\left(e^{3\mathsf{T}} - 1 + 2e^{\mathsf{T}} - 2e^{2\mathsf{T}} - (e^{2t} - e^{3\mathsf{T}})\right) = \frac{1}{2}\left(-e^{2t} - 1 + 2e^{\mathsf{T}} - 2e^{2\mathsf{T}} + 2e^{3\mathsf{T}}\right), \ldots.\end{aligned}$$

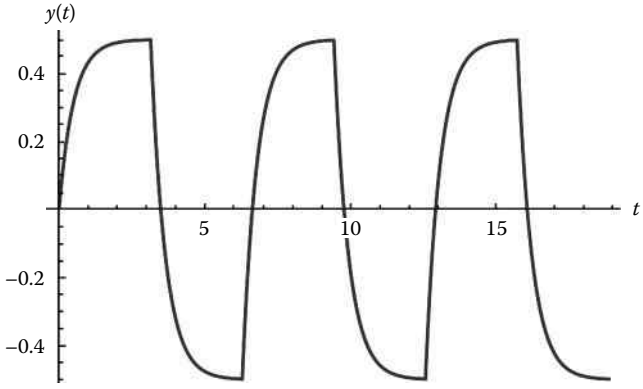


FIGURE 4.19
Solution of Example 4.34.

The solution of the IVP is

$$y(t) = \frac{1}{2}e^{-2t} \cdot \begin{cases} e^{2t} - 1, & 0 \le t \le \frac{\mathsf{T}}{2} \\ -e^{2t} - 1 + 2e^{\mathsf{T}}, & \frac{\mathsf{T}}{2} \le t \le \mathsf{T} \\ e^{2t} - 1 + 2e^{\mathsf{T}} - 2e^{2\mathsf{T}}, & \mathsf{T} \le t \le \frac{3\mathsf{T}}{2} \\ -e^{2t} - 1 + 2e^{\mathsf{T}} - 2e^{2\mathsf{T}} + 2e^{3\mathsf{T}}, & \frac{3\mathsf{T}}{2} \le t \le 2\mathsf{T} \\ \vdots & \vdots \end{cases}.$$

In general, for $k\frac{\mathsf{T}}{2} \le t \le (k+1) \cdot \frac{\mathsf{T}}{2}$,

$$y(t) = \frac{1}{2}e^{-2t}\left((-1)^k e^{2t} - 1 - 2\sum_{j=1}^{k}(-1)^j e^{j\mathsf{T}}\right).$$

For $\mathsf{T} = 2\pi$, the solution is graphed in Figure 4.19. By the way, the solution graphed in Figure 4.19 is *not* a periodic solution, as we will see in Example 5.41 in Section 5.8. Nevertheless, we can see a steady-state solution hiding in Figure 4.19.

4.5.5 Delta "Functions"

While $step(t - c)$ is used to model turning on or off a function at $t = c$, there is another related but more exotic "function," $\delta(t - c)$, called a **delta function**. We assume c is a positive constant.

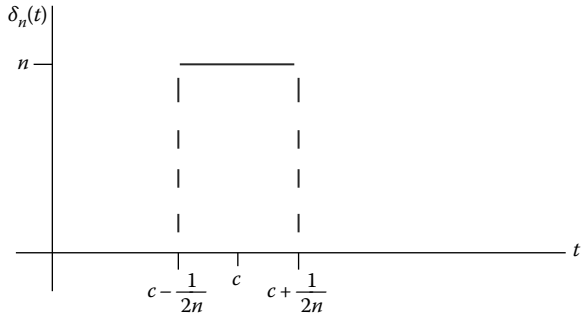


FIGURE 4.20
Approximate delta function.

Intuitively, a delta function is an "infinitesimally brief but infinitely strong input." We can think of $\delta(t-c)$ as the limit, as $n \to \infty$, of "approximate delta functions" defined by

$$\delta_n(t) = \begin{cases} 0, & t < c - \frac{1}{2n} \\ n, & c - \frac{1}{2n} \le t < c + \frac{1}{2n} \\ 0, & t \ge c + \frac{1}{2n} \end{cases} = n \cdot \left(step\left(t - \left(c - \frac{1}{2n}\right)\right) - step\left(t - \left(c + \frac{1}{2n}\right)\right)\right).$$

Physically, an approximate delta function can model a strong but brief input to a system. Figure 4.20 shows an approximate delta function. For each n, the area under the graph of $\delta_n(t)$ is 1.

If $f(t)$ is continuous at $t = c$, for n large enough that $c - \frac{1}{2n} > 0$, that is, $n > \frac{1}{2c}$, then

$$\int_0^\infty f(t)\delta_n(t)dt = \int_{c-\frac{1}{2n}}^{c+\frac{1}{2n}} f(t) \cdot n dt \approx nf(c) \cdot \left(\text{length of interval } \left[c - \frac{1}{2n}, c + \frac{1}{2n}\right]\right)$$
$$= nf(c) \cdot \frac{1}{n} = f(c).$$

Definition 4.6

$\delta(t-c)$ is *defined to satisfy* $\int_0^\infty f(t)\delta(t-c)\,dt = f(c)$ for any function f that is continuous at $t = c$.

Theorem 4.12

(Property of Laplace transform of a delta function) Assume c is a nonnegative constant. Then
(15) $\mathcal{L}[\,\delta(t-c)](s) = e^{-cs}$.

Example 4.35

Solve the IVP $\dot{y} + 4y = 2\delta(t-1),\ y(0) = 3$.

Method: Take the Laplace transform of both sides of the ODE to get

$$(s+4)Y(s) - 3 = 2e^{-s},$$

where $Y(s) = \mathcal{L}[y(t)](s)$. Using property (11) backward, the solution of the IVP is given by

$$y(t) = \mathcal{L}^{-1}\left[\frac{3}{s+4} + \frac{2}{s+4}e^{-s}\right] = 3e^{-4t} + \mathcal{L}^{-1}\left[\frac{2}{s+4}\right]\Big|_{t\mapsto(t-1)} \cdot step(t-1)$$
$$= 3e^{-4t} + (2e^{-4t})\Big|_{t\mapsto(t-1)} \cdot step(t-1) = 3e^{-4t} + 2e^{-4(t-1)}step(t-1). \bigcirc$$

The graph of the solution,

$$y(t) = 3e^{-4t} + \begin{Bmatrix} 0, & 0 \le t < 1 \\ 2e^{-4(t-1)}, & t \ge 1 \end{Bmatrix} = \begin{Bmatrix} 3e^{-4t}, & 0 \le t < 1 \\ 3e^{-4t} + 2e^{-4(t-1)}, & t \ge 1 \end{Bmatrix},$$

in Figure 4.21 has a vertical jump, by two units, at $t = 1$: a horizontal translate of $2e^{-4t}$ is switched on at $t = 1$.

In some sense, the derivative of a step function is a delta function because of properties (4), (10), and (15):

$$\mathcal{L}\left[\frac{d}{dt}\left[\,step(t-c)\right]\right] = s\mathcal{L}\left[\,step(t-c)\right] - step(0-c) = s\cdot\frac{1}{s}e^{-cs} - 0$$
$$= e^{-cs} = \mathcal{L}[\delta(t-c)].$$

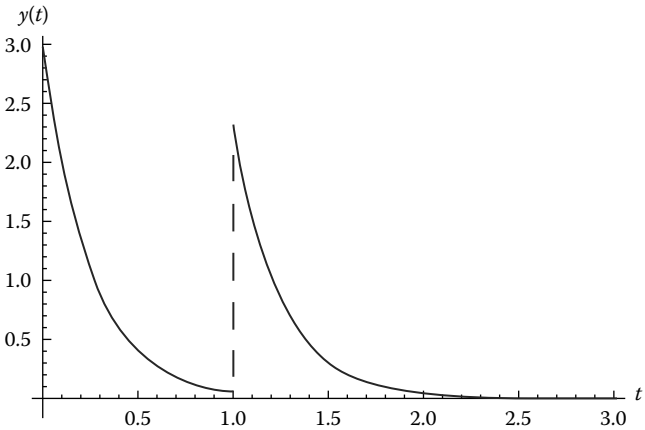


FIGURE 4.21
Example 4.25.

4.5.6 Laplace Transform of a Periodic Function

Theorem 4.13

Property of Laplace transform of a periodic function
(16) $\mathcal{L}[f(t)](s) = \frac{1}{1-e^{-s\mathsf{T}}} \int_0^{\mathsf{T}} f(t)e^{-st}\,dt$.

Example 4.36

Find the Laplace transform of the square wave defined in (4.53) over one period, $0 < t < T$, and simplify the result.

Method:

$$\mathcal{L}[f(t)](s) = \frac{1}{1-e^{-s\mathsf{T}}} \int_0^{\mathsf{T}} f(t)e^{-st}\,dt = \frac{1}{1-e^{-s\mathsf{T}}} \left(\int_0^{\mathsf{T}/2} f(t)e^{-st}dt + \int_{\mathsf{T}/2}^{\mathsf{T}} f(t)e^{-st}dt \right)$$

$$= \frac{1}{1-e^{-s\mathsf{T}}} \left(\int_0^{\mathsf{T}/2} (1)e^{-st}dt + \int_{\mathsf{T}/2}^{\mathsf{T}} (-1)e^{-st}dt \right) = \frac{1}{1-e^{-s\mathsf{T}}} \left(\left[\frac{e^{-st}}{-s} \right]_0^{\mathsf{T}/2} - \left[\frac{e^{-st}}{-s} \right]_{\mathsf{T}/2}^{\mathsf{T}} \right)$$

$$= \frac{1}{1-e^{-s\mathsf{T}}} \cdot \frac{1}{s} \cdot \left(\left(-e^{-s\mathsf{T}/2} + 1 \right) + \left(e^{-s\mathsf{T}} - e^{-s\mathsf{T}/2} \right) \right) = \frac{e^{-s\mathsf{T}} - 2e^{-s\mathsf{T}/2} + 1}{s(1-e^{-s\mathsf{T}})}$$

$$= \frac{(1-e^{-s\mathsf{T}/2})^2}{s(1-(e^{-s\mathsf{T}/2})^2)} = \frac{(1-e^{-s\mathsf{T}/2})^2}{s(1-e^{-s\mathsf{T}/2})(1+e^{-s\mathsf{T}/2})} = \frac{1-e^{-s\mathsf{T}/2}}{s(1+e^{-s\mathsf{T}/2})} \cdot \bigcirc$$

4.5.7 Remarks

1. The result of Example 4.36 could be used in an attempt to solve Example 4.34; however, finding $\mathcal{L}^{-1}[Y(t)]$ would be difficult.
2. Another way to approach Example 4.34 would be to rewrite the periodic function $f(t)$ as an infinite series of step functions, specifically

$$f(t) = 1 - 2\,step\left(t - \frac{\mathsf{T}}{2}\right) + 2\,step\,(t - \mathsf{T}) - 2\,step\left(t - \frac{3\mathsf{T}}{2}\right) \pm \cdots .$$

This would turn out to be a fairly reasonable approach even though the solution, $y(t)$, would also be an infinite series involving step functions. It would be reasonable because at any finite time, only finitely many of the terms in $y(t)$ would be nonzero, so convergence of the infinite series would not be an issue.

4.5.8 Problems

In problems 1–5, find the Laplace transform of the given function.

1. $7 + 3e^{-2t} - t\,step(t-4)$
2. $(t-2)step(t-2) - 5t\,step(t-3)$

3. $f(t) = \begin{Bmatrix} -t+4, & 0 \le t < 2 \\ t, & t \ge 2 \end{Bmatrix}$
4. $f(t) = \begin{Bmatrix} t, & 0 \le t < 1 \\ 1, & 1 \le t < 3 \\ -1, & t \ge 3 \end{Bmatrix}$
5. $h(t) = \int_0^t e^{-(t-u)} \cos(2u) du$
6. Which of the following, if any, is the Laplace transform of $t\, step(t-3)$?
   (a) $\frac{1}{s^2} e^{-3s}$
   (b) $\frac{1}{s^2} e^{3s}$
   (c) $(\frac{1}{s^2} + \frac{3}{s}) e^{3s}$
   (d) $(\frac{1}{s^2} + \frac{3}{s}) e^{-3s}$
   (e) $(\frac{1}{s^2} + \frac{1}{s}) e^{-3s}$
7. Which of the following, if any, is the Laplace transform of $\frac{1}{\sqrt{3}} e^{-t/2} \sin(\frac{\sqrt{3}t}{2})$? It might be true that more than one is correct.
   (a) $\frac{1}{2(s^2+s+1)}$
   (b) $\frac{1}{2(s^2-s+1)}$
   (c) $\frac{\sqrt{3}/2}{(s+\frac{1}{2})(s^2+\frac{3}{4})}$
   (d) $\frac{\sqrt{3}/2}{(s-\frac{1}{2})(s^2+\frac{3}{4})}$
   (e) $\frac{1}{2((s-\frac{1}{2})^2+\frac{3}{4})}$

In problems 8–13, find the inverse Laplace transform of the given function.

8. $\frac{e^{-\pi s/4}}{s^2+9}$
9. $\frac{e^{-s}}{s^2-2s}$
10. $\frac{e^{-2s}}{s^3+4s}$
11. $\frac{s}{(s^2+4)^2}$
12. $\frac{s^2}{(s^2+4)^2}$
13. $\frac{4}{(s^2+4)^2}$
14. Express $\mathcal{L}^{-1}\left[\frac{s}{(s^2+6)(s^2+4)}\right]$ as a specific definite integral. Do use specific functions and variables in your integral, but do not evaluate or simplify the integral.

In problems 15–21, use Laplace transforms to solve the given IVP. Also, graph the solution.

15. $\dot{y} + 3y = 2 - step(t-1),\ y(0) = 4$
16. $\ddot{y} + 9y = 2 - step(t-2),\ y(0) = -1, \dot{y}(0) = 6$

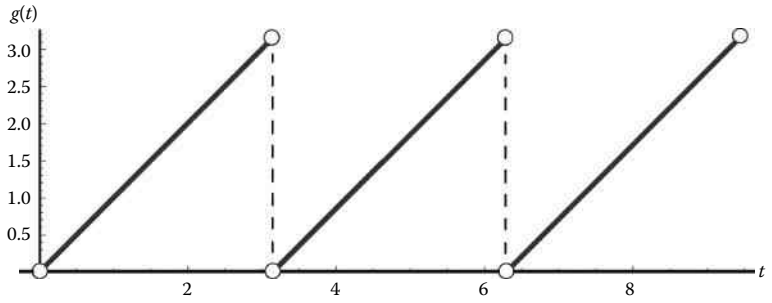


FIGURE 4.22
Sawtooth function with period $\mathsf{T} = \pi$.

17. $\ddot{y} + 4y = step(t - c),\ y(0) = 0, \dot{y}(0) = -5$, where c is an unspecified constant
18. $\ddot{y} + y = f(t),\ y(0) = 1, \dot{y}(0) = 0$, where $f(t) = \begin{Bmatrix} t, & 0 \le t < \pi \\ 0, & t \ge \pi \end{Bmatrix}$
19. $\dot{y} + 3y = \delta(t - 2),\ y(0) = -1$
20. $\ddot{y} + 4y = \delta(t - c),\ y(0) = 0, \dot{y}(0) = 0$, where c is an unspecified constant
21. $\ddot{y} + y = \cos t$

In problems 22–24, use Laplace transforms to solve the given IVP *and* to express the solution in the form of a convolution. Do not replace $f(t)$ by a specific function, that is, your final conclusion should be in terms of $f(t)$.

22. $\dot{y} + 5y = f(t),\ y(0) = 0$
23. $\ddot{y} + 4y = f(t),\ y(0) = 0,\ \dot{y}(0) = 0$
24. $\ddot{y} + 4\dot{y} + 5y = f(t),\ y(0) = 0,\ \dot{y}(0) = 0$

Definition 4.7

Define a sawtooth function over one period by $g(t) \triangleq t, 0 < t < \mathsf{T}$.

In problems 25 and 26, use Laplace transforms to solve the given IVP and to express the solution in the form of a convolution. Here, $g(t)$ is the **sawtooth function** shown in Figure 4.22 and has period $\mathsf{T} = \pi$. Analyze the solution as in Example 4.34.

25. $\dot{y} + 5y = g(t),\ y(0) = 0$
26. $\ddot{y} + 4y = g(t),\ y(0) = 0,\ \dot{y}(0) = 0$

4.6 Scalar Difference Equations

A "difference equation" is a relationship involving a sequence of values, say $\{y_k\}_{k=0}^{N}$. For example, $y_k = ry_{k-1}$ says that we have a geometric sequence, $y_0, ry_0, r^2y_0, \ldots$.

A difference equation can have its origins in the time evolution of a system. Suppose we partition an interval $[a, b]$ into N equal subintervals, that is,

$$a = t_0, \quad a + h = t_1, \quad a + 2h = t_2, \quad \cdots, \quad a + Nh = b = t_N$$

and define

$$y_k \triangleq y(t_k) = y(a + kh), \ k = 0, 1, 2, \ldots, N.$$

You saw this when approximating definite integrals in calculus I or II, for example, using the trapezoidal rule. Another notation for h is Δt, the "time-step size" or "grid size."

The difference equation

$$y_{k+2} = y_{k+1} + y_k \tag{4.54}$$

is famous because it produces the "Fibonacci numbers": starting with initial data $y_0 = 0$, $y_1 = 1$, the sequence is

$$\{y_k\}_{k=0}^{\infty} = 0, 1, 1, 2, 3, 5, 8, 13, 21, \ldots.$$

Fibonacci numbers show up in an amazing variety of scientific problems, including a simple model for the growth of a population of rabbits, where $y_k, k = 1, 2, 3, \ldots$, is the number of female rabbits at successive, discrete times.

Difference equations also naturally show up in digital computation, for example, discrete-time signal theory, discrete-time control theory, and in approximating solutions of differential equations.

In general, an **n-th order difference equation** is given by

$$y_{k+n} = f(k, y_k, y_{k+1}, \ldots, y_{k+n-1}), \ k = 0, 1, 2, \ldots. \tag{4.55}$$

So, for example, (4.54) is a second-order difference equation. A **solution** of an n-th order difference equation is a sequence $\{y_k\}_{k=0}^{N}$ or $\{y_k\}_{k=0}^{\infty}$. We may also refer to y_k, without the curly brackets, as a solution. Substituting in $k = 0$, (4.55) yields

$$y_n = f(0, y_0, y_1, \ldots, y_{n-1});$$

substituting in $k = 1$, (4.55) implies

$$y_{n+1} = f(1, y_1, y_2, \ldots, y_n);$$

etc. So, in order to solve an n-order difference equation, we need initial data

$$y_0, y_1, \ldots, y_{n-1}, \tag{4.56}$$

after which we can "recursively" calculate y_n, then y_{n+1}, then y_{n+2}, etc. Of course, how far ahead in time we can calculate the successive values in the recursive sequence depends upon the computing machinery, upon n, and upon the function f. Sometimes the function f is not known explicitly as a formula in $k, y_k, y_{k+1}, \ldots, y_{k+n-1}$ but is instead a subroutine that itself involves recursive processes of length not initially known.

For an n-th order linear constant coefficients homogeneous difference equation (also known as LCCHΔE)

$$y_{k+n} = a_1 y_{k+n-1} + a_2 y_{k+n-2} + \cdots + a_n y_k, \tag{4.57}$$

we will assume that, in addition, $a_n \neq 0$, so that the LCCHΔE is truly of n-th order. Fortunately, we will see that there is a solution method like the solution methods for LCCHODEs and for Cauchy–Euler equations: try solutions of the form

$$y_k = r^k,$$

where r is a constant to be determined. When we substitute this into (4.57), we get

$$r^{n+k} = a_1 r^{n+k-1} + a_2 r^{n+k-2} + \cdots + a_n r^k,$$

that is,

$$r^k r^n = r^k (a_1 r^{n-1} + a_2 r^{n-2} + \cdots + a_n r^0),$$

that is,

$$r^k (r^n - a_1 r^{n-1} - a_2 r^{n-2} - \cdots - a_{n-1} r - a_n) = 0.$$

So, r should satisfy a characteristic equation,

$$r^n - a_1 r^{n-1} - a_2 r^{n-2} - \cdots - a_{n-1} r - a_n = 0. \tag{4.58}$$

Example 4.37

Solve the Fibonacci LCCHΔE (4.54) with ICs $y_0 = 0, y_1 = 1$.

Method: With $y_k = r^k$, this second-order difference equation has characteristic equation

$$r^2 - r - 1 = 0,$$

whose solutions are

$$r = \frac{1 \pm \sqrt{(-1)^2 - 4(1)(-1)}}{2} = \frac{1 \pm \sqrt{5}}{2}.$$

The Fibonacci difference (4.54) has solutions

$$y_k = c_1 \left(\frac{1+\sqrt{5}}{2} \right)^k + c_2 \left(\frac{1-\sqrt{5}}{2} \right)^k, \tag{4.59}$$

where c_1, c_2 are arbitrary constants.

We substitute (4.59) into the ICs to get

$$0 = y_0 = c_1 + c_2 \quad \text{and} \quad 1 = y_1 = \left(\frac{1+\sqrt{5}}{2} \right) c_1 + \left(\frac{1-\sqrt{5}}{2} \right) c_2.$$

So, we get

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \frac{1+\sqrt{5}}{2} & \frac{1-\sqrt{5}}{2} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \frac{1}{-\sqrt{5}} \begin{bmatrix} \frac{1-\sqrt{5}}{2} & -1 \\ -\frac{1+\sqrt{5}}{2} & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{5}} \\ -\frac{1}{\sqrt{5}} \end{bmatrix}.$$

The Fibonacci sequence is given by

$$y_k = \frac{1}{\sqrt{5}} \left(\left(\frac{1+\sqrt{5}}{2} \right)^k - \left(\frac{1-\sqrt{5}}{2} \right)^k \right). \bigcirc \tag{4.60}$$

Because it looks overly complicated to have all of those $\sqrt{5}$s, let's check by substituting in, say, $k = 4$:

$$\begin{aligned} 3 = y_4 &=^{?} \frac{1}{\sqrt{5}} \left(\left(\frac{1+\sqrt{5}}{2} \right)^4 - \left(\frac{1-\sqrt{5}}{2} \right)^4 \right) \\ &= \frac{1}{16\sqrt{5}} \left(1 + 4\sqrt{5} + 6(5) + 4(5\sqrt{5}) + 1(25) - 1 + 4\sqrt{5} - 6(5) + 4(5\sqrt{5}) - 1(25) \right) \\ &= \frac{1}{16\sqrt{5}} (8\sqrt{5} + 40\sqrt{5}) = \frac{1}{16\sqrt{5}} 48\sqrt{5} = 3. \bigcirc \end{aligned}$$

For any second-order LCCHΔE

$$y_{k+2} = a_1 y_{k+1} + a_2 y_k, \tag{4.61}$$

we will assume the coefficients a_1, a_2 are real. It is possible to allow them to be complex, but that would alter some of the results that follow. In addition, we assume that $a_2 \neq 0$ so that the LCCHΔE is truly of second order. The characteristic polynomial is

$$r^2 - a_1 r - a_2, \tag{4.62}$$

whose roots are

$$r = \frac{a_1 \pm \sqrt{a_1^2 - 4a_2}}{2}.$$

If the roots are real, distinct r_1, r_2, then the solutions of (4.61) are given by

$$y_k = c_1 r_1^k + c_2 r_2^k,$$

where c_1, c_2 are arbitrary constants.

Analogously to the critically damped case of Section 3.3, it turns out that if the roots of the characteristic polynomial are real but equal, that is, $r_1 = r_2$, then the solutions of (4.61) are given by

$$y_k = c_1 r_1^k + c_2 k r_2^k,$$

where c_1, c_2 are arbitrary constants.

If the roots of the characteristic polynomial are a complex conjugate pair $r = \alpha \pm i\nu$, where α, ν are real and $\nu > 0$, then the solutions of (4.61),

$$y_k = \tilde{c}_1(\alpha + i\nu)^k + \tilde{c}_2(\alpha - i\nu)^k,$$

can be rewritten in real form: First, we need to mention the "polar form" of a complex number:

$$\alpha + i\nu = \rho e^{i\omega} = \rho(\cos\omega + i\sin\omega), \tag{4.63}$$

where ρ is real and nonnegative and $-\pi < \omega \leq \pi$. This is just another example of using polar coordinates/trigonometry because

$$\alpha + i\nu = \rho\cos\omega + i\rho\sin\omega$$

is solvable for real, nonnegative ρ and real ω in terms of α, ν:

$$\rho = \sqrt{\alpha^2 + \nu^2}, \quad \tan\omega = \frac{\nu}{\alpha}.$$

As usual, it helps to draw a picture, Figure 4.23, in the (α, ν)-plane. Note that $\nu > 0$ implies that $\omega \neq 0$.

We see that with $-\pi < \omega \leq \pi$, $\alpha + i\nu = \rho\cos\omega + i\rho\sin\omega = \rho e^{i\omega}$; the latter is called the **polar form** of the complex number. This implies that

$$\alpha - i\nu = \rho\cos\omega - i\rho\sin\omega = \rho\cos(-\omega) + i\rho\sin(-\omega) = \rho e^{-i\omega}.$$

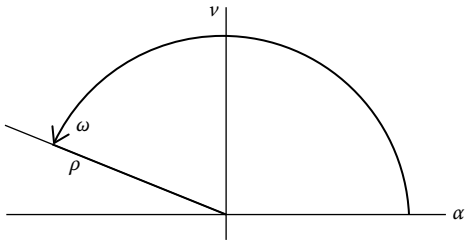


FIGURE 4.23
ρ and ω picture.

So, we can rewrite the solutions of (4.61) in the complex roots case:

$$y_k = \tilde{c}_1(\alpha + i\nu)^k + \tilde{c}_2(\alpha - i\nu)^k = \tilde{c}_1(\rho e^{i\omega})^k + \tilde{c}_2(\rho e^{-i\omega})^k = \rho^k\left(\tilde{c}_1 e^{i\omega k} + \tilde{c}_2 e^{-i\omega k}\right)$$

$$= \rho^k\left(\tilde{c}_1(\cos\omega k + i\sin\omega k) + \tilde{c}_2(\cos\omega k - i\sin\omega k)\right) = \rho^k(c_1\cos\omega k + c_2\sin\omega k).$$

So, when there is a complex conjugate pair of roots, the solution is

$$y_k = c_1\rho^k\cos\omega k + c_2\rho^k\sin\omega k, \tag{4.64}$$

where c_1, c_2 are arbitrary constants.

Example 4.38

For $k \geq 3$, let A_k be the $k \times k$ "tri-diagonal" matrix

$$A_k = \begin{bmatrix} -1.9 & 1 & 0 & . & . & . & & & & 0 \\ 1 & -1.9 & 1 & 0 & & & & & & . \\ 0 & 1 & -1.9 & 1 & & & & & & . \\ . & & & . & . & . & & & & . \\ . & & & & . & . & . & & & \\ . & & & & & . & . & -1.9 & 1 & 0 \\ & & & & & & & 1 & -1.9 & 1 \\ 0 & . & . & . & & & & 0 & 1 & -1.9 \end{bmatrix}.$$

Find a formula for $|A_k|$, the determinant of A_k in terms of k.

Method: At first, this problem seems to have nothing to do with difference equations, except perhaps the "subscript k" reminds us of the solution of a difference equation. But, as we will see, we can find an LCCHΔE satisfied by

$$y_k \triangleq |A_k|.$$

But first, it will help if we look at some small-sized examples of the matrix A_k and their determinants.

For $k = 3$,

$$A_3 = \begin{bmatrix} -1.9 & 1 & 0 \\ 1 & -1.9 & 1 \\ 0 & 1 & -1.9 \end{bmatrix}.$$

By expanding along the first row,

$$y_3 = |A_3| = (-1.9)\begin{vmatrix} -1.9 & 1 \\ 1 & -1.9 \end{vmatrix} - (1)\begin{vmatrix} 1 & 1 \\ 0 & -1.9 \end{vmatrix} + (0)\begin{vmatrix} 1 & -1.9 \\ 0 & 1 \end{vmatrix} = -3.059.$$

For $k = 4$,

$$y_4 = |A_4| = \begin{vmatrix} -1.9 & 1 & 0 & 0 \\ 1 & -1.9 & 1 & 0 \\ 0 & 1 & -1.9 & 1 \\ 0 & 0 & 1 & -1.9 \end{vmatrix}$$

$$= (-1.9)\begin{vmatrix} -1.9 & 1 & 0 \\ 1 & -1.9 & 1 \\ 0 & 1 & -1.9 \end{vmatrix} - (1)\begin{vmatrix} 1 & 1 & 0 \\ 0 & -1.9 & 1 \\ 0 & 1 & -1.9 \end{vmatrix} + (0) + (0) = -1.9y_3 - \begin{vmatrix} -1.9 & 1 \\ 1 & -1.9 \end{vmatrix}.$$

This suggests that it would be good to also define

$$A_2 \triangleq \begin{bmatrix} -1.9 & 1 \\ 1 & -1.9 \end{bmatrix},\ y_2 \triangleq |A_2| = 2.61.$$

Then we see that

$$y_4 = -1.9y_3 - y_2.$$

In fact, in general, by expanding the determinant along the first row, we have

$$y_k = \begin{vmatrix} -1.9 & 1 & 0 & 0 & \cdots & & 0 \\ 1 & -1.9 & 1 & 0 & & & \vdots \\ 0 & 1 & -1.9 & 1 & & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & \ddots & \ddots & \vdots \\ \vdots & & & & & \ddots & 0 \\ & & & & & 1 & -1.9 & 1 \\ 0 & \cdots & \cdots & \cdots & & 0 & 1 & -1.9 \end{vmatrix}$$

$$= (-1.9)\begin{vmatrix} -1.9 & 1 & 0 & \cdots & & 0 \\ 1 & -1.9 & 1 & & & \vdots \\ 0 & 1 & -1.9 & & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 & -1.9 \end{vmatrix} - \begin{vmatrix} 1 & 1 & 0 & \cdots & & 0 \\ 0 & -1.9 & 1 & & & \vdots \\ \vdots & 1 & -1.9 & & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 0 & 1 & -1.9 \end{vmatrix}.$$

Now, expand the second term along its first column to get

$$y_k = -1.9y_{k-1} - (1)\begin{vmatrix} -1.9 & 1 & 0 & \cdots & & 0 \\ 1 & -1.9 & 1 & & & \vdots \\ 0 & 1 & -1.9 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & -1.9 & 1 \\ 0 & \cdots & \cdots & 0 & 1 & -1.9 \end{vmatrix} = -1.9y_{k-1} - y_{k-2}.$$

So, y_k satisfies the second-order difference equation

$$y_k = -1.9y_{k-1} - y_{k-2},\ \text{for } k \geq 3,$$

that is,

$$y_{k+2} = -1.9y_{k+1} - y_k,\ \text{for } k \geq 1,$$

whose characteristic polynomial, $r^2 + 1.9r + 1$, has a complex conjugate pair of roots

$$r = -0.95 \pm i\sqrt{0.0975}.$$

According to our previous theoretical result in (4.64), here $\rho = \sqrt{(-0.95)^2 + (0.0975)^2} = 1$ and $\tan\omega = \frac{\sqrt{0.0975}}{-0.95}$ and ω in the second quadrant of the (α, ν)-plane give

$$y_k = c_1 \cos\omega k + c_2 \sin\omega k,$$

as the solutions of the difference equation, where c_1, c_2 are arbitrary constants. Just as we did for the amplitude-phase form of solutions in Section 3.3, we get

$$\omega = \pi + \tan^{-1}\left(\frac{\sqrt{0.0975}}{-0.95}\right) = \pi - \tan^{-1}\left(\frac{\sqrt{0.0975}}{0.95}\right).$$

To satisfy the IC,

$$\left\{\begin{array}{l} 2.61 = y_2 = c_1 \cos 2\omega + c_2 \sin 2\omega \\ \\ -3.059 = y_3 = c_1 \cos 3\omega + c_2 \sin 3\omega \end{array}\right\},$$

we get

$$\begin{bmatrix} c_1 \\ \\ c_2 \end{bmatrix} = \begin{bmatrix} \cos 2\omega & \sin 2\omega \\ \\ \cos 3\omega & \sin 3\omega \end{bmatrix}^{-1} \begin{bmatrix} 2.61 \\ \\ -3.059 \end{bmatrix}$$

$$= \frac{1}{\cos 2\omega \sin 3\omega - \sin 2\omega \cos 3\omega} \begin{bmatrix} \sin 3\omega & -\sin 2\omega \\ \\ -\cos 3\omega & \cos 2\omega \end{bmatrix} \begin{bmatrix} 2.61 \\ \\ -3.059 \end{bmatrix}$$

$$= \cdots = \begin{bmatrix} 1 \\ \\ -0.95/\sqrt{0.0975} \end{bmatrix}.$$

So,

$$det(A_k) = \cos\omega - \frac{0.95}{\sqrt{0.0975}} \sin\omega k, k \geq 3. \bigcirc$$

By the way, we can conclude that for all $k \geq 3$,

$$\left|det(A_k)\right| = \left|\cos\omega k + \left(-\frac{0.95}{\sqrt{0.0975}}\right)\sin\omega k\right| \leq |\cos\omega k| + \left|-\frac{0.95}{\sqrt{0.0975}} \sin\omega k\right|$$

$$\leq 1 + \left|-\frac{0.95}{\sqrt{0.0975}}\right| \approx 4.042434922.$$

4.6.1 General Solution and the Casorati Determinant

Analogous to the differential operator, $D = \frac{d}{dt}$, is the **forward difference operator** Δ defined by

$$(\Delta y)_k \triangleq y_{k+1} - y_k, \quad \text{for} \quad k \geq 0. \tag{4.65}$$

TABLE 4.2

The Zoo of Solutions of Linear Homogeneous Constant Coefficients Difference Equations

Roots of Characteristic Polynomial	**Solutions y_k**
$1, 1, \ldots, 1$ [m times]	$1, k, \ldots, k^{m-1}$
$\alpha, \alpha, \ldots, \alpha$ [m times]	$\alpha^k, k\alpha^k, \ldots, k^{m-1}\alpha^k$
$\alpha \pm i\nu, \ldots, \alpha \pm i\nu$ [m times]	$\rho^k \cos \omega k,\ k\rho^k \cos \omega k,\ \ldots,\ k^{m-1}\rho^k \cos \omega k$ and $\rho^k \sin \omega k,\ k\rho^k \sin \omega k,\ \ldots,\ k^{m-1}\rho^k \sin \omega k$

Recall from calculus I that if $y_k = y(t_k)$, then

$$\dot{y}(t_k) \approx \frac{y_{k+1} - y_k}{t_{k+1} - t_k} = \frac{\Delta y_k}{\Delta t_k}.$$

We can also define higher order difference operators by

$$(\Delta^2 y)_k \triangleq (\Delta(\Delta y))_k = \Delta(y_{k+1} - y_k) = (y_{k+2} - y_{k+1}) - (y_{k+1} - y_k) = y_{k+2} - 2y_{k+1} + y_k, \quad \text{for } k \geq 0,$$

that is, $\Delta^2 \triangleq \Delta\Delta$. Similarly, we can define $\Delta^3 \triangleq \Delta\Delta^2$, etc.

In Table 4.2, the first line is a special case of the second line because $\alpha^k \equiv 1$ when $\alpha = 1$. In the last line, the point $(\alpha, \nu) = \rho \cdot (\cos \omega, \sin \omega)$, where $\rho = \sqrt{\alpha^2 + \nu^2} > 0$.

If $\{y_k^{(1)}, y_k^{(2)}, \ldots, y_k^{(n)}\}$ is a set of n sequences, the **Casorati determinant** plays a role analogous to that of the Wronskian determinant for linear ODEs. We define the Casorati by

$$C\left(y_k^{(1)}, y_k^{(2)}, \ldots, y_k^{(n)}\right) = \begin{vmatrix} y_k^{(1)} & y_k^{(2)} & \cdots & y_k^{(n)} \\ y_{k+1}^{(1)} & y_{k+1}^{(2)} & \cdots & y_{k+1}^{(n)} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ y_{k+n-1}^{(1)} & y_{k+n-1}^{(2)} & \cdots & y_{k+n-1}^{(n)} \end{vmatrix}.$$

For example,

$$C\left(r_1^k, r_2^k\right) = \begin{vmatrix} r_1^k & r_2^k \\ r_1^{k+1} & r_2^{k+1} \end{vmatrix} = r_1^k r_2^{k+1} - r_2^k r_1^{k+1} = r_1^k r_2^k (r_1 - r_2) \neq 0,$$

as long as $r_1 \neq r_2$.

In Problem 4.6.4.20, you will explain why the Casorati determinant can be rewritten in a form that looks more analogous to the Wronskian determinant.

An n-th order linear homogeneous difference equation has the form

$$y_{k+n} = a_{1,k}y_{k+n-1} + a_{2,k}y_{k+n-2} + \cdots + a_{n,k}y_k \triangleq (Ly)_k, \text{ for } k \geq 0, \tag{4.66}$$

where we assume that

$$a_{n,k} \neq 0, \quad \text{for all} \quad k \geq 0. \tag{4.67}$$

We assume (4.67) so that (4.66) is truly of order n.

Definition 4.8

The **general solution** of an n-th order linear homogeneous difference (4.66) has the form

$$y_k^{(h)} = c_1 y_k^{(1)} + c_2 y_k^{(2)} + \cdots + c_n y_k^{(n)}$$

if for every solution y_k^* of (4.66), there are values of constants $c_1, c_2, \ldots, c_n$ giving $y_k^* = c_1 y_k^{(1)} + c_2 y_k^{(2)} + \cdots + c_n y_k^{(n)}$. In this case, we call the set of sequences $\{y_k^{(1)}, y_k^{(2)}, \ldots, y_k^{(n)}\}$ a **complete set of basic solutions** for linear homogeneous difference (4.66). Each of the n sequences $y_k^{(1)}, y_k^{(2)}, \ldots, y_k^{(n)}$ is called a **basic solution** of (4.66).

Theorem 4.14

The linear homogeneous n-th order linear homogeneous difference (4.66) has a general solution, that is, a complete set of n basic solutions.

Why? The explanation is similar to that for Theorem 3.15 in Section 3.4: Each of the n IVPs

$$\left\{ \begin{array}{c} y_{k+n}^{(1)} = (Ly^{(1)})_k \\ y_0^{(1)} = 1, y_1^{(1)} = 0, \ldots, y_{n-1}^{(1)} = 0 \end{array} \right\}, \left\{ \begin{array}{c} y_{k+n}^{(2)} = (Ly^{(2)})_k \\ y_0^{(2)} = 0, y_1^{(2)} = 1, \ldots, y_{n-1}^{(2)} = 0 \end{array} \right\}, \cdots,$$
$$\left\{ \begin{array}{c} y_{k+n}^{(n)} = (Ly^{(n)})_k \\ y_0^{(n)} = 0, y_1^{(n)} = 0, \ldots, y_{n-1}^{(n)} = 1 \end{array} \right\}$$

has a solution. The rest of the explanation is also similar to that given for Theorem 3.9 in Section 3.3. □

Theorem 4.15

(Abel's theorem) Suppose $\{y_k^{(1)}, y_k^{(2)}, \ldots, y_k^{(n)}\}$ is a set of n solutions of the same n-th order linear homogeneous difference (4.66). Then the Casorati determinant satisfies

$$C\left(y_k^{(1)}, y_k^{(2)}, \ldots, y_k^{(n)}\right) = \left(\prod_{\ell=0}^{k-1}\left((-1)^{n-1}a_{n,\ell}\right)\right) C\left(y_0^{(1)}, y_0^{(2)}, \ldots, y_0^{(n)}\right) \tag{4.68}$$

for any $k \geq 1$.

Why? Analogous to the explanation of Theorem 3.12 in Section 3.3, first, we claim that

$$C_k \triangleq C\left(y_k^{(1)}, y_k^{(2)}, \ldots, y_k^{(n)}\right)$$

satisfies the first-order linear homogeneous difference equation

$$C_{k+1} = (-1)^{n-1}a_{n,k}C_k. \tag{4.69}$$

Here we will give an explanation of (4.69) for $n = 2$. In Problem 4.6.4.21, you will use the same techniques to explain (4.69) for $n = 3$ or for *all* $n \geq 2$. For $n = 2$, using the effects of elementary row operations on determinants, we have

$$\begin{aligned} C_{k+1} &= \begin{vmatrix} y_{k+1}^{(1)} & y_{k+1}^{(2)} \\ y_{k+2}^{(1)} & y_{k+2}^{(2)} \end{vmatrix} = \begin{vmatrix} y_{k+1}^{(1)} & y_{k+1}^{(2)} \\ a_{1,k}y_{k+1}^{(1)} + a_{2,k}y_k^{(1)} & a_{1,k}y_{k+1}^{(2)} + a_{2,k}y_k^{(2)} \end{vmatrix} \\ &\underset{-a_{1,k}R_1+R_2\to R_2}{=} \begin{vmatrix} y_{k+1}^{(1)} & y_{k+1}^{(2)} \\ a_{2,k}y_k^{(1)} & a_{2,k}y_k^{(2)} \end{vmatrix} \underset{R_2\leftarrow a_{2,k}R_2}{=} a_{2,k}\begin{vmatrix} y_{k+1}^{(1)} & y_{k+1}^{(2)} \\ y_k^{(1)} & y_k^{(2)} \end{vmatrix} \underset{R_1\leftrightarrow R_2}{=} -a_{2,k}\begin{vmatrix} y_k^{(1)} & y_k^{(2)} \\ y_{k+1}^{(1)} & y_{k+1}^{(2)} \end{vmatrix} \\ &= -a_{2,k}C_k, \end{aligned}$$

as we desired.

It is easy to solve (4.69). Continuing in the case $n = 2$,

$$C_{k+1} = \left(\prod_{\ell=0}^{k}\left((-1)^{2-1}a_{n,\ell}\right)\right) C_0 = (-1)^{k+1}\left(\prod_{\ell=0}^{k} a_{n,\ell}\right) C_0,$$

as desired. □

Theorem 4.16

Suppose $y_k^{(1)}, y_k^{(2)}, \ldots, y_k^{(n)}$ are solutions of the same n-th order linear homogeneous LCCHΔE (4.66). Then $C(y_k^{(1)}, y_k^{(2)}, \ldots, y_k^{(n)}) \neq 0$ for all $k \geq 0$ if, and only if, $\{y_k^{(1)}, y_k^{(2)}, \ldots, y_k^{(n)}\}$ is a complete set of basic solutions of LCCHΔE (4.66).

Why? In order for this to be truly an n-th order linear homogeneous difference equation (4.66), we assumed in (4.67) that $a_{n,k} \neq 0$, for all $k \geq 0$. It follows that for all $k \geq 1$, $\left(\prod_{\ell=0}^{k} a_{n,\ell}\right) \neq 0$; hence, the Casorati determinant C_{k+1} is never zero as long as $C_0 \neq 0$. □

4.6.2 Nonhomogeneous Linear Difference Equation

Definition 4.9

(a) A **particular solution** of an *n*-**th order nonhomogeneous linear difference equation**

$$y_{k+n} = f_k + a_{1,k}y_{k+n-1} + a_{2,k}y_{k+n-2} + \cdots + a_{n,k}y_k, \text{ for } k \geq 0, \tag{4.70}$$

is any sequence $y_k^{(p)}$ that satisfies (4.70).

(b) $y_k = y_k^{(p)} + c_1 y_k^{(1)} + c_2 y_k^{(2)} + \cdots + c_n y_k^{(n)}$, for $k \geq 0$, is the **general solution** of nonhomogeneous linear difference (4.70) if for every solution y_k^* of (4.70), there are values for scalar constants $c_1, \ldots, c_n$ giving $y^*(t) = y_k^{(p)} + c_1 y_k^{(1)} + c_2 y_k^{(2)} + \cdots + c_n y_k^{(n)}$, for $k \geq 0$.

Analogous to Theorem 3.1 in Section 3.1, we have

Theorem 4.17

If $y_k^{(p)}$ is any particular solution of *n*-th order linear nonhomogeneous difference (4.70) and $\{y_k^{(1)}, y_k^{(2)}, \ldots, y_k^{(n)}\}$ is any complete set of basic solutions of the corresponding linear homogeneous difference (4.66), that is,

$$y_{k+n} = a_{1,k}y_{k+n-1} + a_{2,k}y_{k+n-2} + \cdots + a_{n,k}y_k, \text{ for } k \geq 0,$$

then $y_k = y_k^{(p)} + c_1 y_k^{(1)} + c_2 y_k^{(2)} + \cdots + c_n y_k^{(n)}$, for $k \geq 0$, is a general solution of the nonhomogeneous linear difference (4.70), where $c_1, \ldots, c_n$ are arbitrary constants.

4.6.3 The Method of Undetermined Coefficients

It turns out that for difference equations, the method of undetermined coefficients is completely analogous to that for differential equations. Suppose an *n*-th order linear *constant coefficients* nonhomogeneous difference (4.70) has forcing term f_k that is a linear combination of solutions from the zoo of solutions of LCCHΔEs. Construct the list L_1 of roots of the corresponding LCCHΔEs characteristic polynomial, $r^n - a_1 r^{n-1} + a_2 r^{n-2} + \cdots + a_n$, and the list L_2 of roots corresponding to the forcing term(s) in f_k. The superlist $\mathsf{L} \triangleq \mathsf{L}_1, \mathsf{L}_2$ will show us the form of a particular solution, $y_k^{(p)}$, involving constants to be determined after we substitute it into the original, nonhomogeneous difference equation. The general solution of the nonhomogeneous difference equation will be $y_k = y_k^{(h)} + y_k^{(p)}$, where $y_k^{(h)}$ is the general solution of the corresponding linear homogeneous difference equation.

Example 4.39

Find the form of a particular solution of the difference equation

$$y_{k+2} = 2(1-\gamma)y_{k+1} - y_k + 3k - \delta^k,$$

where δ and γ are real constants, $\delta \neq 0$, and $0 < \gamma < 2$.

Method: The corresponding linear homogeneous difference equation has characteristic polynomial $r^2 - 2(1-\gamma)r + 1 = (r-(1-\gamma))^2 + (1-(1-\gamma)^2)$, which has roots

$$r = (1-\gamma) \pm i\sqrt{1-(1-\gamma)^2} = \alpha \pm i\nu = \rho\cos\omega \pm i\rho\sin\omega.$$

We have

$$\rho = \sqrt{\alpha^2+\nu^2} = \sqrt{(1-\gamma)^2 + \left(\sqrt{1-(1-\gamma)^2}\right)^2} = \cdots = 1$$

and

$$\tan\omega = \frac{\sqrt{1-(1-\gamma)^2}}{1-\gamma}.$$

The forcing term(s) in $f_k \triangleq 3k - r^k$ is (a) $3k$, which comes from roots $r = 1, 1$, and (b) δ^k, which comes from root $r = \delta$. We have list $\mathsf{L}_2 = 1, 1, \delta$, so the superlist is $1 \cdot e^{\pm i\omega}, 1, 1, \delta$. This gives $y_k = c_1\cos\omega k + c_2\sin\omega k + c_3 1^k + c_4 k\cdot 1^k + c_5\delta^k$, so the correct form of a particular solution is

$$y_k^{(p)} = A + Bk + C\delta^k,$$

where A, B, C are constants to be determined. ◯

Example 4.40

Professor Sitsin Ivory Tower contemplates taking out a loan in the amount of P dollars. The letter P stands for the "principal" of the loan. The First Federal Bank would charge her interest of $100R$ % per month, and The Bank will take an automatic loan repayment of M dollars per month from her checking account.

(a) Derive a linear constant coefficients nonhomogeneous difference equation IVP model for this situation.

(b) Solve that model.

(c) If the professor wishes the loan to be exactly paid off after the Nth monthly payment, find formulas for
   (i) The monthly payment amount in terms of the loan principal, the monthly interest rate, and the number for payments.
   (ii) The number of payments, in terms of the loan principal, the monthly payment, and the monthly interest rate.

(d) Suppose the annual percentage interest rate on her "subprime" loan is 11%. If the professor can only afford to make monthly payments of \$123.45 and needs to finish paying off the loan in seven years, what is the largest loan that she can afford to take out?

Method:

(a) Let y_k be the monthly outstanding balance on her loan, which satisfies the initial condition $y_0 = P$. The Bank charges interest on the outstanding balance before crediting her monthly payment, so

$$y_{k+1} = y_k + Ry_k - M,$$

that is, $y_{k+1} = (1 + R)y_k - M$.

(b) The corresponding linear homogeneous difference equation has characteristic polynomial $(r - (1 + R))$, so $y_k^{(h)} = c_1(1 + R)^k$. We have list $\mathsf{L}_2 = 1$, so the superlist is $\mathsf{L} = \mathsf{L}_1, \mathsf{L}_2 = 1 + R, 1$, and the correct form of a particular solution is

$$y_k^{(p)} = A \cdot (1)^k \equiv A,$$

where A is a constant to be determined after substituting $y_k^{(p)}$ into the original, linear nonhomogeneous difference equation:

$$A = y_{k+1}^{(p)} = (1 + R)y_k^{(p)} - M = (1 + R)A - M.$$

We get $A = \frac{M}{R}$, so the general solution of the difference equation is

$$y_k = y_k^{(p)} + y_k^{(h)} = \frac{M}{R} + c_1(1 + R)^k,$$

where c_1 is an arbitrary constant. Substitute this into the IC to get

$$P = y_0 = \frac{M}{R} + c_1(1 + R)^0 = \frac{M}{R} + c_1;$$

hence, $c_1 = P - \frac{M}{R}$, and the solution of the model is

$$y_k = \frac{M}{R} + \left(P - \frac{M}{R}\right)(1 + R)^k.$$

(c) Having the loan exactly paid off after the Nth-monthly payment means that

$$y_N = 0.$$

This is exactly like having a boundary condition for an ODE, except here we are dealing with discrete time measured in months. Substitute the general solution of the difference equation into the boundary condition to get

$$0 = y_N = \frac{M}{R} + \left(P - \frac{M}{R}\right)(1 + R)^N.$$

or, after multiplying through by R,

$$0 = M + (PR - M)(1 + R)^N. \tag{4.71}$$

(i) To solve for M in terms of the other parameters, put all terms involving M on one side of (4.71) to get

$$\left((1 + R)^N - 1\right)M = PR(1 + R)^N;$$

hence, the desired monthly payment is

$$M = P \cdot \frac{R(1 + R)^N}{(1 + R)^N - 1}.$$

(ii) To solve, instead, for N in terms of the other parameters, put all terms involving N on one side of (4.71) to get

$$(1+R)^N = \frac{-M}{PR-M} = \frac{M}{M-PR}. \tag{4.72}$$

The monthly payment, M, must exceed PR, the interest charged during the first month of the loan; otherwise it is impossible to pay of the loan by making monthly payments of M dollars! So, $M - PR > 0$.

So, taking the natural logarithm of both sides of (4.72), we get

$$N\,\ln(1+R) = \ln\left((1+R)^N\right) = \ln\left(\frac{M}{M-PR}\right);$$

hence, the number of monthly payments is

$$N = \frac{\ln\left(\frac{M}{M-PR}\right)}{\ln(1+R)}.$$

(d) The annual percentage interest rate is 11%, that is, 0.11, so the monthly interest rate is $R = \frac{0.11}{12}$. Substituting R, $N = 84$, and $M = \$123.45$ into (4.71), we get

$$0 = 123.45 + \left(P \cdot \frac{0.11}{12} - 123.45\right)\left(1 + \frac{0.11}{12}\right)^{84};$$

hence,

$$P = 123.45\left(1 - \frac{1}{(1+(0.11/12))^{84}}\right)\Big/\left(\frac{0.11}{12}\right) \approx \$7209.838417.$$

Because each month the bank may do some rounding off (or up) to the nearest (or higher) cent, to be on the safe side, she should limit her loan to, say, \$7208.99. ○

Over the life of the loan, she made 84 equal payments of \$123.45, that is, she gave the Bank a total of \$10369.80, of which about \$10369.80−7208.99 = \$3160.81 was interest. We say approximately because the 84th, last loan payment will be a little less than \$123.45. This is because the loan amount was hopefully lower than the actual amount paid when allowing for some cents lost due to rounding.

Learn More About It

Some good references for scalar difference equations are *Introduction to Difference Equations*, by Samuel Goldberg, Dover Publications, © 1986; *Finite Difference Equations*, by H. Levy and F. Lessman, Dover Publications, © 1992; and *Calculus of Finite Differences and Difference Equations*, by Murray R. Spiegel, Schaum's Outline Series, McGraw-Hill Book Company, © 1971. In addition, there is an interesting chapter on difference equations in *Lectures on Applications-Oriented Mathematics*, by Bernard Friedman, edited by Victor Twersky, Holden-Day, Inc., © 1969 [reprinted by Wiley-Interscience, © 1991].

4.6.4 Problems

In problems 1–7, solve the difference equation.

1. $y_{k+2} = y_k$
2. $y_{k+2} = -5y_{k+1} - 6y_k$
3. $y_{k+2} = -y_k$
4. $y_{k+2} = 2y_{k+1} - 10y_k$
5. $y_{k+3} = y_k$
6. $y_{k+1} = y_k + \left(\frac{1}{2}\right)^k$
7. $y_{k+2} = y_k - k$

In problems 8–12, solve the IVP.

8. $y_{k+2} = 4y_k,\ k \geq 0;\ y_0 = 2,\ y_1 = 1$
9. $y_{k+2} = -y_{k+1} - 2y_k,\ k \geq 0; y_0 = 2, y_1 = 1$
10. $y_{k+2} = -9y_k,\ k \geq 0;\ y_0 = -1,\ y_1 = 2$
11. $y_{k+2} = -4y_{k+1} - 13y_k,\ k \geq 0; y_0 = -2, y_1 = 1$
12. $y_{k+2} = y_{k+1} + y_k + \left(\frac{1}{2}\right)^k, k \geq 0; y_0 = -2, y_1 = 1$

In problems 13–17, the $k \times k$ matrix A_k is given. Find a formula for $\det(A_k)$ in terms of k.

13. $A_k = \begin{bmatrix} 1 & 1 & 0 & \cdots & & & 0 \\ 1 & 1 & 1 & 0 & & & \vdots \\ 0 & 1 & 1 & 1 & & & \vdots \\ \vdots & & & \ddots & \ddots & & \vdots \\ \vdots & & & & \ddots & \ddots & 0 \\ & & & & & 1 & 1 \\ 0 & \cdots & & & 0 & 1 & 1 \end{bmatrix}$

14. $A_k = \begin{bmatrix} -1.8 & 1 & 0 & \cdots & & & 0 \\ 1 & -1.8 & 1 & 0 & & & \vdots \\ 0 & 1 & -1.8 & 1 & & & \vdots \\ \vdots & & & \ddots & \ddots & & \vdots \\ \vdots & & & & \ddots & \ddots & 0 \\ & & & & & 1 & -1.8 \end{bmatrix}$

15. $A_k = \begin{bmatrix} -2 & 1 & 0 & \cdots & & 0 \\ 1 & -2 & 1 & 0 & & \vdots \\ 0 & 1 & -2 & 1 & & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ \vdots & & & & \ddots & 0 \\ & & & & 1 & -2 \end{bmatrix}$

[Note: The A_k of Problem 15 is among the most famous of matrices. It appears repeatedly when studying numerical methods for differential equations, as we will see in Chapter 8.]

16. $A_k = \begin{bmatrix} 2 & 1 & 0 & . & . & . & & & 0 \\ 1 & 2 & 1 & 0 & & & & & . \\ 0 & 1 & 2 & 1 & & & & & . \\ . & & & . & . & . & & & . \\ . & & & & . & . & . & & \\ . & & & & & . & 2 & 1 & 0 \\ & & & & & & 1 & 2 & 1 \\ 0 & . & . & . & & & 0 & 1 & 2 \end{bmatrix}$

17. $A_k = \begin{bmatrix} -3 & 2 & 0 & . & . & . & & & 0 \\ 2 & -3 & 2 & 0 & & & & & . \\ 0 & 2 & -3 & 2 & & & & & . \\ . & & & . & . & . & & & . \\ . & & & & . & . & . & & \\ . & & & & & . & -3 & 2 & 0 \\ & & & & & & 2 & -3 & 2 \\ 0 & . & . & . & & & 0 & 2 & -3 \end{bmatrix}$

18. Find a formula for y_k if $y_{k+2} = 4y_{k+1} + y_k = 0, k = 0, 1, 2, \ldots; y_0 = 3$; and $\lim_{k\to\infty} y_k = 0$.

19. Suppose that uranium hexafluoride comes in two forms, $U^{238}F_6$ and $U^{235}F_6$. Passing a mixture of $U^{238}F_6$ and $U^{235}F_6$ through a single centrifuge can increase the fraction that is $U^{235}F_6$ from f to rf, where $r = 1.0014$. Let y_k be the fraction of $U^{235}F_6$ after the mixture has been run through the *kth*-centrifuge. If the initial fraction is 1%, how enriched is the mixture after running through such a *cascade* of 1024 centrifuges in series?

20. Use Theorem 1.29(a) in Section 1.6 to explain why the Casorati determinant can be rewritten as

$$C\left(y_k^{(1)}, y_k^{(2)}, \ldots, y_k^{(n)}\right) = \begin{vmatrix} y_k^{(1)} & y_k^{(2)} & . & . & . & y_k^{(n)} \\ (\Delta y^{(1)})_k & (\Delta y^{(2)})_k & . & . & . & (\Delta y^{(n)})_k \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ (\Delta^{n-1} y^{(1)})_k & (\Delta^{n-1} y^{(2)})_k & . & . & . & (\Delta^{n-1} y^{(n)})_k \end{vmatrix}.$$

21. Use the effects of elementary row operations on determinants to explain why (4.69) is true:
    (a) For $n = 3$, that is, $C_{k+1} = a_{3,k} C_k$.
    (b) For all n, that is, $C_{k+1} = (-1)^{n-1} a_{n,k} C_k$.

22. For the difference equation $z_{k+1} = z_k + f_k$, we get

$$\begin{aligned} z_1 &= z_0 + f_0, \\ z_2 &= z_1 + f_1 = z_0 + (f_0 + f_1), \\ &\text{etc.} \end{aligned}$$

Come up with a formula for the solution, z_k in terms of z_0, and a sum.

23. Project: Come up with a general solution method for first-order linear difference equations, that is, of the form $y_{k+1} - p_k y_k = f_k$. Hint: Define $\mu_k = (p_0 \cdot p_1 \cdots p_{k-1})$ and $z_k = \mu_k y_k$. An equation of the form $\Delta z_k = f_k$, that is, $z_{k+1} = z_k + f_k$, is solvable using sums.
24. Explain why the method of undetermined coefficients works for linear constant coefficients nonhomogeneous difference equations (4.70) whose forcing term f_k is a linear combination of solutions from the zoo of solutions of LCCHΔEs.
25. Project: Come up with a variation of parameters solution method for second-order linear difference equations, that is, of the form $y_{k+2} = p_{k+1} y_{k+1} + q_k y_k + f_k$.

4.7 Short Take: z-Transforms

Suppose $x = x(t)$ is a function of time, for example, a voltage measurement. An **analog-to-digital converter** (A/D or ADC) may sample the voltage at times $t = \ldots, -T, 0, T, 2T, \ldots$. In other words, the A/D produces a sequence $\mathbf{x} = \{x_n\}_{n=-\infty}^{\infty}$ defined by

$$x_n \triangleq x(nT), \text{ for } n = \ldots, -1, 0, 1, 2, \ldots. \tag{4.73}$$

The sequence $\{x_n\}_{n=-\infty}^{\infty}$ is called the **discrete-time signal**, or **signal**, for short. Another notation commonly used by electrical engineers is

$$x[n] \triangleq x_n \text{ for } n = \ldots, -1, 0, 1, 2, \ldots.$$

Note that $x[n]$ is $x(nT)$. For example, if the voltage comes from a microphone used to record an audio compact disk, $T = \frac{1}{44{,}100}$ s, that is, the sampling is done 44,100 times each second.*

By definition, the Laplace transform uses $x(t)$ for $t \geq 0$ to produce

$$F(s) = \mathcal{L}[f(t)](s) \triangleq \int_0^\infty f(t) e^{-st}\, dt,$$

an improper integral that is a function of s. We have seen in Sections 4.4 and 4.5 many examples of Laplace transforms, which usually involve rational functions of s such as $F(s) = \frac{s}{(s+1)(s^2+4)}$.

The discrete-time version of this is called the **unilateral** or **one-sided** z**-transform** and is defined by

$$X(z) = \mathcal{Z}\big[\, x[n] \,\big] \triangleq \sum_{n=0}^{\infty} x_n z^{-n}. \tag{4.74}$$

* This sampling rate was chosen according to the Nyquist Shannon sampling theorem so as to allow reproduction of sounds with frequencies up to 20,000 Hz.

Note that, like the Laplace transform, $X(z)$ is calculated using only measurements at times $t \geq 0$, even if the discrete-time signal $x[n]$ is defined for $n < 0$.

We will usually be brief by omitting the words "unilateral" or "one sided" and refer to $X(z)$ as the z-transform of the signal $x[n]$. There is another kind of z-transform, using all of the signal values including $x[n]$ for $n < 0$, called the "bilateral z-transform," but we will not study it.

The z- transform sum is like the left endpoints Riemann sum approximating the Laplace transform:

$$\mathcal{L}[f(t)](s) \triangleq \int_0^\infty f(t)e^{-st}\,dt \approx_\triangle t \sum_{n=0}^{\infty} f(t_n)e^{-st_n} = T\sum_{n=0}^{\infty} f(t_n)z^{-n}.$$

This uses sampling points $0, T, 2T, 3T, \ldots$ and with

$$e^{-st_n} = e^{-s(nT)} = e^{(-n)(sT)} = \left(e^{sT}\right)^{-n} = z^{-n}, \quad \text{after defining} \quad z = e^{sT}.$$

We will see how the z- transform can be used to solve the linear difference equations we saw in Section 4.6. So, all solutions in the zoo of solutions of linear difference equations give particularly relevant discrete-time signals. Also, we will need results concerning operations on signals similar to the operations we saw in the study of Laplace transforms. For example, instead of taking the Laplace transform of a time derivative, $\frac{d}{dt}$, of a function, we could take the z-transform of a time difference, $x[n+1] - x[n]$, of a signal.

We will follow the conventional notation that signals have index n instead of using Section 4.6's conventional index k.

Clearly, the signal $x[n] \triangleq \alpha^n$ will be relevant because it is in the zoo of solutions of LCCHΔEs. Also, $x[n] \triangleq \alpha^n$ is the discrete-time signal for the function $x(t) = e^{\beta t/T}$ because $x(nT) = e^{\beta n} = \left(e^{\beta}\right)^n = \alpha^n$, if $\beta = \ln\alpha$ and $\alpha > 0$. Using the sum of a geometric series, we calculate

$$\mathcal{Z}\big[\,x[n]\,\big] \triangleq \sum_{n=0}^{\infty} x[n]z^{-n} = \sum_{n=0}^{\infty} \alpha^n z^{-n} = \sum_{n=0}^{\infty} \left(\frac{\alpha}{z}\right)^n = \frac{1}{1-(\alpha/z)} = \frac{z}{z-\alpha} = z\cdot\frac{1}{z-\alpha}. \tag{4.75}$$

Implicitly, we are assuming $|z| > |\alpha|$ so that the series converges. In this section, we will ignore this technicality.

After the transform of an exponential, the next ingredients are the transforms of a time shift and a forward difference. Recall from (4.65) in Section 4.6 the notation that

$$(\Delta x)[n] \triangleq x[n+1] - x[n], \quad \text{for all} \quad n. \tag{4.76}$$

This suggests that we also need the **shift operator** S defined by

$$(Sx)[n] \triangleq x[n+1], \quad \text{for all} \quad n. \tag{4.77}$$

Using the change of index $n' = n + 1$ and denoting $X(z) \triangleq \mathcal{Z}\big[\, x[n] \,\big]$, we calculate

$$\mathcal{Z}\big[\, Sx[n] \,\big] = \mathcal{Z}\big[\, x[n+1] \,\big] \triangleq \sum_{n=0}^{\infty} x[n+1]z^{-n} = \sum_{n=1}^{\infty} x[n']z^{-(n'-1)} = \sum_{n=1}^{\infty} z \cdot x[n']z^{-n'};$$

hence,

$$\mathcal{Z}\big[\, x[n+1] \,\big] = z \sum_{n=1}^{\infty} x[n']z^{-n'} = z\left(-x[0] + \sum_{n=0}^{\infty} x[n']z^{-n'}\right).$$

So,

$$\mathcal{Z}\big[\, x[n+1] \,\big] = zX(z) - z\,x[0]. \tag{4.78}$$

This is reminiscent of the Laplace transforms property (4) in Theorem 4.6 in Section 4.4, namely, that $\mathcal{L}[\,\dot{y}](s) = sY(s) - y(0)$.

Based on (4.78), we calculate that

$$\mathcal{Z}\big[\, x[n+2] \,\big] = \mathcal{Z}\big[\, S^2x[n] \,\big] = z\mathcal{Z}\big[\, Sx[n] \,\big](z) - z(Sx)[0] = z\,(zX(z) - z\,x[0]) - z\,x[0+1];$$

hence,

$$\mathcal{Z}\big[\, x[n+2] \,\big] = z^2X(z) - z^2x[0] - z\,x[1]. \tag{4.79}$$

The usual linearity properties apply to z-transforms:

(1) $\mathcal{Z}\big[\, x[n] + y[n] \,\big](z) = X(z) + Y(z)$.

(2) $\mathcal{Z}\big[\, cx[n] \,\big](z) = c\,X(z)$, any constant c.

Example 4.41

Solve the LCCHΔE $x[n+1] = x[n] + \alpha^n$, with IC $x[0] = 3$, assuming the constant $\alpha \neq 1$.

Method: Take the z-transforms of both sides of the difference equation $Sx[n] = x[n] + \alpha^n$ and denote $X(z) \triangleq \mathcal{Z}\big[\, x[n] \,\big]$. Using (4.78), we calculate

$$zX(z) - z\,x[0] = \mathcal{Z}\big[\, Sx[n] \,\big](z) = \mathcal{Z}\big[\, x[n] + \alpha^n \,\big](z) = X(z) + z \cdot \frac{1}{z-\alpha};$$

hence,

$$X(z) = z \cdot \frac{3}{z-1} + z \cdot \frac{1}{(z-1)(z-\alpha)}.$$

Use partial fractions in the form*

$$\frac{1}{(z-1)(z-\alpha)} = \frac{A}{z-1} + \frac{B}{z-\alpha},$$

* Note that we leave off the factor of z.

where A and B are constants to be determined. Multiply through by $(z-1)(z-\alpha)$ to get

$$1 = A(z-\alpha) + B(z-1).$$

Substitute in $z = 1$ and $z = \alpha$ to get $1 = A(1-\alpha)$ and $1 = B(\alpha-1)$, respectively. So,

$$X(z) = z\left(\frac{3}{z-1} + \frac{1}{\alpha-1}\left(-\frac{1}{z-1} + \frac{1}{z-\alpha}\right)\right) = z\cdot\left(\left(3 - \frac{1}{\alpha-1}\right)\frac{1}{z-1} + \frac{1}{\alpha-1}\cdot\frac{1}{z-\alpha}\right).$$

Using (4.75) in reverse, that is, taking $x[n] = \mathcal{Z}^{-1}\big[\,X(z)\,\big]$, gives that the solution is

$$x[n] = \left(3 - \frac{1}{\alpha-1}\right)\cdot 1^n + \frac{1}{\alpha-1}\cdot\alpha^n, \quad \text{for } n \geq 0,$$

that is,

$$x[n] = 3 + \frac{1}{\alpha-1}\cdot(\alpha^n - 1) = 3 + \frac{1-\alpha^n}{1-\alpha}, \quad \text{for } n \geq 0. \bigcirc$$

We could have found the solution by direct work with the difference equation to get the pattern of a geometric series, but the method of Example 4.41 was useful for illustrating use of the z-transform. Also, the methods of Section 4.6 might have been faster and easier for Example 4.41, although that might be a matter of personal preference.

Similar to work on ODE -IVPs using Laplace transforms, using the z-transform to solve a LCCHΔE makes use of the IC(s) right away rather than Section 4.6's method of first finding the general solution of the LCCHΔE. Also, the z-transform makes uses of an inhomogeneity right away as opposed to using the method of undetermined coefficients, as in Section 4.6.

Example 4.42

(Example 4.37 in Section 4.6 again) Solve the Fibonacci LCCHΔE (4.37) in Section 4.6, that is, $x[n+2] = x[n+1] + x[n]$, with ICs $x[0] = 0$ and $x[1] = 1$.

Method: Take the z-transforms of both sides of the difference equation $S^2x[n] = Sx[n] + x[n]$ and denote $X(z) \triangleq \mathcal{Z}\big[\,x[n]\,\big]$. Using (4.78) and (4.79), we calculate that the Fibonacci difference equation implies

$$z^2X(z) - z^2x[0] - z\,x[1] = (zX(z) - z\,x[0]) + X(z),$$

that is, using the ICs,

$$(z^2 - z - 1)X(z) = z\,(z\,x[0] + x[1] - x[0]) = z^2\cdot 0 + z(1-0) = z\cdot 1.$$

So,

$$X(z) = z\cdot\frac{1}{z^2 - z - 1}.$$

Using the quadratic formula, as in Example 4.37 in Section 4.6 , we factor $z^2 - z - 1 = \left(z - \frac{1+\sqrt{5}}{2}\right)\left(z - \frac{1-\sqrt{5}}{2}\right)$.

Use partial fractions in the form

$$\frac{1}{z^2 - z - 1} = \frac{A}{z - \frac{1+\sqrt{5}}{2}} + \frac{B}{z - \frac{1-\sqrt{5}}{2}}$$

where A and B are constants to be determined. Multiply through by $\left(z - \frac{1+\sqrt{5}}{2}\right)\left(z - \frac{1-\sqrt{5}}{2}\right)$ to get

$$1 = A\left(z - \frac{1-\sqrt{5}}{2}\right) + B\left(z - \frac{1+\sqrt{5}}{2}\right).$$

Substitute in $z = \frac{1+\sqrt{5}}{2}$ and $z = \frac{1-\sqrt{5}}{2}$ to get, respectively,

$$1 = \left(\frac{1+\sqrt{5}}{2} - \frac{1-\sqrt{5}}{2}\right)A, \quad \text{hence} \quad A = \frac{1}{\sqrt{5}},$$

and

$$1 = \left(\frac{1-\sqrt{5}}{2} - \frac{1+\sqrt{5}}{2}\right)B, \quad \text{hence} \quad B = -\frac{1}{\sqrt{5}}.$$

So,

$$X(z) = z \cdot \left(\frac{1}{\sqrt{5}} \cdot \frac{1}{z - \frac{1+\sqrt{5}}{2}} - \frac{1}{\sqrt{5}} \cdot \frac{1}{z - \frac{1-\sqrt{5}}{2}}\right).$$

Using (4.75) in reverse, that is, taking $x[n] = \mathcal{Z}^{-1}\big[\,X(z)\big]$, gives that the solution is

$$x[n] = \frac{1}{\sqrt{5}} \cdot \left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}} \cdot \left(\frac{1-\sqrt{5}}{2}\right)^n. \bigcirc$$

This conclusion agrees with (4.60) in Section 4.6.

4.7.1 Sinusoidal Signals

Again, from our study of difference equations, we know of other signals whose z-transforms should be relevant: First, using Euler's formula, we note that

$$\mathcal{Z}\big[\,\cos[\omega n]\big] = \mathcal{Z}\Big[\frac{1}{2}\left(\left(e^{i\omega n}\right) + \left(e^{-i\omega n}\right)\right)\Big] = \mathcal{Z}\Big[\frac{1}{2}\left(\left(e^{i\omega}\right)^n + \left(e^{-i\omega}\right)^n\right)\Big];$$

hence, using (4.75),

$$\begin{aligned}\mathcal{Z}\big[\,\cos[\omega n]\,\big] &= \frac{\frac{1}{2}z}{z - e^{i\omega}} + \frac{\frac{1}{2}z}{z - e^{-i\omega}} = z \cdot \frac{\frac{1}{2}(z - e^{-i\omega}) + \frac{1}{2}(z - e^{i\omega})}{(z - e^{i\omega})(z - e^{-i\omega})} \\ &= z \cdot \frac{z - \dfrac{e^{i\omega} + e^{-i\omega}}{2}}{z^2 - (e^{i\omega} + e^{-i\omega})z + 1},\end{aligned}$$

so

$$\mathcal{Z}\big[\,\cos[\omega n]\big] = z \cdot \frac{z - (\cos\omega)}{z^2 - 2(\cos\omega)z + 1}. \tag{4.80}$$

Similarly,

$$\mathcal{Z}\big[\,\sin[\omega n]\,\big] = \frac{1}{2i}\,\mathcal{Z}\Big[\,\left(e^{i\omega}\right)^n - \left(e^{-i\omega}\right)^n\,\Big];$$

hence, using (4.75),

$$\begin{aligned}\mathcal{Z}\big[\,\sin[\omega n]\,\big] &= \frac{\frac{1}{2i}\,z}{z-e^{i\omega}} - \frac{\frac{1}{2i}\,z}{z-e^{-i\omega}} = z\cdot\frac{\frac{1}{2i}\,(z-e^{-i\omega}) - \frac{1}{2i}\,(z-e^{i\omega})}{(z-e^{i\omega})(z-e^{-i\omega})}\\ &= z\cdot\frac{\dfrac{e^{i\omega}-e^{-i\omega}}{2i}}{z^2-(e^{i\omega}+e^{-i\omega})z+1},\end{aligned}$$

so

$$\mathcal{Z}\big[\,\sin[\omega n]\,\big] = z\cdot\frac{\sin\omega}{z^2-2(\cos\omega)z+1}. \tag{4.81}$$

4.7.2 Steady-State Solution

Analogous to the situation for functions $x(t)$, we have the concepts of transient and steady-state signal solutions.

Definition 4.10

A **transient solution** of an LCCHΔE is a signal $x[n]$ that (a) is either a particular solution or a solution of the corresponding homogeneous LCCHΔE and (b) has $\lim_{n\to\infty} x[n] = 0$.

Let $x_T[n]$ denote a transient solution. Intuitively, a transient solution is that part of a solution that becomes insignificant *eventually*.

The definition does *not* say that a transient solution must be a solution, by itself, of the LCCHΔE. As we will see, often, but not always, a transient solution turns out to be a solution of the corresponding linear homogeneous LCCHΔE.

Definition 4.11

A signal $f[n]$ is **bounded as** $n\to\infty$ if there is *no* subsequence n_m for which $\lim_{m\to\infty}|f[n_m]| = \infty$.

For example, $f[n] \triangleq e^{n\pi}\cos\frac{n\pi}{2}$ is *not* bounded as $n\to\infty$ because, using the subsequence $n_m = 2m$, $f[2m] = |e^{2m\pi}\cos m\pi| = e^{2m\pi}\to\infty$ as $m\to\infty$.

Definition 4.12

(a) If a solution of an LCCHΔE can be written as $x[n] = x_T[n] + x_S[n]$, where

- $x_T[n]$ is a transient solution, and
- $x_S[n]$ is bounded as $t \to \infty$ and does *not* have $\lim_{n\to\infty} x_S[n] = 0$,

then we say $x_S[n]$ is a **steady-state solution**.

(b) Alternatively, if a solution $x[n]$ of a LCCHΔE is bounded as $n \to \infty$ and does *not* have $\lim_{n\to\infty} x[n] = 0$, then we say that $x[n] = x_S[n]$ is **itself a steady-state solution**.

Example 4.43

Find the steady-state solution of $x[n+1] = \frac{1}{3}\,x[n] + 7\cos\frac{\pi n}{3}$, with IC $x[0] = 0$.

Method: Take the z-transforms of both sides of the difference equation $Sx[n] = \frac{1}{3}\,x[n] + 7\cos\frac{\pi n}{3}$ and denote $X(z) \triangleq \mathcal{Z}\big[\,x[n]\,\big]$. Using (4.78), and (4.80) with $\omega = \frac{\pi}{3}$, hence $\cos\omega = \frac{1}{2}$ and $\sin\omega = \frac{\sqrt{3}}{2}$, we calculate that the difference equation implies

$$zX(z) - z\,x[0] = \frac{1}{3}\,X(z) + z\cdot\frac{7\left(z-\frac{1}{2}\right)}{z^2-z+1};$$

hence, using the IC,

$$\left(z-\frac{1}{3}\right)X(z) = z\cdot\frac{7\left(z-\frac{1}{2}\right)}{z^2-z+1}.$$

So,

$$X(z) = z\cdot\frac{7\left(z-\frac{1}{2}\right)}{(z^2-z+1)\left(z-\frac{1}{3}\right)}.$$

Use partial fractions in the form

$$\frac{7\left(z-\frac{1}{2}\right)}{(z^2-z+1)\left(z-\frac{1}{3}\right)} = \frac{A}{z-\frac{1}{3}} + \frac{Bz+C}{z^2-z+1}$$

where $A, B,$ and C are constants to be determined. We multiply through by $(z^2-z+1)\left(z-\frac{1}{2}\right)$ to get

$$(\star)\quad 7\left(z-\frac{1}{2}\right) = A(z^2-z+1) + (Bz+C)\left(z-\frac{1}{3}\right).$$

Substitute in $z = \frac{1}{3}$ to get

$$-\frac{7}{6} = A\left(\left(\frac{1}{3}\right)^2 - \frac{1}{3} + 1\right);\quad \text{hence,}\quad A = -\frac{3}{2}.$$

Substitute this value of A into $(\star)$ to get

$$7\left(z-\frac{1}{2}\right) = -\frac{3}{2}(z^2-z+1) + (Bz+C)\left(z-\frac{1}{3}\right);$$

hence,

$$7\left(z-\frac{1}{2}\right) + \frac{3}{2}(z^2-z+1) = (Bz+C)\left(z-\frac{1}{3}\right),$$

that is,

$$\frac{3}{2}z^2 + \frac{11}{2}z - 2 = (Bz+C)\left(z-\frac{1}{3}\right).$$

So,

$$(Bz+C) = \frac{\frac{3}{2}z^2 + \frac{11}{2}z - 2}{\left(z-\frac{1}{3}\right)} = \frac{3}{2}z + 6.$$

This gives

$$X(z) = z\cdot\left(-\frac{3}{2}\cdot\frac{1}{z-\frac{1}{3}} + \frac{\frac{3}{2}z+6}{z^2-z+1}\right).$$

We want to use (4.75), (4.80), and (4.81) in reverse in order to find $x[n] = \mathcal{Z}^{-1}\big[\,X(z)\,\big]$.

We already know that

$$\mathcal{Z}\left[\cos\frac{\pi n}{3}\right] = z\cdot\frac{\left(z-\frac{1}{2}\right)}{z^2-z+1},$$

and (4.81) with $\omega = \frac{\pi}{3} = \frac{1}{2}$ implies

$$\mathcal{Z}\left[\sin\frac{\pi n}{3}\right] = z\cdot\frac{\frac{\sqrt{3}}{2}}{z^2-z+1}.$$

So, concerning the $\dfrac{\frac{3}{2}z+6}{z^2-z+1}$ part of $X(z)$, we calculate

$$\begin{aligned}
\mathcal{Z}^{-1}\Big[\,z\cdot\frac{\frac{3}{2}z+6}{z^2-z+1}\,\Big] &= \mathcal{Z}^{-1}\left[\,z\cdot\frac{\frac{3}{2}\left(z-\frac{1}{2}+\frac{1}{2}\right)+6}{z^2-z+1}\,\right] = \mathcal{Z}^{-1}\left[\,z\cdot\frac{\frac{3}{2}\left(z-\frac{1}{2}\right)+\frac{27}{4}}{z^2-z+1}\,\right] \\
&= \mathcal{Z}^{-1}\left[\,z\cdot\frac{\frac{3}{2}\left(z-\frac{1}{2}\right)}{z^2-z+1}\,\right] + \mathcal{Z}^{-1}\left[\,z\cdot\frac{\frac{27}{4}}{z^2-z+1}\,\right] \\
&= \frac{3}{2}\cdot\cos\frac{\pi n}{3} + \frac{27}{4}\cdot\frac{2}{\sqrt{3}}\cdot\sin\frac{\pi n}{3}.
\end{aligned}$$

The solution of the LCCHΔE and the IC is

$$\begin{aligned}
x[n] &= \mathcal{Z}^{-1}\Big[\,-\frac{3}{2}\cdot\frac{1}{z-\frac{1}{3}}\,\Big] + \frac{3}{2}\cdot\cos\frac{\pi n}{3} + \frac{9\sqrt{3}}{2}\cdot\sin\frac{\pi n}{3} \\
&= -\frac{3}{2}\cdot\left(\frac{1}{3}\right)^n + \frac{3}{2}\cdot\cos\frac{\pi n}{3} + \frac{9\sqrt{3}}{2}\cdot\sin\frac{\pi n}{3}.
\end{aligned}$$

The transient solution is $x_T[n] = -\frac{3}{2} \cdot \left(\frac{1}{3}\right)^n$. The steady-state solution is

$$x_S[n] = \frac{3}{2}\cos\frac{\pi n}{3} + \frac{9\sqrt{3}}{2}\sin\frac{\pi n}{3}. \bigcirc$$

In Problem 4.7.5.10, you will express this steady-state solution in the amplitude-phase form.

4.7.3 Convolution and *z*-Transforms

The convolution of two signals is defined by

$$(f * g)[n] \triangleq \sum_{m=0}^{n} f[m]g[n-m]. \tag{4.82}$$

and satisfies the following:

Theorem 4.18

(Convolution)

$$\mathcal{Z}\big[(f * g)[n]\big] = F(z) \cdot G(z). \tag{4.83}$$

4.7.4 Transfer Function

Example 4.44

The signal $x[n]$ is the solution of a **digital filter**

$$x[n+2] - a_1 x[n+1] + a_2 x[n] = y[n+1] + \beta y[n],$$

with ICs $x[0] = x[1] = 0$ and $y[0] = 0$. Find the input–output relationship between the signals $x[n]$ and $y[n]$.

Method: Using (4.78) and (4.79), we have

$$z^2X(z) - z^2x[0] - z\,x[1] - a_1\,(zX(z) - z\,x[0]) - a_2X(z) = zY(z) - z\,y[0] + \beta Y(z).$$

The ICs reduce this to

$$(z^2 - a_1 z - a_2)X(z) = (z + \beta)Y(z);$$

hence,

$$X(z) = \frac{z+\beta}{z^2 - a_1 z - a_2}Y(z).$$

Define the **transfer function** by

$$T(z) = \frac{z+\beta}{z^2 - a_1 z + a_2},$$

so $X(z) = T(z) \cdot Y(z)$. The convolution theorem implies

$$x[n] = \left(\mathcal{Z}^{-1}\big[\, T(z) \,\big] * y \right)[n]. \bigcirc$$

Learn More About It

Complex Variables for Mathematics and Engineering, 5th edn., by John H. Mathews and Russell W. Howell, Jones and Bartlett Publisher, Inc., © 2006, has an extensive chapter on z transforms and applications. Chapter 4 of *Lectures on Applications-Oriented Mathematics*, by Bernard Friedman, ed. by Victor Twersky, Holden-Day Inc., © 1969 (reprinted by Wiley Classics Library, © 1991) has useful material on summation by parts and on use of the "generating function," which involves nonnegative powers of z rather the nonpositive powers of z used in the z transform.

4.7.5 Problems

In problems 1 and 2, solve the difference equation with the IC(s).

1. $x[n+1] = -x[n] + \alpha^n$, with IC $x[0] = 1$, assuming the constant $\alpha \neq 1$.
2. $x[n+1] = \frac{1}{2}\, x[n] + \alpha^n$, with IC $x[0] = 0$, assuming the constant $\alpha \neq 1$.

In problems 3–5, find the steady-state solution of the difference equation with the IC(s).

3. $x[n+2] = -\frac{1}{2}\, x[n+1] + \frac{1}{2}\, x[n] + \cos \frac{\pi n}{3}$, with IC $x[0] = x[1] = 0$.
4. $x[n+1] = \frac{1}{2}\, x[n] + 12 \cos \frac{\pi n}{3}$, with IC $x[0] = 1$.
5. $x[n+2] = -x[n] + \alpha^n$, with IC $x[0] = x[1] = 0$, assuming the constant α satisfies $|\alpha| < 1$.
6. Use table entry Z.11 in Table 4.3 to explain why $\mathcal{Z}\big[\, \alpha^n \cos \omega n \,\big] = z \cdot \dfrac{z - (\alpha \cos \omega)}{z^2 - 2z \cos \omega + 1}$.
7. Use table entry Z.11 in Table 4.3 to explain why $\mathcal{Z}\big[\, \alpha^n \sin \omega n \,\big] = z \cdot \dfrac{a \sin \omega}{z^2 - 2z(\cos \omega) + 1}$.
8. Find the general solution of the Fibonacci LCCHΔE (4.54) in Section 4.6, that is, $x[n+2] = x[n+1] + x[n]$ in terms of the unspecified initial values $x[0]$ and $x[1]$.
9. Solve $x[n+2] = x[n+1] + 2x[n] + \cos \frac{\pi n}{3}$, with IC $x[0] = x[1] = 0$.
10. Express the steady-state solution of Example 4.43 in the amplitude-phase form.

Key Terms

analog-to-digital converter: beginning of Section 4.7
basic solution: Definition 4.8 in Section 4.6
beats phenomenon: after (4.23) in Section 4.2
Casorati determinant: before Definition 4.8 in Section 4.6
complete set of basic solutions: Definition 4.8 in Section 4.6
convolution: Definition 4.4 in Section 4.5, (4.82) in Section 4.7

TABLE 4.3

Table of Unilateral z-Transforms, Where c, ω Are Positive Constants and ℓ Is a Nonnegative Integer

Formula	$f[n]$	$X(z) = \mathcal{L}[f[n]]$
Z.1	$x[n] + y[n]$	$X(z) + Y(z)$
Z.2	$cx[n]$	$cX(z)$
Z.3	α^n	$z \cdot \frac{1}{z-\alpha}$
Z.4	$x[n+1]$	$zX(z) - z\,x[0]$
Z.5	$x[n+2]$	$z^2X(z) - z^2x[0] - z\,x[1]$
Z.6	$step[n]$	$\frac{1}{z-1}$
Z.7	$n\,step[n]$	$\frac{z}{(z-1)^2}$
Z.8	$n^2\,step[n]$	$\frac{z(z+1)}{(z-1)^3}$
Z.9	$\sin \omega n$	$z \cdot \frac{\sin \omega}{z^2 - 2z\cos\omega + 1}$
Z.10	$\cos \omega n$	$z \cdot \frac{z - (\cos \omega)}{z^2 - 2z\cos\omega + 1}$
Z.11	$\alpha^n y[n]$	$Y\left(\frac{z}{\alpha}\right)$
Z.12	$(f * g)[n]$	$F(z) \cdot G(z)$
LZ.13	$step[n-\ell]\,y[n-\ell]$	$z^{-\ell}\,Y(z)$
Z.14	$\delta[n-\ell]\,y[n-\ell]$	$z^{-\ell}\,y[\ell]$

delta function: Definition 4.6 in Section 4.5
de-tuning parameter: before Example 4.14 in Section 4.2
digital filter: Example 4.44 in Section 4.7
discrete-time signal: beginning of Section 4.7
forward difference operator: beginning of Section 4.6.1
frequency response: (4.13) in Section 4.2
general solution: Definition 4.1 in Section 4.1, Definitions 4.8 and 4.9 in Section 4.6
method of undetermined coefficients: Section 4.1.2
***n*-th order difference equation, solution**: (4.55) in Section 4.6
***n*-th order nonhomogeneous linear difference equation**: Definition 4.9 in Section 4.6
one-sided z-transform: (4.74) in Section 4.7
particular solution: Definition 4.1 in Section 4.1, Definition 4.9 in Section 4.6
periodic: Definition 4.2 in Section 4.2 and before (4.53) in Section 4.5
polar form: before Example 4.38 in Section 4.6
poles: Definition 4.5 in Section 4.5
practical resonance: before Example 4.14 in Section 4.2
pulse: before (4.50) in Section 4.5
quasiperiodic: Definition 4.2 in Section 4.2
resonance: Section 4.2.1
sawtooth function: Definition 4.7 in Section 4.5
shift operator S: (4.77) in Section 4.7
signal: beginning of Section 4.7
solution in the frequency domain: after Example 4.20 in Section 4.4
square wave: (4.53) in Section 4.5
steady-state solution: before Example 4.12 in Section 4.2, Definition 4.12 in Section 4.7

transfer function: Definition 4.5 in Section 4.5, Example 4.44 in Section 4.7
transient solution: Definition 4.10 in Section 4.7
unilateral *z*-transform: (4.74) in Section 4.7
unit step function: Definition 4.3 in Section 4.5
variation of constants, **variation of parameters**: Example 4.16 in Section 4.3

References

Bearman, P.W., Vortex shedding from Oscillating bluff bodies, *Ann. Rev. Fluid Mech.*, 16, 195–222, 1984.

Farlow, J., Hall, J.E., Mc Dill, J.M., and West, B.H., *Differential Equations and Linear Algebra*, 2nd edn., Prentice Hall, Inc., 2007.

Ugural, A.C. and Fenster, S.K., *Advanced Strength and Applied Elasticity*, Pearson, 2003, Section 8.6.

5

Linear Systems of ODEs

5.1 Systems of ODEs

In a sense, Chapter 5 equals Chapter 2 "plus" Chapter 3, in the sense that Chapter 5 combines use of matrix theory and ordinary differential equation (ODE) methods. When we have more than one linear ODE, results from matrix theory turn out to be useful.

Example 5.1

For the circuit shown in Figure 5.1, let $v(t)$ be the voltage drop across the capacitor and $I(t)$ be the loop current. The input $V(t)$ is a given function. Assume, as usual, that L, R, and C are constants. Write down a system of ODEs in $\mathbb{R}^2$ that models this circuit.

Method: The series *RLC* circuit shown in Figure 5.1 is analogous to the *DC* series *RLC* circuit discussed near the end of Section 3.3. The first ODE models the voltage drop across the capacitor being $v(t) = \frac{1}{C}\, q(t)$, where $q(t)$ is the charge on the capacitor and $\dot{q}(t) = I(t)$. The second ODE in the system is Kirchhoff's voltage law, $L\dot{I}(t) + RI(t) + v(t) = V(t)$, after dividing through by L. The system is

$$\left\{ \begin{array}{l} \dot{v}(t) = \frac{1}{C}\, I(t) \\ \\ \dot{I}(t) = \frac{1}{L}\, (V(t) - RI(t) - v(t)) \end{array} \right\} . \bigcirc \tag{5.1}$$

More generally, consider a system of two ODEs in unknowns $x_1(t), x_2(t)$:

$$\left\{ \begin{array}{l} \dot{x}_1(t) = F_1\big(t, x_1(t), x_2(t)\big) \\ \\ \dot{x}_2(t) = F_2\big(t, x_1(t), x_2(t)\big) \end{array} \right\} . \tag{5.2}$$

A special case is

$$\left\{ \begin{array}{l} \dot{x}_1(t) = a_{11}(t)x_1 + a_{12}(t)x_2 + f_1(t) \\ \\ \dot{x}_2(t) = a_{21}(t)x_1 + a_{22}(t)x_2 + f_2(t) \end{array} \right\} , \tag{5.3}$$

which is called a **linear system**.

In (5.3), we write x_1 instead of $x_1(t)$ even though x_1 is a function of t; we call this "suppressing the dependence on t" from the unknowns x_1, x_2. We will not suppress dependence on t in the coefficients $a_{ij}(t)$ or the right-hand sides $f_i(t)$.

5

Linear Systems of ODEs

5.1 Systems of ODEs

In a sense, Chapter 5 equals Chapter 2 "plus" Chapter 3, in the sense that Chapter 5 combines use of matrix theory and ordinary differential equation (ODE) methods. When we have more than one linear ODE, results from matrix theory turn out to be useful.

Example 5.1

For the circuit shown in Figure 5.1, let $v(t)$ be the voltage drop across the capacitor and $I(t)$ be the loop current. The input $V(t)$ is a given function. Assume, as usual, that L, R, and C are constants. Write down a system of ODEs in $\mathbb{R}^2$ that models this circuit.

Method: The series *RLC* circuit shown in Figure 5.1 is analogous to the *DC* series *RLC* circuit discussed near the end of Section 3.3. The first ODE models the voltage drop across the capacitor being $v(t) = \frac{1}{C}\, q(t)$, where $q(t)$ is the charge on the capacitor and $\dot{q}(t) = I(t)$. The second ODE in the system is Kirchhoff's voltage law, $L\dot{I}(t) + RI(t) + v(t) = V(t)$, after dividing through by L. The system is

$$\left\{ \begin{array}{l} \dot{v}(t) = \frac{1}{C}\, I(t) \\ \\ \dot{I}(t) = \frac{1}{L}\, (V(t) - RI(t) - v(t)) \end{array} \right\} . \bigcirc \tag{5.1}$$

More generally, consider a system of two ODEs in unknowns $x_1(t), x_2(t)$:

$$\left\{ \begin{array}{l} \dot{x}_1(t) = F_1\big(t, x_1(t), x_2(t)\big) \\ \\ \dot{x}_2(t) = F_2\big(t, x_1(t), x_2(t)\big) \end{array} \right\} . \tag{5.2}$$

A special case is

$$\left\{ \begin{array}{l} \dot{x}_1(t) = a_{11}(t)x_1 + a_{12}(t)x_2 + f_1(t) \\ \\ \dot{x}_2(t) = a_{21}(t)x_1 + a_{22}(t)x_2 + f_2(t) \end{array} \right\} , \tag{5.3}$$

which is called a **linear system**.

In (5.3), we write x_1 instead of $x_1(t)$ even though x_1 is a function of t; we call this "suppressing the dependence on t" from the unknowns x_1, x_2. We will not suppress dependence on t in the coefficients $a_{ij}(t)$ or the right-hand sides $f_i(t)$.

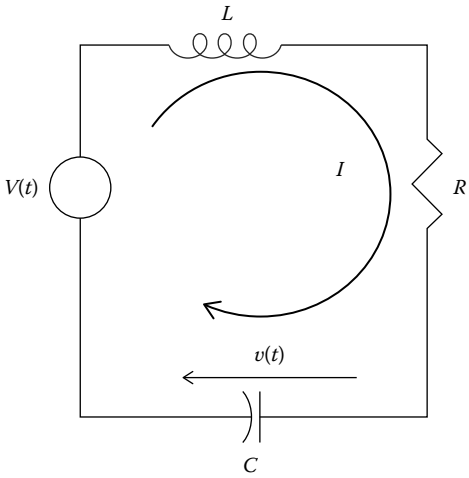


FIGURE 5.1
RLC series circuit.

The simplest such system is

$$\left.\begin{aligned}\dot{x}_1 &= a_{11}x_1 + a_{12}x_2\\ \dot{x}_2 &= a_{21}x_1 + a_{22}x_2\end{aligned}\right\}, \tag{5.4}$$

where $a_{11}, a_{12}, a_{21}, a_{22}$ are constants.

Chapter 5 will focus on linear constant coefficients homogeneous systems (**LCCHS**) and their linear nonhomogeneous analogues; in Chapter 18, we will look at nonlinear systems of ODEs, including how they relate to linear homogeneous systems of ODEs.

More generally, we can study systems involving n unknowns, $x_1(t), x_2(t), \ldots, x_n(t)$. The system is in $\mathbb{R}^n$ if n is the number of unknown functions whose derivatives appear, assuming no derivatives higher than the first appear. So, (5.1) through (5.4) are all systems in $\mathbb{R}^2$.

We can write system (5.2) compactly in vector form as

$$\dot{\mathbf{x}}(t) = \mathbf{F}\big(t, \mathbf{x}(t)\big),$$

where we define

$$\mathbf{x}(t) \triangleq \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}, \quad \dot{\mathbf{x}}(t) \triangleq \frac{d}{dt}\mathbf{x}(t) \triangleq \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix}, \quad \text{and} \quad \mathbf{F}\big(t, \mathbf{x}(t)\big) \triangleq \begin{bmatrix} F_1\,(t, x_1(t), x_2(t)) \\ F_2\,(t, x_1(t), x_2(t)) \end{bmatrix}.$$

We can rewrite linear system (5.3) compactly in matrix–vector form as

$$\dot{\mathbf{x}}(t) = A(t)\mathbf{x}(t) + \mathbf{f}(t),$$

where we define

$$A(t) \triangleq \begin{bmatrix} a_{11}(t) & a_{12}(t) \\ a_{21}(t) & a_{22}(t) \end{bmatrix} \quad \text{and} \quad \mathbf{f}(t) \triangleq \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix}.$$

Here we extend the definition of multiplication of matrix times vector to functions of t:

$$A(t)\mathbf{x}(t) = \begin{bmatrix} a_{11}(t) & a_{12}(t) \\ a_{21}(t) & a_{22}(t) \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \triangleq \begin{bmatrix} a_{11}(t)x_1(t) + a_{12}(t)x_2(t) \\ a_{21}(t)x_1(t) + a_{22}(t)x_2(t) \end{bmatrix}.$$

In particular, the system of ODEs (5.4) can be written compactly as

$$\dot{\mathbf{x}} = A\mathbf{x},$$

where the constant coefficient matrix is $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$.

All of (5.1) through (5.4) can be generalized to systems in $\mathbb{R}^n$, for example, $\dot{\mathbf{x}} = A\mathbf{x}$ can be short for

$$\begin{bmatrix} \dot{x}_1 \\ . \\ . \\ . \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} a_{11} & . & . & . & a_{1n} \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ a_{n1} & . & . & . & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ . \\ . \\ . \\ x_n \end{bmatrix}.$$

Definition 5.1

A **solution** of an ODE system in $\mathbb{R}^n$,

$$\dot{\mathbf{x}} = \mathbf{F}(t, \mathbf{x}), \tag{5.5}$$

is an n-vector of functions $\mathbf{x}(t)$ defined on an open interval $\mathcal{I}$ for which the derivative also exists on $\mathcal{I}$ and satisfies (5.5), that is, $\dot{\mathbf{x}}(t) = \mathbf{F}\big(t, \mathbf{x}(t)\big)$, for t in $\mathcal{I}$.

Theorem 5.1

(Existence and uniqueness for solution of a linear system) If the matrix of functions $A(t)$ and the vector of functions $f(t)$ are continuous for t in an open interval $\mathcal{I}$ and t_0 is inside $\mathcal{I}$, then the initial value problem (IVP) for the linear system

$$\left\{ \begin{array}{l} \dot{\mathbf{x}} = A(t)\mathbf{x} + \mathbf{f}(t) \\ \mathbf{x}(t_0) = \mathbf{x}_0 \end{array} \right\}$$

has exactly one solution on I.

Example 5.2

For the circuit shown in Figure 5.2, let $v_1(t), v_2(t)$ be the voltage drops across the capacitors whose capacitances are C_1, C_2 and let $I_1(t), I_2(t)$ be the loop currents. Write down a system of ODEs in $\mathbb{R}^3$ that models this circuit, assuming L, R, C_1, and C_2 are, as usual, constants.

Method: In the first loop, Kirchhoff's voltage law gives

$$L\dot{I}_1(t) + v_1(t) = V(t). \tag{5.6}$$

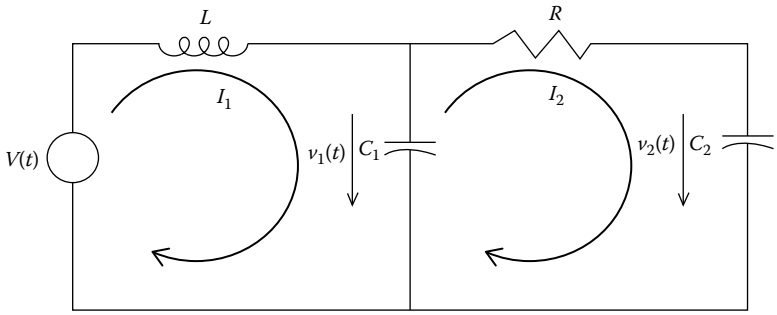


FIGURE 5.2
RLC two-loop circuit.

The input $V(t)$ is a given function. In the second loop, Kirchhoff's voltage law gives the algebraic equation

$$RI_2(t) + v_2(t) - v_1(t) = 0,$$

which we can solve for I_2 in terms of $v_1(t), v_2(t)$ to get

$$I_2(t) = \frac{1}{R}(v_1(t) - v_2(t)). \tag{5.7}$$

In terms of the loop currents, the voltages across the capacitors satisfy

$$\dot{v}_1(t) = \frac{1}{C_1}(I_1 - I_2) \quad \text{and} \quad \dot{v}_2(t) = \frac{1}{C_2}I_2. \tag{5.8}$$

Together, (5.6) through (5.8) give a linear system in $\mathbb{R}^3$, that is, a linear system of three ODEs in three unknowns, $I_1(t)$, $v_1(t)$, and $v_2(t)$:

$$\frac{d}{dt}\begin{bmatrix} I_1(t) \\ v_1(t) \\ v_2(t) \end{bmatrix} = \begin{bmatrix} 0 & -\frac{1}{L} & 0 \\ \frac{1}{C_1} & -\frac{1}{C_1 R} & \frac{1}{C_1 R} \\ 0 & \frac{1}{C_2 R} & -\frac{1}{C_2 R} \end{bmatrix} \begin{bmatrix} I_1(t) \\ v_1(t) \\ v_2(t) \end{bmatrix} + \begin{bmatrix} \frac{1}{L}V(t) \\ 0 \\ 0 \end{bmatrix}. \bigcirc \tag{5.9}$$

Example 5.3

Rewrite as a linear system the ODE that modeled the spring–mass–damper system at the beginning of Section 3.3.

Method: The spring–mass–damper system is modeled by ODE $m\ddot{y} + b\dot{y} + ky = 0$. If we define the velocity of the mass by $v = \dot{y}$, the physical situation is modeled by the system of two ODEs:

$$\left\{ \begin{array}{c} \dot{y} = v \\ m\dot{v} = \Sigma Forces = -bv - ky \end{array} \right\}.$$

We can rewrite these ODEs in the matrix–vector form of a linear system in $\mathbb{R}^2$:

$$\frac{d}{dt}\begin{bmatrix} y \\ v \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{b}{m} \end{bmatrix} \begin{bmatrix} y \\ v \end{bmatrix}. \bigcirc \tag{5.10}$$

Example 5.4

Suppose an object has temperature T and it is in a medium whose temperature is M. In Section 3.1, we used Newton's law of cooling,

$$\dot{T}(t) = -k_T(T - M),$$

where k_T is a constant dependent on the object's material nature, in units of $\frac{1}{\text{s}}$.

Unlike Section 3.1, now let us assume that the temperature of the medium is affected by the object. Find a system of ODEs that models the whole situation.

Method: Let us apply Newton's law of cooling to the medium. We get

$$\dot{M}(t) = -k_M(M - T),$$

where k_M is a constant dependent on the medium's material nature.

So, the temperature of the medium affects the temperature of the object, which in turn affects the temperature of the medium: the temperatures of the object and the medium are intertwined. They satisfy the system of ODEs:

$$\frac{d}{dt}\begin{bmatrix} T \\ M \end{bmatrix} = \begin{bmatrix} -k_T & k_T \\ k_M & -k_M \end{bmatrix}\begin{bmatrix} T \\ M \end{bmatrix}. \tag{5.11}$$

We'll assume that k_T, k_M are constants, which is reasonable as long as the temperatures are not changing too much and the materials are not changing their phases. ○

5.1.1 Systems of Second-Order Equations

We saw that a second-order scalar ODE can be rewritten as a system of two first-order scalar ODEs. Newton's second law of motion relating the acceleration of an object to the sum of the forces naturally leads to a second-order ODE.

Similarly, if there are several objects, Newton's law will apply to each of them, giving a system of second-order scalar ODEs.

Just as for a single second-order scalar ODE, we can rewrite a system of m second-order scalar ODEs as a system of $2m$ first-order scalar ODEs. We will see that for certain systems of second-order scalar ODEs, it is simpler to leave them as first-order ODEs.

Example 5.5

Describe the motion of the two objects, whose masses are m_1 and m_2, in the physical system depicted in Figure 5.3. Assume that the system is in equilibrium when

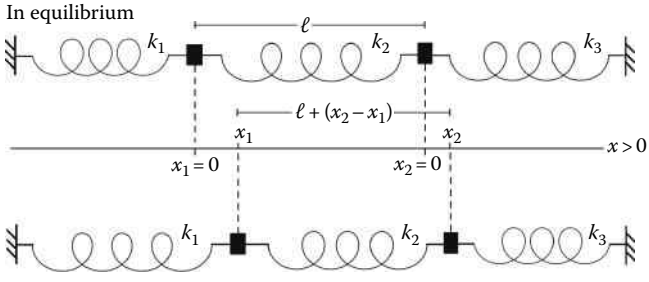


FIGURE 5.3
Two masses and three horizontal springs.

$x_1 = x_2 = 0$. As depicted in the picture, k_1, k_2, k_3 are the spring constants of the three horizontal springs. Assume there are no damping forces.

Method: Assume $x_1 > 0$ when the first object is to the right of its equilibrium position and similarly for $x_2 > 0$. The first spring is stretched a distance of x_1, if $x_1 > 0$, and conversely, the first spring is compressed a distance of $-x_1$, if $x_1 < 0$. The first spring exerts a force of $-k_1 x_1$ on the first object, so the first spring acts to bring the first object back to equilibrium.

The third spring is compressed by a distance of x_2 if $x_2 > 0$, and conversely, the third spring is stretched by a distance of $-x_2$, if $x_2 < 0$. The third spring exerts a force of $-k_3 x_2$ on the second object, so the third spring acts to bring the second object back to equilibrium.

The second, middle spring is compressed by a distance of x_1 *and* compressed by a distance of $-x_2$. In the picture, $x_2 > 0$, and conversely, the position of the second object contributes a negative compression, that is, a positive stretch, to the length of the middle spring. So, the middle spring has (*net compression*) $= x_1 + (-x_2) = (x_1 - x_2)$, that is, the middle spring has (*net stretch*) $= -$(*net compression*) $= (x_2 - x_1)$. The middle spring exerts on the first object a force of k_2(*net stretch*), that is, $k_2(x_2 - x_1)$. [For example, the picture has $x_1 > x_2$, so the middle spring pulls the first object to the right.] The middle spring exerts on the second object a force of $k_2(x_1 - x_2)$. In the picture, $x_1 > x_2$, so the middle spring pushes the second object to the right.

Newton's second law of motion gives us the ODEs

$$m_1 \ddot{x}_1 = \Sigma\text{Forces on first object} = -k_1 x_1 + k_2(x_2 - x_1)$$

and

$$m_2 \ddot{x}_2 = \Sigma\text{Forces on second object} = k_2(x_1 - x_2) - k_3 x_3.$$

Recall that we assumed this system has no damping forces.

We can write this system of second-order ODEs in terms of the vector $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$:

$$\ddot{\mathbf{x}} = \begin{bmatrix} -\dfrac{k_1 + k_2}{m_1} & \dfrac{k_2}{m_1} \\ \dfrac{k_2}{m_2} & -\dfrac{k_2 + k_3}{m_2} \end{bmatrix} \mathbf{x} \triangleq A\mathbf{x}. \bigcirc \tag{5.12}$$

In Problem 5.1.3.2, you will choose specific values for the physical parameters in Example 5.5.

5.1.2 Compartment Models

In many biological and chemical systems, there is one or several *species* or locations of matter. For example, some matter may transmutate from one isotope into another isotope. In another example, one type of organism may utilize other organisms to survive or increase its population.

In Example 5.6 in the following, iodine moves among several locations or categories in the human body and also leaves the body. Those locations are called **compartments**. In Problem 3.1.4.32, we had a one compartment model for the amount of glucose in the bloodstream. Aside from a basic scientific interest, the study of iodine in the body is relevant to the prevention of radioactive contamination of the thyroid gland.

Example 5.6

(Compartmental model of iodine metabolism) (Riggs, 1952) Iodide compounds containing iodine are absorbed from food by the digestive system, circulate in the bloodstream, accumulate in and are used by the thyroid gland and other body tissues including the organs, and are excreted from the body in urine and, to usually a lesser extent, in feces. The thyroid gland uses iodine to produce and store thyroid hormone, which is essential to health. As body tissues use the hormone, it sends iodine, a breakdown product, back into the bloodstream.

Write down a system of ordinary differential equations modeling the amounts of iodine in the bloodstream, the thyroid, other body tissues (including other organs), the urine, and the feces, assuming that the rate of iodine flow out of a compartment to another is proportional to the amount of iodine in the compartment.

Method: As Riggs (see Riggs, 1952) put it, "$\cdots$ these so-called compartments do not exist within the body as actual physical entities with clearly defined boundaries, but are merely convenient abstractions." The amounts of iodine in the five compartments are defined by

$x_1(t)$ = the amount of iodine in the bloodstream
$x_2(t)$ = the amount of iodine in the thyroid
$x_3(t)$ = the amount of iodine in other body tissues
$x_4(t)$ = the amount of iodine excreted in feces
$x_5(t)$ = the amount of iodine in urine

We will ignore time delays in the movements of iodine due to nonuniform spatial distributions. Also, we will not distinguish between the many iodide compounds in which iodine is found in the body.

The flows of iodine are depicted in Figure 5.4.

The rate of change of $x_1(t)$ includes flows *into* the bloodstream from the digestive system at a rate f_1 and from the other body tissues from breakdown of hormone. We will ignore flow of iodine into the bloodstream from the thyroid because we assume that hormone moves very quickly from the bloodstream to the other body tissues.

The rate of change of $x_1(t)$ includes flows *out* of the bloodstream as the thyroid absorbs iodine, as hormone is absorbed by the other body tissues, and as iodine is excreted in urine:

$$\dot{x}_1 = a_{11}x_1 + a_{13}x_3 + f_1,$$

with constant $a_{11} < 0$ and constants $f_1, a_{13} > 0$.

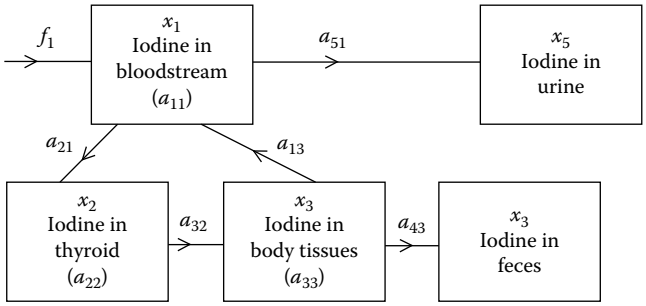


FIGURE 5.4
Example 5.6: Iodine model.

The rate of change of $x_2(t)$ includes flows into the thyroid from the bloodstream and flow out in the form of hormone:

$$\dot{x}_2 = a_{21}x_1 + a_{22}x_2,$$

with constant $a_{22} < 0$ and constant $a_{21} > 0$.

The rate of change of $x_3(t)$ includes flows into the other body tissues "directly" from the thyroid and flow out as a breakdown product of hormone:

$$\dot{x}_3 = a_{32}x_2 + a_{33}x_3,$$

with constant $a_{33} < 0$ and constant $a_{31} > 0$.

The rate of change of $x_4(t)$ includes flows into the feces from the other body tissues, specifically the liver. The rate of change of $x_5(t)$ includes flows into the urine from the bloodstream via the kidney(s):

$$\dot{x}_4 = a_{43}x_3$$

$$\dot{x}_5 = a_{51}x_1,$$

with constants $a_{41}, a_{51} > 0$.

By the conservation of iodine, we have $0 = a_{11} + a_{21} + a_{51}$, $0 = a_{22} + a_{32}$, and $0 = a_{33} + a_{13} + a_{43}$.

Altogether, the system of ODEs is

$$\dot{\mathbf{x}} = \begin{bmatrix} a_{11} & 0 & a_{13} & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 & 0 \\ 0 & a_{32} & a_{33} & 0 & 0 \\ 0 & 0 & a_{43} & 0 & 0 \\ a_{51} & 0 & 0 & 0 & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} f_1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

We can solve the first three ODEs together by themselves because the amounts x_4 and x_5 do not affect x_1, x_2, or x_3. Thus, the system can be reduced to

$$(\star) \quad \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} a_{11} & 0 & a_{13} \\ a_{21} & a_{22} & 0 \\ 0 & -a_{22} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} f_1 \\ 0 \\ 0 \end{bmatrix}.$$

After solving $(\star)$, we integrate $x_1(t)$ and $x_3(t)$ to find $x_4(t)$ and $x_5(t)$, which modelers can use as measurable outputs from the body in order to estimate the other parameters in the system.

According to Killough and Eckerman (1984), appropriate values for the constants are

$$a_{11} = -2.773,\ a_{13} = 5.199 \times 10^{-2},\ a_{21} = 0.832,\ a_{22} = -8.664 \times 10^{-3},$$

$$a_{32} = 8.664 \times 10^{-3}, a_{33} = -5.776 \times 10^{-2},\ a_{43} = 5.770 \times 10^{-3},\ a_{51} = 1.941. \bigcirc$$

We'll explain how to solve all six of Examples 5.1 through 5.6 in the next four sections.

5.1.3 Problems

1. Modify the iodine metabolism model of Example 5.6 to include the assumption that iodine also flows into the bloodstream from the thyroid in the form of hormone and from there flows into the other body tissues, that is, the flow of iodine from the thyroid to other body tissues is indirect.

2. Write a specific example of the system of two second-order ODEs in (5.12) after choosing specific values for the physical parameters.
3. Write a general model for a system of three masses and four springs that generalizes the system of two second-order ODEs in (5.12).
4. Rewrite the system of two second-order ODEs in (5.12) as a system of four first-order ODEs in a manner similar to what was done in Example 5.3.
5. In each of the two tanks depicted in Figure 5.5, there is a mixture containing a dye. Write down a system of two first-order ODEs specifying the amount of dye in tanks #1 and #2. The numbers in the tanks specify the volumes of mixture in the tanks. Each inflow arrow comes with two pieces of information: a flow rate, in gallons per minute, and a concentration of dye, in pounds per gallon; if a concentration is not specified, assume that the mixture in the tank is well-mixed and the concentration in the outflow equals the concentration in the tank.
6. For the circuit shown in Figure 5.6, let $v_1(t)$ be the voltage drop across the first resistor and $v_2(t)$ be the voltage drop across the capacitor, and let $I_1(t), I_2(t)$ be the loop currents. Write down a system of ODEs in $\mathbb{R}^3$ that models this circuit, assuming L, R_1, R_2, and C are, as usual, constants.
7. Suppose two objects have temperatures T_1 and T_2 and they are in a medium whose temperature is M. Assuming the two objects are far apart from each other, find a system of three ODEs that models the whole situation.

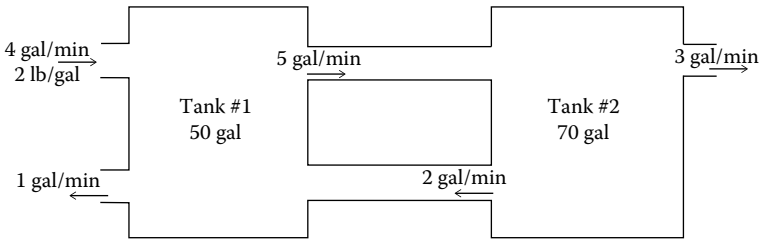


FIGURE 5.5
Problem 5.1.3.5.

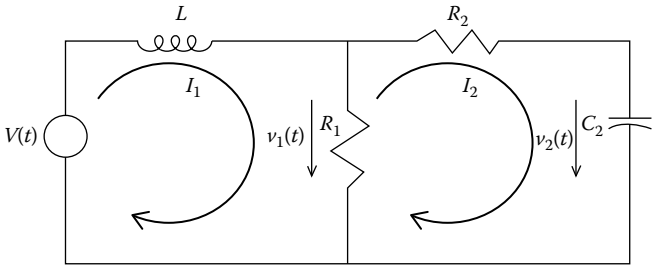


FIGURE 5.6
Problem 5.1.3.6.

5.2 Solving Linear Homogenous Systems of ODEs

While most of our attention will be devoted to solving LCCHS

$$\dot{\mathbf{x}} = A\mathbf{x}, \tag{5.13}$$

we will also discuss general systems of linear homogeneous ODEs whose coefficients are not necessarily constant.

What we will learn in Sections 5.2 and 5.3 for systems of linear homogeneous ODEs will also be useful in Sections 5.4 and 5.5 for solving systems of linear nonhomogeneous ODEs.

Example 5.7

Use eigenvalues and eigenvectors to solve the LCCHS:

$$\dot{\mathbf{x}} = \begin{bmatrix} -4 & -2 \\ 6 & 3 \end{bmatrix} \mathbf{x}. \tag{5.14}$$

Method: Let $A = \begin{bmatrix} -4 & -2 \\ 6 & 3 \end{bmatrix}$. Because A is 2×2, $\mathbf{x}$ must be a vector in $\mathbb{R}^2$.

In Chapter 3, we tried solutions of scalar linear, constant coefficients homogeneous ordinary differential equations (LCCHODEs) of the form $y(t) = ce^{st}$, where c and s were constants. Now let's try solutions of (5.14) in the form

$$\mathbf{x}(t) = e^{\lambda t}\mathbf{v},$$

where λ is a constant and $\mathbf{v}$ is a constant vector in $\mathbb{R}^2$. First, we note that $\frac{d}{dt}\Big[\, e^{\lambda t}\mathbf{v} \,\Big] = \lambda e^{\lambda t}\mathbf{v}$; you will explain why in Problem 5.2.5.16. So, substituting $\mathbf{x}(t)$ into LCCHS (5.14), we want $\dot{\mathbf{x}} = A\mathbf{x}$, that is,

$$\lambda e^{\lambda t}\mathbf{v} = \frac{d}{dt}\Big[\, e^{\lambda t}\mathbf{v} \,\Big] = \frac{d}{dt}\Big[\, \mathbf{x}(t) \,\Big] = A\mathbf{x}(t) = A(e^{\lambda t}\mathbf{v}) = e^{\lambda t}A\mathbf{v}.$$

Multiplying through by $e^{-\lambda t}$ gives us

$$\lambda\mathbf{v} = A\mathbf{v}.$$

So, we want $\mathbf{v}$ to be an eigenvector of A corresponding to eigenvalue λ.

In Example 2.4 in Section 2.1, we found eigenvalues and eigenvectors of this matrix A:

$$\lambda_1 = -1 \text{ with } \mathbf{v}^{(1)} = \begin{bmatrix} 2 \\ -3 \end{bmatrix} \quad \text{and} \quad \lambda_2 = 0 \text{ with } \mathbf{v}^{(2)} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

Using the principle of linearity superposition,

$$\mathbf{x}(t) = c_1 e^{\lambda_1 t}\mathbf{v}^{(1)} + c_2 e^{\lambda_2 t}\mathbf{v}^{(2)} = c_1 e^{(-1)t}\begin{bmatrix} 2 \\ -3 \end{bmatrix} + c_2 e^{0\cdot t}\begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

solves (5.14) for arbitrary constants c_1, c_2. Theorem 5.2 will explain why

$$\mathbf{x}(t) = c_1 e^{-t}\begin{bmatrix} 2 \\ -3 \end{bmatrix} + c_2 \begin{bmatrix} 1 \\ -2 \end{bmatrix},$$

where c_1, c_2 are arbitrary constants, gives *all* of the solutions of (5.14). ◯

Example 5.8

Solve the IVP for Example 5.4 in Section 5.1 model of the temperatures of the object and medium, that is,

$$\left\{ \begin{bmatrix} \dot{T} \\ \dot{M} \end{bmatrix} = \begin{bmatrix} -k_T & k_T \\ k_M & -k_M \end{bmatrix} \begin{bmatrix} T \\ M \end{bmatrix} \right\}, \quad \left\{ \begin{bmatrix} T(0) \\ M(0) \end{bmatrix} = \begin{bmatrix} T_0 \\ M_0 \end{bmatrix} \right\} \tag{5.15}$$

and interpret the results physically. Assume k_T, k_M are positive constants.

Method: Let $A = \begin{bmatrix} -k_T & k_T \\ k_M & -k_M \end{bmatrix}$. First, find the eigenvalues:

$$0 = |A - \lambda I| = \begin{vmatrix} -k_T - \lambda & k_T \\ k_M & -k_M - \lambda \end{vmatrix} = (-k_T - \lambda)(-k_M - \lambda) - k_T k_M = \lambda^2 + (k_T + k_M)\lambda$$

$$= \lambda(\lambda + k_T + k_M).$$

The eigenvalues are $\lambda_1 = -(k_T + k_M)$ and $\lambda_2 = 0$. To find the corresponding eigenvectors, we do two different row reductions: First,

$$\left[A - \left(-\left(k_T + k_M \right) \right) I \mid \mathbf{0} \right] = \begin{bmatrix} k_M & k_T & \mid 0 \\ k_M & k_T & \mid 0 \end{bmatrix} \sim \cdots \sim \begin{bmatrix} \textcircled{1} & k_T/k_M & \mid 0 \\ 0 & 0 & \mid 0 \end{bmatrix},$$

so M is the only free variable and the first eigenvalue's eigenvectors are

$$\mathbf{v}^{(1)} = c_1 \begin{bmatrix} -k_T \\ k_M \end{bmatrix}, \; c_1 \neq 0.$$

Second,

$$\left[A - (0)I \mid \mathbf{0} \right] = \begin{bmatrix} -k_T & k_T & \mid 0 \\ k_M & -k_M & \mid 0 \end{bmatrix} \sim \quad \cdots \quad \sim \begin{bmatrix} \textcircled{1} & -1 & \mid 0 \\ 0 & 0 & \mid 0 \end{bmatrix},$$

so M is the only free variable and the second eigenvalue's eigenvectors are

$$\mathbf{v}^{(2)} = c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \; c_1 \neq 0.$$

The solutions of LCCHS (5.15) are

$$\mathbf{x}(t) = c_1 e^{-(k_T + k_M)t} \begin{bmatrix} -k_T \\ k_M \end{bmatrix} + c_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

where c_1, c_2 are arbitrary constants, which we use to satisfy the initial conditions (ICs). Using Lemma 1.3 in Section 1.7 and defining $\mathbf{c} = [c_1 \quad c_2]^T$, we have

$$\begin{bmatrix} T_0 \\ M_0 \end{bmatrix} = \begin{bmatrix} T(0) \\ M(0) \end{bmatrix} = c_1 \begin{bmatrix} -k_T \\ k_M \end{bmatrix} + c_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -k_T & 1 \\ k_M & 1 \end{bmatrix} \mathbf{c},$$

which has unique solution

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} -k_T & 1 \\ k_M & 1 \end{bmatrix}^{-1} \begin{bmatrix} T_0 \\ M_0 \end{bmatrix} = \frac{1}{-k_T - k_M} \begin{bmatrix} 1 & -1 \\ -k_M & -k_T \end{bmatrix} \begin{bmatrix} T_0 \\ M_0 \end{bmatrix}$$

$$= \frac{-1}{k_T + k_M} \begin{bmatrix} T_0 - M_0 \\ -k_M T_0 - k_T M_0 \end{bmatrix}.$$

After some algebraic manipulations, we see that the solution of the IVP is

$$\begin{bmatrix} T(t) \\ M(t) \end{bmatrix} = \frac{1}{k_T + k_M} \left((M_0 - T_0) e^{-(k_T + k_M)t} \begin{bmatrix} -k_T \\ k_M \end{bmatrix} + (k_M T_0 + k_T M_0) \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right).$$

Because $e^{-(k_T+k_M)t} \to 0$ as $t \to \infty$, we have

$$\lim_{t \to \infty} \begin{bmatrix} T(t) \\ M(t) \end{bmatrix} = \frac{1}{k_T + k_M} \begin{bmatrix} k_M T_0 + k_T M_0 \\ k_M T_0 + k_T M_0 \end{bmatrix}.$$

Physically, this means that as t gets larger and larger, the temperatures of the object, $T(t)$, and the surrounding medium, $M(t)$, both approach the steady-state value of

$$\frac{k_M T_0 + k_T M_0}{k_T + k_M}.$$

This is what we think of as "common sense," specifically that the temperatures $T(t), M(t)$ should approach thermal equilibrium, as $t \to \infty$. But, our analysis establishes this *and* tells us the equilibrium temperature *value*, which depends on the constants k_T and k_M *and* the initial temperatures. Models whose solutions make quantitative predictions are very useful in engineering and science. ◯

The system of differential equations (5.11) in Section 5.1, that is, the ODEs in Example 5.8, has constant solutions $T(t) \equiv M(t) = T_\infty$ for *any* value of the constant T_∞, but we need to know the initial temperatures in order to find *what* constant value T_∞ gives physical equilibrium.

Analogous to Definition 3.10 in Section 3.4, we have

Definition 5.2

The **general solution** of a linear homogeneous system of ODEs

$$\dot{\mathbf{x}} = A(t)\mathbf{x} \tag{5.16}$$

in $\mathbb{R}^n$ has the form

$$\mathbf{x}_h(t) = c_1\mathbf{x}_1(t) + c_2\mathbf{x}_2(t) + \cdots + c_n\mathbf{x}_n(t)$$

if for every solution $\mathbf{x}^*(t)$ of (5.16) there are values of constants $c_1, c_2, \ldots, c_n$ giving $\mathbf{x}^*(t) = c_1\mathbf{x}_1(t) + c_2\mathbf{x}_2(t) + \cdots + c_n\mathbf{x}_n(t)$. In this case, we call the set of functions $\{\mathbf{x}_1(t), \ldots, \mathbf{x}_n(t)\}$ a **complete set of basic solutions**. Each of the vector-valued functions $\mathbf{x}_1(t), \ldots, \mathbf{x}_n(t)$ is called a **basic solution** of the linear homogeneous system (5.16).

For an LCCHS (5.13), we can say a lot:

Theorem 5.2

For an LCCHS (5.13), that is, $\dot{\mathbf{x}} = A\mathbf{x}$ in $\mathbb{R}^n$, suppose the $n \times n$ constant matrix A has a set of eigenvectors $\left\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}\right\}$ that is a basis for $\mathbb{R}^n$. If the corresponding eigenvalues are $\lambda_1, \ldots, \lambda_n$, then

$$\left\{e^{\lambda_1 t}\mathbf{v}^{(1)}, \ldots, e^{\lambda_n t}\mathbf{v}^{(n)}\right\}$$

is a complete set of basic solutions of $\dot{\mathbf{x}} = A\mathbf{x}$.

Why? To be very brief, similar to the explanation of Theorem 3.15 in Section 3.4, this follows from the existence and uniqueness Theorem 5.1 in Section 5.1 combined with Lemma 1.3 in Section 1.7. The next example will illustrate why Theorem 5.1 in Section 5.1 makes sense. □

Complex eigenvalues and eigenvectors will be discussed in Section 5.3.

Example 5.9

Solve the IVP

$$\left\{\begin{array}{c}\dot{\mathbf{x}} = \begin{bmatrix}-4 & -2\\ 6 & 3\end{bmatrix}\mathbf{x}\\ \\ \mathbf{x}(0) = \begin{bmatrix}5\\7\end{bmatrix}\end{array}\right\}. \tag{5.17}$$

Method: Using Theorem 5.2 and the result of Example 5.7, the general solution of LCCHS (5.17) is

$$\mathbf{x}(t) = c_1 e^{-t}\begin{bmatrix}2\\-3\end{bmatrix} + c_2\begin{bmatrix}1\\-2\end{bmatrix},$$

where c_1, c_2 are arbitrary constants. Substitute this into the ICs:

$$\begin{bmatrix}5\\7\end{bmatrix} = \mathbf{x}(0) = c_1\begin{bmatrix}2\\-3\end{bmatrix} + c_2\begin{bmatrix}1\\-2\end{bmatrix}.$$

By Lemma 1.3 in Section 1.7, this is the same as $\begin{bmatrix}5\\7\end{bmatrix} = \begin{bmatrix}2 & 1\\-3 & -2\end{bmatrix}\mathbf{c}$, which is solved by

$$\mathbf{c} = \begin{bmatrix}2 & 1\\-3 & -2\end{bmatrix}^{-1}\begin{bmatrix}5\\7\end{bmatrix} = \begin{bmatrix}2 & 1\\-3 & -2\end{bmatrix}\begin{bmatrix}5\\7\end{bmatrix} = \begin{bmatrix}17\\-29\end{bmatrix} = \begin{bmatrix}c_1\\c_2\end{bmatrix}.$$

The solution of the IVP is

$$\mathbf{x}(t) = 17e^{-t}\begin{bmatrix}2\\-3\end{bmatrix} - 29\begin{bmatrix}1\\-2\end{bmatrix} = \begin{bmatrix}-29 + 34e^{-t}\\ 58 - 51e^{-t}\end{bmatrix}. \bigcirc$$

For a quick check of part of the work, substitute in $t = 0$ to verify that $\mathbf{x}(0) = \begin{bmatrix}5\\7\end{bmatrix}$.

Example 5.10

Find the general solution of

$$\dot{\mathbf{x}} = \begin{bmatrix}2 & 2 & 4\\ 2 & -1 & 2\\ 4 & 2 & 2\end{bmatrix}\mathbf{x}. \tag{5.18}$$

Method: In Example 2.5 in Section 2.1, we found that the matrix

$$A \triangleq \begin{bmatrix} 2 & 2 & 4 \\ 2 & -1 & 2 \\ 4 & 2 & 2 \end{bmatrix}$$

has eigenvalues $\lambda_1 = \lambda_2 = -2,\ \lambda_3 = 7$, with corresponding eigenvectors

$$\mathbf{v}^{(1)} = \begin{bmatrix} 1 \\ -2 \\ 0 \end{bmatrix}, \quad \mathbf{v}^{(2)} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \quad \mathbf{v}^{(3)} = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix},$$

and that set of three vectors is a basis for $\mathbb{R}^3$. By Theorem 5.2, the general solution of LCCHS (5.18) is

$$\mathbf{x}(t) = c_1 e^{-2t} \begin{bmatrix} 1 \\ -2 \\ 0 \end{bmatrix} + c_2 e^{-2t} \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} + c_3 e^{7t} \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix},$$

where c_1, c_2, c_3 are arbitrary constants. ◯

Example 5.11

Solve the IVP

$$\left\{ \begin{array}{l} \dot{\mathbf{x}} = \begin{bmatrix} -4 & 0 & 3 \\ 0 & -4 & 0 \\ 1 & 0 & -2 \end{bmatrix} \mathbf{x} \\ \\ \mathbf{x}(0) = \begin{bmatrix} 1 \\ -2 \\ -1 \end{bmatrix} \end{array} \right\}. \tag{5.19}$$

Method: First, we find the eigenvalues using the characteristic equation, by expanding along the second row:

$$0 = |A - \lambda I| = \begin{vmatrix} -4-\lambda & 0 & 3 \\ 0 & -4-\lambda & 0 \\ 1 & 0 & -2-\lambda \end{vmatrix} = (-4-\lambda) \begin{vmatrix} -4-\lambda & 3 \\ 1 & -2-\lambda \end{vmatrix}$$

$$= (-4-\lambda)(\lambda^2 + 6\lambda + 5) = (-4-\lambda)(\lambda+5)(\lambda+1).$$

The eigenvalues are

$$\lambda_1 = -5, \lambda_2 = -4, \lambda_3 = -1.$$

To find the corresponding eigenvectors, we do three different but easy row reductions. The first is

$$\left[A - (-5)I \mid \mathbf{0} \right] = \begin{bmatrix} 1 & 0 & 3 & \mid 0 \\ 0 & 1 & 0 & \mid 0 \\ 1 & 0 & 3 & \mid 0 \end{bmatrix} \sim \begin{bmatrix} \textcircled{1} & 0 & 3 & \mid 0 \\ 0 & \textcircled{1} & 0 & \mid 0 \\ 0 & 0 & 0 & \mid 0 \end{bmatrix},$$

so v_3 is the only free variable and the first eigenvalue's eigenvectors are

$$\mathbf{v}^{(1)} = c_1 \begin{bmatrix} -3 \\ 0 \\ 1 \end{bmatrix},\ c_1 \neq 0.$$

The second is

$$\left[\,A-(-4)I \mid \mathbf{0}\,\right]=\left[\begin{array}{ccc|c} 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 2 & 0 \end{array}\right]\sim \quad \cdots \quad \sim\left[\begin{array}{ccc|c} \textcircled{1} & 0 & 0 & 0 \\ 0 & 0 & \textcircled{1} & 0 \\ 0 & 0 & 0 & 0 \end{array}\right],$$

so v_2 is the only free variable and the second eigenvalue's eigenvectors are

$$\mathbf{v}^{(2)}=c_1\begin{bmatrix}0\\1\\0\end{bmatrix},\ c_1\neq 0.$$

The third is

$$\left[\,A-(-1)I \mid \mathbf{0}\,\right]=\left[\begin{array}{ccc|c} -3 & 0 & 3 & 0 \\ 0 & -3 & 0 & 0 \\ 1 & 0 & -1 & 0 \end{array}\right]\sim \quad \cdots \quad \sim\left[\begin{array}{ccc|c} \textcircled{1} & 0 & -1 & 0 \\ 0 & \textcircled{1} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right],$$

so v_3 is the only free variable and the third eigenvalue's eigenvectors are

$$\mathbf{v}^{(3)}=c_1\begin{bmatrix}1\\0\\1\end{bmatrix},\ c_1\neq 0.$$

By Theorems 5.2 and 2.7(c) in Section 2.2, the general solution of LCCHS (5.19) is

$$\mathbf{x}(t)=c_1e^{-5t}\begin{bmatrix}-3\\0\\1\end{bmatrix}+c_2e^{-4t}\begin{bmatrix}0\\1\\0\end{bmatrix}+c_3e^{-t}\begin{bmatrix}1\\0\\1\end{bmatrix}, \tag{5.20}$$

where c_1, c_2, c_3 are arbitrary constants.

To satisfy the ICs, that is,

$$\begin{bmatrix}1\\-2\\-1\end{bmatrix}=c_1\begin{bmatrix}-3\\0\\1\end{bmatrix}+c_2\begin{bmatrix}0\\1\\0\end{bmatrix}+c_3\begin{bmatrix}1\\0\\1\end{bmatrix}=\begin{bmatrix}-3&0&1\\0&1&0\\1&0&1\end{bmatrix}\mathbf{c},$$

we solve for $\mathbf{c}$:

$$\mathbf{c}=\begin{bmatrix}-3&0&1\\0&1&0\\1&0&1\end{bmatrix}^{-1}\begin{bmatrix}1\\-2\\-1\end{bmatrix}=\begin{bmatrix}-0.25&0&0.25\\0&1&0\\0.25&0&0.75\end{bmatrix}\begin{bmatrix}1\\-2\\-1\end{bmatrix}=\begin{bmatrix}-0.5\\-2\\-0.5\end{bmatrix}.$$

The solution of the IVP is

$$\mathbf{x}(t)=-\frac{1}{2}\,e^{-5t}\begin{bmatrix}-3\\0\\1\end{bmatrix}-2e^{-4t}\begin{bmatrix}0\\1\\0\end{bmatrix}-\frac{1}{2}\,e^{-t}\begin{bmatrix}1\\0\\1\end{bmatrix}=\begin{bmatrix}1.5e^{-5t}-0.5e^{-t}\\-2e^{-4t}\\-0.5e^{-5t}-0.5e^{-t}\end{bmatrix}.\ \bigcirc \tag{5.21}$$

We could have used other choices of eigenvectors and thus had a different looking general solution. But the *final* conclusion would still agree with the final conclusion of (5.21). You will explore this in Problem 5.2.5.19.

5.2.1 Fundamental Matrix and e^{tA}

Example 5.12

Recall that for Example 5.11 the general solution was (5.20), that is,

$$\mathbf{x}(t) = c_1 e^{-5t}\begin{bmatrix}-3\\0\\1\end{bmatrix} + c_2 e^{-4t}\begin{bmatrix}0\\1\\0\end{bmatrix} + c_3 e^{-t}\begin{bmatrix}1\\0\\1\end{bmatrix},$$

where c_1, c_2, c_3 are arbitrary constants. If we define three vector-valued functions of t by

$$\mathbf{x}^{(1)}(t) \triangleq e^{-5t}\begin{bmatrix}-3\\0\\1\end{bmatrix}, \quad \mathbf{x}^{(2)}(t) \triangleq e^{-4t}\begin{bmatrix}0\\1\\0\end{bmatrix}, \quad \mathbf{x}^{(3)}(t) \triangleq e^{-t}\begin{bmatrix}1\\0\\1\end{bmatrix},$$

Lemma 1.3 in Section 1.7 allows us to rewrite the general solution as

$$\mathbf{x}(t) = \left[\,\mathbf{x}^{(1)}(t) \;\vdots\; \mathbf{x}^{(2)}(t) \;\vdots\; \mathbf{x}^{(3)}(t)\,\right]\mathbf{c} \triangleq X(t)\mathbf{c}. \tag{5.22}$$

This defines the 3×3 matrix

$$X(t) = \left[e^{-5t}\begin{bmatrix}-3\\0\\1\end{bmatrix} \vdots\; e^{-4t}\begin{bmatrix}0\\1\\0\end{bmatrix} \vdots\; e^{-t}\begin{bmatrix}1\\0\\1\end{bmatrix}\right] = \begin{bmatrix}-3e^{-5t} & 0 & e^{-t}\\ 0 & e^{-4t} & 0\\ e^{-5t} & 0 & e^{-t}\end{bmatrix}. \bigcirc$$

This is an example of our next definition.

Definition 5.3

A **fundamental matrix of solutions**, or **fundamental matrix**, for a linear homogeneous system of ODEs (5.16) in $\mathbb{R}^n$, that is, $\dot{\mathbf{x}} = A(t)\mathbf{x}$, is an $n \times n$ matrix $X(t)$ satisfying

- Each of its n columns is a solution of the same system (5.16).
- $X(t)$ is invertible for all t in an open time interval of existence.

Definition 5.4

Given n solutions $\mathbf{x}^{(1)}(t), \ldots, \mathbf{x}^{(n)}(t)$ of the same linear homogeneous system $\dot{\mathbf{x}} = A(t)\mathbf{x}(t)$ in $\mathbb{R}^n$, their **Wronskian** determinant is defined by

$$W\left(\mathbf{x}^{(1)}(t), \ldots, \mathbf{x}^{(n)}(t)\right) \triangleq \left|\,\mathbf{x}^{(1)}(t) \;\vdots\; \cdots \;\vdots\; \mathbf{x}^{(n)}(t)\,\right|.$$

Theorem 5.3

Suppose $Z(t)$ is a fundamental matrix for $\dot{\mathbf{x}} = A(t)\mathbf{x}$. Then the unique solution of the IVP

$$\left\{\begin{matrix}\dot{\mathbf{x}} = A(t)\mathbf{x}\\ \mathbf{x}(t_0) = \mathbf{x}_0\end{matrix}\right\} \tag{5.23}$$

is given by

$$\mathbf{x}(t) = Z(t)\Big(Z(t_0)\Big)^{-1}\mathbf{x}_0. \tag{5.24}$$

Why? For all constant vectors

$$\mathbf{c} = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix},$$

the vector-valued function

$$Z(t)\,\mathbf{c} = \Big[\ \mathbf{x}^{(1)}(t)\ \big|\ \cdots\ \big|\ \mathbf{x}^{(n)}(t)\ \Big]\mathbf{c} = c_1\mathbf{x}^{(1)}(t) + \cdots + c_n\mathbf{x}^{(n)}(t) \triangleq \mathbf{x}(t),$$

by Lemma 1.3 in Section 1.7. We assumed that the columns of $Z(t)$, that is, the vector-valued functions $\mathbf{x}^{(1)}(t), \ldots, \mathbf{x}^{(n)}(t)$, are all solutions of $\dot{\mathbf{x}} = A(t)\mathbf{x}$, so the principle of linear superposition tells us that $\mathbf{x}(t)$ is a solution of $\dot{\mathbf{x}} = A(t)\mathbf{x}$.

Also, to solve the ICs, we want

$$\mathbf{x}_0 = \mathbf{x}(t_0) = Z(t_0)\mathbf{c},$$

and this can be accomplished by choosing $\mathbf{c} \triangleq (Z(t_0))^{-1}\,\mathbf{x}_0$. This leads to the solution of the IVP being

$$\mathbf{x}(t) = Z(t)\mathbf{c} = Z(t)\,(Z(t_0))^{-1}\,\mathbf{x}_0. \ \square$$

Theorem 5.4

Suppose an $n \times n$ matrix A has a set of n real eigenvectors $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}$ that is a basis for $\mathbb{R}^n$, corresponding to real eigenvalues $\lambda_1, \ldots, \lambda_n$. Then

$$Z(t) \triangleq \Big[\ e^{\lambda_1 t}\mathbf{v}^{(1)}\ \big|\ \cdots\ \big|\ e^{\lambda_n t}\mathbf{v}^{(n)}\ \Big]$$

is a fundamental matrix for LCCHS (5.13), that is, $\dot{\mathbf{x}} = A\mathbf{x}$.

Theorem 5.5

Suppose $Z(t)$ is an $n \times n$-valued differentiable function of t and is invertible for all t in an open time interval. Then $Z(t)$ is a fundamental matrix of $\dot{\mathbf{x}} = A(t)\mathbf{x}$ if, and only if, $\dot{Z}(t) = A(t)Z(t)$.

Why? Suppose $Z(t)$ is any fundamental matrix of $\dot{\mathbf{x}} = A(t)\mathbf{x}$. Denote the columns of $Z(t)$ by $\mathbf{z}^{(1)}(t), \ldots, \mathbf{z}^{(n)}(t)$. Then

$$\dot{Z}(t) = \left[\, \dot{\mathbf{z}}^{(1)}(t) \;\vdots\; \cdots \;\vdots\; \dot{\mathbf{z}}^{(n)}(t) \,\right] = \left[\, A(t)\mathbf{z}^{(1)}(t) \;\vdots\; \cdots \;\vdots\; A(t)\mathbf{z}^{(n)}(t) \,\right]$$

$$= A(t)\left[\, \mathbf{z}^{(1)}(t) \;\vdots\; \cdots \;\vdots\; \mathbf{z}^{(n)}(t) \right] = A(t)Z(t),$$

using Theorem 1.9 in Section 1.2.

In Problem 5.2.5.23, you will explain why the statement "if $\dot{Z}(t) = A(t)Z(t)$, then the columns of $Z(t)$ are all solutions of the same linear homogeneous system" is true. That, and the assumed invertibility of $Z(t)$, would imply that $Z(t)$ is a fundamental matrix of $\dot{\mathbf{x}} = A(t)\mathbf{x}$. □

Definition 5.5

If $X(t)$ is a fundamental matrix for LCCHS (5.13) in $\mathbb{R}^n$ and $X(t)$ satisfies the matrix-valued initial condition $X(0) = I_n$, then we define

$$e^{tA} \triangleq X(t).$$

Theorem 5.6

For LCCHS (5.13), that is, $\dot{\mathbf{x}} = A\mathbf{x}$,

(a) e^{tA} is unique.

(b) If $Z(t)$ is any fundamental matrix of that LCCHS, then $e^{tA} = Z(t)\big(Z(0)\big)^{-1}$.

Why?

(a) Uniqueness of e^{tA} follows from uniqueness of solutions of LCCHS (5.13), which follows from Theorem 5.1 in Section 5.1.

(b) Suppose $Z(t)$ is *any* fundamental matrix of that LCCHS. Denote $X(t) = Z(t)\big(Z(0)\big)^{-1}$. By Theorem 5.5, $X(t)$ is also a fundamental matrix of $\dot{\mathbf{x}} = A\mathbf{x}$, because $\big(Z(0)\big)^{-1}$ being a constant matrix implies that

$$\dot{X}(t) = \Big(\dot{Z}(t)\Big)\Big(Z(0)\Big)^{-1} = \Big(AZ(t)\Big)\Big(Z(0)\Big)^{-1} = A\Big(Z(t)\big(Z(0)\big)^{-1}\Big) = AX(t).$$

In addition, $X(0) = Z(0)\big(Z(0)\big)^{-1} = I_n$. By the definition of e^{tA}, it follows that $e^{tA} = X(t) = Z(t)\big(Z(0)\big)^{-1}$, as desired. □

One of the nice things about the uniqueness of e^{tA} is that different people may come up with radically different*-looking fundamental matrices $Z(t)$, but they should still agree on e^{tA}.

For the next result, we need another definition from matrix theory: if $B = \left[\, b_{ij} \,\right]$ is an $n \times n$ matrix, then the **trace** of B is defined by $tr(B) \triangleq b_{11} + b_{22} + \cdots + b_{nn}$, that is, the sum of the diagonal elements.

Theorem 5.7

(Abel's theorem) Suppose $\mathbf{x}^{(1)}(t), \ldots, \mathbf{x}^{(n)}(t)$ are n solutions of the same system of linear homogeneous system of ODEs $\dot{\mathbf{x}} = A(t)\mathbf{x}$. Then

$$W\left(\mathbf{x}^{(1)}(t), \ldots, \mathbf{x}^{(n)}(t)\right) = \exp\left(-\int_{t_0}^{t} tr\big(A(\tau)\big)d\tau\right) \quad W\left(\mathbf{x}^{(1)}(t_0), \ldots, \mathbf{x}^{(n)}(t_0)\right). \tag{5.25}$$

Why? This requires work with determinants that is more sophisticated than we want to present here. A reference will be given at the end of the chapter. □

Theorem 5.7 is also known as Liouville's theorem.

Example 5.13

Find e^{tA} for $A = \begin{bmatrix} 0 & 1 \\ -6 & -5 \end{bmatrix}$.

Method: It's easy to find that the eigenvalues of A are $\lambda_1 = -3, \lambda_2 = -2$ and that

$$\mathbf{v}^{(1)} = \begin{bmatrix} 1 \\ -3 \end{bmatrix}, \quad \mathbf{v}^{(2)} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

are corresponding eigenvectors. Theorem 5.4 says that

$$Z(t) \triangleq \left[\, e^{-3t}\begin{bmatrix} 1 \\ -3 \end{bmatrix} \,\middle|\, e^{-2t}\begin{bmatrix} 1 \\ -2 \end{bmatrix} \,\right] = \begin{bmatrix} e^{-3t} & e^{-2t} \\ -3e^{-3t} & -2e^{-2t} \end{bmatrix}$$

is a fundamental matrix for $\dot{\mathbf{x}} = A\mathbf{x}$. Then Theorem 5.6(b) says that

$$e^{tA} = Z(t)\big(Z(0)\big)^{-1} = \begin{bmatrix} e^{-3t} & e^{-2t} \\ -3e^{-3t} & -2e^{-2t} \end{bmatrix}\begin{bmatrix} 1 & 1 \\ -3 & -2 \end{bmatrix}^{-1} = \begin{bmatrix} e^{-3t} & e^{-2t} \\ -3e^{-3t} & -2e^{-2t} \end{bmatrix}\begin{bmatrix} -2 & -1 \\ 3 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} -2e^{-3t} + 3e^{-2t} & -e^{-3t} + e^{-2t} \\ 6e^{-3t} - 6e^{-2t} & 3e^{-3t} - 2e^{-2t} \end{bmatrix}. \bigcirc$$

Lemma 5.1

(Law of exponents) $e^{tA+uA} \triangleq e^{tA}\, e^{uA}$, for any real numbers t, u.

* For example, by using different choices of eigenvectors and a different order of listing the eigenvalues.

Theorem 5.8

(a) $e^{-tA} = (e^{tA})^{-1}$, and (b) the unique solution of the IVP

$$\dot{\mathbf{x}} = A\mathbf{x},\ \mathbf{x}(t_0) = \mathbf{x}_0$$

is

$$\mathbf{x}(t) = e^{(t-t_0)A}\mathbf{x}_0. \tag{5.26}$$

We will apply this theorem in Example 5.14.

5.2.2 Equivalence of Second-Order LCCHODE and LCCHS in $\mathbb{R}^2$

Definition 5.6

A 2×2 real, constant matrix is in **companion form** if it has the form

$$\begin{bmatrix} 0 & 1 \\ * & * \end{bmatrix},$$

where the $*$'s can be any numbers.

Given that a second-order LCCHODE

$$\ddot{y} + p\dot{y} + qy = 0 \tag{5.27}$$

has a solution $y(t)$, let us define

$$x_1(t) \triangleq y(t), \text{ and}$$
$$x_2(t) \triangleq \dot{y}(t).$$

Physically, if $y(t)$ is the position, then $x_2(t)$ is the velocity, $v(t)$. We calculate that

$$\dot{x}_1(t) = \dot{y}(t) = x_2(t)$$

and

$$\dot{x}_2(t) = \ddot{y}(t) = -qy(t) - p\dot{y}(t) = -qx_1(t) - px_2(t).$$

So,

$$\mathbf{x}(t) \triangleq \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

satisfies the LCCHS

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ -q & -p \end{bmatrix} \mathbf{x}, \tag{5.28}$$

which we call an **LCCHS in companion form** in $\mathbb{R}^2$.

On the other hand, in Problem 5.2.5.22, you will explain why $y(t) \triangleq x_1(t)$ satisfies LCCHODE (5.27) if $\mathbf{x}(t)$ satisfies LCCHS (5.28). So, we say that LCCHODE (5.27) and LCCHS (5.28) in companion form in $\mathbb{R}^2$ are *equivalent* in the sense that there is a natural correspondence between their solutions.

Example 5.14

For the IVP $\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ -8 & -6 \end{bmatrix} \mathbf{x},\ \mathbf{x}(t_0) = \mathbf{x}_0,$

(a) Use eigenvalues and eigenvectors to find e^{tA}.

(b) Use the equivalent LCCHODE to find e^{tA}.

(c) Solve the IVP.

Method:

(a) First, solve

$$0 = |\, A - \lambda I \,| = \begin{vmatrix} -\lambda & 1 \\ -8 & -6-\lambda \end{vmatrix} = -\lambda(-6-\lambda) + 8 = \lambda^2 + 6\lambda + 8 = (\lambda+2)(\lambda+4),$$

so the eigenvalues are $\lambda_1 = -4, \lambda_2 = -2$. Corresponding eigenvectors are found by

$$\left[\, A - (-4)I \mid \mathbf{0} \,\right] = \left[\begin{array}{cc|c} 4 & 1 & 0 \\ -8 & -2 & 0 \end{array}\right] \sim \left[\begin{array}{cc|c} 4 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right],$$

after row operation $2R_1 + R_2 \to R_2$, so corresponding to eigenvalue $\lambda_1 = -4$, we have an eigenvector $\mathbf{v}^{(1)} = \begin{bmatrix} 1 \\ -4 \end{bmatrix}$. Similarly,

$$\left[\, A - (-2)I \mid \mathbf{0} \,\right] = \left[\begin{array}{cc|c} 2 & 1 & 0 \\ -8 & -4 & 0 \end{array}\right] \sim \left[\begin{array}{cc|c} 2 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right],$$

after row operation $4R_1 + R_2 \to R_2$, so corresponding to eigenvalue $\lambda_1 = -2$, we have an eigenvector $\mathbf{v}^{(2)} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$. Theorem 5.4 says that

$$Z(t) \triangleq \left[\, e^{-4t}\begin{bmatrix} 1 \\ -4 \end{bmatrix} \,\middle|\, e^{-2t}\begin{bmatrix} 1 \\ -2 \end{bmatrix} \right]$$

is a fundamental matrix for $\dot{\mathbf{x}} = A\mathbf{x}$. Then Theorem 5.6(b) says that

$$e^{tA} = Z(t)\big(Z(0)\big)^{-1} = \begin{bmatrix} e^{-4t} & e^{-2t} \\ -4e^{-4t} & -2e^{-2t} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -4 & -2 \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} e^{-4t} & e^{-2t} \\ -4e^{-4t} & -2e^{-2t} \end{bmatrix} \left(\frac{1}{2} \begin{bmatrix} -2 & -1 \\ 4 & 1 \end{bmatrix} \right) = \begin{bmatrix} -e^{-4t} + 2e^{-2t} & -\frac{1}{2}e^{-4t} + \frac{1}{2}e^{-2t} \\ 4e^{-4t} - 4e^{-2t} & 2e^{-4t} - e^{-2t} \end{bmatrix}.$$

(b) First, write the equivalent scalar second-order ODE, $\ddot{y}+6\dot{y}+8y=0$. Its characteristic polynomial, $s^2+6s+8=(s+4)(s+2)$, has roots $s_1=-4, s_2=-2$. The scalar ODE has general solution

$$y(t)=c_1e^{-4t}+c_2e^{-2t},$$

where c_1, c_2 are arbitrary constants. Correspondingly, the solutions of the original system are

$$\mathbf{x}(t)=\begin{bmatrix}y(t)\\ \dot{y}(t)\end{bmatrix}=\begin{bmatrix}c_1e^{-4t}+c_2e^{-2t}\\ -4c_1e^{-4t}-2c_2e^{-2t}\end{bmatrix}=c_1e^{-4t}\begin{bmatrix}1\\-4\end{bmatrix}+c_2e^{-2t}\begin{bmatrix}1\\-2\end{bmatrix}$$

$$=\begin{bmatrix}e^{-4t} & e^{-2t}\\ -4e^{-4t} & -2e^{-2t}\end{bmatrix}\begin{bmatrix}c_1\\c_2\end{bmatrix},$$

so

$$Z(t)\triangleq\left[\,e^{-4t}\begin{bmatrix}1\\-4\end{bmatrix}\;\vdots\;e^{-2t}\begin{bmatrix}1\\-2\end{bmatrix}\right]$$

is a fundamental matrix for the original 2×2 system. To find e^{tA}, proceed as in part (a):

$$e^{tA}=Z(t)\big(Z(0)\big)^{-1}=\cdots=\begin{bmatrix}-e^{-4t}+2e^{-2t} & -\frac{1}{2}e^{-4t}+\frac{1}{2}e^{-2t}\\ 4e^{-4t}-4e^{-2t} & 2e^{-4t}-e^{-2t}\end{bmatrix}.$$

(c) Note that t_0 and $\mathbf{x}_0$ were not specified. Using Theorem 5.8(b), the solution of the IVP is

$$\mathbf{x}(t)=e^{(t-t_0)A}\mathbf{x}_0=\begin{bmatrix}-e^{-4(t-t_0)}+2e^{-2(t-t_0)} & -\frac{1}{2}e^{-4(t-t_0)}+\frac{1}{2}e^{-2(t-t_0)}\\ 4e^{-4(t-t_0)}-4e^{-2(t-t_0)} & 2e^{-4(t-t_0)}-e^{-2(t-t_0)}\end{bmatrix}\mathbf{x}_0.\ \bigcirc$$

The eigenvalues of an LCCHS in companion form equal the roots of the characteristic equation for the equivalent LCCHODE.

It turns out that the Wronskian for a (possibly time-varying) second-order scalar linear homogeneous ODE and the Wronskian for the corresponding system of ODEs in $\mathbb{R}^2$ are equal! Here's why: if $y(t)$ satisfies a second-order scalar ODE $\ddot{y}+p(t)\dot{y}+q(t)y=0$, define

$$\mathbf{x}(t)\triangleq\begin{bmatrix}y(t)\\ \dot{y}(t)\end{bmatrix}.$$

Then $\ddot{y}=-p(t)\dot{y}-q(t)y$ implies that $\mathbf{x}(t)$ satisfies the system:

$$\dot{\mathbf{x}}(t)=\begin{bmatrix}0 & 1\\ -q(t) & -p(t)\end{bmatrix}\mathbf{x}(t).$$

The Wronskian for two solutions, $y_1(t), y_2(t)$, for the second-order scalar ODE $\ddot{y}+p(t)\dot{y}+q(t)y=0$ is

$$W\left(y_1(t),y_2(t)\right)=\begin{vmatrix}y_1(t) & y_2(t)\\ \dot{y}_1(t) & \dot{y}_2(t)\end{vmatrix}.$$

The Wronskian for two solutions, $\mathbf{x}^{(1)}(t), \mathbf{x}^{(2)}(t)$, for the linear homogeneous system in $\mathbb{R}^2$,

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & 1 \\ -q(t) & -p(t) \end{bmatrix} \mathbf{x}(t),$$

is

$$W\left(\mathbf{x}^{(1)}(t), \mathbf{x}^{(2)}(t)\right) = \left| \mathbf{x}^{(1)}(t) \;\vdots\; \mathbf{x}^{(2)}(t) \right|.$$

But, for the system, solutions $\mathbf{x}_1(t), \mathbf{x}_2(t)$ are of the form

$$\mathbf{x}^{(1)}(t) = \begin{bmatrix} y_1(t) \\ \dot{y}_1(t) \end{bmatrix}, \quad \mathbf{x}^{(2)}(t) = \begin{bmatrix} y_2(t) \\ \dot{y}_2(t) \end{bmatrix},$$

so

$$\begin{aligned} W\left(\mathbf{x}^{(1)}(t), \mathbf{x}^{(2)}(t)\right) = \left| \mathbf{x}^{(1)}(t) \;\vdots\; \mathbf{x}^{(2)}(t) \right| &= \left\| \begin{bmatrix} y_1(t) \\ \dot{y}_1(t) \end{bmatrix} \;\vdots\; \begin{bmatrix} y_2(t) \\ \dot{y}_2(t) \end{bmatrix} \right\| \\ &= \begin{vmatrix} y_1(t) & y_2(t) \\ \dot{y}_1(t) & \dot{y}_2(t) \end{vmatrix} = W\left(y_1(t), y_2(t)\right), \end{aligned}$$

so the two types of Wronskian are equal. This is another aspect of the relationship between the solutions of a linear homogeneous second-order scalar ODE and a linear homogeneous system of two first-order ODEs.

5.2.3 Maclaurin Series for e^{tA}

If A is a constant matrix, we can also define e^{tA} using the Maclaurin series for e^{θ} by replacing θ by tA:

$$e^{tA} \triangleq I + tA + \frac{t^2}{2!} A^2 + \frac{t^3}{3!} A^3 + \cdots.$$

From this, it follows that

$$Ae^{tA} = e^{tA}A. \tag{5.29}$$

It's even possible to use the Maclaurin series to calculate e^{tA}, especially if A is diagonalizable and $A = PDP^{-1}$ where D is a real diagonal matrix:

$$\begin{aligned} e^{tA} &\triangleq I + tPDP^{-1} + \frac{t^2}{2!} PD\cancel{P^{-1}}\cancel{P}DP^{-1} + \frac{t^3}{3!} PD\cancel{P^{-1}}\cancel{P}D\cancel{P^{-1}}\cancel{P}DP^{-1} + \cdots \\ &= P\left(I + tD + \frac{t^2}{2!} D^2 + \frac{t^3}{3!} D^3 + \cdots\right) P^{-1} = Pe^{tD}P^{-1}. \end{aligned}$$

Also, if $D = diag(d_{11}, \ldots, d_{nn})$, then $e^{tD} = diag(e^{d_{11}t}, \ldots, e^{d_{nn}t})$. So, in this special case,

$$e^{tA} = P\, diag(e^{d_{11}t}, \ldots, e^{d_{nn}t})\, P^{-1}.$$

5.2.4 Nonconstant Coefficients

One might ask. "What if A is not constant? Can we use e^{tA} as a fundamental matrix?" Unfortunately, "No," although some numerical methods use it as the first step in an approximation process.

Recall that in Section 3.5, we saw how to solve the Cauchy–Euler ODE $r^2y''(r) + pry'(r) + qy = 0$, where p, q are constants and $' = \dfrac{d}{dr}$: try solutions in the form $y(r) = r^n$.

Example 5.15

For

$$r^2y''(r) - 4ry'(r) + 6y(r) = 0, \tag{5.30}$$

(a) Define $x_1(r) = y(r)$, $x_2(r) = y'(r)$ and convert (5.30) into a system of the form

$$\mathbf{x}'(r) = A(r)\mathbf{x}(r). \tag{5.31}$$

(b) Find a fundamental matrix for system (5.31).

(c) Explain why $e^{rA(r)}$ is not a fundamental matrix for your system (5.31).

Method: We have $x_1'(r) = y'(r) = x_2(r)$, so

$$x_2'(r) = y''(r) = r^{-2}\left(4ry'(r) - 6y(r)\right) = -\frac{6}{r^2}\, y(r) + \frac{4}{r}\, y'(r).$$

So,

$$\mathbf{x}(r) \triangleq \begin{bmatrix} x_1(r) \\ x_2(r) \end{bmatrix}$$

satisfies the system

$$\mathbf{x}'(r) = \begin{bmatrix} 0 & 1 \\ -6r^{-2} & 4r^{-1} \end{bmatrix} \triangleq A(r)\mathbf{x}(r). \tag{5.32}$$

(b) In Example 3.30, in Section 3.5, we saw that the solution of the Cauchy–Euler ODE $r^2y''(r) - 4ry'(r) + 6y(r) = 0$ is $y(r) = c_1r^2 + c_2r^3$; hence,

$$\mathbf{x}(r) = \begin{bmatrix} x_1(r) \\ x_2(r) \end{bmatrix} = \begin{bmatrix} y(r) \\ y'(r) \end{bmatrix} = \begin{bmatrix} c_1r^2 + c_2r^3 \\ 2c_1r + 3c_2r^2 \end{bmatrix} = c_1 \begin{bmatrix} r^2 \\ 2r \end{bmatrix} + c_2 \begin{bmatrix} r^3 \\ 3r^2 \end{bmatrix},$$

where c_1, c_2 are arbitrary constants. Using Lemma 1.3 in Secdtion 1.7, we rewrite this as

$$\mathbf{x}(r) = \begin{bmatrix} r^2 & r^3 \\ 2r & 3r^2 \end{bmatrix} \mathbf{c}, \quad \text{where} \quad \mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}.$$

So,

$$Z(r) = \begin{bmatrix} r^2 & r^3 \\ 2r & 3r^2 \end{bmatrix}$$

is a fundamental matrix for (5.32).

(c) If $e^{rA(r)}$ were a fundamental matrix for (5.32), Theorem 5.5 would require that

$$\frac{d}{dr}\left[\, e^{rA(r)} \,\right] = A(r)e^{rA(r)}.$$

The chain rule and then the product rule imply

$$\frac{d}{dr}\left[\, e^{rA(r)} \,\right] = e^{rA(r)}\frac{d}{dr}\left[\, rA(r) \,\right] = e^{rA(r)}\left(A(r) + rA'(r)\right) \neq A(r)e^{rA(r)}$$

because

$$A'(r) = \begin{bmatrix} 0 & 0 \\ 12r^{-3} & -4r^{-2} \end{bmatrix} \neq O.$$

So $e^{rA(r)}$ is not a fundamental matrix for (5.32), a system with nonconstant coefficients. ◯

So, in general, we should not bother mentioning $e^{tA(t)}$ unless $A(t)$ is actually constant. But see Problem 5.2.5.30 for a *special* circumstance where we can use a matrix exponential to get a fundamental matrix.

5.2.5 Problems

Use exact values wherever possible, that is, do not use decimal approximations of square roots.

In problems 1–4, find the general solution of the LCCHS.

1. $\dot{\mathbf{x}} = \begin{bmatrix} 5 & 4 \\ 4 & -1 \end{bmatrix}\mathbf{x}$
2. $\dot{\mathbf{x}} = \begin{bmatrix} -3 & \sqrt{5} \\ \sqrt{5} & 1 \end{bmatrix}\mathbf{x}$
3. $\dot{\mathbf{x}} = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 3 & 1 \\ 0 & 0 & -1 \end{bmatrix}\mathbf{x}$
4. $\dot{\mathbf{x}} = \begin{bmatrix} -6 & 5 & -5 \\ 0 & -1 & 2 \\ 0 & 7 & 4 \end{bmatrix}\mathbf{x}$

In problems 5 and 6, find the general solution of the LCCHS. Determine the time constant, if all solutions have $\lim_{t\to\infty}\mathbf{x}(t) = 0$.

5. $\dot{\mathbf{x}} = \begin{bmatrix} -3 & \sqrt{2} \\ \sqrt{2} & -2 \end{bmatrix}\mathbf{x}$

6. $A = \begin{bmatrix} -3 & 0 & -1 \\ -1 & -4 & 1 \\ -1 & 0 & -3 \end{bmatrix}$
7. $\dot{\mathbf{x}} = \begin{bmatrix} a & 0 \\ b & c \end{bmatrix} \mathbf{x}$
   Suppose a, b, c are unspecified constants, that is, do not use specific values for them, but do assume that $a \neq c$.
8. Solve the IVP

$$\left\{ \begin{array}{l} \dot{\mathbf{x}} = \begin{bmatrix} 1 & 1 \\ 4 & 1 \end{bmatrix} \mathbf{x} \\ \\ \mathbf{x}(0) = \begin{bmatrix} 0 \\ -2 \end{bmatrix} \end{array} \right\}.$$

9. Find a fundamental matrix for

$$\left\{ \begin{array}{rcrcrcr} \dot{x}_1 & = & 2x_1 & + & x_2 & & \\ \dot{x}_2 & = & & & 3x_2 & + & x_3 \\ \dot{x}_3 & = & & & & - & x_3 \end{array} \right\}.$$

In problems 10–14, find e^{tA}.

10. $A = \begin{bmatrix} -1 & 1 \\ 0 & -2 \end{bmatrix}$
11. $A = \begin{bmatrix} \sqrt{3} & -\sqrt{3} \\ -2\sqrt{3} & -\sqrt{3} \end{bmatrix}$
12. $A = \begin{bmatrix} 1 & \sqrt{5} \\ \sqrt{5} & -3 \end{bmatrix}$
13. $A = \begin{bmatrix} 1 & -1 & 0 \\ 0 & -1 & 3 \\ -1 & 1 & 0 \end{bmatrix}$
14. Suppose that A is a real, 3×3, constant matrix,

$$[A + 2I \mid \mathbf{0}] = \left[\begin{array}{ccc|c} -2 & -2 & 2 & 0 \\ 1 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right]$$

and

$$[A + 3I \mid \mathbf{0}] = \left[\begin{array}{ccc|c} -1 & -2 & 2 & 0 \\ 1 & 2 & -1 & 0 \\ 0 & 0 & 1 & 0 \end{array} \right].$$

Without finding the matrix A, solve the IVP

$$\left\{\begin{array}{c}\dot{\mathbf{x}} = A\mathbf{x} \\ \mathbf{x}(0) = \begin{bmatrix}0\\0\\1\end{bmatrix}\end{array}\right\}.$$

15. Find a fundamental matrix for

$$\dot{\mathbf{x}} = \begin{bmatrix}-a & b\\ b & -a\end{bmatrix}\mathbf{x},$$

where a, b are unspecified positive constants, that is, do not give specific values for them.

16. Suppose $\mathbf{v} = \begin{bmatrix}v_1\\ \vdots \\ v_n\end{bmatrix}$ is a constant vector and λ is a constant. Explain why $\frac{d}{dt}\left[e^{\lambda t}\mathbf{v}\right] = \lambda e^{\lambda t}\mathbf{v}$. [Hint: First multiply through to get

$$e^{\lambda t}\mathbf{v} = \begin{bmatrix}v_1 e^{\lambda t}\\ v_2 e^{\lambda t}\\ \vdots \\ v_n e^{\lambda t}\end{bmatrix}.]$$

17. Let $A = \begin{bmatrix}\blacksquare & 0 & 0\\ \star & 0 & 0\\ \star & \star & \blacksquare\end{bmatrix}$.
    (a) Replace the two ■'s by different positive integers and the three ⋆'s by different negative integers. Write down your A.
    (b) For the matrix A you wrote in part (a), solve $\dot{\mathbf{x}} = A\mathbf{x}$.
18. For the matrix A of Example 4.14 in Section 4.2, find e^{tA}:
    (a) Using the eigenvectors found in Example 4.14 in Section 4.2.
    (b) Using eigenvectors $\begin{bmatrix}1\\-2\\0\end{bmatrix}, \begin{bmatrix}1\\0\\-1\end{bmatrix}, \begin{bmatrix}1\\\frac{1}{2}\\1\end{bmatrix}$.
19. Suppose that in Example 5.13 you had used instead eigenvectors:

$$\mathbf{v}^{(1)} = \begin{bmatrix}-\frac{1}{2}\\ \\ 1\end{bmatrix}, \mathbf{v}^{(2)} = \begin{bmatrix}-\frac{1}{3}\\ \\ 1\end{bmatrix}.$$

    (a) Find a fundamental matrix using those eigenvectors.
    (b) Use your result from part (a) to find e^{tA}. Does it equal what we found in Example 5.13? If it is, why should it be the same?

20. Find a fundamental matrix for

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ -4t^{-2} & -t^{-1} \end{bmatrix} \mathbf{x}.$$

[Hint: The system is equivalent to a Cauchy–Euler ODE for $x_1(t)$, after using the fact that $\dot{x}_1(t) = x_2(t)$ follows from the first ODE in the system.]

21. Find a fundamental matrix for

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ -2t^{-2} & 2t^{-1} \end{bmatrix} \mathbf{x}.$$

[Hint: The system is equivalent to a Cauchy–Euler ODE for $x_1(t)$, after using the fact that $\dot{x}_1(t) = x_2(t)$ follows from the first ODE in the system.]

22. If $\mathbf{x}(t)$ satisfies LCCHS (5.28), explain why $y(t) \triangleq x_1(t)$ satisfies LCCHODE (5.27).
23. Suppose $Z(t)$ is an $n \times n$-valued differentiable function of t and is invertible for all t in an open time interval. If $\dot{Z}(t) = A(t)Z(t)$, explain why $Z(t)$ is a fundamental matrix of $\dot{\mathbf{x}} = A(t)\mathbf{x}$, that is, the columns of $Z(t)$ are all solutions of the same linear homogeneous system.
24. Suppose a system of ODEs $(\star)$ $\dot{\mathbf{x}} = A(t)\mathbf{x}$ has two fundamental matrices $X(t)$ and $Y(t)$. Explain why there is a constant matrix B such that $Y(t) = X(t)B$. [Hint: Use initial conditions $X(0)$ and $Y(0)$ to discover what the matrix B should be.]
25. Suppose a system of ODEs $(\star)$ $\dot{\mathbf{x}} = A(t)\mathbf{x}$ has a fundamental matrix $X(t)$ and B is an invertible constant matrix and define $Y(t) = X(t)B$. Must $Y(t)$ be a fundamental matrix for $(\star)$? Why, or why not? If the former, explain; if the latter, give a specific counterexample, that is, a specific choice of $A(t), X(t)$, and B for which $X(t)$ is a fundamental matrix but $X(t)B$ isn't.
26. Must $e^{\gamma t}e^{tA}$ be a fundamental matrix for $\dot{\mathbf{x}} = (\gamma I + A)\mathbf{x}$?
27. Suppose $X(t)$ is a fundamental matrix for a system of ODEs $(\star)$ $\dot{\mathbf{x}} = A(t)\mathbf{x}$ and $X(0) = I$. Suppose also that $A(-t) \equiv -A(t)$, that is, $A(t)$, is an *odd* function. Explain why $X(t)$ is an *even* function, that is, satisfies $X(-t) \equiv X(t)$. [Hint: Define $Y(t) \triangleq X(-t)$ and use uniqueness of solutions of linear systems of ODEs.]
28. Suppose $X(t)$ is a fundamental matrix for a system $(\star)$ $\dot{\mathbf{x}} = A(t)\mathbf{x}$. Explain why $Y(t) \triangleq \left(X(t)^T\right)^{-1}$ is a fundamental matrix for the system $(\star\star)$ $\dot{\mathbf{x}} = -A(t)^T\mathbf{x}$, by using the steps in the following:
    (a) Explain why $\frac{d}{dt}[X(t)^T] = (\dot{X}(t))^T$.
    (b) Use the product rule for matrices to calculate the time derivatives of both sides of $I = X(t)^T \left(X(t)^T\right)^{-1}$.
    (c) Explain why $Y(t)$ satisfies $\dot{Y} = -A(t)^T Y(t)$. [By the way, system $(\star\star)$ is called the **adjoint system** for system $(\star)$.]
29. For $A = \begin{bmatrix} -3 & 1 \\ 1 & -3 \end{bmatrix}$, calculate the improper integral

$$\int_0^\infty e^{tA^T} e^{tA}\, dt.$$

30. Suppose $A(t)$ is given and we define $B(t) \triangleq \int_0^t A(s)ds$. Suppose $A(t)B(t) \triangleq B(t)A(t)$. Explain why $e^{B(t)}$ is a fundamental matrix for the system $\dot{\mathbf{x}} = A(t)\mathbf{x}$. [Hint: First, find $\dot{B}(t)$ using the chain rule for matrix exponentials $\frac{d}{dt}\left[e^{B(t)}\right] = e^{B(t)}\frac{d}{dt}\left[\, B(t)\,\right]$.] [By the way, if $A(t)$ is constant, then $B(t) = tA$.]
31. Suppose $X(t)$ is a fundamental matrix for a system of ODEs $(\star)\ \dot{\mathbf{x}} = A(t)\mathbf{x}$ and $X(0) = I$. Suppose also that $A(t)^T \equiv -A(t)$. Explain why $X(t)^T = (X(t))^{-1}$.
[Hint: Define $Y(t) = (X(t))^{-1}$ and find the ODE that $Y(t)$ satisfies. How? Begin by noting that $I = X(t)\,(X(t))^{-1} = X(t)\,Y(t)$ and differentiate both sides with respect to t using the product rule.]
32. Solve the *homogeneous* system that corresponds to the model of iodine metabolism found in Example 5.6 in Section 5.1.
33. Find the generalization to (a) $\mathbb{R}^3$ and (b) $\mathbb{R}^n$ for the concept of companion form given in Definition 5.6.
    By the way, the MATLAB® command roots finds the roots of an n-th degree polynomial by rewriting it as the characteristic polynomial of an $n \times n$ matrix in companion form, and then MATLAB exploits its excellent methods for finding the eigenvalues of that matrix.
34. Suppose $A^T = -A$ is a real, $n \times n$ matrix. Explain why e^{itA} is a Hermitian matrix. [Hint: The matrix exponential can also be defined by the infinite series $e^B = I + B + \frac{1}{2!}B^2 + \frac{1}{3!}B^3 + \cdots$.]
35. If A is a real, symmetric $n \times n$ matrix, must e^{tA} be real and symmetric? If so, why? If not, give a specific counterexample.
36. Solve the system of Problem 5.1.3.7 and discuss the long-term behavior of the solutions.

5.3 Complex or Deficient Eigenvalues

5.3.1 Complex Eigenvalues

Recall that for a second-order LCCHODE $\ddot{y} + p\dot{y} + qy = 0$, if the characteristic polynomial has a complex conjugate pair of roots $s = \alpha \pm i\nu$, where α, ν are real and $\nu > 0$, then

$$\{\, e^{\alpha t}\cos \nu t, \quad e^{\alpha t}\sin \nu t\,\} = \{\, \mathcal{R}e(e^{(\alpha+i\nu)t}), \quad \mathcal{I}m(e^{(\alpha+i\nu)t})\,\}$$

gives a complete set of basic solutions for the ODE. Similar to that result is the following:

Theorem 5.9

Suppose the characteristic polynomial of a real $n \times n$ matrix A has a complex conjugate pair of roots $\lambda = \alpha \pm i\nu$, where α, ν are real and $\nu > 0$, and corresponding eigenvectors

are $\mathbf{v}, \overline{\mathbf{v}}$. Then LCCHS (5.13) in Section 5.2, that is, $\dot{\mathbf{x}} = A\mathbf{x}$, has a pair of solutions given by

$$\mathbf{x}^{(1)}(t) \triangleq \mathcal{R}e(e^{(\alpha+i\nu)t}\mathbf{v}), \quad \mathbf{x}^{(2)}(t) \triangleq \mathcal{I}m(e^{(\alpha+i\nu)t})\mathbf{v}).$$

In addition, if A is 2×2, then $\{\mathbf{x}^{(1)}(t), \mathbf{x}^{(2)}(t)\}$ is a complete set of basic solutions of the LCCHS in $\mathbb{R}^2$.

As in Section 2.1, for a complex conjugate pair of eigenvalues, we don't need the eigenvector $\overline{\mathbf{v}}$!

Caution: Usually $\mathcal{R}e(e^{(\alpha+i\nu)t}\mathbf{v}) \neq \mathcal{R}e(e^{(\alpha+i\nu)t})\mathcal{R}e(\mathbf{v})$.

Example 5.16

Find e^{tA} for the LCCHS $\dot{\mathbf{x}} = \begin{bmatrix} -1 & 2 \\ -2 & -1 \end{bmatrix}\mathbf{x}$.

Method: First, solve $0 = |A - \lambda I| = \begin{vmatrix} -1-\lambda & 2 \\ -2 & -1-\lambda \end{vmatrix} = (-1-\lambda)^2 + 4$, so the eigenvalues are $\lambda = -1 \pm i2$. Corresponding to eigenvalue $\lambda_1 = -1 + i2$, eigenvectors are found by

$$\left[A - (-1+i2)I \mid \mathbf{0}\right] = \begin{bmatrix} -i2 & 2 & \mid 0 \\ -2 & -i2 & \mid 0 \end{bmatrix} \sim \begin{bmatrix} \textcircled{1} & i & \mid 0 \\ 0 & 0 & \mid 0 \end{bmatrix},$$

after row operations $\frac{i}{2}R_1 \to R_1, 2R_1 + R_2 \to R_2$. Corresponding to eigenvalue $\lambda_1 = -1 + i2$, we have an eigenvector $\mathbf{v}^{(1)} = \begin{bmatrix} -i \\ 1 \end{bmatrix}$. This gives two solutions of the LCCHS: the first is

$$\mathbf{x}^{(1)}(t) = \mathcal{R}e\left(e^{(-1+i2)t}\begin{bmatrix} -i \\ 1 \end{bmatrix}\right) = \mathcal{R}e\left(e^{-t}(\cos 2t + i\sin 2t)\begin{bmatrix} -i \\ 1 \end{bmatrix}\right)$$

$$= \mathcal{R}e\left(e^{-t}\begin{bmatrix} \sin 2t - i\cos 2t \\ \cos 2t + i\sin 2t \end{bmatrix}\right) = e^{-t}\begin{bmatrix} \sin 2t \\ \cos 2t \end{bmatrix}.$$

For the second, we don't have to do all of the algebra steps again:

$$\mathbf{x}^{(2)}(t) = \mathcal{I}m\left(e^{(-1+i2)t}\begin{bmatrix} -i \\ 1 \end{bmatrix}\right) = \mathcal{I}m\left(e^{-t}\begin{bmatrix} \sin 2t - i\cos 2t \\ \cos 2t + i\sin 2t \end{bmatrix}\right) = e^{-t}\begin{bmatrix} -\cos 2t \\ \sin 2t \end{bmatrix}.$$

So, a fundamental matrix is given by

$$Z(t) = \left[\mathbf{x}^{(1)}(t) \;\vdots\; \mathbf{x}^{(2)}(t)\right] = e^{-t}\left[\begin{bmatrix} \sin 2t \\ \cos 2t \end{bmatrix} \quad \begin{bmatrix} -\cos 2t \\ \sin 2t \end{bmatrix}\right] = e^{-t}\begin{bmatrix} \sin 2t & -\cos 2t \\ \cos 2t & \sin 2t \end{bmatrix}.$$

That gives us

$$e^{tA} = Z(t)\big(Z(0)\big)^{-1} = \left(e^{-t}\begin{bmatrix} \sin 2t & -\cos 2t \\ \cos 2t & \sin 2t \end{bmatrix}\right)\left(\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}^{-1}\right)$$

$$= e^{-t}\begin{bmatrix} \sin 2t & -\cos 2t \\ \cos 2t & \sin 2t \end{bmatrix}\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = e^{-t}\begin{bmatrix} \cos 2t & \sin 2t \\ -\sin 2t & \cos 2t \end{bmatrix}. \bigcirc$$

Example 5.17

(Short-cut if A is in companion form) Find e^{tA} for the LCCHS

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ -10 & -2 \end{bmatrix} \mathbf{x}.$$

Method: First, write the equivalent scalar second-order ODE, $\ddot{y} + 2\dot{y} + 10y = 0$. Its characteristic polynomial, $s^2 + 2s + 10 = (s+1)^2 + 9$, has roots $s = -1 \pm i3$. The scalar ODE has general solution

$$y(t) = c_1 e^{-t} \cos 3t + c_2 e^{-t} \sin 3t,$$

where c_1, c_2 are arbitrary constants. Correspondingly, the solutions of the original system are, after using the product rule,

$$\begin{aligned} \mathbf{x}(t) = \begin{bmatrix} y(t) \\ \dot{y}(t) \end{bmatrix} &= \begin{bmatrix} c_1 e^{-t} \cos 3t + c_2 e^{-t} \sin 3t \\ c_1 e^{-t}(-\cos 3t - 3\sin 3t) + c_2 e^{-t}(-\sin 3t + 3\cos 3t) \end{bmatrix} \\ &= c_1 e^{-t} \begin{bmatrix} \cos 3t \\ -\cos 3t - 3\sin 3t \end{bmatrix} + c_2 e^{-t} \begin{bmatrix} \sin 3t \\ -\sin 3t + 3\cos 3t \end{bmatrix} \\ &= e^{-t} \begin{bmatrix} \cos 3t & \sin 3t \\ -\cos 3t - 3\sin 3t & 3\cos 3t - \sin 3t \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}, \\ &\triangleq Z(t) \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}. \end{aligned}$$

This implicitly defines $Z(t)$, a fundamental matrix for the original 2×2 system. So,

$$\begin{aligned} e^{tA} = Z(t)\big(Z(0)\big)^{-1} &= \left(e^{-t} \begin{bmatrix} \cos 3t & \sin 3t \\ -\cos 3t - 3\sin 3t & 3\cos 3t - \sin 3t \end{bmatrix} \right) \left(\begin{bmatrix} 1 & 0 \\ -1 & 3 \end{bmatrix}^{-1} \right) \\ &= e^{-t} \begin{bmatrix} \cos 3t & \sin 3t \\ -\cos 3t - 3\sin 3t & 3\cos 3t - \sin 3t \end{bmatrix} \left(\frac{1}{3} \begin{bmatrix} 3 & 0 \\ 1 & 1 \end{bmatrix} \right) \\ &= \frac{1}{3} e^{-t} \begin{bmatrix} 3\cos 3t + \sin 3t & \sin 3t \\ -10\sin 3t & 3\cos 3t - \sin 3t \end{bmatrix}. \bigcirc \end{aligned}$$

Example 5.18

Find the general solution of the LCCHS for the circuit in Example 5.2 in Section 5.1, assuming $V(t) \equiv 0$, $L = 1$, $R = \frac{8}{3}$, $C_1 = \frac{1}{8}$, and $C_2 = \frac{3}{8}$.

Method: With these parameter values, the LCCHS is $\dot{\mathbf{x}} = \begin{bmatrix} 0 & -1 & 0 \\ 8 & -3 & 3 \\ 0 & 1 & -1 \end{bmatrix} \mathbf{x}$. First, solve

$$\begin{aligned} 0 = |A - \lambda I| &= \begin{vmatrix} -\lambda & -1 & 0 \\ 8 & -3-\lambda & 3 \\ 0 & 1 & -1-\lambda \end{vmatrix} = -\lambda \begin{vmatrix} -3-\lambda & 3 \\ 1 & -1-\lambda \end{vmatrix} + \begin{vmatrix} 8 & 3 \\ 0 & -1-\lambda \end{vmatrix} \\ &= \cdots = -(\lambda^3 + 4\lambda^2 + 8\lambda + 8). \end{aligned}$$

Standard advice says to *try* $\lambda = \pm\dfrac{\text{factors of 4}}{\text{factors of 1}}$, that is, $\lambda = \pm 1, \pm 2, \pm 4$. We are lucky in this example, as $\lambda = -2$ is a root of the characteristic polynomial. We factor to get

$$0 = |\, A - \lambda I \,| = -(\lambda + 2)(\lambda^2 + 2\lambda + 4).$$

The eigenvalues are $\lambda_1 = -2$ and the complex conjugate pair $\lambda = -1 \pm i\sqrt{3}$. Corresponding to eigenvalue λ_1, eigenvectors are found by

$$\left[\, A - (-2)I \mid \mathbf{0} \,\right] = \begin{bmatrix} 2 & -1 & 0 & \mid 0 \\ 8 & -1 & 3 & \mid 0 \\ 0 & 1 & 1 & \mid 0 \end{bmatrix} \sim \begin{bmatrix} 2 & 0 & 1 & \mid 0 \\ 0 & 1 & 1 & \mid 0 \\ 0 & 0 & 0 & \mid 0 \end{bmatrix},$$

after row operations $-4R_1 + R_2 \to R_2$, $\frac{1}{3}R_2 \to R_2$, $-R_2 + R_3 \to R_3$, $R_2 + R_1 \to R_1$, so corresponding to eigenvalue $\lambda_1 = -2$, we have an eigenvector $\mathbf{v}^{(1)} = \begin{bmatrix} -1 \\ -2 \\ 2 \end{bmatrix}$.

Corresponding to eigenvalue $\lambda = -1 + i\sqrt{3}$, eigenvectors are found by

$$\left[\, A - \left(-1 + i\sqrt{3}\right)I \mid \mathbf{0} \,\right] = \begin{bmatrix} 1 - i\sqrt{3} & -1 & 0 & \mid 0 \\ 8 & -2 - i\sqrt{3} & 3 & \mid 0 \\ 0 & 1 & -i\sqrt{3} & \mid 0 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & \frac{3 - i\sqrt{3}}{4} & \mid 0 \\ 0 & 1 & -i\sqrt{3} & \mid 0 \\ 0 & 0 & 0 & \mid 0 \end{bmatrix},$$

after row operations $R_1 \leftrightarrow R_2$, $\frac{1}{8}R_1 \to R_1$, $-(1 - i\sqrt{3})R_1 + R_2 \to R_2$, $R_2 \leftrightarrow R_3$, $\dfrac{3 + i\sqrt{3}}{8}R_2 + R_3 \to R_3$, $\dfrac{2 + i\sqrt{3}}{8}R_2 + R_1 \to R_1$. Corresponding to eigenvalue $\lambda_1 = -1 + i\sqrt{3}$, we have an eigenvector $\mathbf{v}^{(1)} = \begin{bmatrix} -3 + i\sqrt{3} \\ i4\sqrt{3} \\ 4 \end{bmatrix}$.

This gives two solutions of the LCCHS: The first is

$$\begin{aligned}
\mathbf{x}^{(1)}(t) &= \mathcal{R}e\left(e^{(-1+i\sqrt{3})t} \begin{bmatrix} -3 + i\sqrt{3} \\ i4\sqrt{3} \\ 4 \end{bmatrix} \right) \\
&= e^{-t}\mathcal{R}e\left(\cos(\sqrt{3}\,t) + i\sin(\sqrt{3}\,t) \begin{bmatrix} -3 + i\sqrt{3} \\ i4\sqrt{3} \\ 4 \end{bmatrix} \right) \\
&= e^{-t}\,\mathcal{R}e\left(\begin{bmatrix} -3\cos(\sqrt{3}\,t) - \sqrt{3}\sin(\sqrt{3}\,t) + i\left(\sqrt{3}\cos(\sqrt{3}\,t) - 3\sin(\sqrt{3}\,t)\right) \\ -4\sqrt{3}\sin(\sqrt{3}\,t) + i4\sqrt{3}\cos(\sqrt{3}\,t) \\ 4\cos(\sqrt{3}\,t) + i\,4\sin(\sqrt{3}\,t) \end{bmatrix} \right) \\
&= e^{-t} \begin{bmatrix} -3\cos(\sqrt{3}\,t) - \sqrt{3}\sin(\sqrt{3}\,t) \\ -4\sqrt{3}\sin(\sqrt{3}\,t) \\ 4\cos(\sqrt{3}\,t) \end{bmatrix}.
\end{aligned}$$

For the second, we don't have to do all of the algebra steps again:

$$\mathbf{x}^{(2)}(t) = \mathcal{I}m\left(e^{(-1+i\sqrt{3})t}\begin{bmatrix}-3+i\sqrt{3}\\ i4\sqrt{3}\\ 4\end{bmatrix}\right) = e^{-t}\begin{bmatrix}\sqrt{3}\cos(\sqrt{3}\,t) - 3\sin(\sqrt{3}\,t)\\ 4\sqrt{3}\cos(\sqrt{3}\,t)\\ 4\sin(\sqrt{3}\,t)\end{bmatrix}.$$

The general solution of the circuit is

$$\begin{bmatrix}I_1(t)\\ v_1(t)\\ v_2(t)\end{bmatrix}$$
$$= c_1e^{-2t}\begin{bmatrix}-1\\ -2\\ 2\end{bmatrix} + c_2e^{-t}\begin{bmatrix}-3\cos(\sqrt{3}\,t)-\sqrt{3}\sin(\sqrt{3}\,t)\\ -4\sqrt{3}\sin(\sqrt{3}\,t)\\ 4\cos(\sqrt{3}\,t)\end{bmatrix}$$
$$+ c_3e^{-t}\begin{bmatrix}\sqrt{3}\cos(\sqrt{3}\,t)-3\sin(\sqrt{3}\,t)\\ 4\sqrt{3}\cos(\sqrt{3}\,t)\\ 4\sin(\sqrt{3}\,t)\end{bmatrix},$$

where c_1, c_2, c_3 are arbitrary constants. Finally, (5.7) in Section 5.1 yields $I_2(t) = \frac{1}{R}(v_1(t) - v_2(t))$. ◯

5.3.2 Solving Homogeneous Systems of Second-Order Equations

We saw in Example 5.5 in Section 5.1 that a physical system of three horizontal springs and two masses can be modeled by a system of second-order scalar ODEs.

Just as for a single second-order scalar ODE, we can rewrite a system of m second-order scalar ODEs as a system of $2m$ first-order scalar ODEs. As we will see, for certain systems of second-order scalar ODEs, it is usually simpler to *not* rewrite them as first-order ODEs.

To solve a system of two second-order ODEs of the special form

$$\ddot{\mathbf{x}} = A\mathbf{x},$$

where A is a real matrix, it helps to try solutions in the form

$$\mathbf{x}(t) = e^{\sigma t}\mathbf{v}.$$

When we substitute that into the system, we get

$$\sigma^2 e^{\sigma t}\mathbf{v} = \ddot{\mathbf{x}} = A\mathbf{x} = e^{\sigma t}A\mathbf{v},$$

that is,

$$A\mathbf{x} = \sigma^2\mathbf{v}.$$

So, we want $\mathbf{v}$ to be an eigenvector of A corresponding to eigenvalue $\lambda \triangleq \sigma^2$. Note that σ is *not* necessarily an eigenvalue of A. *In the following, we will assume that* $\mathbf{v}$ *is a* real *eigenvector of* A.

If A has an eigenvalue $\lambda < 0$, then setting $\sigma^2 = \lambda < 0$ would give $\sigma = \pm i\sqrt{-\lambda} \triangleq \pm i\nu$, and thus, the original system would have two solutions:

$$\mathbf{x}^{(1)}(t) = \mathcal{R}e\left(e^{i\nu t}\mathbf{v}\right) = \cos(\nu t)\mathbf{v},$$
$$\mathbf{x}^{(2)}(t) = \mathcal{I}m\left(e^{i\nu t}\mathbf{v}\right) = \sin(\nu t)\mathbf{v}.$$

For example, if $\mathbf{x}$ is in $\mathbb{R}^2$ and A has two distinct *negative* eigenvalues λ_1, λ_2 and corresponding eigenvectors $\mathbf{v}_1, \mathbf{v}_2$, denote $\nu_1 = \sqrt{-\lambda_1}$, $\nu_2 = \sqrt{-\lambda_2}$. Then the system $\ddot{\mathbf{x}} = A\mathbf{x}$ has general solution

$$\mathbf{x} = \left(c_1 \cos \nu_1 t + d_1 \sin \nu_1 t\right)\mathbf{v}_1 + \left(c_2 \cos \nu_2 t + d_2 \sin \nu_2 t\right)\mathbf{v}_2, \tag{5.33}$$

where c_1, c_2, d_1, d_2 are arbitrary constants.

Example 5.19

Find the general solution of the system of second-order ODEs (5.12) in Section 5.1 for the parameter values $m_1 = 1, m_2 = 2, k_1 = 5, k_2 = 6, k_3 = 8$.

Method: With these parameter values, (5.12) in Section 5.1 is $\ddot{\mathbf{x}} = A\mathbf{x}$, where $A = \begin{bmatrix} -11 & 6 \\ 3 & -7 \end{bmatrix}$. First, find the eigenvalues $\lambda = \sigma^2$ of A by solving

$$0 = |A - \lambda I| = \begin{vmatrix} -11 - \lambda & 6 \\ 3 & -7 - \lambda \end{vmatrix} = (-11 - \lambda)(-7 - \lambda) - 18 = \lambda^2 + 18\lambda + 59 :$$

$$\lambda = \frac{-18 \pm \sqrt{18^2 - 4(1)(59)}}{2} = \frac{-18 \pm \sqrt{88}}{2} = -9 \pm \sqrt{22}.$$

Denote $\lambda_1 = -9 - \sqrt{22} = \sigma_1^2$, $\lambda_2 = -9 + \sqrt{22} = \sigma_2^2$. The two frequencies of vibration are

$$\nu_1 = \sqrt{-\lambda_1} = \sqrt{9 + \sqrt{22}}, \quad \nu_2 = \sqrt{-\lambda_2} = \sqrt{9 - \sqrt{22}}.$$

Next, find $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}$, eigenvectors of A corresponding to the eigenvalues λ_1, λ_2 of A, respectively:

$$\left[A - (-9 - \sqrt{22})I \mid \mathbf{0}\right] = \begin{bmatrix} -2 + \sqrt{22} & 6 & \mid 0 \\ 3 & 2 + \sqrt{22} & \mid 0 \end{bmatrix} \sim \begin{bmatrix} 3 & 2 + \sqrt{22} & \mid 0 \\ 0 & 0 & \mid 0 \end{bmatrix},$$

after row operations $R_1 \leftrightarrow R_2, -\left(\frac{-2+\sqrt{22}}{3}\right)R_1 + R_2 \to R_2$. Corresponding to eigenvalue $\lambda_1 = -9 - \sqrt{22}$, we have an eigenvector $\mathbf{v}^{(1)} = \begin{bmatrix} -2 - \sqrt{22} \\ 3 \end{bmatrix}$.

Second,

$$\left[A - (-9 + \sqrt{22})I \mid \mathbf{0}\right] = \begin{bmatrix} -2 - \sqrt{22} & 6 & \mid 0 \\ 3 & 2 - \sqrt{22} & \mid 0 \end{bmatrix} \sim \begin{bmatrix} 3 & 2 - \sqrt{22} & \mid 0 \\ 0 & 0 & \mid 0 \end{bmatrix},$$

after row operations $R_1 \leftrightarrow R_2, -\left(\frac{-2-\sqrt{22}}{3}\right)R_1 + R_2 \to R_2$. Corresponding to eigenvalue $\lambda_1 = -9 + \sqrt{22}$, we have an eigenvector $\mathbf{v}^{(2)} = \begin{bmatrix} -2 + \sqrt{22} \\ 3 \end{bmatrix}$.

Using formula (5.33), we have that the solutions of the two mass and three horizontal spring systems are given by

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \left(c_1 \cos\left(\sqrt{9+\sqrt{22}}\, t\right) + d_1 \sin\left(\sqrt{9+\sqrt{22}}\, t\right)\right) \begin{bmatrix} -2-\sqrt{22} \\ 3 \end{bmatrix} + \left(c_2 \cos\left(\sqrt{9-\sqrt{22}}\, t\right) + d_2 \sin\left(\sqrt{9-\sqrt{22}}\, t\right)\right) \begin{bmatrix} -2+\sqrt{22} \\ 3 \end{bmatrix}$$

where c_1, c_2, d_1, d_2 are arbitrary constants. ○

The ratio of the two frequencies of vibration is

$$\frac{\sqrt{9+\sqrt{22}}}{\sqrt{9-\sqrt{22}}} = \frac{\sqrt{9+\sqrt{22}}}{\sqrt{9-\sqrt{22}}}\frac{\sqrt{9-\sqrt{22}}}{\sqrt{9-\sqrt{22}}} = \frac{\sqrt{(9+\sqrt{22})(9-\sqrt{22})}}{9-\sqrt{22}} = \frac{\sqrt{9^2-(\sqrt{22})^2}}{9-\sqrt{22}}$$

$$= \frac{\sqrt{59}}{9-\sqrt{22}} = \frac{\sqrt{59}}{(9-\sqrt{22})}\frac{(9+\sqrt{22})}{(9+\sqrt{22})} = \frac{(9+\sqrt{22})}{\sqrt{59}},$$

hence is not a rational number, so the motion of the positions of the two masses is quasi-periodic and *not* periodic, except in the special case when the initial conditions are satisfied by either $c_1 = d_1 = 0$ or $c_2 = d_2 = 0$.

From an engineering point of view, the quasiperiodic case is more likely to happen than the periodic case because it is unusual for the ratio of two randomly chosen real numbers to be a rational number.

5.3.3 Deficient Eigenvalues

Recall from Example 2.16 in Section 2.2 that

$$A = \begin{bmatrix} 29 & 18 \\ -50 & -31 \end{bmatrix}$$

has only one *distinct* eigenvalue, $\lambda = -1$, and it is *deficient* because its algebraic multiplicity is two, but its geometric multiplicity is one, that is, there is only one linearly independent eigenvector.

As for complex eigenvalues, it helps to first consider an easier example of a system in companion form.

Example 5.20

Find the general solution of the LCCHS $\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ -9 & -6 \end{bmatrix} \mathbf{x}$.

Method: First, write the equivalent scalar second-order ODE, $\ddot{y} + 6\dot{y} + 9y = 0$ and solve its characteristic equation, $0 = s^2 + 6s + 9 = (s+3)^2$: $s = -3, -3$. The scalar ODE has general solution

$$y(t) = c_1 e^{-3t} + c_2 t e^{-3t},$$

where c_1, c_2 are arbitrary constants. Correspondingly, the solutions of the original system are

$$\mathbf{x}(t) = \begin{bmatrix} y(t) \\ \dot{y}(t) \end{bmatrix} = \begin{bmatrix} c_1 e^{-3t} + c_2 t e^{-3t} \\ c_1(-3e^{-3t}) + c_2(-3t+1)e^{-3t} \end{bmatrix}.$$

The general solution of the system of ODEs is

$$\mathbf{x}(t) = c_1 e^{-3t} \begin{bmatrix} 1 \\ -3 \end{bmatrix} + c_2 e^{-3t} \begin{bmatrix} t \\ -3t+1 \end{bmatrix},$$

where c_1, c_2 are arbitrary constants, because the Wronskian is

$$\begin{vmatrix} e^{-3t} & te^{-3t} \\ -3e^{-3t} & (-3t+1)e^{-3t} \end{vmatrix} = e^{-6t} \neq 0. \quad \bigcirc$$

If we study the conclusion of this example, we see that one solution is

$$\mathbf{x}^{(1)}(t) = e^{-3t} \begin{bmatrix} 1 \\ -3 \end{bmatrix}$$

and the second solution is

$$\mathbf{x}^{(2)}(t) = e^{-3t} \left(t \begin{bmatrix} 1 \\ -3 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right). \tag{5.34}$$

We note that $\begin{bmatrix} 1 \\ -3 \end{bmatrix}$ is an eigenvector of the matrix $A = \begin{bmatrix} 0 & 1 \\ -9 & -6 \end{bmatrix}$.

Example 5.21

Find the general solution of the LCCHS $\dot{\mathbf{x}} = \begin{bmatrix} 29 & 18 \\ -50 & -31 \end{bmatrix} \mathbf{x}$ and the corresponding e^{tA}.

Method: Denote $A = \begin{bmatrix} 29 & 18 \\ -50 & -31 \end{bmatrix}$. In Example 2.16 in Section 2.2, we found that $\lambda = -1$ is the only eigenvalue of A, with corresponding eigenvector

$$\mathbf{v}^{(1)} = \begin{bmatrix} -0.6 \\ 1 \end{bmatrix};$$

hence, $A\mathbf{v}^{(1)} = (-1)\mathbf{v}^{(1)}$. So,

$$\mathbf{x}^{(1)}(t) = e^{-t} \begin{bmatrix} -0.6 \\ 1 \end{bmatrix}$$

gives one solution of the system. Similar to (5.34) in the previous example, let's try a second solution of the form

$$\mathbf{x}^{(2)}(t) = e^{-t} \left(t\mathbf{v}^{(1)} + \mathbf{w} \right).$$

Substitute it into the LCCHS: by the product rule, we need

$$-e^{-t} \left(t\mathbf{v}^{(1)} + \mathbf{w} \right) + e^{-t} \left(\mathbf{v}^{(1)} + \mathbf{0} \right) = \dot{\mathbf{x}}^{(2)}(t) = A\mathbf{x}^{(2)}(t) = e^{-t} A \left(t\mathbf{v}^{(1)} + \mathbf{w} \right).$$

After multiplying through by e^t, this becomes

$$\cancel{-t\mathbf{v}^{(1)}} + \mathbf{v}^{(1)} - \mathbf{w} = \cancel{tA\mathbf{v}^{(1)}} + A\mathbf{w},$$

where we canceled terms because $A\mathbf{v}^{(1)} = (-1)\mathbf{v}^{(1)}$. So we need

$$(A - (-1)I)\,\mathbf{w} = \mathbf{v}^{(1)}. \tag{5.35}$$

Such a vector $\mathbf{w}$ is called a **generalized eigenvector** of A corresponding to the eigenvalue $\lambda = -1$. We can solve for $\mathbf{w}$ using row reduction of an augmented matrix:

$$\left[A - (-1)I \mid \mathbf{v}^{(1)}\right] = \left[\begin{array}{cc|c} 30 & 18 & -0.6 \\ -50 & -30 & 1 \end{array}\right] \sim \left[\begin{array}{cc|c} \textcircled{1} & 0.6 & -0.02 \\ 0 & 0 & 0 \end{array}\right]$$

after $\frac{5}{3}R_1 + R_2 \to R_2$, $\frac{1}{30}R_1 \to R_1$. The solutions are

$$\mathbf{w} = \begin{bmatrix} -0.02 \\ 0 \end{bmatrix} + c\begin{bmatrix} -0.6 \\ 1 \end{bmatrix},$$

where c is an arbitrary constant. For convenience, we can take $c = 0$, as we shall see later, so our second solution of the LCCHS is

$$\mathbf{x}^{(2)}(t) = e^{-t}\left(t\mathbf{v}^{(1)} + \mathbf{w}\right) = e^{-t}\left(t\begin{bmatrix} -0.6 \\ 1 \end{bmatrix} + \begin{bmatrix} -0.02 \\ 0 \end{bmatrix}\right).$$

We check that this gives us a complete set of basic solutions by calculating the Wronskian:

$$\left|\mathbf{x}^{(1)}(t) \;\vdots\; \mathbf{x}^{(2)}(t)\right| = \begin{vmatrix} -0.6e^{-t} & e^{-t}(-0.6t - 0.02) \\ e^{-t} & te^{-t} \end{vmatrix} = 0.02e^{-2t} \neq 0.$$

The general solution is

$$\mathbf{x}(t) = c_1 e^{-t}\begin{bmatrix} -0.6 \\ 1 \end{bmatrix} + c_2 e^{-t}\left(t\begin{bmatrix} -0.6 \\ 1 \end{bmatrix} + \begin{bmatrix} -0.02 \\ 0 \end{bmatrix}\right),$$

where c_1, c_2 are arbitrary constants.

To find e^{tA}, first rewrite the general solution as

$$\mathbf{x}(t) = c_1 e^{-t}\begin{bmatrix} -0.6 \\ 1 \end{bmatrix} + c_2 e^{-t}\begin{bmatrix} -0.6t - 0.02 \\ t \end{bmatrix} = e^{-t}\begin{bmatrix} -0.6 & -0.6t - 0.02 \\ 1 & t \end{bmatrix}\begin{bmatrix} c_1 \\ c_2 \end{bmatrix},$$

so

$$X(t) \triangleq e^{-t}\begin{bmatrix} -0.6 & -0.6t - 0.02 \\ 1 & t \end{bmatrix}$$

is a fundamental matrix. We calculate that

$$e^{tA} = X(t)\,(X(0))^{-1} = e^{-t}\begin{bmatrix} -0.6 & -0.6t - 0.02 \\ 1 & t \end{bmatrix}\begin{bmatrix} -0.6 & -0.02 \\ 1 & 0 \end{bmatrix}^{-1}$$

$$= e^{-t}\begin{bmatrix} -0.6 & -0.6t - 0.02 \\ 1 & t \end{bmatrix}\begin{bmatrix} 0 & 1 \\ -50 & -30 \end{bmatrix} = e^{-t}\begin{bmatrix} 30t + 1 & 18t \\ -50t & -30t + 1 \end{bmatrix}. \bigcirc$$

The reason we *could* take $c = 0$ in finding $\mathbf{x}^{(2)}(t)$ is because it succeeded in finding a complete set of basic solutions. The reasons why we *would want to* take $c = 0$ are because usually "simpler is better" and also because if we had kept the c in $\mathbf{x}^{(2)}(t)$, then it would include the redundant term $c\mathbf{x}^{(1)}(t)$.

5.3.4 Laplace Transforms and e^{tA}

If A is a constant matrix, the unique solution of the IVP $\dot{\mathbf{x}} = A\mathbf{x}, \mathbf{x}(0) = \mathbf{x}_0$ is

$$\mathbf{x}(t) = e^{tA}\mathbf{x}_0.$$

On the other hand, if we take the Laplace transform of the LCCHS $\dot{\mathbf{x}} - A\mathbf{x} = \mathbf{0}$, we get

$$s\mathcal{L}[\,\mathbf{x}(t)\,] - \mathbf{x}_0 - A\mathbf{x} = \mathbf{0}.$$

So, $e^{tA}\mathbf{x}_0 = \mathcal{L}^{-1}[\,(sI - A)^{-1}\,]\,\mathbf{x}_0$. It follows that

Theorem 5.10

$\mathcal{L}^{-1}[\,(sI - A)^{-1}\,] = e^{tA}$.

Example 5.22

Find the general solution of the LCCHS of Example 5.21.

Method: We have

$$e^{tA} = \mathcal{L}^{-1}\,[(sI - A)^{-1}\,] = \mathcal{L}^{-1}\left[\left(sI - \begin{bmatrix} 29 & 18 \\ -50 & -31 \end{bmatrix}\right)^{-1}\right]$$

$$= \mathcal{L}^{-1}\left[\begin{bmatrix} s-29 & -18 \\ 50 & s+31 \end{bmatrix}^{-1}\right] = \mathcal{L}^{-1}\left[\frac{1}{(s-29)(s+31)+900}\begin{bmatrix} s+31 & 18 \\ -50 & s-29 \end{bmatrix}\right]$$

$$= \mathcal{L}^{-1}\left[\begin{bmatrix} \dfrac{s+31}{(s^2+2s+1)} & \dfrac{18}{(s^2+2s+1)} \\ -\dfrac{50}{(s^2+2s+1)} & \dfrac{s-29}{(s^2+2s+1)} \end{bmatrix}\right] = \mathcal{L}^{-1}\left[\begin{bmatrix} \dfrac{(s+1)+30}{(s+1)^2} & \dfrac{18}{(s+1)^2} \\ -\dfrac{50}{(s+1)^2} & \dfrac{(s+1)-30}{(s+1)^2} \end{bmatrix}\right]$$

$$= \mathcal{L}^{-1}\left[\begin{bmatrix} \dfrac{1}{(s+1)} + \dfrac{30}{(s+1)^2} & \dfrac{18}{(s+1)^2} \\ -\dfrac{50}{(s+1)^2} & \dfrac{1}{(s+1)} - \dfrac{30}{(s+1)^2} \end{bmatrix}\right]$$

$$= \begin{bmatrix} e^{-t} + 30te^{-t} & 18te^{-t} \\ -50te^{-t} & e^{-t} - 30te^{-t} \end{bmatrix}.$$

The general solution of the LCCHS is

$$\mathbf{x}(t) = \begin{bmatrix} e^{-t} + 30te^{-t} & 18te^{-t} \\ -50te^{-t} & e^{-t} - 30te^{-t} \end{bmatrix} \mathbf{x}_0.$$

This agrees with the second conclusion of Example 5.21. ◯

5.3.5 Stability

Definition 5.7

LCCHS (5.13) in Section 5.2, that is, $\dot{\mathbf{x}} = A\mathbf{x}$, is

(a) **Asymptotically stable** if all its solutions have $\lim_{t\to\infty} \mathbf{x}(t) = \mathbf{0}$
(b) **Neutrally stable** if it is not asymptotically stable, but all its solutions are bounded on $[0, \infty)$, that is, for each component $x_j(t)$ of $\mathbf{x}(t) = [x_1(t) \quad x_2(t) \quad \cdots \quad x_n(t)]^T$, there exists M_j such that for all $t \geq 0$ we have $|x_j(t)| \leq M_j$
(c) **Unstable** if it is neither asymptotically stable nor neutrally stable, that is, there is at least one solution $\mathbf{x}(t)$ that is not bounded on $[0, \infty)$

We have,

Theorem 5.11

LCCHS (5.13) in Section 5.2, that is, $\dot{\mathbf{x}} = A\mathbf{x}$, is

(a) Asymptotically stable if all of A's eigenvalues λ satisfy $\mathcal{R}e(\lambda) < 0$
(b) Neutrally stable if all of A's eigenvalues λ satisfy $\mathcal{R}e(\lambda) \leq 0$ and no deficient eigenvalue λ has real part equal to zero
(c) Unstable if A has an eigenvalue whose real part is positive or if it has a deficient eigenvalue whose real part is 0

Why? (a) Suppose λ is a real, negative eigenvalue of A with corresponding eigenvector $\mathbf{v}$. Then $\mathbf{x}(t) = e^{\lambda t}\mathbf{v}$ will be a solution of the LCCHS and will have $\lim_{t\to\infty} \mathbf{x}(t) = \mathbf{0}$. If $\lambda = \alpha \pm i\nu$ is a nonreal eigenvalue of A with negative real part α and corresponding eigenvector $\mathbf{v}$, then solutions

$$e^{\alpha t}\mathcal{R}e\left(e^{i\nu t}\mathbf{v}\right), \quad e^{\alpha t}\mathcal{I}m\left(e^{i\nu t}\mathbf{v}\right)$$

will have limit $\mathbf{0}$ as $t \to \infty$ because $\alpha = \mathcal{R}e(\lambda) < 0$. The explanation for (b) is similar by again using the form of solutions in the two cases λ real versus non-real. The explanation for (c) is similar, although requires some care in the deficient eigenvalue case. □

For an ODE, the time constant indicates how long it takes for a solution to decay to $1/e$ of its initial value. Suppose an LCCHS is asymptotically stable. The time constant τ for that system of ODEs can be *defined* by

$$\tau = \frac{1}{r_{min}},$$

where r_{min} is the slowest decay rate. Because each solution $\mathbf{x}(t)$ may include many different decaying exponential functions, "weighted" by constant vectors, we *can't* guarantee that $\mathbf{x}(\tau) = \frac{1}{e}\mathbf{x}(0)$. Nevertheless, for physical intuition, it is still useful to think of the time constant as being about how long it takes for the solution to decay in a standard way.

5.3.6 Problems

In your final conclusions, the symbol "i" should not appear. If you are asked for a fundamental matrix, do explain why it is invertible.

In problems 1–6, find the general solution of the LCCHS.

1. $\dot{\mathbf{x}} = \begin{bmatrix} -2 & -5 \\ 1 & 0 \end{bmatrix} \mathbf{x}$
2. $\dot{\mathbf{x}} = \begin{bmatrix} 0 & 2 \\ -3 & -2 \end{bmatrix} \mathbf{x}$
3. $\dot{\mathbf{x}} = \begin{bmatrix} 3 & 2 \\ -4 & -1 \end{bmatrix} \mathbf{x}$
4. $\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ -\frac{9}{4} & -1 \end{bmatrix} \mathbf{x}$
5. $\dot{\mathbf{x}} = \begin{bmatrix} 5 & 0 & -10 \\ 0 & -2 & 0 \\ 4 & 0 & -7 \end{bmatrix} \mathbf{x}$
6. $\dot{\mathbf{x}} = \begin{bmatrix} -12 & -25 \\ 4 & 8 \end{bmatrix} \mathbf{x}$

In problems 7–13, find a fundamental matrix for the LCCHS.

7. $\dot{\mathbf{x}} = \begin{bmatrix} 1 & -5 \\ 1 & -3 \end{bmatrix} \mathbf{x}$
8. $\dot{\mathbf{x}} = \begin{bmatrix} -3 & 2 \\ -5 & 3 \end{bmatrix} \mathbf{x}$
9. The system of Problem 5.3.6.6
10. $A = \begin{bmatrix} \sqrt{3} & \sqrt{3} \\ -2\sqrt{3} & -\sqrt{3} \end{bmatrix}$
11. $\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix} \mathbf{x}$
12. $\dot{\mathbf{x}} = \begin{bmatrix} -4 & -1 \\ 9 & 2 \end{bmatrix} \mathbf{x}$

13. $\dot{\mathbf{x}} = \begin{bmatrix} a & b \\ -b & a \end{bmatrix} \mathbf{x}$, where a, b are unspecified constants, that is, do not give specific values for them. Do assume that $b \neq 0$.
14. $\dot{\mathbf{x}} = \begin{bmatrix} -a & b \\ 0 & -a \end{bmatrix} \mathbf{x}$, where a, b are unspecified constants, that is, do not give specific values for them. Do assume that $b \neq 0$.
15. Find a fundamental matrix and e^{tA} for $\dot{\mathbf{x}} = \begin{bmatrix} -2 & -3 \\ 2 & -4 \end{bmatrix} \mathbf{x}$.

In problems 16–18, find e^{tA}.

16. $A = \begin{bmatrix} 1 & -4 \\ 2 & -3 \end{bmatrix}$
17. $A = \begin{bmatrix} 3 & 0 & -2 \\ 0 & -1 & 0 \\ 4 & 0 & 3 \end{bmatrix}$
18. $A = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & -2 \\ 2 & 2 & 1 \end{bmatrix}$

In problems 19 and 20, find e^{tA} (a) using eigenvalues and eigenvectors, and (b) using Laplace transforms.

19. $A = \begin{bmatrix} 10 & 11 \\ -11 & -12 \end{bmatrix}$
20. $A = \begin{bmatrix} 3 & 3 & -1 \\ 0 & -1 & 0 \\ 4 & 4 & -1 \end{bmatrix}$
21. You may assume that the matrix

$$A = \begin{bmatrix} -1 & -1 & 0 \\ 2 & -1 & 1 \\ 0 & 1 & -1 \end{bmatrix}$$

has eigenvalues $-1, -1 \pm i$. Find e^{tA}.

22. For the system

$$\dot{\mathbf{x}} = \begin{bmatrix} -1 & 4 \\ -1 & 1 \end{bmatrix} \mathbf{x},$$

(a) Find a fundamental matrix.

(b) Find the solution that passes through the point $(x_1(0), x_2(0)) = (2, -3)$.

23. Solve the IVP

$$\left\{ \begin{array}{c} \dot{\mathbf{x}} = \begin{bmatrix} -1 & 2 \\ -2 & -1 \end{bmatrix} \mathbf{x} \\ \\ \mathbf{x}(0) = \begin{bmatrix} \pi \\ 2 \end{bmatrix} \end{array} \right\}.$$

24. Solve the IVP

$$\left\{ \begin{array}{c} \begin{Bmatrix} \dot{x}_1 & = & -2x_1 & + & x_2 \\ \dot{x}_2 & = & -2x_1 & - & 4x_3 \end{Bmatrix} \\ x_1(0) = 1,\ x_2(0) = -2 \end{array} \right\}.$$

25. Solve the IVP

$$\left\{ \begin{array}{c} \begin{Bmatrix} \dot{y} & = & v \\ \dot{v} & = & -5y - 2v \end{Bmatrix} \\ y(0) = 1,\ v(0) = 0 \end{array} \right\}.$$

26. Find the exact frequencies of vibration for

$$\ddot{\mathbf{x}} = \begin{bmatrix} -4 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 2 & -3 \end{bmatrix} \mathbf{x}.$$

27. Find the general solution of

$$\ddot{\mathbf{x}} = \begin{bmatrix} -3 & 1 \\ 2 & -2 \end{bmatrix} \mathbf{x}.$$

28. For the matrix $A = \begin{bmatrix} 0 & 2 & -1 \\ -3 & -5 & 2 \\ -2 & -2 & 0 \end{bmatrix}$,
    (a) Find an eigenvector and a generalized eigenvector corresponding to eigenvalue $\lambda = -2$.
    (b) Use your results for part (a) to help find e^{tA}.

In problems 29–35, determine if the system $\dot{\mathbf{x}} = A\mathbf{x}$ is asymptotically stable, neutrally stable, or unstable. Try to do as little work as is necessary to give a fully explained conclusion.

29. $A = \begin{bmatrix} -1 & 1 \\ 0 & -2 \end{bmatrix}$
30. $A = \begin{bmatrix} \sqrt{2} & \sqrt{2} \\ -3\sqrt{2} & -\sqrt{2} \end{bmatrix}$
31. The LCCHS of Problem 5.3.6.6

32. $A = \begin{bmatrix} -1 & 0 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & 0 \end{bmatrix}$

33. $A = \begin{bmatrix} -3 & 0 & -1 \\ -1 & -4 & 1 \\ -1 & 0 & -3 \end{bmatrix}$

34. $A = \begin{bmatrix} -5 & 2 & 4 \\ 2 & -8 & 2 \\ 4 & 2 & -5 \end{bmatrix}$

35. Assume A is a constant, real, 5×5, matrix and has eigenvalues $i, -i, i, -i, -1$, including repetitions. Consider the LCCHS $(\star)$ $\dot{\mathbf{x}} = A\mathbf{x}$. For each of (a) through (e), decide whether it must be true, must be false, or may be true and may be false:
    (a) The system is asymptotically stable.
    (b) The system is neutrally stable.
    (c) The system may be neutrally stable, depending upon more information concerning A.
    (d) $(\star)$ has solutions that are periodic with period 2π.
    (e) $(\star)$ has solutions of the form $t\mathbf{p}(t) + \mathbf{q}(t)$ where $\mathbf{p}(t)$ is periodic with period 2π.

36. If the matrix A has an eigenvalue λ with $\mathcal{R}e(\lambda) = 0$ that *is* deficient, explain why LCCHS $\dot{\mathbf{x}} = A\mathbf{x}$ is not neutrally stable.

37. (Small project) Analogous to the Cauchy–Euler ODE, consider systems of the form $\dot{\mathbf{x}} = t^{-1}A\mathbf{x}$, where A is a real, constant matrix. Create a method to find all solutions of such systems, using solutions of the form $\mathbf{x}(t) = t^r\mathbf{v}$ where $\mathbf{v}$ is constant. Do consider at least these three cases of roots: real, complex, and real but deficient.

5.4 Nonhomogeneous Linear Systems

Here we will explain how to solve

$$\dot{\mathbf{x}} = A(t)\mathbf{x}(t) + \mathbf{f}(t) \tag{5.36}$$

using a fundamental matrix $X(t)$ for the corresponding homogeneous linear system $\dot{\mathbf{x}} = A(t)\mathbf{x}$. Recall that $X(t)$ satisfies

$$\dot{X}(t) = A(t)X(t), \tag{5.37}$$

that is, each column of $X(t)$ is a solution of $\dot{\mathbf{x}} = A(t)\mathbf{x}$. In the special case of a linear constant coefficients system $\dot{\mathbf{x}} = A\mathbf{x} + \mathbf{f}(t)$, we will especially use the fundamental matrix e^{tA}.

It turns out that the method developed in the following is a generalization of the method of variation of parameters that we used in Section 4.3.

We try a solution of (5.36) in the form

$$\mathbf{x}(t) \triangleq X(t)\mathbf{v}(t). \tag{5.38}$$

Using a generalization of the product rule, we have

$$\frac{d}{dt}\left[X(t)\mathbf{v}(t)\right] = \dot{X}(t)\mathbf{v}(t) + X(t)\dot{\mathbf{v}}(t).$$

So for $\mathbf{x}(t) = X(t)\mathbf{v}(t)$, (5.37) implies that

$$\dot{\mathbf{x}}(t) = \dot{X}(t)\mathbf{v}(t) + X(t)\dot{\mathbf{v}}(t) = A(t)X(t)\mathbf{v}(t) + X(t)\dot{\mathbf{v}}(t).$$

So, for $\mathbf{x}(t)$ to solve the original, nonhomogeneous system (5.36), we need

$$\cancel{A(t)X(t)\mathbf{v}(t)} + X(t)\dot{\mathbf{v}}(t) = \dot{\mathbf{x}}(t) = A(t)\mathbf{x}(t) + \mathbf{f}(t) = \cancel{A(t)X(t)\mathbf{v}(t)} + \mathbf{f}(t),$$

that is,

$$X(t)\dot{\mathbf{v}}(t) = \mathbf{f}(t).$$

But, one requirement of a fundamental matrix is that it should be invertible at all t, or at least all t in an interval of existence. The earlier equation is equivalent to

$$\dot{\mathbf{v}}(t) = (X(t))^{-1}\,\mathbf{f}(t). \tag{5.39}$$

Using indefinite integration we have

$$\mathbf{v}(t) = \int (X(t))^{-1}\,\mathbf{f}(t)\,dt,$$

or, using definite integration we have

$$\mathbf{v}(t) = \int_{t_0}^{t} (X(t))^{-1}\,\mathbf{f}(\tau)\,d\tau,$$

where t_0 is a constant; either gives a formula for a particular solution, $\mathbf{x}_p(t) = X(t)\mathbf{v}(t)$, for the original, nonhomogeneous system (5.36).

To get the general solution of (5.36), we add in $\mathbf{x}_h(t) = X(t)\mathbf{c}$, where $\mathbf{c}$ is a vector of arbitrary constants. The general solution of (5.36) can be written in either of the forms

$$\mathbf{x}(t) = X(t)\left(\mathbf{c} + \int (X(t))^{-1}\,\mathbf{f}(t)\,dt\right) \tag{5.40}$$

or, using definite integration,

$$\mathbf{x}(t) = X(t)\left(\mathbf{c} + \int_{t_0}^{t} (X(\tau))^{-1}\,\mathbf{f}(\tau)\,d\tau\right), \tag{5.41}$$

where t_0 is a constant.

In the *special* but often occurring case when A is a constant matrix, we can use $X(t) = e^{tA}$: Recalling that $\left(e^{sA}\right)^{-1} = e^{-sA}$, we can rewrite these two formulas as

$$\mathbf{x}(t) = e^{tA}\left(\mathbf{c} + \int e^{-tA}\mathbf{f}(t)\,dt\right) \tag{5.42}$$

and, using the law of exponents $e^{tA}e^{-\tau A} = e^{(t+(-\tau))A}$,

$$\mathbf{x}(t) = e^{tA}\mathbf{c} + \int_{t_0}^{t} e^{(t-\tau)A}\mathbf{f}(\tau)\,d\tau, \tag{5.43}$$

respectively. Any one of (5.40) through (5.43) is called a variation of parameters, or variation of constants, formula for the solutions of the nonhomogeneous system (5.36).

Evaluate (5.43) at $t = t_0$ to get

$$\mathbf{x}(t_0) = e^{t_0A}\mathbf{c} + \int_{t_0}^{t_0} e^{(t_0-\tau)A}\mathbf{f}(\tau)d\tau = e^{t_0A}\mathbf{c};$$

hence,

$$\mathbf{c} = \left(e^{t_0A}\right)^{-1}\mathbf{x}(t_0) = e^{-t_0A}\mathbf{x}(t_0).$$

The solution of an IVP can be written in the form

$$\mathbf{x}(t) = e^{tA}e^{-t_0A}\mathbf{x}(t_0) + \int_{t_0}^{t} e^{(t-\tau)A}\mathbf{f}(\tau)\,d\tau = e^{(t-t_0)A}\mathbf{x}(t_0) + \int_{t_0}^{t} e^{(t-\tau)A}\mathbf{f}(\tau)\,d\tau. \tag{5.44}$$

Example 5.23

Solve the IVP

$$\left\{\begin{array}{l} \left\{\begin{array}{rcrcrl} \dot{x}_1 &=& 2x_1 &+& x_2 & \\ \dot{x}_2 &=& 7x_1 &-& 4x_2 & -e^{-t} \end{array}\right\} \\ \\ \left\{\begin{array}{rcl} x_1(0) &=& 3 \\ x_2(0) &=& -2 \end{array}\right\}. \end{array}\right\}. \tag{5.45}$$

Method: First, we find e^{tA} using the eigenvalues and eigenvectors of $A = \begin{bmatrix} 2 & 1 \\ 7 & -4 \end{bmatrix}$: The characteristic polynomial,

$$\mathcal{P}(\lambda) \triangleq |\, A - \lambda I \,| = \begin{vmatrix} 2-\lambda & 1 \\ 7 & -4-\lambda \end{vmatrix} = \lambda^2 + 2\lambda - 15 = (\lambda+5)(\lambda-3)$$

has roots $\lambda_1 = -5, \lambda_2 = 3$. We find corresponding eigenvectors:

$$\left[A - \lambda_1 I \mid \mathbf{0}\right] = \left[\begin{array}{cc|c} 7 & 1 & 0 \\ 7 & 1 & 0 \end{array}\right] \quad \text{yields} \quad \mathbf{v}^{(1)} = \begin{bmatrix} 1 \\ -7 \end{bmatrix}$$

and

$$\left[A-\lambda_2 I \mid \mathbf{0}\right]=\begin{bmatrix}-1 & 1 & \mid 0\\ 7 & -7 & \mid 0\end{bmatrix} \quad \text{yields} \quad \mathbf{v}^{(2)}=\begin{bmatrix}1\\1\end{bmatrix}.$$

So,

$$Z(t)=\begin{bmatrix}e^{-5t} & e^{3t}\\ -7e^{-5t} & e^{3t}\end{bmatrix}.$$

We use the fundamental matrix to calculate

$$e^{tA}=Z(t)\big(Z(0)\big)^{-1}=\begin{bmatrix}e^{-5t} & e^{3t}\\ -7e^{-5t} & e^{3t}\end{bmatrix}\begin{bmatrix}1 & 1\\ -7 & 1\end{bmatrix}^{-1}=\begin{bmatrix}e^{-5t} & e^{3t}\\ -7e^{-5t} & e^{3t}\end{bmatrix}\left(\frac{1}{8}\begin{bmatrix}1 & -1\\ 7 & 1\end{bmatrix}\right)$$

$$=\frac{1}{8}\begin{bmatrix}e^{-5t}+7e^{3t} & -e^{-5t}+e^{3t}\\ -7e^{-5t}+7e^{3t} & 7e^{-5t}+e^{3t}\end{bmatrix}.$$

Applying (5.44) with $t_0=0$, the solution of the IVP is

$$\mathbf{x}(t)=e^{tA}\mathbf{x}(0)+\int_0^t e^{(t-\tau)A}\mathbf{f}(\tau)d\tau=\frac{1}{8}\begin{bmatrix}e^{-5t}+7e^{3t} & -e^{-5t}+e^{3t}\\ -7e^{-5t}+7e^{3t} & 7e^{-5t}+e^{3t}\end{bmatrix}\begin{bmatrix}3\\-2\end{bmatrix}+$$

$$+\int_0^t \frac{1}{8}\begin{bmatrix}e^{-5(t-\tau)}+7e^{3(t-\tau)} & -e^{-5(t-\tau)}+e^{3(t-\tau)}\\ -7e^{-5(t-\tau)}+7e^{3(t-\tau)} & 7e^{-5(t-\tau)}+e^{3(t-\tau)}\end{bmatrix}\begin{bmatrix}0\\-e^{-\tau}\end{bmatrix}d\tau$$

$$=\frac{1}{8}\begin{bmatrix}5e^{-5t}+19e^{3t}\\ -35e^{-5t}+19e^{3t}\end{bmatrix}+\int_0^t \frac{1}{8}\begin{bmatrix}e^{-5t}e^{4\tau}-e^{3t}e^{-4\tau}\\ -7e^{-5t}e^{4\tau}-e^{3t}e^{-4\tau}\end{bmatrix}d\tau.$$

When integrating with respect to τ, functions of t are treated as if they were constants. The solution of the IVP is

$$\mathbf{x}(t)=\frac{1}{8}\begin{bmatrix}5e^{-5t}+19e^{3t}\\ -35e^{-5t}+19e^{3t}\end{bmatrix}+\frac{1}{8}\begin{bmatrix}e^{-5t}\left[\frac{1}{4}e^{4\tau}\right]_0^t-e^{3t}\left[-\frac{1}{4}e^{-4\tau}\right]_0^t\\ -7e^{-5t}\left[\frac{1}{4}e^{4\tau}\right]_0^t-e^{3t}\left[-\frac{1}{4}e^{-4\tau}\right]_0^t\end{bmatrix}$$

$$=\frac{1}{8}\begin{bmatrix}5e^{-5t}+19e^{3t}\\ -35e^{-5t}+19e^{3t}\end{bmatrix}+\frac{1}{32}\begin{bmatrix}e^{-5t}\left(e^{4t}-1\right)+e^{3t}\left(e^{-4t}-1\right)\\ -7e^{-5t}\left(e^{4t}-1\right)+e^{3t}\left(e^{-4t}-1\right)\end{bmatrix}$$

$$=\frac{1}{8}\begin{bmatrix}5e^{-5t}+19e^{3t}\\ -35e^{-5t}+19e^{3t}\end{bmatrix}+\frac{1}{32}\begin{bmatrix}e^{-t}+e^{-t}\\ -7e^{-t}+e^{-t}\end{bmatrix}-\frac{1}{32}\begin{bmatrix}e^{-5t}+e^{3t}\\ -7e^{-5t}+e^{3t}\end{bmatrix}.$$

The solution of the IVP is

$$\mathbf{x}(t)=\frac{1}{32}\begin{bmatrix}19e^{-5t}+75e^{3t}\\ -133e^{-5t}+75e^{3t}\end{bmatrix}+\frac{1}{16}\begin{bmatrix}e^{-t}\\ -3e^{-t}\end{bmatrix}.\ \bigcirc$$

Example 5.24

Solve the IVP

$$\left\{\begin{array}{l} \dot{\mathbf{x}} = \begin{Bmatrix} \dot{x}_1 &=& 29x_1 &+& 18x_2 &+t \\ \dot{x}_2 &=& -50x_1 &-& 31x_2 & \end{Bmatrix} \\ \\ \qquad\qquad \begin{Bmatrix} x_1(0) &=& 3 \\ x_2(0) &=& -2 \end{Bmatrix}. \end{array}\right\}. \tag{5.46}$$

Method: In Example 5.21 in Section 5.3, we found a complete set of basic solutions $\{\mathbf{x}^{(1)}(t), \mathbf{x}^{(2)}(t)\}$. Using them, a fundamental matrix for the corresponding linear LCCHS is given by

$$Z(t) = \left[\, \mathbf{x}^{(1)}(t) \;\vdots\; \mathbf{x}^{(2)}(t) \,\right] = \begin{bmatrix} -0.6e^{-t} & e^{-t}(-0.6t - 0.02) \\ e^{-t} & te^{-t} \end{bmatrix}.$$

So, we calculate

$$\begin{aligned} e^{tA} = Z(t)\big(Z(0)\big)^{-1} &= \begin{bmatrix} -0.6e^{-t} & e^{-t}(-0.6t - 0.02) \\ e^{-t} & te^{-t} \end{bmatrix} \begin{bmatrix} -0.6 & -0.02 \\ 1 & 0 \end{bmatrix}^{-1} \\ &= \begin{bmatrix} -0.6e^{-t} & e^{-t}(-0.6t - 0.02) \\ e^{-t} & te^{-t} \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -50 & -30 \end{bmatrix} \\ &= \begin{bmatrix} (30t+1)e^{-t} & 18te^{-t} \\ -50te^{-t} & (1-30t)e^{-t} \end{bmatrix}. \end{aligned}$$

Applying (5.44) with $t_0 = 0$, the solution of the IVP is

$$\begin{aligned} \mathbf{x}(t) &= \begin{bmatrix} (30t+1)e^{-t} & 18te^{-t} \\ -50te^{-t} & (1-30t)e^{-t} \end{bmatrix} \begin{bmatrix} 3 \\ -2 \end{bmatrix} + \\ &\quad + \int_0^t \begin{bmatrix} (30(t-\tau)+1)e^{-(t-\tau)} & 18te^{-(t-\tau)} \\ -50(t-\tau)e^{-(t-\tau)} & (1-30(t-\tau))\,e^{-(t-\tau)} \end{bmatrix} \begin{bmatrix} \tau \\ 0 \end{bmatrix} d\tau \\ &= \begin{bmatrix} (54t+3)e^{-t} \\ -(90t+2)e^{-t} \end{bmatrix} + e^{-t} \int_0^t \begin{bmatrix} \left((30t+1)\tau - 30\tau^2\right)e^{\tau} \\ \left(-50t\tau + 50\tau^2\right)e^{\tau} \end{bmatrix} d\tau. \end{aligned}$$

We calculate on the side that

$$\begin{aligned} \int_0^t \begin{bmatrix} \left((30t+1)\tau - 30\tau^2\right)e^{\tau} \\ \left(-50t\tau + 50\tau^2\right)e^{\tau} \end{bmatrix} &= \begin{bmatrix} (30t+1)\int_0^t \tau e^{\tau} d\tau - 30\int_0^t \tau^2 e^{\tau} d\tau \\ \\ -50t\int_0^t \tau e^{\tau} d\tau + 50\int_0^t \tau^2 e^{\tau} d\tau \end{bmatrix} \\ &= \begin{bmatrix} (30t+1)\left[(\tau-1)e^{\tau}\right]_0^t - 30\left[(\tau^2 - 2\tau + 2)e^{\tau}\right]_0^t \\ \\ -50t\left[(\tau-1)e^{\tau}\right]_0^t + 50\left[(\tau^2 - 2\tau + 2)e^{\tau}\right]_0^t \end{bmatrix} \\ &= \begin{bmatrix} (30t+1)\left((t-1)e^t + 1\right) - 30\left((t^2 - 2t + 2)e^t - 2\right) \\ \\ -50t\left((t-1)e^t + 1\right) + 50\left((t^2 - 2t + 2)e^t - 2\right) \end{bmatrix}. \end{aligned}$$

Returning to the full expression for the solution, we have

$$\mathbf{x}(t) = \begin{bmatrix} (54t+3)e^{-t} \\ -(90t+2)e^{-t} \end{bmatrix} + e^{-t}\begin{bmatrix} (31t-61)e^{t}+30t+61 \\ (-50t+100)e^{t}-50t-100 \end{bmatrix},$$

that is,

$$\mathbf{x}(t) = \begin{bmatrix} 31t-61 \\ -50t+100 \end{bmatrix} + e^{-t}\begin{bmatrix} 84t+64 \\ -140t-102 \end{bmatrix}. \bigcirc$$

Example 5.25

Suppose that the system $\dot{\mathbf{x}} = \begin{bmatrix} -t^{-1} & t^{-1} \\ -2-3t^{-1} & 1+3t^{-1} \end{bmatrix}\mathbf{x}$ has solutions

$$\mathbf{x}^{(1)}(t) = \begin{bmatrix} t+1 \\ 2t+1 \end{bmatrix}, \quad \mathbf{x}^{(2)}(t) = \begin{bmatrix} e^{t} \\ (t+1)e^{t} \end{bmatrix}.$$

Solve ODE system

$$\dot{\mathbf{x}} = \begin{bmatrix} -t^{-1} & t^{-1} \\ -2-3t^{-1} & 1+3t^{-1} \end{bmatrix}\mathbf{x} + \begin{bmatrix} 1 \\ 2 \end{bmatrix} \tag{5.47}$$

on all open intervals not containing $t = 0$.

Method: We were given two solutions of the corresponding linear homogeneous system, so we hope that

$$Z(t) = \left[\mathbf{x}^{(1)}(t) \;\vdots\; \mathbf{x}^{(2)}(t)\right] = \begin{bmatrix} t+1 & e^{t} \\ 2t+1 & (t+1)e^{t} \end{bmatrix}$$

is a fundamental matrix. To affirm this, all we need to do is check its invertibility, a calculation that will be useful when we need the inverse later:

$$|Z(t)| = \begin{vmatrix} t+1 & e^{t} \\ 2t+1 & (t+1)e^{t} \end{vmatrix} = (t+1)(t+1)e^{t} - (2t+1)e^{t} = t^{2}e^{t},$$

which is never zero on any open interval not containing $t = 0$.

In this problem, the matrix $A(t)$ is not constant, so $e^{tA(t)}$ is not a fundamental matrix.

We try to use (5.40) to find the solution of the nonhomogeneous system (5.47):

$$\begin{aligned}
\mathbf{x}(t) &= Z(t)\left(\mathbf{c} + \int (Z(t))^{-1}\,\mathbf{f}(t)dt\right) \\
&= \begin{bmatrix} t+1 & e^{t} \\ 2t+1 & (t+1)e^{t} \end{bmatrix}\left(\mathbf{c} + \int \frac{1}{t^{2}e^{t}}\begin{bmatrix} (t+1)e^{t} & -e^{t} \\ -(2t+1) & t+1 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \end{bmatrix}dt\right) \\
&= \begin{bmatrix} t+1 & e^{t} \\ 2t+1 & (t+1)e^{t} \end{bmatrix}\left(\mathbf{c} + \int \frac{1}{t^{2}e^{t}}\begin{bmatrix} (t-1)e^{t} \\ 1 \end{bmatrix}dt\right) \\
&= \begin{bmatrix} t+1 & e^{t} \\ 2t+1 & (t+1)e^{t} \end{bmatrix}\left(\mathbf{c} + \begin{bmatrix} \int \frac{(t-1)}{t^{2}}dt \\ \int t^{-2}e^{-t}dt \end{bmatrix}\right).
\end{aligned}$$

Unfortunately, the second integral has no "closed form" solution, unless one uses the Maclaurin (infinite) series for e^{-t}. Instead, we can use (5.41), a definite integral version of

the variation of parameters formula: For any $t_0 \neq 0$, the solution of the nonhomogeneous system (5.47) is, after using the earlier work,

$$\mathbf{x}(t) = Z(t)\left(\mathbf{c} + \int_{t_0}^{t} (Z(\tau))^{-1}\,\mathbf{f}(\tau)d\tau\right)$$

$$= \cdots = \begin{bmatrix} t+1 & e^t \\ 2t+1 & (t+1)e^t \end{bmatrix}\left(\mathbf{c} + \begin{bmatrix} \int_{t_0}^{t} \frac{(\tau-1)}{\tau^2}d\tau \\ \int_{t_0}^{t} \tau^{-2}e^{-\tau}d\tau \end{bmatrix}\right)$$

$$= \begin{bmatrix} t+1 & e^t \\ 2t+1 & (t+1)e^t \end{bmatrix}\left(\mathbf{c} + \begin{bmatrix} \ln\left|\frac{t}{t_0}\right| + \frac{1}{t} - \frac{1}{t_0} \\ \int_{t_0}^{t} \tau^{-2}e^{-\tau}d\tau \end{bmatrix}\right),$$

where $\mathbf{c}$ is a vector of arbitrary constants. ◯

5.4.1 Problems

In problems 1–3, a fundamental matrix is given for $\dot{\mathbf{x}} = A\mathbf{x}$ for some real, 2×2, constant matrix A. Without finding A, solve the given nonhomogeneous ODE system:

1. $X(t) = \begin{bmatrix} e^{-2t} & e^{-3t} \\ -2e^{-2t} & -3e^{-3t} \end{bmatrix}$, $\dot{\mathbf{x}} = A\mathbf{x} + \begin{bmatrix} 1 \\ 0 \end{bmatrix}$
2. $X(t) = \begin{bmatrix} \cos 3t & \sin 3t \\ -3\sin 3t & 3\cos 3t \end{bmatrix}$, $\dot{\mathbf{x}} = A\mathbf{x} + \begin{bmatrix} \cos t \\ 0 \end{bmatrix}$
3. $X(t) = e^{-3t}\begin{bmatrix} \cos t - \sin t & \cos t + \sin t \\ 2\cos t & 2\sin t \end{bmatrix}$, $\dot{\mathbf{x}} = A\mathbf{x} + \begin{bmatrix} 0 \\ e^{-3t} \end{bmatrix}$
4. Suppose $X(t) = \begin{bmatrix} 2t + t^2 & 3t^2 + t^3 \\ t^2 & t^3 \end{bmatrix}$ is a fundamental matrix for $\dot{\mathbf{x}} = A(t)\mathbf{x}$. Without finding $A(t)$, find all solutions of

$$\dot{\mathbf{x}} = A(t)\mathbf{x} + \begin{bmatrix} 0 \\ t^3e^{-t} \end{bmatrix}.$$

5. Solve

$$\dot{\mathbf{x}} = \begin{bmatrix} 3 & 0 & -2 \\ 0 & -1 & 0 \\ 4 & 0 & 3 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 0 \\ e^{-t} \\ 7 \end{bmatrix}.$$

6. Solve the IVP

$$\left\{\begin{array}{c}\dot{\mathbf{x}}=\begin{bmatrix}-1 & -1\\ 1 & 1\end{bmatrix}\mathbf{x}+\begin{bmatrix}t\\ 0\end{bmatrix}\\ \\ \mathbf{x}(0)=\begin{bmatrix}0\\ 1\end{bmatrix}\end{array}\right\}.$$

7. Solve the two-compartment model

$$\left\{\begin{array}{rcrcl}\dot{A}_1 &=& 5 &-& \frac{A_1}{10}\\ \\ \dot{A}_2 &=& &\frac{A_1}{10} &- \frac{A_2}{6}\end{array}\right\}.$$

8. (a) Find a fundamental matrix for

$$\dot{\mathbf{x}}=\begin{bmatrix}0 & 1\\ -3t^{-2} & 3t^{-1}\end{bmatrix}\mathbf{x}.$$

[Hint: The system is equivalent to a Cauchy–Euler ODE for $x_1(t)$, after using the fact that $\dot{x}_1(t)=x_2(t)$ follows from the first ODE in the system.]

(b) Solve the IVP

$$\left\{\begin{array}{c}\dot{\mathbf{x}}=\begin{bmatrix}0 & 1\\ -3t^{-2} & 3t^{-1}\end{bmatrix}\mathbf{x}+\begin{bmatrix}0\\ 1\end{bmatrix}\\ \\ \mathbf{x}(1)=\begin{bmatrix}-5\\ 1\end{bmatrix}\end{array}\right\}.$$

9. (a) Find a fundamental matrix for

$$\dot{\mathbf{x}}=\begin{bmatrix}0 & 1\\ -8t^{-2} & 5t^{-1}\end{bmatrix}\mathbf{x}.$$

[Hint: The system is equivalent to a Cauchy–Euler ODE for $x_1(t)$, after using the fact that $\dot{x}_1(t)=x_2(t)$ follows from the first ODE in the system.]

(b) Solve the IVP

$$\left\{\begin{array}{c}\dot{\mathbf{x}}=\begin{bmatrix}0 & 1\\ -8t^{-2} & 5t^{-1}\end{bmatrix}\mathbf{x}+\begin{bmatrix}3\\ 0\end{bmatrix}\\ \\ \mathbf{x}(1)=\begin{bmatrix}0\\ -1\end{bmatrix}\end{array}\right\}.$$

10. Solve

$$\dot{\mathbf{x}}=\begin{bmatrix}2 & 1\\ -3 & 6\end{bmatrix}\mathbf{x}+\begin{bmatrix}e^{-5t}\\ 4e^{-t}\end{bmatrix}.$$

11. Explain why there is a particular solution of ODE $\ddot{y} + y = f(t)$ given by $y(t) = \int_0^t \sin(t-u)f(u)du$ by using variation of parameters formula (5.43) for the equivalent system in $\mathbb{R}^2$. [Hint: Use a trigonometric identity for the sin(difference of angles).] [Note: The final conclusion in this problem agrees with the result of Example 4.33 in Section 4.5.]
12. Explain why there is a particular solution of ODE $\ddot{y} + \omega^2 y = f(t)$ given by $y(t) = \frac{1}{\omega}\int_0^t \sin(\omega(t-\tau))f(\tau)d\tau$ by using variation of parameters formula (5.43) for the equivalent system in $\mathbb{R}^2$.
13. Explain why (4.39) in Section 4.3, a formula for all solutions of a second-order ODE $\ddot{y} + p(t)\dot{y} + q(t)y = f(t)$, follows from (5.41).
14. If A is a constant matrix, α is a positive constant, and $\mathbf{w}$ is a constant vector, use Laplace transforms to solve the IVP system

$$\left\{\begin{array}{ll} \dot{\mathbf{x}} & = A\mathbf{x} + \delta(t-\alpha)\mathbf{w} \\ \mathbf{x}(0) & = \mathbf{x}_0 \end{array}\right\}$$

in terms of $A, e^{tA}, \alpha, \mathbf{w}$, and $\mathbf{x}_0$. [Recall that $\mathcal{L}[\,\delta(t-\alpha)] = e^{-\alpha s}$.]
15. Use Laplace transforms to solve the IVP system

$$\left\{\begin{array}{ll} \dot{\mathbf{x}} & = \begin{bmatrix} 2 & -5 \\ 1 & -2 \end{bmatrix}\mathbf{x} + \begin{bmatrix} -\cos 2t \\ \sin 3t \end{bmatrix} \\ \mathbf{x}(0) & = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \end{array}\right\}.$$

5.5 Nonresonant Nonhomogeneous Systems

Here we use the method of undetermined coefficients to find a particular solution of a nonhomogeneous system of ODEs. If the nonhomogeneous term(s) is simple and the matrix of coefficients is constant, this can be a quicker and easier method than a variation of parameters formula requiring integration.

Example 5.26

Solve

$$\dot{\mathbf{x}} = \begin{bmatrix} -1 & 3 \\ 1 & 1 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 3e^{-t} \\ 2e^{-t} \end{bmatrix}. \tag{5.48}$$

Method: Because the nonhomogeneous term is of the form

$$\mathbf{f}(t) = e^{-t}\mathbf{w}$$

for a constant vector $\mathbf{w}$, let's try a particular solution in the form

$$\mathbf{x}_p(t) = e^{-t}\mathbf{a}$$

where $\mathbf{a}$ is a constant vector to be determined. We substitute $\mathbf{x}_p(t)$ into the nonhomogeneous system $\dot{\mathbf{x}} = A\mathbf{x} + \mathbf{f}(t)$ to get

$$-e^{-t}\mathbf{a} = \dot{\mathbf{x}}_p = A\mathbf{x}_p + \mathbf{f}(t) = A\left(e^{-t}\mathbf{a}\right) + e^{-t}\mathbf{w} = e^{-t}\left(A\mathbf{a} + \mathbf{w}\right),$$

that is,

$$-\mathbf{a} = A\mathbf{a} + \mathbf{w},$$

that is,

$$-\mathbf{w} = \left(A - (-1)I\right)\mathbf{a}.$$

The solution is

$$\mathbf{a} = -\left(A - (-1)I\right)^{-1}\mathbf{w},$$

as long as $(A - (-1)I)$ is invertible.

In this specific example, we have $\mathbf{w} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$, and $A = \begin{bmatrix} -1 & 3 \\ 1 & 1 \end{bmatrix}$ is implicitly given earlier. So, we have

$$\mathbf{a} = -\left(A - (-1)\right)^{-1}\mathbf{w} = -\left(\begin{bmatrix} -1 & 3 \\ 1 & 1 \end{bmatrix} - (-1)I\right)^{-1}\begin{bmatrix} 3 \\ 2 \end{bmatrix} = -\begin{bmatrix} 0 & 3 \\ 1 & 2 \end{bmatrix}^{-1}\begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

$$= \frac{1}{3}\begin{bmatrix} 2 & -3 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

The general solution of the original, nonhomogeneous problem is

$$\mathbf{x}(t) = \mathbf{x}_h(t) + \mathbf{x}_p(t) = Z(t)\mathbf{c} + e^{-t}\begin{bmatrix} 0 \\ -1 \end{bmatrix},$$

where $Z(t)$ is a fundamental matrix for the corresponding LCCHS and is $\mathbf{c}$ is a vector of arbitrary constants. Unfortunately, our nice method for finding a particular solution does not help us find $Z(t)$, except we do know that (-1) is not an eigenvalue of A because we were able to calculate $(A - (-1)I)^{-1}$!

As usual, we construct $Z(t)$ using eigenvalues and eigenvectors of A:

$$0 = |\, A - \lambda I \,| = \begin{vmatrix} -1 - \lambda & 3 \\ 1 & 1 - \lambda \end{vmatrix} = \lambda^2 - 4,$$

so the eigenvalues are $\lambda = \pm 2$. For $\lambda_1 = -2$,

$$\left[\, A - (-2)I \mid \mathbf{0} \,\right] = \left[\begin{array}{cc|c} 1 & 3 & 0 \\ 1 & 3 & 0 \end{array}\right] \quad \text{yields} \quad \mathbf{v}^{(1)} = \begin{bmatrix} -3 \\ 1 \end{bmatrix}.$$

For $\lambda_2 = 2$,

$$\left[\, A - 2I \mid \mathbf{0} \,\right] = \left[\begin{array}{cc|c} -3 & 3 & 0 \\ 1 & -1 & 0 \end{array}\right] \quad \text{yields} \quad \mathbf{v}^{(2)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

So,

$$Z(t) = \begin{bmatrix} -3e^{-2t} & e^{2t} \\ e^{-2t} & e^{2t} \end{bmatrix}$$

is a fundamental matrix for the corresponding LCCHS. The general solution of problem (5.48) is

$$\mathbf{x}(t) = \begin{bmatrix} -3e^{-2t} & e^{2t} \\ e^{-2t} & e^{2t} \end{bmatrix} \mathbf{c} + e^{-t} \begin{bmatrix} 0 \\ -1 \end{bmatrix},$$

where $\mathbf{c}$ is a vector of arbitrary constants. ◯

In general, consider a problem of the form

$$\dot{\mathbf{x}} = A\mathbf{x} + e^{\alpha t}\mathbf{w}.$$

We try to find a particular solution of the form

$$\mathbf{x}_p(t) = e^{\alpha t}\mathbf{a}.$$

Theorem 5.12

Suppose that α is not an eigenvalue of the constant matrix A and $\mathbf{w}$ is a constant vector. Then

$$\mathbf{x}_p(t) = -e^{\alpha t}\,(A - \alpha I)^{-1}\,\mathbf{w} \tag{5.49}$$

is a particular solution of the nonhomogeneous system:

$$\dot{\mathbf{x}} = A\mathbf{x} + e^{\alpha t}\mathbf{w}. \tag{5.50}$$

Why? As you will explain in Problem 5.5.2.15, by calculations similar to those in Example 5.27,

$$\mathbf{x}_p(t) = e^{\alpha t}\mathbf{a} \tag{5.51}$$

will guarantee that (5.49) is a particular solution of (5.50). □

We say that "α is not an eigenvalue of A" is a **nonresonance assumption**. We will explore this idea further in Example 5.27.

Theorem 5.12 says in the special case of $\alpha = 0$ that if 0 is not an eigenvalue of A, hence A is invertible, and A and $\mathbf{w}$ are constant, then $\mathbf{x}_p(t) = -A^{-1}\mathbf{w}$ is a particular solution of the system:

$$\dot{\mathbf{x}} = A\mathbf{x} + e^{\alpha t}\mathbf{w}.$$

Example 5.27

Here is a model for obsolescence of computer hardware at a company. Assume that their hardware is sorted into three categories:

(I) The latest models

(II) Not the latest models but very useful

(III) Worth keeping in use but far from the most useful

In general, the categories correspond to the age of the equipment, so we expect that the rates at which pieces of equipment will fail depend on their categories.

Assume also that with the passage of time, some equipment in category I will move into category II, some equipment in category II will move into category III, and some equipment in category III will be disposed of because it becomes obsolete.

Assume that equipment that fails or becomes obsolete will be replaced immediately by new equipment in category I. (This is probably the most unrealistic assumption in the model because it may take some time for equipment to be replaced.)

Let $x_j(t)$, $j = 1, 2, 3$ be the fractions of the company's computer hardware in the three categories I, II, III, respectively. Note that $0 \leq x_j(t) \leq 1$ for all time.

Corresponding to failure of equipment will be "death" rates δ_2 and δ_3. Because equipment in category I that fails is immediately replaced by equipment in category I, we don't need to know that failure rate.

The earlier assumptions lead to the system of ODEs

$$\left.\begin{array}{rcccc} \dot{x}_1 &=& -a_{11}x_1 & +\delta_2 x_2 & +(a_{33}+\delta_3)x_3 \\ \dot{x}_2 &=& a_{11}x_1 & -(a_{22}+\delta_2)x_2 & \\ \dot{x}_3 &=& & a_{22}x_{22} & -(a_{33}+\delta_3)x_3 \end{array}\right\},$$

where $a_{11}, a_{22}, a_{33}, \delta_2, \delta_3 > 0$. Note that $\dfrac{d}{dt}[x_1 + x_1 + x_1] \equiv 0$, so $x_1(t) + x_2(t) + x_3(t)$ will be constant in time. Indeed, all of the equipment is in one of the three categories, so $x_1(t) + x_2(t) + x_3(t) \equiv 1$.

Assume the company is just starting up and they estimate the rates as $a_{11} = 0.3, a_{22} = 0.2, a_{33} = 0.5, \delta_2 = 0.1, \delta_3 = 0.2$, assuming time is measured in years. Solve the system and describe the behavior of the amounts of equipment in the three categories.

Method: Because $x_1(t) + x_2(t) + x_3(t) \equiv 1$, we can reduce the size of the system by substituting $x_1 = 1 - x_2 - x_3$ into the second and third ODEs of the system. This gives a system in $\mathbb{R}^2$:

$$\begin{bmatrix} \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} -(a_{11}+a_{22}+\delta_2) & -a_{11} \\ a_{22} & -(a_{33}+\delta_3) \end{bmatrix} \begin{bmatrix} x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} a_{11} \\ 0 \end{bmatrix}. \tag{5.52}$$

This system has constant coefficients and a constant forcing function.

The method of undetermined coefficients suggests we try a particular solution of the form

$$\begin{bmatrix} x_{2,p} \\ x_{3,p} \end{bmatrix} = \begin{bmatrix} w_2 \\ w_3 \end{bmatrix},$$

where $\mathbf{w} = [w_2 \quad w_3]^T$ is a constant vector. Denoting $A = \begin{bmatrix} -(a_{11}+a_{22}+\delta_2) & -a_{11} \\ a_{22} & -(a_{33}+\delta_3) \end{bmatrix}$, and substituting $\mathbf{x} = [x_{2,p} \quad x_{3,p}]^T = \mathbf{w}$ into the system (5.52), we get

$$\mathbf{0} = \dot{\mathbf{w}} = A\mathbf{w} + \begin{bmatrix} a_{11} \\ 0 \end{bmatrix},$$

whose solution is

$$\begin{aligned}\mathbf{w} &= -A^{-1}\begin{bmatrix} a_{11} \\ 0 \end{bmatrix} \\ &= \frac{-1}{(a_{11}+a_{22}+\delta_2)(a_{33}+\delta_3)+a_{11}a_{22}}\begin{bmatrix} -(a_{33}+\delta_3) & a_{11} \\ -a_{22} & -(a_{11}+a_{22}+\delta_2) \end{bmatrix}\begin{bmatrix} a_{11} \\ 0 \end{bmatrix} \\ &= \frac{-1}{(a_{11}+a_{22}+\delta_2)(a_{33}+\delta_3)+a_{11}a_{22}}\begin{bmatrix} -a_{11}(a_{33}+\delta_3) \\ -a_{11}a_{22} \end{bmatrix} = \begin{bmatrix} 7/16 \\ 1/8 \end{bmatrix}\end{aligned}$$

after substituting in the specific parameter values given in the narrative of the problem.

The general solution of the nonhomogeneous system of ODEs is

$$\begin{bmatrix} x_2 \\ x_3 \end{bmatrix} = e^{tA}\mathbf{c} + \mathbf{w}, \tag{5.53}$$

where $\mathbf{c}$ is a vector of arbitrary constants and $\mathbf{w} = \begin{bmatrix} 7/16 \\ 1/8 \end{bmatrix}$.

We will suppose that initially there is only hardware in category I, because the company is just starting up. (This may not be an appropriate assumption in a struggling economy.) We solve

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} x_2(0) \\ x_3(0) \end{bmatrix} = \mathbf{c} + \mathbf{w}$$

for $\mathbf{c}$ to get $\mathbf{c} = -\mathbf{w}$. So, the solution of this model is

$$\begin{bmatrix} x_2(t) \\ x_3(t) \end{bmatrix} = -(e^{tA} - I)\begin{bmatrix} 7/16 \\ 1/8 \end{bmatrix} \tag{5.54}$$

and $x_1(t) = 1 - x_2(t) - x_3(t)$. ◯

We used *Mathematica*™ to find approximate eigenvalues $-0.65 \pm i0.239792$ and approximate eigenvector $\mathbf{v} = [1 \;\; (0.166667 - i0.799305)]^T$ for the 2×2 matrix A. After using this to find two real solutions, we found a fundamental matrix and then e^{tA}, the details of which are routine, so we will omit. This gives explicit solution of the model

$$\begin{aligned} x_2(t) \approx \frac{7}{16} + e^{-0.65t}\Bigg(&-\frac{7(0.834058)}{16}\cos(0.239792t) \\ &-\frac{7(0.208514)}{16}\sin(0.239792t) + \frac{1}{8}\sin(0.239792t)\Bigg)\end{aligned}$$

and

$$\begin{aligned} x_3(t) \approx \frac{1}{8} + e^{-0.65t}\Bigg(&-\frac{7(0.834058)}{16}\sin(0.239792t) \\ &-\frac{1}{8}\cos(0.239792t) + \frac{0.208514}{8}\sin(0.239792t)\Bigg).\end{aligned}$$

We graphed $x_1(t)$ as a dotted curve, $x_2(t)$ as a dashed curve, and $x_3(t)$ as a solid curve in Figure 5.7. Notice that even though the solution has oscillatory factors with

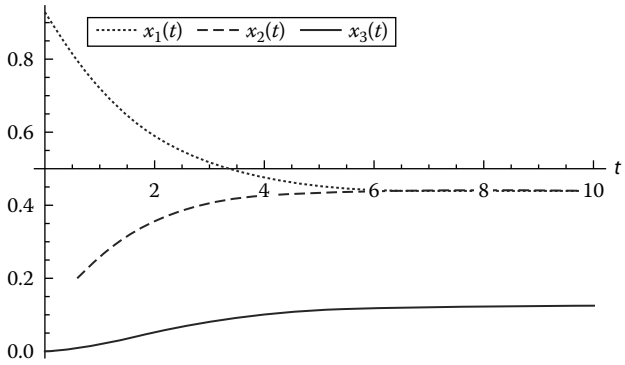


FIGURE 5.7
Computer hardware obsolescence model.

period $\dfrac{2\pi}{0.239792}$, they oscillate so slowly that the relatively rapid decay of $e^{-0.65t}$ makes the solution appear not to have oscillatory factors.

The model suggests that within about four years the amounts of computer hardware in the categories I and II will predominate.

5.5.1 Sinusoidal Forcing

Here we consider systems whose nonhomogeneous term(s) is a sinusoidal function times a constant vector.

Theorem 5.13

Suppose that $\pm i\omega$ is not an eigenvalue of the real, constant matrix A, $\mathbf{w}$ is a constant vector, and $g(t)$ is either $\cos \omega t$ or $\sin \omega t$, where the constant ω is nonzero. Then

$$\dot{\mathbf{x}} = A\mathbf{x} + g(t)\mathbf{w} \tag{5.55}$$

has a particular solution of the form

$$\mathbf{x}_p(t) = (\cos \omega t)\mathbf{a_1} + (\sin \omega t)\mathbf{a_2}, \tag{5.56}$$

where $\mathbf{a_1}, \mathbf{a_2}$ are constant vectors to be determined.

The next example will illustrate why this theorem is true.

Example 5.28

Find a particular solution of

$$\dot{\mathbf{x}} = \begin{bmatrix} -1 & 3 \\ 1 & 1 \end{bmatrix}\mathbf{x} + \begin{bmatrix} \cos 2t \\ 0 \end{bmatrix}. \tag{5.57}$$

Method: Rather than solve a system of the form

$$\dot{\mathbf{x}} = A\mathbf{x} + (\cos 2t)\mathbf{w},$$

where

$$A = \begin{bmatrix} -1 & 3 \\ 1 & 1 \end{bmatrix}, \mathbf{w} = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

we will solve its **complexification**:

$$\dot{\mathbf{x}} = A\mathbf{x} + e^{i2t}\mathbf{w}. \tag{5.58}$$

The relationship between $\widetilde{\mathbf{x}}_p(t)$, the solution of (5.58), and $\mathbf{x}_p(t)$, the solution of (5.57), is

$$\mathbf{x}_p(t) = \mathcal{R}e\left(\widetilde{\mathbf{x}}_p(t)\right),$$

because $\cos 2t = \mathcal{R}e\left(e^{i2t}\right)$.

We try a solution of (5.58) in the form

$$\widetilde{\mathbf{x}}_p(t) = e^{i2t}\widetilde{\mathbf{a}}, \tag{5.59}$$

where $\widetilde{\mathbf{a}}$ is constant vector, possibly complex. We substitute (5.59) into (5.58):

$$i2e^{i2t}\widetilde{\mathbf{a}} = \dot{\widetilde{\mathbf{x}}}_p(t) = A\widetilde{\mathbf{x}}_p(t) + e^{i2t}\mathbf{w} = A\left(e^{i2t}\widetilde{\mathbf{a}}\right) + e^{i2t}\mathbf{w} = e^{i2t}\left(A\widetilde{\mathbf{a}} + \mathbf{w}\right),$$

that is,

$$-\mathbf{w} = (A - i2I)\,\widetilde{\mathbf{a}}.$$

Here we see where the nonresonance condition comes in: if $\pm i2$ is not an eigenvalue of A, then $(A - i2I)$ is invertible and we can solve for $\widetilde{\mathbf{a}}$. Here, that is,

$$\widetilde{\mathbf{a}} = -\left(A - i2I\right)^{-1}\mathbf{w} = -\left(\begin{bmatrix} -1 & 3 \\ 1 & 1 \end{bmatrix} - i2I\right)^{-1}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = -\begin{bmatrix} -1 - i2 & 3 \\ 1 & 1 - i2 \end{bmatrix}^{-1}\begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$= \frac{1}{8}\begin{bmatrix} 1 - i2 & -3 \\ -1 & -1 - i2 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{1}{8}\begin{bmatrix} 1 - i2 \\ -1 \end{bmatrix},$$

so

$$\widetilde{\mathbf{x}}_p(t) = e^{i2t}\widetilde{\mathbf{a}} = \frac{1}{8}\left(\cos 2t + i\sin 2t\right)\begin{bmatrix} 1 - i2 \\ -1 \end{bmatrix} = \frac{1}{8}\begin{bmatrix} \cos 2t + 2\sin 2t - i(2\cos 2t - \sin 2t) \\ -\cos 2t - i\sin 2t \end{bmatrix}.$$

A particular solution is given by

$$\mathbf{x}_p(t) = \mathcal{R}e\left(\widetilde{\mathbf{x}}_p(t)\right) = \frac{1}{8}\begin{bmatrix} \cos 2t + 2\sin 2t \\ -\cos 2t \end{bmatrix}. \bigcirc$$

One of the nice things about the complexification method is that it easily deals with sinusoidal functions that have phase other than zero.

Example 5.29

Find a particular solution of

$$\dot{\mathbf{x}} = \begin{bmatrix} -1 & 3 \\ 1 & 1 \end{bmatrix}\mathbf{x} + \begin{bmatrix} \sin\left(2t - \dfrac{\pi}{4}\right) \\ 0 \end{bmatrix}. \tag{5.60}$$

Method: Because $\sin(2t-\frac{\pi}{4}) = \mathcal{I}m\left(e^{i\left(2t-\frac{\pi}{4}\right)}\right)$, we try

$$\mathbf{x}_p(t) = \mathcal{I}m\left(\widetilde{\mathbf{x}}_p(t)\right),$$

where $\widetilde{\mathbf{x}}_p(t)$ should satisfy

$$\dot{\mathbf{x}} = A\mathbf{x} + e^{i\left(2t-\frac{\pi}{4}\right)}\mathbf{w}. \tag{5.61}$$

The A and $\mathbf{w}$ are the same as in Example 5.28. The problem that $\widetilde{\mathbf{x}}_p(t)$ should satisfy is almost the same as (5.58). So, we try

$$\widetilde{\mathbf{x}}_p(t) = e^{i\left(2t-\frac{\pi}{4}\right)}\widetilde{\mathbf{a}}. \tag{5.62}$$

When we substitute (5.62) into (5.61), we get the same equation for $\widetilde{\mathbf{a}}$ as in Example 5.28:

$$-\mathbf{w} = (A - i2I)\,\widetilde{\mathbf{a}}.$$

So, without repeating all of the steps of Example 5.28, we have

$$\begin{aligned}
\mathbf{x}_p(t) &= \mathcal{I}m\left(\widetilde{\mathbf{x}}_p(t)\right) \\
&= \cdots = \mathcal{I}m\left(\frac{1}{8}\left(\cos\left(2t-\frac{\pi}{4}\right) + i\sin\left(2t-\frac{\pi}{4}\right)\right)\begin{bmatrix}1-i2\\-1\end{bmatrix}\right) \\
&= \cdots = \frac{1}{8}\mathcal{I}m\begin{bmatrix}\cos\left(2t-\frac{\pi}{4}\right) + 2\sin\left(2t-\frac{\pi}{4}\right) - i\left(2\cos\left(2t-\frac{\pi}{4}\right) - \sin\left(2t-\frac{\pi}{4}\right)\right)\\ -\cos\left(2t-\frac{\pi}{4}\right) + i\sin\left(2t-\frac{\pi}{4}\right)\end{bmatrix},
\end{aligned}$$

that is,

$$\mathbf{x}_p(t) = \frac{1}{8}\begin{bmatrix}-2\cos\left(2t-\frac{\pi}{4}\right) + \sin\left(2t-\frac{\pi}{4}\right)\\ \sin\left(2t-\frac{\pi}{4}\right)\end{bmatrix}. \tag{5.63}$$

Using trigonometric identities, we have

$$\cos\left(2t-\frac{\pi}{4}\right) = \frac{1}{\sqrt{2}}(\cos 2t + \sin 2t) \quad \text{and} \quad \sin\left(2t-\frac{\pi}{4}\right) = \frac{1}{\sqrt{2}}(\sin 2t - \cos 2t),$$

so (5.63), the desired particular solution, simplifies to be

$$\mathbf{x}_p(t) = -\frac{1}{8\sqrt{2}}\begin{bmatrix}3\cos 2t + \sin 2t\\ \cos 2t - \sin 2t\end{bmatrix}. \bigcirc$$

At this point, we can better understand why a condition such as

$$\alpha \text{ is not an eigenvalue of the constant matrix } A$$

or

$$\pm i\omega \text{ is not an eigenvalue of the constant matrix } A$$

is referred to as a "nonresonance" condition. For example, in the conclusion of Theorem 5.13, we have a particular solution

$$\mathbf{x}_p(t) = (\cos\omega t)\mathbf{a_1} + (\sin\omega t)\mathbf{a_2},$$

where $\mathbf{a_1}, \mathbf{a_2}$ are constant vectors. If $i\omega$ were an eigenvalue of the constant matrix A, then the algebraic system of equations

$$-\mathbf{w} = (A - i\omega I)\,\widetilde{\mathbf{a}}$$

may or may not have a solution for $\widetilde{\mathbf{a}}$. In the latter case, we should try instead a particular solution of the form

$$\widetilde{\mathbf{x}}_p(t) = e^{i\omega t}\,(t\widetilde{\mathbf{v}} + \widetilde{\mathbf{u}})\,,$$

which fits our understanding of the word "resonance" that we used in Section 4.2.

5.5.2 Problems

1. Redo Example 5.23 in Section 5.4 using the methods of Section 5.5.
2. Redo Example 5.24 in Section 5.4 using the methods of Section 5.5.
3. Solve $\dot{\mathbf{x}} = \begin{bmatrix} -4 & 3 \\ -2 & 1 \end{bmatrix}\mathbf{x} + \begin{bmatrix} e^{-3t} \\ 0 \end{bmatrix}$.
4. Solve $\dot{\mathbf{x}} = \begin{bmatrix} 3 & 0 & -2 \\ 0 & 1 & 0 \\ 4 & 0 & -3 \end{bmatrix}\mathbf{x} + \begin{bmatrix} -5 \\ 0 \\ e^{-2t} \end{bmatrix}$.
5. Solve $\dot{\mathbf{x}} = \begin{bmatrix} -1 & 2 \\ -2 & -1 \end{bmatrix}\mathbf{x} + \begin{bmatrix} \cos t \\ 0 \end{bmatrix}$.
6. Solve $\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 0 \\ \cos t \end{bmatrix}$
   (a) By converting the system into a linear nonhomogeneous second-order scalar ODE
   (b) By a nonresonance method from Section 5.5
   (c) By a method of variation of parameters from Section 5.4
7. Solve $\left\{ \begin{array}{c} \dot{\mathbf{x}} = \begin{bmatrix} 0 & 3 \\ 1 & -2 \end{bmatrix}\mathbf{x} - \begin{bmatrix} e^{-t} \\ 0 \end{bmatrix} \\ \\ \mathbf{x}(0) = \begin{bmatrix} 4 \\ 5 \end{bmatrix} \end{array} \right\}$.
8. Solve $\left\{ \begin{array}{c} \dot{\mathbf{x}} = \begin{bmatrix} 2 & 1 \\ -3 & -2 \end{bmatrix}\mathbf{x} + \begin{bmatrix} -1 \\ \cos t \end{bmatrix} \\ \\ \mathbf{x}(0) = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \end{array} \right\}$.
9. Solve $\dot{\mathbf{x}} = \begin{bmatrix} -4 & 3 \\ -2 & 1 \end{bmatrix}\mathbf{x} + \begin{bmatrix} e^{-t} \\ 0 \end{bmatrix}$.
10. Solve $\dot{\mathbf{x}} = \begin{bmatrix} -6 & 5 & -5 \\ 0 & -1 & 2 \\ 0 & 7 & 4 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 0 \\ 0 \\ e^{-3t} \end{bmatrix}$.

11. Suppose a constant matrix A has an eigenvalue λ and corresponding eigenvector $\mathbf{v}$. Find a particular solution of $\dot{\mathbf{x}} = A\mathbf{x} + e^{\lambda t}\mathbf{v}$ by assuming a solution of the form $\mathbf{x}_p(t) = e^{\lambda t}(t\mathbf{v} + \mathbf{w})$, analogous to one of the more complicated cases of the method of undetermined coefficients for scalar ODEs.
12. Suppose that the system

$$\left\{\begin{array}{l} 3I_1 + 2\dot{I}_1 - 2\dot{I}_2 = 2 \\ 3I_2 + 2\dot{I}_2 - 2\dot{I}_1 = -3 \end{array}\right\}$$

models an electrical circuit. Find the general solution of the system of ODEs. [Hint: First, write the system in the form $B\dot{\mathbf{x}} = A\mathbf{x} + \mathbf{f}$.]
13. (Small project) Develop a method of undetermined coefficients, analogous to those for scalar ODEs and scalar difference equations, for systems of ODEs, that includes resonant cases.
14. Solve the system that models iodine metabolism found in Example 5.6 in Section 5.1. For the value of the input of iodine from the digestive system, find a value for the minimum daily requirement for iodine according to U.S. government nutritional guidelines. Do cite the source of your value.
15. Explain why Theorem 5.12 is true: suppose that α is not an eigenvalue of the constant matrix A and $\mathbf{w}$ is a constant vector. Substitute into (5.50) a solution in the form (5.51), that is, $\mathbf{x}_p(t) = e^{\alpha t}\mathbf{a}$. Find $\mathbf{a}$, in order to see that (5.49), that is, $\mathbf{x}_p(t) = -e^{\alpha t}(A - \alpha I)^{-1}\mathbf{w}$, solves (5.50).

5.6 Linear Control Theory: Complete Controllability

There are many *types* of control problems. Some examples are to drive a process from a given initial condition to a desired end condition, do so in the shortest time, do so with least cost of fuel or other expense, model or "realize" the dynamics of a "black box" process, and stabilize or decouple a process using feedback control. In this section, we will study the first problem.

In this section, we will work only with real numbers, vectors, and matrices.

Suppose $\mathbf{x}(t)$ in $\mathbb{R}^n$ satisfies a **single input** control system of ODEs, that is,

$$\dot{\mathbf{x}} = A\mathbf{x} + u(t)\mathbf{b}, \tag{5.64}$$

where A and $\mathbf{b}$ are constants. The vector-valued function $\mathbf{x}(t)$ is called the **state** of the system, and the control process takes place in **state space** $\mathbb{R}^n$. The scalar-valued function $u(t)$ is called the **control** or **control function**. The system is called "single input" because there is only one scalar control function.

Given the initial condition $\mathbf{x}(0) = \mathbf{x}_0$ and the desired end condition $\mathbf{x}(t_e) = \mathbf{x}_e$, can we choose or construct a scalar control function $u(t)$ so that $\mathbf{x}(t)$ solves the boundary value problem consisting of

$$\left\{\begin{array}{l} \dot{\mathbf{x}} = A\mathbf{x} + u(t)\mathbf{b} \\ \mathbf{x}(0) = \mathbf{x}_0 \end{array}\right\} \tag{5.65}$$

and

$$\mathbf{x}(t_e) = \mathbf{x}_e, \tag{5.66}$$

for some time $t_e > 0$? If so, we will say that we can **drive** $\mathbf{x}(t)$ from $\mathbf{x}_0$ to $\mathbf{x}_e$. We can think of $\mathbf{x}_e$ as a destination or target that we want to reach. The control that accomplishes the driving may depend on $\mathbf{x}_0$, $\mathbf{x}_e$, and t_e, and there may be more than one scalar control function that can accomplish the driving.

More generally, we can ask whether the system has the property defined in the following:

Definition 5.8

System (5.64) is **completely controllable** if for every $\mathbf{x}_0$ and $\mathbf{x}_e$ there is at least one scalar control function $u(t)$ and time t_e for which the boundary value problem (5.65) through (5.66) has a solution.

Example 5.30

Study whether we can drive $\mathbf{x}(t)$ from $\mathbf{x}_0$ to $\mathbf{0}$.

Method: Using variation of parameters formula (5.43) in Section 5.4, the solution of IVP (5.65) is

$$\mathbf{x}(t) = e^{tA}\left(\mathbf{x}_0 + \int_0^t e^{-\tau A}u(\tau)\mathbf{b}\,d\tau\right). \tag{5.67}$$

Substituting $t = t_e$ and $\mathbf{x}(t_e)$ into (5.67) yields

$$\mathbf{0} = e^{t_e A}\left(\mathbf{x}_0 + \int_0^{t_e} e^{-tA}u(t)\mathbf{b}\,dt\right). \tag{5.68}$$

Multiplying on the left by $e^{-t_e A}$, solving for $\mathbf{x}_0$, and changing the variable of definite integration from τ to t result in

$$\mathbf{x}_0 = -\int_0^{t_e} e^{-tA}u(t)\mathbf{b}\,dt. \tag{5.69}$$

Can we find a final time t_e and a scalar control function $u(t)$ to satisfy (5.69)? We will see in Theorem 5.14 that "The answer is yes only if $\mathbf{x}_0$ is in $\mathcal{V}(A, \mathbf{b})$," where

$$\mathcal{V}(A, \mathbf{b}) \triangleq Span\{\,\mathbf{b}, A\mathbf{b}, A^2\mathbf{b}, \ldots, A^{n-1}\mathbf{b}\,\}. \bigcirc \tag{5.70}$$

Note that, as defined in Section 1.7, the span of a set of vectors is a vector subspace.

Example 5.31

Is there a scalar control function $u(t)$ that can drive $\mathbf{x}(t)$ from $\begin{bmatrix} 0 \\ -1 \\ 3 \end{bmatrix}$ to $\mathbf{0}$, if $\mathbf{x}(t)$ solves $\dot{\mathbf{x}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{x} + u(t) \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}$?

Method:

$$\mathcal{V}(A, \mathbf{b}) = Span \left\{ \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}^2 \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix} \right\}$$

$$= Span \left\{ \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \right\}.$$

Because

$$\begin{bmatrix} 0 \\ -1 \\ 3 \end{bmatrix} = \frac{3}{2} \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix} + \left(-\frac{1}{2} \right) \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix} \quad \text{is in} \quad \mathcal{V}(A, \mathbf{b}),$$

yes, there is a scalar control function that drives $\mathbf{x}(t)$ from $\begin{bmatrix} 0 \\ -1 \\ 3 \end{bmatrix}$ to $\mathbf{0}$. ◯

We don't want to be technical about what kind of function the scalar control can be, but we do need $-\int_0^{t_e} e^{-tA} u(t) \mathbf{b}\, dt$ to make sense. For simplicity, let's assume that

> on every finite time interval, $u(t)$ is continuous except at possibly finitely many values of t where it has finite jumps.

that is, that $u(t)$ is **piecewise continuous**.

Theorem 5.14

$\mathbf{x}(t)$ can be driven from $\mathbf{x}_0$ to $\mathbf{0}$ only if $\mathbf{x}_0$ is in $\mathcal{V}(A, \mathbf{b})$.

In order to explain why this is true, we will need a few other results. To avoid getting bogged down in somewhat technical details, we will note those results and explain them later, that is, we will first keep our eye on explaining why Theorem 5.14 is true.

Our explanation process will seem like patiently peeling layers from an onion, eventually getting to basic results that explain earlier results. We will use the ▷ symbol to indicate the end of one explanation that is followed by another Lemma or explanation that was just used.

Lemma 5.2

For all t_e, $\int_0^{t_e} u(t)e^{-tA}\,\mathbf{b}\,dt$ is in $\mathcal{V}(A,\mathbf{b})$.

If we can explain why Lemma 5.2 is true, then we will have established Theorem 5.14 because of (5.69). ▷

In order to explain why Lemma 5.2 is true, we will use

Lemma 5.3

For all t, $e^{-tA}\,\mathbf{b}$ is in $\mathcal{V}(A,\mathbf{b})$.

Assume for the moment that Lemma 5.3 is true. We will use it to explain why Lemma 5.2 is true: for any fixed t_e, $\int_0^{t_e} u(t)e^{-tA}\,\mathbf{b}\,dt$ is the *limit* of Riemann sums

$$\sum_{j=1}^{N} u(t_j)e^{-t_jA}\,\mathbf{b}\,\Delta t.$$

These sums are linear combinations of terms $e^{-t_jA}\mathbf{b}$, each of which is in $\mathcal{V}(A,\mathbf{b})$ by Lemma 5.3. The result of Lemma 5.2 will be true as long as we know that a limit of vectors in $\mathcal{V}(A,\mathbf{b})$ will be in $\mathcal{V}(A,\mathbf{b})$, and this will follow from Lemma 5.4. ▷

Lemma 5.4

$\mathcal{V}(A,\mathbf{b})$ has the property that whenever $\{\mathbf{S}_N\}$ is a sequence in $\mathcal{V}(A,\mathbf{b})$ and there exists $\overline{\mathbf{S}} \triangleq \lim_{N\to\infty} \mathbf{S}_N$, then it follows that $\overline{\mathbf{S}}$ is in $\mathcal{V}(A,\mathbf{b})$.

We will not explain why Lemma 5.4 is true as it involves the subject of analysis sometimes called "real variables." But we will mention that Lemma 5.4 says that $\mathcal{V}(A,\mathbf{b})$ is a **closed** vector subspace.

The concept of "closed" is familiar to us from calculus where we talk about $[\,a,b\,]$ being a "closed" interval. Another example of a "closed" interval is $(-\infty,b]$: if a sequence of real numbers $\{s_N\}$ has $-\infty < s_N \le b$ for all N *and* there exists $\bar{s} \triangleq \lim_{N\to\infty} s_N$, then $\bar{s}$ will also be in $(-\infty,b]$. Note that we say a limit exists only if the limit is a finite number.

So, we have reduced the explanations to Lemma 5.4, which we will not explain, and Lemma 5.3, which we will discuss now: to explain it, we will use two other results.

Theorem 5.15

(**Cayley–Hamilton theorem**) If A is $n\times n$ and has characteristic polynomial $\mathcal{P}(\lambda) \triangleq |\,A-\lambda I\,|$, then $\mathcal{P}(A) = O$, a zero matrix.

To be explicit, the Cayley–Hamilton theorem says that if we use the eigenvalues of A to factor its characteristic polynomial, that is,

$$\mathcal{P}(\lambda) \triangleq |\, A - \lambda I \,| = (\lambda_1 - \lambda)(\lambda_2 - \lambda)\cdots(\lambda_n - \lambda),$$

then

$$O = \mathcal{P}(A) \triangleq (\lambda_1 I - A)(\lambda_2 I - A)\cdots(\lambda_n I - A).$$

For example, in Example 5.23 in Section 5.4, we saw that $A = \begin{bmatrix} 2 & 1 \\ 7 & -4 \end{bmatrix}$ has characteristic polynomial $\mathcal{P}(\lambda) = \lambda^2 + 2\lambda - 15$, so the Cayley–Hamilton theorem says that this matrix A satisfies the matrix equation

$$O = A^2 + 2A - 15I.$$

We will not explain the Cayley–Hamilton theorem in the most general case, but we can explain it in the special case that A is diagonalizable, that is, there is an invertible matrix P such that $A = PDP^{-1}$ for a diagonal matrix $D = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$, whose diagonal entries are the eigenvalues of A, possibly including repetitions. In this case, we calculate that

$$\begin{aligned}
\mathcal{P}(A) &= (\lambda_1 I - A)(\lambda_2 I - A)\cdots(\lambda_n I - A) \\
&= \left(P\lambda_1 IP^{-1} - PDP^{-1}\right)\left(P\lambda_2 IP^{-1} - PDP^{-1}\right)\cdots\left(P\lambda_n IP^{-1} - PDP^{-1}\right) \\
&= \left(P(\lambda_1 I - D)\cancel{P^{-1}}\right)\left(\cancel{P}(\lambda_1 I - D)\cancel{P^{-1}}\right)\cdots\left(\cancel{P}(\lambda_n I - D)P^{-1}\right) \\
&= P\,diag(0, \lambda_1 - \lambda_2, \ldots, \lambda_1 - \lambda_n)\,diag(\lambda_2 - \lambda_1, 0, \ldots, \lambda_2 - \lambda_n) \\
&\quad \cdots diag(\lambda_n - \lambda_1, \ldots, \lambda_n - \lambda_{n-1}, 0)P^{-1} \\
&= P\,diag(0, 0, \ldots, 0)P^{-1} = O. \triangleright
\end{aligned}$$

Corollary 5.1

If A is $n \times n$, then for every integer $k \geq 0$, we can rewrite A^k as a linear combination of $I, A, \ldots, A^{n-1}$.

Why? The Cayley–Hamilton theorem enables us to express A^n in terms of $I, A, \ldots, A^{n-1}$, and then we can use that to find $A^{n+1} = AA^n$ first in terms of $A, \ldots, A^n$ and then in terms of $I, A, \ldots, A^{n-1}$. One can proceed likewise for A^{n+2}, A^{n+3}, etc.

For example, if $A^2 + 2A - 15I = O$ and A is 2×2, then we have

$A^0 \triangleq I, \quad A^1 = A, \quad A^2 = -2A + 15I,$

$A^3 = AA^2 = A(-2A + 15I) = -2A^2 + 15A = -2(-2A + 15I) + 15A = 19A - 30I, etc.$ ▷

This completes the explanation of Theorem 5.14, that is, explains why we can drive $\mathbf{x}(t)$ from $\mathbf{x}_0$ to $\mathbf{0}$ only if $\mathbf{x}_0$ is in $\mathcal{V}(A, \mathbf{b})$. □

More generally, we have

Theorem 5.16

System (5.64) is completely controllable if, and only if,

$$\left| \mathbf{b} \;\vdots\; A\mathbf{b} \;\vdots\; A^2\mathbf{b} \;\vdots\; \cdots \;\vdots\; A^{n-1}\mathbf{b} \right| \neq 0. \tag{5.71}$$

Theorem 1.43 in Section 1.7 implies that the *controllability condition* (5.71) is equivalent to having $\mathcal{V}(A, \mathbf{b}) = \mathbb{R}^n$ and also equivalent to the $n \times n$ matrix $\left[\mathbf{b} \;\vdots\; A\mathbf{b} \;\vdots\; \cdots \;\vdots\; A^{n-1}\mathbf{b} \right]$ having rank equal to n, by Theorem 1.30 in Section 1.6.

We can explain part of this powerful result by giving a formula for control functions! Suppose that the controllability condition (5.71) is true. We will construct a scalar control that drives $\mathbf{x}(t)$ from a given $\mathbf{x}_0$ to a given $\mathbf{x}_e$:

$$u(t) \triangleq \mathbf{b}^T e^{-tA^T} \mathbf{a}, \tag{5.72}$$

where $\mathbf{a}$ is a constant vector to be chosen later.

First, we can rewrite

$$u(t)\,\mathbf{b} = \mathbf{b}\,u(t)$$

if we interpret $u(t)$ on the left-hand side as a scalar-valued function of t and interpret $u(t)$ on the right-hand side as a 1×1 matrix-valued function of t. Indeed, toward the end of this section, we will consider control systems $\dot{\mathbf{x}} = A\,\mathbf{x} + B\,\mathbf{u}(t)$ with "multivariable control," where $\mathbf{u}(t)$ will be a vector of control functions and B will be a constant matrix.

Substitute (5.72) into (5.67), the variation of parameters formula applied to our problem, to get

$$\mathbf{x}(t) = e^{tA}\left(\mathbf{x}_0 + \int_0^t e^{-\tau A}\,\mathbf{b}\,u(\tau)d\tau\right) = e^{tA}\left(\mathbf{x}_0 + \int_0^t e^{-\tau A}\,\mathbf{b}\,\mathbf{b}^T e^{-\tau A^T}\mathbf{a}\,d\tau\right). \tag{5.73}$$

Define, for any fixed t_e, the matrix

$$M \triangleq \int_0^{t_e} e^{-\tau A}\,\mathbf{b}\,\mathbf{b}^T e^{-\tau A^T}\mathbf{a}\,d\tau. \tag{5.74}$$

With this definition, (5.65) and (5.66) can be restated as

$$\mathbf{x}_e = e^{t_e A}(\mathbf{x}_0 + M\mathbf{a}),$$

that is,

$$M\mathbf{a} = e^{-t_e A}\mathbf{x}_e - \mathbf{x}_0. \tag{5.75}$$

Lemma 5.5

If the system satisfies the controllability condition (5.71) and $t_e > 0$, then M is invertible.

We'll explain why Lemma 5.5 is true later, but for now, because of the lemma, we can solve (5.75):

$$\mathbf{a} = M^{-1}\left(e^{-t_e A}\mathbf{x}_e - \mathbf{x}_0\right)$$

and thus that we can find a control that drives $\mathbf{x}(t)$ from $\mathbf{x}_0$ to $\mathbf{x}_e$. □

What may seem a little surprising is that the controllability condition implies $\mathbf{x}(t)$ can be driven from $\mathbf{x}_0$ to $\mathbf{x}_e$ in time t_e, no matter how small t_e is! At first, this seems to be unrealistic. But, intuitively, if t_e is chosen to be very small, then the matrix $M = \int_0^{t_e} e^{-tA}\,\mathbf{b}\,\mathbf{b}^T e^{-tA^T}\mathbf{a}\,dt$ will be very small because it is an integral of a continuous (matrix-valued) function on a very small time interval. Thus, M^{-1} will be very "large," and so $\mathbf{a} = M^{-1}\left(e^{-t_e A}\mathbf{x}_e - \mathbf{x}_0\right)$ will be very large. Thus, in this case, the control function given by (5.72), that is, $u(t) = \mathbf{b}^T e^{-tA^T}\mathbf{a}$, would be very large.

So, if you want to drive $\mathbf{x}(t)$ from $\mathbf{x}_0$ to $\mathbf{x}_e$ in an unrealistically small amount of time, "all" that you need is an unrealistically large control function requiring an unrealistically large amount of resources to produce! In real life, there would be physical and/or economic limitations on the amount of control we can exert, so we *cannot* drive $\mathbf{x}(t)$ from $\mathbf{x}_0$ to $\mathbf{x}_e$ in an arbitrarily small amount of time.

Now let's explain why Lemma 5.5 is true. Suppose that

$$\left|\,\mathbf{b}\;\vdots\;A\mathbf{b}\;\vdots\;\cdots\;\vdots\;A^{n-1}\mathbf{b}\,\right| \neq 0$$

but that M is not invertible. Eventually we will reach a contradiction. There exists an $\mathbf{a}$ for which $M\mathbf{a} = \mathbf{0}$; hence,

$$0 = \mathbf{a}^T\mathbf{0} = \mathbf{a}^T M\mathbf{a} = \int_0^{t_e} \mathbf{a}^T e^{-tA}\,\mathbf{b}\,\mathbf{b}^T e^{-tA^T}\mathbf{a}\,dt;$$

hence,

$$0 = \int_0^{t_e} \left(\mathbf{b}^T e^{-tA^T}\mathbf{a}\right)^T \left(\mathbf{b}^T e^{-tA^T}\mathbf{a}\right) dt = \int_0^{t_e} w(t)w(t)\,dt = \int_0^{t_e} |\,w(t)\,|^2\,dt, \tag{5.76}$$

where we define the scalar-valued function

$$w(t) \triangleq \mathbf{b}^T e^{-tA^T}\mathbf{a}.$$

Because $|w(t)|^2 \geq 0$ for t in the interval $[0, t_e]$, (5.76) implies that $w(t) = 0$ for all t in $[0, t_e]$, that is,

$$0 \equiv w(t) = \mathbf{a}^T e^{-tA}\mathbf{b},\ 0 \leq t \leq t_e. \tag{5.77}$$

If we take the first through $(n-1)$-st derivatives of (5.77), we get

$$\begin{aligned} 0 &= \mathbf{a}^T e^{-tA}(-A)\mathbf{b} \\ &\vdots \\ 0 &= \mathbf{a}^T e^{-tA}(-A)^{n-1}\mathbf{b}. \end{aligned}$$

After that, substitute $t = 0$ into those equations and into (5.77) to get

$$\begin{aligned} 0 &= \mathbf{a}^T\mathbf{b} \\ 0 &= \mathbf{a}^T(-A)\mathbf{b} \\ &\vdots \\ 0 &= \mathbf{a}^T(-A)^{n-1}\mathbf{b}. \end{aligned}$$

By Theorem 1.9 in Section 1.2, we have

$$\mathbf{a}^T\left[\,\mathbf{b}\ \vdots\ A\mathbf{b}\ \vdots\ \cdots\ \vdots\ A^{n-1}\mathbf{b}\,\right] = \left[\,\mathbf{a}^T\mathbf{b}\ \vdots\ \mathbf{a}^TA\mathbf{b}\ \vdots\ \cdots\ \vdots\ \mathbf{a}^TA^{n-1}\mathbf{b}\,\right] = \left[\,0\ \vdots\ 0\ \cdots\ \vdots\ 0\,\right] = \mathbf{0}^T. \tag{5.78}$$

Take the transpose of both sides of (5.78) to get

$$\left[\,\mathbf{b}\ \vdots\ A\mathbf{b}\ \vdots\ \cdots\ \vdots\ A^{n-1}\mathbf{b}\,\right]^T\mathbf{a} = \mathbf{0}. \tag{5.79}$$

But we assumed the controllability condition, that is, that the determinant

$$\left|\,\mathbf{b}\ \vdots\ A\mathbf{b}\ \vdots\ A^2\mathbf{b}\ \vdots\ \cdots\ \vdots\ A^{n-1}\mathbf{b}\,\right| \neq 0;$$

hence, (5.79) implies that $\mathbf{a} = \mathbf{0}$, contradicting the assumption that $\mathbf{a} \neq \mathbf{0}$. So, by contradiction, we have the desired result that M is invertible. □

5.6.1 Some Other Control Problems

In establishing the controllability criterion, we used the specific form for control functions (5.72), that is,

$$u(t) = \mathbf{b}^T e^{-tA^T}\mathbf{a}.$$

There might be other control functions that also drive $\mathbf{x}(t)$ from $\mathbf{x}_0$ to $\mathbf{x}_e$ in a finite amount of time. The subject of **optimal control** is concerned with choosing which control does the driving in such a way that it optimizes a design criterion, such as taking the least amount of time to arrive at $\mathbf{x}_e$ or such as minimizing the cost of the driving (e.g., in a specified amount of time).

System (5.64) has a single control function. One generalization is to have several control functions, for example, in a **multi variable** control system

$$\dot{\mathbf{x}} = A\mathbf{x} + B\mathbf{u}, \tag{5.80}$$

where B is a constant $n \times m$ matrix and $\mathbf{u}(t)$ is an $\mathbb{R}^m$-valued function of t. It turns out that the controllability criterion for (5.80) is a nice generalization of that for the single input control system (5.64).

Theorem 5.17

System (5.80) is completely controllable, that is, for every $\mathbf{x}_e, \mathbf{x}_0$, there is at least one control function that drives $\mathbf{x}(t)$ from $\mathbf{x}_0$ to $\mathbf{x}_e$ in a finite amount of time, if and only if

$$n = rank\left(\left[\, B \;\vdots\; AB \;\vdots\; \cdots \;\vdots\; A^{n-1}B \,\right]\right).$$

The rank of the $n \times nm$ matrix $\left(\left[\, B \;\vdots\; AB \;\vdots\; \cdots \;\vdots\; A^{n-1}B \,\right]\right.$ can be at most n, so we can say that the controllability criterion is that that matrix should have "full rank."

It turns out that we can use a control function that is a nice generalization of (5.72):

$$\mathbf{u}(t) \triangleq B^T e^{-tA^T}\mathbf{a},$$

where $\mathbf{a}$ is a constant vector. Note that $B^T e^{-tA^T}$ is an $m \times n$ matrix for all t.

Another basic problem in control theory is that of "observing" the initial state $\mathbf{x}_0$ of a system. The simplest context where we can discuss this is the single input control system (5.64). Assume that we can measure only some linear combination of the state components, that is, one can measure

$$y(t) \triangleq c_1 x_1(t) + \cdots + c_n x_n(t) = \mathbf{c}^T\mathbf{x}(t),$$

where $\mathbf{c}$ is a constant vector in $\mathbb{R}^n$. Of course, if we could measure each of the state components $x_1(t), \ldots, x_n(t)$ at all times, then we could just read off $\mathbf{x}(0) = \mathbf{x}_0$.

Altogether, our system is

$$\left\{\begin{array}{rcl} \dot{\mathbf{x}} & = & A\mathbf{x} + B\mathbf{u} \\ y(t) & = & \mathbf{c}^T\mathbf{x}(t) \\ \mathbf{x}(0) & = & \mathbf{x}_0 \end{array}\right\}, \tag{5.81}$$

where constant matrix A and constant vectors $\mathbf{b}, \mathbf{c}$ are assumed to be known.*

* Yet another type of problem in control theory is to find a way to estimate the model parameters $A, \mathbf{b}$, for example, by using a "Luenberger observer."

For this problem, there is a precise definition of a concept concerning observing a system:

Definition 5.9

System (5.81) is **completely observable** if for every $\mathbf{x}_0$ there is a finite time t_e, possibly dependent on $\mathbf{x}_0$, such that if we know both $\mathbf{u}(t)$ and $y(t)$ for all t in the interval $[0, t_e]$, then we can calculate what $\mathbf{x}_0$ was.

In a way, observability is like predicting the past. They say that "hindsight has 20-20 vision," but even predicting the past can be difficult to do.

It turns out that there is a complete observability criterion akin to the complete controllability criterion (5.71): the $n \times n$ determinant

$$\begin{vmatrix} \mathbf{c}^T \\ \mathbf{c}^T A \\ \vdots \\ \mathbf{c}^T A^{n-1} \end{vmatrix}$$

should be nonzero. Also, it turns out that this generalizes nicely to systems with multiple observers, that is, where $\mathbf{y}(t) \triangleq C^T \mathbf{x}(t)$ can be measured.

Learn More About It

We have barely scratched the surface of the subject of control theory. Here are some useful references: (1) "Some fundamental control theory I: Controllability, observability, and duality," William J. Terrell, *Am. Math. Mon.* **106** (1999), 705–719, and (2) "Some fundamental control theory II: feedback linearization of single input nonlinear systems," William J. Terrell, *The American Mathematical Monthly* **106** (1999), 812–828.

The first article relates controllability to systems in companion form and shows how controllability and observability are "dual" concepts. Both articles also have useful bibliographical references.

Here are two papers that made foundational contributions to the subject of control theory: (3) "Contributions to the theory of optimal control," R. E. Kalman, *Bol. Soc. Mat. Mex., Series* II **5** (1960), 102–119, and (4) "Observers for multivariable systems," D. G. Luenberger, *IEEE Trans. Auto. Control* **AC-11** (1966), 190–197.

5.6.2 Problems

In problems 1 and 2, determine whether the system can be driven from $\begin{bmatrix} 6 \\ 3 \end{bmatrix}$ to $\mathbf{0}$.

1. $\dot{\mathbf{x}} = \begin{bmatrix} 1 & -2 \\ -2 & 4 \end{bmatrix} \mathbf{x} + u(t) \begin{bmatrix} 2 \\ 1 \end{bmatrix}$
2. $\dot{\mathbf{x}} = \begin{bmatrix} -2 & 7 \\ 1 & 4 \end{bmatrix} \mathbf{x} + u(t) \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

In problems 3–5, determine whether the system is completely controllable.

3. $\dot{\mathbf{x}} = \begin{bmatrix} -2 & 7 \\ 1 & 4 \end{bmatrix} \mathbf{x} + u(t) \begin{bmatrix} 1 \\ 0 \end{bmatrix}$
4. $\dot{\mathbf{x}} = \begin{bmatrix} 1 & 2 \\ 3 & 2 \end{bmatrix} \mathbf{x} + u(t) \begin{bmatrix} 1 \\ 0 \end{bmatrix}$
5. $\dot{\mathbf{x}} = \begin{bmatrix} 1 & -2 \\ -2 & 4 \end{bmatrix} \mathbf{x} + u(t) \begin{bmatrix} 2 \\ 1 \end{bmatrix}$
6. Suppose A is a 2×2 diagonal matrix $diag(a_{11}, a_{22})$ where a_{11}, a_{22} are nonzero. For what vectors $\mathbf{b}$ is system (5.64) completely controllable? For what vectors $\mathbf{c}$ is system (5.81) completely observable?

In problems 7 and 8, determine whether the system is completely controllable.

7. $\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{u}(t)$
8. $\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1 & 1 \\ 2 & 0 \end{bmatrix} \mathbf{u}(t)$

5.7 Linear Systems of Difference Equations

Here we study systems of first-order difference equations

$$\mathbf{x}_{k+1} = A_k \mathbf{x}_k \tag{5.82}$$

where $\mathbf{x}$ is in $\mathbb{R}^n$ and A_k is an $n \times n$ matrix; usually we will assume that A_k is a constant matrix A, that is, does not depend of k, so that we have a linear, constant coefficients, homogeneous system of difference equations (LCCHSΔ)

$$\mathbf{x}_{k+1} = A\mathbf{x}_k, \; k \geq 0. \tag{5.83}$$

For example, in $\mathbb{R}^2$, an LCCHSΔ would have the form

$$\begin{Bmatrix} x_{1,k+1} = a_{11}x_{1,k} + a_{12}x_{2,k} \\ x_{2,k+1} = a_{21}x_{1,k} + a_{22}x_{2,k} \end{Bmatrix}.$$

The solutions of LCCHSΔ (5.83) are easy to find by an inductive process: denoting $\mathbf{x}_0 = \mathbf{c}$, we have

$$\mathbf{x}_1 = A\mathbf{x}_0 = A\mathbf{c}, \; \mathbf{x}_2 = A\mathbf{x}_1 = A(A\mathbf{c}) = A^2\mathbf{c}, \ldots,$$

$$\mathbf{x}_k = A\mathbf{x}_{k-1} = A(A^{k-1}\mathbf{c}) = A^k\mathbf{c}. \tag{5.84}$$

In the aforementioned, the vector of initial values $\mathbf{c}$ is unspecified, hence plays the role of arbitrary constants. Although (5.84) is a nice formula, it doesn't tell us much about the behavior of solutions.

Example 5.32

Use eigenvalues and eigenvectors to write the solution of

$$\begin{Bmatrix} x_{1,k+1} &=& -2x_{1,k} &+& x_{2,k} \\ x_{2,k+1} &=& x_{1,k} &-& 2x_{2,k} \end{Bmatrix}. \tag{5.85}$$

Method: Denote $A = \begin{bmatrix} -2 & 1 \\ 1 & -2 \end{bmatrix}$. If we can find an eigenvector $\mathbf{v}$ of A corresponding to an eigenvalue λ, then $A\mathbf{v} = \lambda\mathbf{v}$ implies $A^2\mathbf{v} = A(A\mathbf{v}) = A(\lambda\mathbf{v}) = \lambda A\mathbf{v} = \lambda^2\mathbf{v}$, etc. We see inductively that

$$A^k\mathbf{v} = \lambda^k\mathbf{v}. \tag{5.86}$$

One can show that this matrix A has eigenvalues and eigenvectors:

$$\lambda_1 = -3, \quad \mathbf{v}^{(1)} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \quad \text{and} \quad \lambda_2 = -1, \quad \mathbf{v}^{(2)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Using these, (5.86), and linear superposition, we know (5.85) has solutions

$$\mathbf{x}_k = \bar{c}_1(-3)^k \begin{bmatrix} -1 \\ 1 \end{bmatrix} + \bar{c}_2(-1)^k \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \; k = 1, 2, \cdots \tag{5.87}$$

where c_1, c_2 are arbitrary constants. ◯

In fact, these are the solutions given by (5.84) by using diagonalization and the explanation of Theorem 2.10 in Section 2.2:

$$\begin{aligned} \mathbf{x}_k = A^k\mathbf{c} = (PDP^{-1})^k\mathbf{c} = P\left(D^k\right)P^{-1}\mathbf{c} = P\begin{bmatrix} -3 & 0 \\ 0 & -1 \end{bmatrix}^k P^{-1}\mathbf{c} \\ = \left[\begin{array}{c|c} \mathbf{v}^{(1)} & \mathbf{v}^{(2)} \end{array}\right]\begin{bmatrix} (-3)^k & 0 \\ 0 & (-1)^k \end{bmatrix} P^{-1}\mathbf{c} = \left[\begin{array}{c|c} (-3)^k\mathbf{v}^{(1)} & (-1)^k\mathbf{v}^{(2)} \end{array}\right] P^{-1}\mathbf{c} \\ = \left[\begin{array}{c|c} (-3)^k\mathbf{v}^{(1)} & (-1)^k\mathbf{v}^{(2)} \end{array}\right]\begin{bmatrix} \bar{c}_1 \\ \bar{c}_2 \end{bmatrix} = \bar{c}_1(-3)^k\mathbf{v}^{(1)} + \bar{c}_2(-1)^k\mathbf{v}^{(2)}, \end{aligned}$$

where the vector $\bar{\mathbf{c}} \triangleq P^{-1}\mathbf{c}$.

5.7.1 Color Blindness

A **gene** is a string of DNA, like a computer file, that can manifest in a characteristic of a living organism. Chromosomes are like folders containing those gene "files." Normally, a human being has a pair of chromosomes, one each inherited from their mother and father. Normally, a female has two X chromosomes and a male has one X chromosome and one Y chromosome, so gender is determined by whether or not one has a Y chromosome.

Some human beings have three, or even more, chromosomes, but such people are rare and will be ignored in the following.

Consider a **sex-linked** gene located on an X chromosome, so females have two (possibly different) copies and males have only one copy. A gene variant is **recessive** if it manifests only if all copies are that variant. So, a male will manifest a sex-linked recessive gene variant if its single copy is that variant, while a female will manifest the variant only if both of its copies are that variant.

Red–green color blindness (deuteranopia), that is, the inability to see the difference between the colors red and green, is an example of a sex-linked recessive gene. Other examples in human beings are hemophilia and Duchenne's muscular dystrophy.

Suppose that in the kth generation $x_{1,k}$ is the proportion of the gene variant for red–green color blindness in human females and $x_{2,k}$ is the proportion of that gene variant in human males. For example, if in the kth generation 1% of females have one copy of that gene variant and 0.02% of females have two copies of that gene variant, then $x_{1,k} = 0.9898 \cdot 0 + 0.01 \cdot 1 + 0.0002 \cdot 2 = 0.0104$. This is obviously going to be a simplified model because human beings, unlike flies in a laboratory experiment, do not live and reproduce in well-defined generations. A more sophisticated model would break down the human populations of males and females by age and take into account how their ages affect reproduction and future life span. Nevertheless, we can learn something significant from our simple model.

A male can only inherit the gene variant from one of his mother's X chromosomes, so $x_{2,k+1} = x_{1,k}$. A female will inherit the average of the proportion of the gene variant in her mother's and father's X chromosomes, so

$$x_{2,k+1} = \frac{x_{1,k} + x_{1,k}}{2}.$$

So, $\mathbf{x}_k = [x_{1,k} \quad x_{2,k}]^T$ satisfies the system of linear constant coefficients difference equations

$$\mathbf{x}_{k+1} = \begin{bmatrix} 0.5 & 0.5 \\ 1 & 0 \end{bmatrix} \mathbf{x}_k. \tag{5.88}$$

It's easy to calculate that the matrix $A \triangleq \begin{bmatrix} 0.5 & 0.5 \\ 1 & 0 \end{bmatrix}$ has eigenvalues 1 and -0.5 and corresponding complete set of corresponding eigenvectors

$$\left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -0.5 \\ 1 \end{bmatrix} \right\},$$

so we can diagonalize

$$A = \begin{bmatrix} 1 & -\frac{1}{2} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ -\frac{2}{3} & \frac{2}{3} \end{bmatrix}.$$

The solutions are

$$\mathbf{x}_k = \begin{bmatrix} 1 & -\frac{1}{2} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -\frac{1}{2} \end{bmatrix}^k \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ -\frac{2}{3} & \frac{2}{3} \end{bmatrix} \mathbf{x}_0 = \begin{bmatrix} 1 & -\frac{1}{2} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (-\frac{1}{2})^k \end{bmatrix} \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ -\frac{2}{3} & \frac{2}{3} \end{bmatrix} \mathbf{x}_0$$

$$= \frac{1}{3} \begin{bmatrix} 2 + (-\frac{1}{2})^k & 1 - (-\frac{1}{2})^k \\ 2 - 2(-\frac{1}{2})^k & 1 + 2(-\frac{1}{2})^k \end{bmatrix} \mathbf{x}_0 .$$

So, there exists the steady state

$$\mathbf{x}_\infty = \lim_{k \to \infty} \mathbf{x}_k = \frac{1}{3} \begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix} \mathbf{x}_0 = \frac{2x_{1,0} + x_{2,0}}{3} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

If $p \triangleq \frac{2x_{1,0}+x_{2,0}}{3}$ is the *steady-state* proportion of the red–green color blindness gene in woman, and hence in men, too, $100p\%$ of the men will be red–green color blind. But, a woman needs two copies of that gene, the probability of which will be about $p \cdot p$, so about $100p^2\%$ of the women will be red–green color blind. Men should be about $\frac{1}{p}$ as likely to manifest red–green color blindness as women. In the United States, $p \approx 0.07$, that is, about 7%, of males manifest red–green blindness, and about 0.004 of females manifest it; the proportions are roughly of the form p and p^2, as predicted by the theory. By the way, the proportion p can vary among different ethnic groups.

For example, about 0.0002, that is, two in 10,000, of men manifest hemophilia A; hence, about $(0.0002)^2$, or about four in a hundred million, of women would manifest hemophilia A. But about two in 10,000 women have one copy of the hemophilia A gene and thus would be "carriers" of this genetic disease.

5.7.2 General Solution and the Casorati Determinant

Analogous to Definition 5.2 in Section 5.2, as well as Definitions 3.2 in Section 3.1, 3.8 in Section 3.3, and 3.10 in Section 3.4, we have

Definition 5.10

The **general solution** of a system of linear homogeneous difference equations (5.82) has the form

$$\mathbf{x}_k^{(h)} = c_1 \mathbf{x}_k^{(1)} + c_2 \mathbf{x}_k^{(2)} + \cdots + c_n \mathbf{x}_k^{(n)}$$

if for every solution $\mathbf{x}_k^*$ of (5.82) there are values of constants $c_1, c_2, \ldots, c_n$ giving $\mathbf{x}_k^* = c_1 \mathbf{x}_k^{(1)} + c_2 \mathbf{x}_k^{(2)} + \cdots + c_n \mathbf{x}_k^{(n)}$. In this case, we call the set of sequences $\{\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}, \ldots, \mathbf{x}_k^{(n)}\}$ a **complete set of basic solutions** for linear homogeneous difference equation (5.82). Each of the n sequences $\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}, \ldots, \mathbf{x}_k^{(n)}$ is called a **basic solution** of (5.82).

Theorem 5.18

The system of linear homogeneous difference equations (5.82) in $\mathbb{R}^n$ has a general solution, that is, a complete set of n basic solutions.

Why? The explanation for this is similar to that for Theorem 5.14 in Section 5.6, as well as those for Theorems 3.9 in Section 3.3 and 3.15 in Section 3.4: Denote the n columns of I_n by $\mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(n)}$. Each of the n IVPs

$$\left\{\mathbf{x}_{k+1}^{(1)} = A_k\mathbf{x}_k^{(1)},\ \mathbf{x}_0^{(1)} = \mathbf{e}^{(1)}\right\}, \left\{\mathbf{x}_{k+1}^{(2)} = A_k\mathbf{x}_k^{(2)},\ \mathbf{x}_0^{(2)} = \mathbf{e}^{(2)}\right\}, \ldots, \left\{\mathbf{x}_{k+1}^{(n)} = A_k\mathbf{x}_k^{(n)},\ \mathbf{x}_0^{(n)} = \mathbf{e}^{(n)}\right\}$$

has a solution. The rest of the explanation is also similar to that given for Theorem 3.9 in Section 3.3. □

If $\{\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}, \ldots, \mathbf{x}_k^{(n)}\}$ is a set of n sequences, the **Casorati determinant** plays a role analogous to that of the Wronskian determinant for systems of linear homogeneous ODEs. We define the Casorati by

$$C(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}, \ldots, \mathbf{x}_k^{(n)}) \triangleq \left| \mathbf{x}_k^{(1)} \,\vdots\, \mathbf{x}_k^{(2)} \,\vdots\, \cdots \,\vdots\, \mathbf{x}_k^{(n)} \right|.$$

Theorem 5.19

(Abel's theorem) Suppose $\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}, \ldots, \mathbf{x}_k^{(n)}$ are n solutions of the same system of linear homogeneous difference equations (5.82) in $\mathbb{R}^n$. Then

$$C\left(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}, \ldots, \mathbf{x}_k^{(n)}\right) = \left(\prod_{\ell=0}^{k-1} |A_\ell|\right) C\left(\mathbf{x}_0^{(1)}, \mathbf{x}_0^{(2)}, \ldots, \mathbf{x}_0^{(n)}\right) \tag{5.89}$$

for any $k \geq 1$.

Why? Analogous to the explanation of Theorems 4.15 in Section 4.6 and 3.12 in Section 3.3, first we claim that

$$C_k \triangleq C\left(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}, \ldots, \mathbf{x}_k^{(n)}\right)$$

satisfies the first-order linear homogeneous difference equation:

$$C_{k+1} = |A_k|\, C_k. \tag{5.90}$$

But this is not difficult to explain, using Theorems 1.9 in Section 1.2 and 1.28(b) in Section 1.6:

$$C_{k+1} = C\left(\mathbf{x}_{k+1}^{(1)}, \mathbf{x}_{k+1}^{(2)}, \ldots, \mathbf{x}_{k+1}^{(n)}\right) = C\left(A_k\mathbf{x}_k^{(1)}, \ldots, A_k\mathbf{x}_k^{(n)}\right) = \left| A_k\mathbf{x}_k^{(1)} \,\vdots\, \cdots \,\vdots\, A_k\mathbf{x}_k^{(n)} \right|$$

$$= \left| A_k \left[\mathbf{x}_k^{(1)} \,\vdots\, \cdots \,\vdots\, \mathbf{x}_k^{(n)} \right] \right| = |A_k| \left| \mathbf{x}_k^{(1)} \,\vdots\, \cdots \,\vdots\, \mathbf{x}_k^{(n)} \right| = |A_k|\, C_k. \ \square$$

Theorem 5.20

Suppose $\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}, \ldots, \mathbf{x}_k^{(n)}$ are solutions of the same system of linear homogeneous difference equations (5.82) in $\mathbb{R}^n$ and $|A_k| \neq 0$ for $k = 0, 1, \cdots$. Then

$$C\left(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}, \ldots, \mathbf{x}_k^{(n)}\right) \neq 0 \text{ for all } k \geq 0$$

if, and only if, $\left\{\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}, \ldots, \mathbf{x}_k^{(n)}\right\}$ is a complete set of basic solutions of system (5.82).

Why? Apply Abel's Theorem 5.19 and the existence and uniqueness conclusions of Theorem 5.18. □

5.7.3 Complex Eigenvalues

Suppose a real matrix A has a complex conjugate pair of eigenvalues $r = \alpha \pm i\nu$, where α, ν are real and $\nu > 0$, and corresponding complex conjugate pair of eigenvectors $\mathbf{v}, \overline{\mathbf{v}}$. Then, as in (4.63) in Section 4.6, it helps to use the polar form of the eigenvalues:

$$\alpha + i\nu = \rho e^{i\omega} = \rho(\cos\omega + i\sin\omega),$$

where ρ is real and nonnegative and $-\pi < \omega \leq \pi$. Note that $\omega \neq 0$ because $\nu > 0$. Then we have

$$A^k\mathbf{v} = \left(\rho^k(\cos\omega k + i\sin\omega k)\right)\mathbf{v}.$$

It follows that two solutions of LCCHSΔ (5.83) are given by

$$\mathbf{x}_k^{(1)} = \rho^k\, \mathcal{R}e\left((\cos\omega k + i\sin\omega k)\mathbf{v}\right), \quad \mathbf{x}_k^{(2)} = \rho^k\, \mathcal{I}m\left((\cos\omega k + i\sin\omega k)\mathbf{v}\right).$$

The same as for LCCHS, we don't need the second eigenvector, $\overline{\mathbf{v}}$.

Example 5.33

Solve

$$\begin{Bmatrix} x_{1,k+1} & = x_{1,k} & -2x_{2,k} \\ x_{2,k+1} & = 4x_{1,k} & -3x_{2,k} \end{Bmatrix}. \tag{5.91}$$

Method: Denote $A = \begin{bmatrix} 1 & -2 \\ 4 & -3 \end{bmatrix}$. In Example 2.9 in Section 2.1, we found that the eigenvalues are $\lambda = -1 \pm i2$ and that corresponding to $\lambda = -1 + i2$, there is an eigenvector

$$\mathbf{v} = \begin{bmatrix} \frac{1}{2} + i\,\frac{1}{2} \\ \\ 1 \end{bmatrix}.$$

We calculate that

$$\alpha + i\nu = -1 + i2 = \rho(\cos\omega + i\sin\omega),$$

where

$$\rho = \sqrt{(-1)^2 + 2^2} = \sqrt{5}, \quad \tan\omega = \frac{2}{-1}.$$

Because $\alpha + i\nu = -1 + i2$ is in the second quadrant, we have $\omega = \pi - \arctan 2$. We calculate

$$(\cos\omega k + i\sin\omega k)\mathbf{v}$$
$$= (\cos\omega k + i\sin\omega k)\begin{bmatrix} \frac{1}{2} + i\,\frac{1}{2} \\ 1 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} \cos\omega k - \sin\omega k + i(\cos\omega k + \sin\omega k) \\ 2\cos\omega k + i2\sin\omega k \end{bmatrix}.$$

By discussion before this example, (5.91) has solutions

$$\mathbf{x}_k^{(1)} = \frac{1}{2}\,5^{k/2}\begin{bmatrix} \cos\omega k - \sin\omega k \\ 2\cos\omega k \end{bmatrix} \quad \text{and} \quad \mathbf{x}_k^{(2)} = \frac{1}{2}\,5^{k/2}\begin{bmatrix} \cos\omega k + \sin\omega k \\ 2\sin\omega k \end{bmatrix}.$$

The general solution of (5.91) is

$$\mathbf{x}_k = c_1\,5^{k/2}\begin{bmatrix} \cos\omega k - \sin\omega k \\ 2\cos\omega k \end{bmatrix} + c_2\,5^{k/2}\begin{bmatrix} \cos\omega k + \sin\omega k \\ 2\sin\omega k \end{bmatrix};$$

the factors of $\frac{1}{2}$ were absorbed in the arbitrary constants c_1, c_2. ◯

5.7.4 Equivalence of Second-Order Scalar Difference Equation and a System in $\mathbb{R}^2$

Given a second-order LCCHΔE

$$y_{k+2} = a_1 y_{k+1} + a_2 y_k, \tag{5.92}$$

we can define

$$\mathbf{x}_k = \begin{bmatrix} x_{1,k} \\ x_{2,k} \end{bmatrix} \triangleq \begin{bmatrix} y_k \\ y_{k+1} \end{bmatrix}.$$

Then we have

$$x_{1,k+1} = y_{k+1} = x_{2,k}$$

and

$$x_{2,k+1} = y_{k+2} = a_2 y_k + a_1 y_{k+1} = a_2 x_{1,k} + a_1 x_{2,k}.$$

So, $\mathbf{x}_k$ should satisfy

$$\mathbf{x}_{k+1} = \begin{bmatrix} x_{2,k} \\ a_2 x_{1,k} + a_1 x_{2,k} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ a_2 & a_1 \end{bmatrix}\mathbf{x}_k.$$

The two types of Casorati determinant that we have defined, one for systems of linear homogeneous difference equations before Theorem 5.19 and the other for scalar linear homogeneous difference equations before Definition 4.8 in Section 4.6, are equal. For example, in $\mathbb{R}^2$,

$$C(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}) = \begin{vmatrix} x_{1,k}^{(1)} & x_{1,k}^{(2)} \\ x_{2,k}^{(1)} & x_{2,k}^{(2)} \end{vmatrix} = \begin{vmatrix} y_k^{(1)} & y_k^{(2)} \\ y_{k+1}^{(1)} & y_{k+1}^{(2)} \end{vmatrix} = C(y_k^{(1)}, y_k^{(2)}).$$

Example 5.34

Solve

$$\mathbf{x}_{k+1} = \begin{bmatrix} 0 & 1 \\ -\frac{1}{4} & 1 \end{bmatrix} \mathbf{x}_k. \tag{5.93}$$

Method: The matrix $A = \begin{bmatrix} 0 & 1 \\ -\frac{1}{4} & 1 \end{bmatrix}$ is in companion form, so it will probably be easier to solve the equivalent second-order scalar LCCHΔE:

$$y_{k+2} = y_{k+1} - \frac{1}{4} y_k.$$

The characteristic polynomial for the latter, $r^2 - r + \frac{1}{4}$, has repeated real root $r = \frac{1}{2}, \frac{1}{2}$, so the general solution is

$$y_k = c_1 \left(\frac{1}{2}\right)^k + c_2 k \left(\frac{1}{2}\right)^k,$$

where c_1, c_2 are arbitrary constants. Correspondingly, the general solution of the original problem, (5.93), is

$$\mathbf{x}_k = \begin{bmatrix} y_k \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} c_1 \frac{1}{2^k} + c_2 k \frac{1}{2^k} \\ c_1 \frac{1}{2^{k+1}} + c_2 (k+1) \frac{1}{2^{k+1}} \end{bmatrix} = \frac{c_1}{2^{k+1}} \begin{bmatrix} 2 \\ 1 \end{bmatrix} + \frac{c_2}{2^{k+1}} \begin{bmatrix} 2k \\ k+1 \end{bmatrix},$$

where c_1, c_2 are arbitrary constants. ◯

5.7.5 Ladder Network Electrical Circuits

One technique electrical engineers use to analyze circuits with a sinusoidal voltage source, $V(t) = \cos \omega t$, is to replace it by its complexification, $\tilde{V}(t) = e^{j\omega t}$. Here we are following the electrical engineering convention of denoting a square root of -1 by j rather than by i as we do in mathematics. A complex exponential current, $\tilde{I} = I_0 e^{j\omega t}$, is the response of a circuit element to a complex exponential voltage source, $\tilde{V}$, and we define their ratio to be the **impedance**:

$$Z \triangleq \frac{\tilde{V}}{\tilde{I}}.$$

Here is a table of impedances corresponding to circuit elements, assuming that L, R, and C are, as usual, constants. Note that the impedances are complex numbers that depend on the frequency of the source:

Element	*Impedance*
Resistor	R
Inductor	$j\omega L$
Capacitor	$(j\omega C)^{-1}$

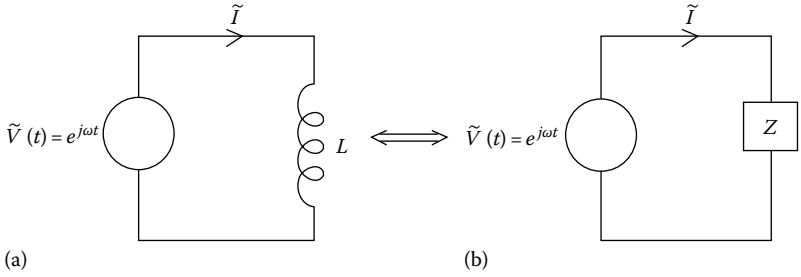


FIGURE 5.8
(a) Inductor in AC circuit and (b) impedance in AC circuit.

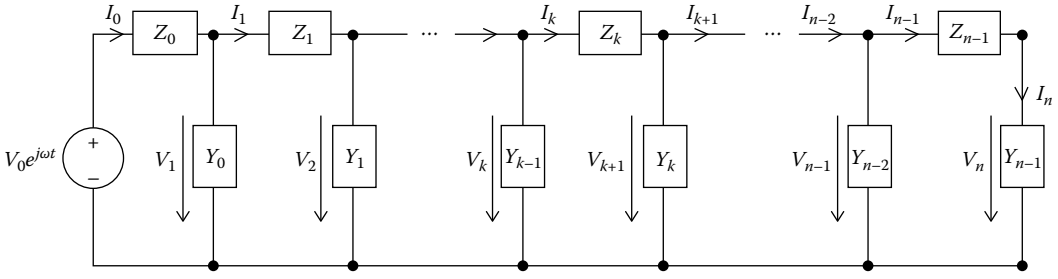


FIGURE 5.9
Ladder network in Example 5.35.

For example, shown in Figure 5.8a is a circuit with only a voltage source and an inductor. Kirchhoff's voltage law implies that the voltage across the inductor, $L\dot{\tilde{I}} = L \cdot (j\omega)I$, equals the voltage source, so $\tilde{V} = j\omega L\tilde{I}$; hence, the impedance is

$$Z = \frac{\tilde{V}}{\tilde{I}} = j\omega L.$$

Figure 5.8b shows an abstract picture of the same circuit as Figure 5.8a, if $Z = j\omega L$.

The **admittance**, Y, is defined to be the reciprocal of the impedance, that is, $Y = 1/Z$.

In the next example and in Figure 5.9, "V_0" is the modulus of the voltage source. If all of the other circuit elements are resistances, then assume the voltage source is constant and we have a DC circuit. If any of the other circuit elements are capacitors or inductors, assume the voltage source is $V_0e^{j\omega t}$ and we have an AC circuit.

Example 5.35

Solve the **ladder network** shown in Figure 5.9. Assume that V_0 is given.

Method: Indicated in the picture are currents into or out of nodes in the circuit. This concept of current is a little "old fashioned" but is convenient for ladder network problems. This is not the loop current concept we used earlier in Examples 1.7 in Section 1.1, 5.1 in Section 5.1, and 5.2 in Section 5.1.

By Kirchhoff's voltage law and the definition of impedance in each of the loops, we have

$$V_0 = V_1 + Z_0 I_0, V_1 = V_2 + Z_1 I_1, \ldots, V_k = V_{k+1} + Z_k I_k, \ldots,$$
$$V_{n-1} = V_n + Z_{n-1} I_{n-1}, \tag{5.94}$$

and by Kirchhoff's current law and the definition of admittance at each of the nodes, we have

$$I_0 = I_1 + Y_0 V_1, I_1 = I_2 + Y_1 V_2, \ldots, I_k = I_{k+1} + Y_k V_{k+1}, \ldots,$$
$$I_{n-2} = I_{n-1} + Y_{n-2} V_{n-1}. \tag{5.95}$$

In addition, we have in the nth loop that

$$I_{n-1} = I_n = Y_{n-1} V_n. \tag{5.96}$$

For $k = 0, \ldots, n-1$, (5.94) yields

$$V_{k+1} = V_k - Z_k I_k. \tag{5.97}$$

Equations (5.97) and (5.95) together imply, for $k = 0, \ldots, n-2$,

$$I_{k+1} = I_k - Y_k V_{k+1} = I_k - Y_k (V_k - Z_k I_k) = -Y_k V_k + (1 + Y_k Z_k) I_k. \tag{5.98}$$

Define $\mathbf{x}_k \triangleq \begin{bmatrix} V_k \\ I_k \end{bmatrix}$. By (5.97) and (5.98), we have

$$\mathbf{x}_{k+1} = A_k \mathbf{x}_k, \tag{5.99}$$

where for $k = 0, \ldots, n-2$,

$$A_k \triangleq \begin{bmatrix} 1 & -Z_k \\ -Y_k & 1 + Y_k Z_k \end{bmatrix}, \tag{5.100}$$

and (5.96) yields

$$A_{n-1} \triangleq \begin{bmatrix} 1 & -Z_{n-1} \\ 0 & 1 \end{bmatrix}.$$

While V_0, the "input" voltage source, is assumed to be known, I_0 is not. This will make our problem more difficult than it just being an LCCHSΔ with given initial conditions.

The solution of (5.99) is, for $k = 1, \ldots, n$,

$$\mathbf{x}_k = A_{k-1} A_{k-2} \cdots A_1 A_0 \mathbf{x}_0,$$

that is,

$$\begin{bmatrix} V_k \\ I_k \end{bmatrix} = A_{k-1}A_{k-2}\cdots A_1A_0 \begin{bmatrix} V_0 \\ I_0 \end{bmatrix}.$$

The solution implies that, after noting that $I_n = Y_{n-1}V_n$,

$$\begin{bmatrix} V_0 \\ I_0 \end{bmatrix} = \mathbf{x}_0 = A_0^{-1}A_1^{-1}\cdots A_{n-1}^{-1}\mathbf{x}_n = A_0^{-1}A_1^{-1}\cdots A_{n-1}^{-1}\begin{bmatrix} V_n \\ Y_{n-1}V_n \end{bmatrix},$$

that is,

$$\begin{bmatrix} V_0 \\ I_0 \end{bmatrix} = V_n \cdot A_0^{-1}A_1^{-1}\cdots A_{n-1}^{-1}\begin{bmatrix} 1 \\ Y_{n-1} \end{bmatrix}. \tag{5.101}$$

From the first component of the vector, it follows that

$$V_0 = \eta V_n,$$

where η is the $(1,1)$ entry of the 2×1 vector $A_0^{-1}A_1^{-1}\cdots A_{n-1}^{-1}\begin{bmatrix} 1 \\ Y_{n-1} \end{bmatrix}$. Since we assumed that V_0 is given, we can solve for

$$V_n = \frac{V_0}{\eta}.$$

When the latter is substituted into (5.101), we can find I_0. Noting that V_n is a scalar, we also see that for $k = 1,\ldots,n-1$,

$$\begin{bmatrix} V_k \\ I_k \end{bmatrix} = A_{k-1}A_{k-2}\cdots A_1A_0\begin{bmatrix} V_0 \\ I_0 \end{bmatrix} = A_{k-1}A_{k-2}\cdots \cancel{A_1A_0}\left(V_n \cancel{A_0^{-1}A_1^{-1}}\cdots A_{n-1}^{-1}\begin{bmatrix} 1 \\ Y_{n-1} \end{bmatrix}\right)$$

$$= V_n\, A_k^{-1}A_{k+1}^{-1}\cdots A_{n-1}^{-1}\begin{bmatrix} 1 \\ Y_{n-1} \end{bmatrix} = \frac{V_0}{\eta}A_k^{-1}A_{k+1}^{-1}\cdots A_{n-1}^{-1}\begin{bmatrix} 1 \\ Y_{n-1} \end{bmatrix}. \bigcirc$$

One interesting concept is that of replacing the whole ladder network by an "equivalent impedance" defined to be

$$Z \triangleq \frac{V_0}{I_n} = \frac{V_0}{Y_{n-1}V_n} = \frac{V_0}{Y_{n-1}\eta^{-1}V_0} = \frac{\eta}{Y_{n-1}}.$$

Example 5.36

For the ladder network shown in Figure 5.10, assume V_0 is an unspecified constant:

(a) Find the equivalent impedance.
(b) Find the voltages V_k for $k = 0,1,\ldots,4$.

Method:

(a) Indicated in the picture are constant source voltage V_0 and impedances $Z_k = 1$ and admittances $Y_k = 1$ for $k = 0,1,2,3$. From (5.100), we have matrices

$$A_k = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix},$$

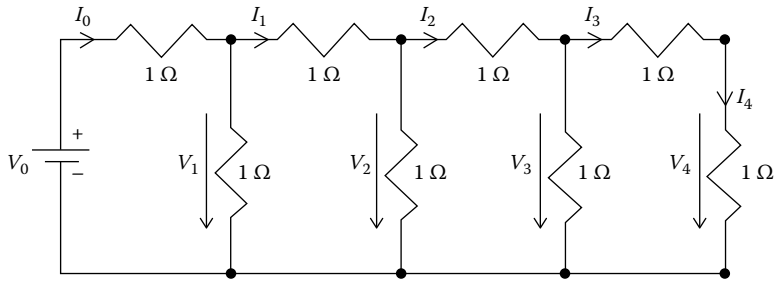


FIGURE 5.10
DC ladder network in Example 5.36.

for $k = 0, 1, 2$ and

$$A_3 = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}.$$

From (5.101) we have

$$\begin{bmatrix} V_0 \\ I_0 \end{bmatrix} = \mathbf{x}_0 = V_4 \, A_0^{-1} A_1^{-1} \cdots A_3^{-1} \begin{bmatrix} 1 \\ Y_3 \end{bmatrix} = V_4 \left(\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \right)^3 \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \cdots = V_4 \begin{bmatrix} 34 \\ 21 \end{bmatrix}.$$

The equivalent impedance is $Z = \eta / Y_3 = 34$.

(b) From the last equation in part (a), we have $I_0 = 21 V_4 = \frac{21}{34} V_0$.
The solution of (5.99) is, for $k = 1, 2, 3$,

$$\mathbf{x}_k = A_{k-1} A_{k-2} \cdots A_1 A_0 \mathbf{x}_0 = \left(\begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \right)^k \begin{bmatrix} V_0 \\ I_0 \end{bmatrix}$$

$$= V_0 \left(\begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \right)^k \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \frac{21}{34} \end{bmatrix}.$$

The eigenvalues and eigenvectors enable the diagonalization

$$\begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 1 - \sqrt{5} & 1 + \sqrt{5} \\ 2 & 2 \end{bmatrix} \begin{bmatrix} \frac{3+\sqrt{5}}{2} & 0 \\ 0 & \frac{3-\sqrt{5}}{2} \end{bmatrix} \left(\frac{1}{4\sqrt{5}} \begin{bmatrix} -2 & 1 + \sqrt{5} \\ 2 & -1 + \sqrt{5} \end{bmatrix} \right).$$

This gives, for $k = 1, 2, 3$,

$$\begin{bmatrix} V_k \\ I_k \end{bmatrix} = P D^k P^{-1} \begin{bmatrix} V_0 \\ I_0 \end{bmatrix}$$

$$= \begin{bmatrix} 1 - \sqrt{5} & 1 + \sqrt{5} \\ 2 & 2 \end{bmatrix} \begin{bmatrix} \left(\frac{3+\sqrt{5}}{2} \right)^k & 0 \\ 0 & \left(\frac{3-\sqrt{5}}{2} \right)^k \end{bmatrix}$$

$$\times \left(\frac{1}{4\sqrt{5}} \begin{bmatrix} -2 & 1 + \sqrt{5} \\ 2 & -1 + \sqrt{5} \end{bmatrix} \right) \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \frac{21}{34} \end{bmatrix}.$$

$$= \cdots = \frac{V_0}{4\sqrt{5}} \begin{bmatrix} -2(1 - \sqrt{5})\lambda_1^k + 2(1 + \sqrt{5})\lambda_2^k & -4\lambda_1^k + 4\lambda_2^k \\ -4\lambda_1^k + 4\lambda_2^k & 2(1 + \sqrt{5})\lambda_1^k + 2(-1 + \sqrt{5})\lambda_2^k \end{bmatrix} \begin{bmatrix} \frac{13}{34} \\ \frac{21}{34} \end{bmatrix},$$

where

$$\lambda_1 = \frac{3+\sqrt{5}}{2}, \quad \lambda_2 = \frac{3-\sqrt{5}}{2}.$$

So, for $k = 0, 1, 2, 3$,

$$V_k = \frac{V_0}{4\sqrt{5}}\left(\frac{13}{34}\left(-2(1-\sqrt{5})\lambda_1^k + 2(1+\sqrt{5})\lambda_2^k\right) + \frac{21}{34}(-4\lambda_1^k + 4\lambda_2^k)\right).$$

Using the facts that $I_4 = I_3$ and that here $V_4 = 1 \cdot I_4$, we have

$$V_4 = \frac{V_0}{4\sqrt{5}}\left(\frac{13}{34}\left(-4\lambda_1^3 + 4\lambda_2^3\right) + \frac{21}{34}\left(2(1+\sqrt{5})\lambda_1^3 + 2(-1+\sqrt{5})\lambda_2^3\right)\right). \bigcirc$$

5.7.6 Stability

Definition 5.11

LCCHSΔ (5.83), that is, $\mathbf{x}_{k+1} = \mathbf{x}_k$ is

(a) **Asymptotically stable** if all its solutions have $\lim_{k\to\infty} \mathbf{x}_k = \mathbf{0}$
(b) **Neutrally stable** if it is not asymptotically stable but all its solutions are bounded for $k = 0, 1, 2, \ldots$, that is,

$$\mathbf{x}_k = \begin{bmatrix} x_{1,k} & x_{2,k} & \cdots & x_{n,k} \end{bmatrix}^T,$$

and for each component $x_{j,k}$, there exists an M_j such that for all $k \geq 0$ we have $|x_{j,k}| \leq M_j$
(c) **Unstable** if it is neither asymptotically stable nor neutrally stable, that is, there is at least one solution that is not bounded for $k = 0, 1, 2, \ldots$

Akin to the stability results for LCCHS in Theorem 5.11 in Section 5.3, we have

Theorem 5.21

LCCHSΔ (5.83), that is, $\mathbf{x}_{k+1} = A\mathbf{x}_k$, is

(a) Asymptotically stable if all of A's eigenvalues μ satisfy $|\mu| < 1$
(b) Neutrally stable if all of A's eigenvalues μ satisfy $|\mu| \leq 1$ and no such μ is both deficient and has modulus equal to 1
(c) Unstable if A has an eigenvalue whose modulus is greater than 1 or is both deficient and has modulus equal to 1

Why? (a) Suppose μ is a real eigenvalue of A with corresponding eigenvector $\mathbf{v}$. Then $\mathbf{x}_k = \mu^k\mathbf{v}$ will be a solution of (5.83), so $|\mu| < 1$ implies $\lim_{k\to\infty} \mathbf{x}_k = \mathbf{0}$. If μ is a nonreal

eigenvalue $\rho e^{i\omega}$ of A with corresponding eigenvector $\mathbf{v}$, then solutions

$$\rho^k \mathcal{R}e\left(e^{i\omega k}\mathbf{v}\right), \quad \rho^k \mathcal{I}m\left(e^{i\omega k}\mathbf{v}\right)$$

will have limit $\mathbf{0}$ as $k \to \infty$ because $\rho = |\mu| < 1$. The explanation for (b) is similar by again using the form of solutions in the two cases μ real and $\mu = \rho e^{i\omega}$. The explanation for (c) is similar but requires more care in the deficient eigenvalue case. □

Example 5.37

Use z**-transforms**, as in Section 4.7, to find the general solution of the homogeneous system of difference equations (LCCHSΔ):

$$\mathbf{x}[n+1] = \begin{bmatrix} 0 & 1 \\ -6 & -5 \end{bmatrix} \mathbf{x}[n],\ n \geq 0.$$

Method: Denote $\mathbf{X}(z) \triangleq \mathcal{Z}\big[\,\mathbf{x}[n]\,\big]$. Take the z-transforms of both sides of the system of difference equations and use (4.77) in Section 4.7 to get

$$z \cdot \mathbf{X}(z) - z\,\mathbf{x}[0] = A\mathbf{X}(z);$$

hence,

$$\mathbf{X}(z) = z \cdot (zI - A)^{-1}\mathbf{x}[0].$$

We do some algebraic calculations:

$$z \cdot (zI - A)^{-1} = z \cdot \left(zI - \begin{bmatrix} 0 & 1 \\ -6 & -5 \end{bmatrix}\right)^{-1} = z \cdot \begin{bmatrix} z & -1 \\ 6 & z+5 \end{bmatrix}^{-1}$$

$$= z \cdot \frac{1}{z^2+5z+6}\begin{bmatrix} z+5 & 1 \\ -6 & z \end{bmatrix}$$

$$= z \cdot \begin{bmatrix} \dfrac{z+5}{z^2+5z+6} & \dfrac{1}{z^2+5z+6} \\ -\dfrac{6}{z^2+5z+6} & \dfrac{z}{z^2+5z+6} \end{bmatrix}.$$

After that, we have to do two partial fractions expansions, from which we can reassemble $z \cdot (zI - A)^{-1}$: first,

$$\frac{1}{z^2+5z+6} = \frac{1}{(z+2)(z+3)} = \frac{A}{z+2} + \frac{B}{z+3} \Rightarrow 1 = A(z+3) + B(z+2).$$

Substituting in $z = -2$ and $z = -3$ gives, respectively, $A = 1$ and $B = -1$.

Second,

$$\frac{z}{z^2+5z+6} = \frac{z}{(z+2)(z+3)} = \frac{C}{z+2} + \frac{D}{z+3} \Rightarrow z = C(z+3) + D(z+2).$$

Substituting in $z = -2$ and $z = -3$ gives, respectively, $C = -2$ and $D = 3$.

Using the partial fraction expansions

$$\frac{1}{z^2+5z+6} = \frac{1}{z+2} - \frac{1}{z+3} \quad \text{and} \quad \frac{z}{z^2+5z+6} = -\frac{2}{z+2} + \frac{3}{z+3},$$

we have

$$z\cdot(zI-A)^{-1}=z\cdot\begin{bmatrix}-\dfrac{2}{z+2}+\dfrac{3}{z+3}+5\Big(\dfrac{1}{z+2}-\dfrac{1}{z+3}\Big) & \dfrac{1}{z+2}-\dfrac{1}{z+3}\\ -6\Big(\dfrac{1}{z+2}-\dfrac{1}{z+3}\Big) & -\dfrac{2}{z+2}+\dfrac{3}{z+3}\end{bmatrix}$$

$$=z\cdot\begin{bmatrix}\dfrac{3}{z+2}-\dfrac{2}{z+3} & \dfrac{1}{z+2}-\dfrac{1}{z+3}\\ -\dfrac{6}{z+2}+\dfrac{6}{z+3} & -\dfrac{2}{z+2}+\dfrac{3}{z+3}\end{bmatrix}.$$

Using (4.75) in Section 4.7, that is, $\mathcal{Z}^{-1}\big[\,z\cdot\dfrac{1}{z-\alpha}\,\big]=\alpha^n$, we find that the solution of the system of homogeneous difference equations is

$$\mathbf{x}[n]=\mathcal{Z}^{-1}\big[\,(zI-A)^{-1}\,\big]\mathbf{x}[0]=\begin{bmatrix}3(-2)^n-2(-3)^n & (-2)^n-(-3)^n\\ -6(-2)^n+6(-3)^n & -2(-2)^n+3(-3)^n\end{bmatrix}\mathbf{x}[0].\ \bigcirc$$

Learn More About It

Our study of ladder networks was influenced by *Circuit Theory, with Computer Applications*, Omar Wing, Holt, Rhinehart and Winston, © 1972 and "Difference equations and their applications," Louis A. Pipes, *Math. Mag.* (**32**), 231–246. The former gives a thorough study of circuits and often includes the use of the fundamental matrix, which it calls a "characteristic matrix." The latter has another interesting network and also an application to coupled oscillators. Another useful reference is *Calculus of Finite Differences and Difference Equations*, by Murray R. Spiegel, Schaum's Outline Series, McGraw-Hill Book Company, © 1971. Section 5.7.1 was inspired by Application 2 in Section 6.3 *Linear Algebra, with Applications, 6th ed.,* by Steven J. Leon, Prentice-Hall, Inc. © 2002.

5.7.7 Problems

In problems 1–6, find the general solution.

1. $\mathbf{x}_{k+1}=\begin{bmatrix}3 & 2\\ 2 & -3\end{bmatrix}\mathbf{x}_k$
2. $\mathbf{x}_{k+1}=\begin{bmatrix}-2 & 7\\ 1 & 4\end{bmatrix}\mathbf{x}_k$
3. $\mathbf{x}_{k+1}=\begin{bmatrix}1 & 2\\ 3 & 2\end{bmatrix}\mathbf{x}_k$
4. $\mathbf{x}_{k+1}=\begin{bmatrix}1 & -2\\ 4 & -3\end{bmatrix}\mathbf{x}_k$
5. $\mathbf{x}_{k+1}=\begin{bmatrix}0 & -1\\ 1 & 0\end{bmatrix}\mathbf{x}_k$

6. $\mathbf{x}_{k+1} = \begin{bmatrix} 1 & -3 & 6 \\ 0 & -2 & 0 \\ 2 & 0 & 0 \end{bmatrix} \mathbf{x}_k$. [Hint: -2 is an eigenvalue]

In problems 7–10, find (a) the equivalent impedance and (b) the voltages V_k indicated in the ladder network shown. Assume V_0 is an unspecified constant.

7. The DC network shown in Figure 5.11
8. The DC network shown in Figure 5.12
9. The AC network shown in Figure 5.13
10. The AC network shown in Figure 5.14

In problems 11 and 12, find the equivalent impedance, where V_0 is an unspecified constant.

11. The DC network shown in Figure 5.15
12. The AC network shown in Figure 5.16

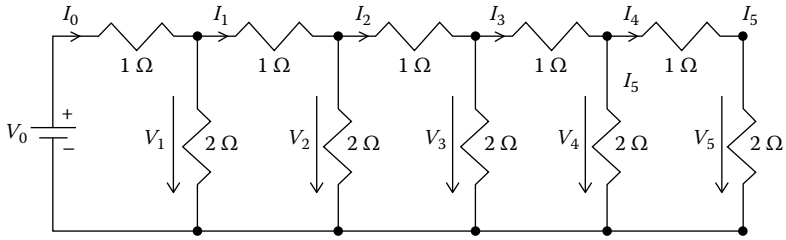


FIGURE 5.11
Problem 5.7.7.7: DC ladder network.

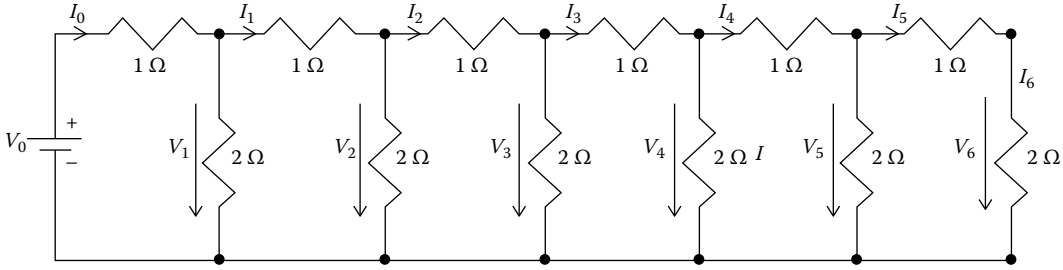


FIGURE 5.12
Problem 5.7.7.8: DC ladder network.

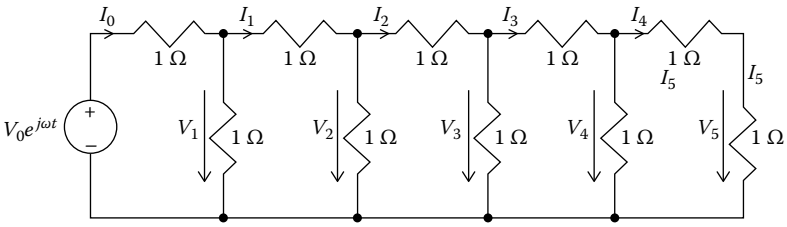


FIGURE 5.13
Problem 5.7.7.9: AC ladder network.

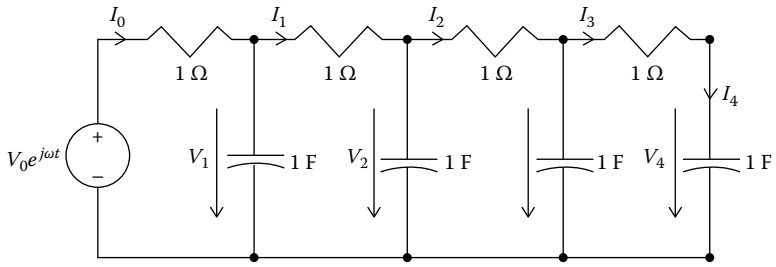


FIGURE 5.14
Problem 5.7.7.10: AC ladder network.

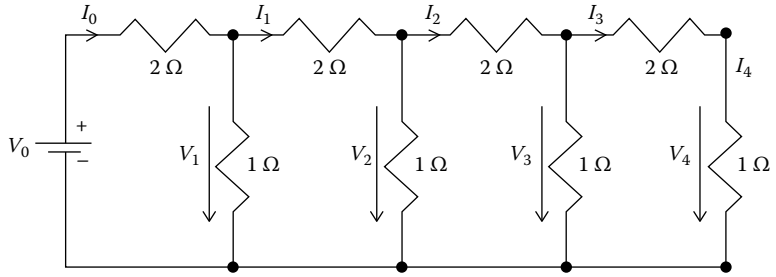


FIGURE 5.15
Problem 5.7.7.11: DC ladder network.

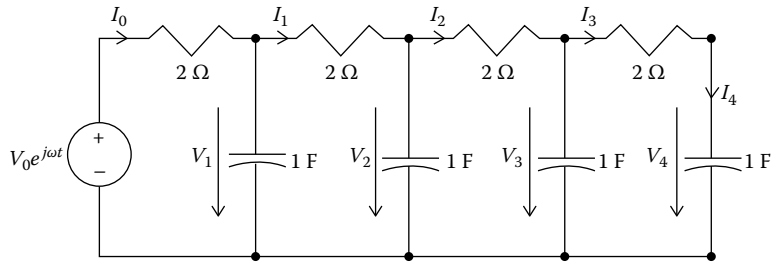


FIGURE 5.16
Problem 5.7.7.12: AC ladder network.

In problems 13–15, solve the system of difference equations using a method of undetermined coefficients analogous to a method for nonresonant nonhomogeneous linear systems of ODEs.

13. $\mathbf{x}_{k+1} = \begin{bmatrix} 1 & 4 \\ 1 & -1 \end{bmatrix} \mathbf{x}_k + \left(\frac{1}{2}\right)^k \begin{bmatrix} 3 \\ 1 \end{bmatrix}$

14. $\mathbf{x}_{k+1} = \begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix} \mathbf{x}_k + \cos(2k) \begin{bmatrix} 3 \\ 1 \end{bmatrix}$

15. $\mathbf{x}_{k+1} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 3 & 1 \\ -1 & -2 & 5 \end{bmatrix} \mathbf{x}_k + \left(\frac{1}{2}\right)^k \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$

16. (Project) Develop a method of undetermined coefficients, analogous to those for scalar ODEs and scalar difference equations and possibly to that for systems of ODEs, which includes resonant cases.
17. Assume A is a real, 5×5, constant matrix and has eigenvalues $\frac{\sqrt{3}}{2} \pm \frac{i}{2}, \frac{\sqrt{3}}{2} \pm \frac{i}{2}, \frac{1}{2}$, including repetitions. Consider the LCCHSΔ $(\star)$ $\mathbf{x}_{k+1} = A\mathbf{x}_k$. For each of (a) through (d), decide whether it must be true, must be false, or may be true and may be false:
    (a) The system is neutrally stable.
    (b) The system is asymptotically stable.
    (c) The system may be neutrally stable, depending upon more information concerning A.
    (d) $(\star)$ has solutions $\mathbf{x}_k$ that are periodic in k with period 6, that is, $\mathbf{x}_{k+6} \equiv \mathbf{x}_k$.
18. If the matrix A has an eigenvalue λ with $|\lambda| = 1$ that *is* deficient, explain why LCCHSΔ $\mathbf{x}_{k+1} = A\mathbf{x}_k$ is not neutrally stable.
19. Explain why Abel's Theorem 4.15 in Section 4.6 for the second-order scalar difference equation $y_{k+2} = a_{1,k}y_{k+1} + a_{2,k}y_k$ follows from Abel's Theorem 5.19 for system of linear homogeneous difference equations (5.82).
20. (Small project) Develop a concept of "fundamental matrix of solutions" for LCCHSΔ $\mathbf{x}_{k+1} = A\mathbf{x}_k$ and implement your concept for a specific example of your choosing.
21. (Small project) Develop a concept analogous to "e^{tA}" for LCCHSΔ $\mathbf{x}_{k+1} = A\mathbf{x}_k$ and implement your concept for a specific example of your choosing.

In problems 22–27, determine if the system $\mathbf{x}_{k+1} = A_k\mathbf{x}_k$ is asymptotically stable, neutrally stable, or unstable. Try to do as little work as is necessary to give a fully explained conclusion.

22. The system of Problem 5.7.7.1
23. The system of Problem 5.7.7.3
24. The system of Problem 5.7.7.4
25. The system of Problem 5.7.7.5
26. $A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$
27. $A = \begin{bmatrix} \frac{1}{2} & \frac{1}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix}$

5.8 Short Take: Periodic Linear Differential Equations

We will study linear periodic ODEs

$$\ddot{y} + p(t)\dot{y} + q(t)y = 0, \tag{5.102}$$

where $p(t), q(t)$ are periodic functions with period T, that is, $p(t + \mathrm{T}) \equiv p(t), q(t + \mathrm{T}) \equiv q(t)$.

First, we will study the general case of systems of first-order ODEs:

$$\dot{\mathbf{x}} = A(t)\mathbf{x}, \tag{5.103}$$

in $\mathbb{R}^n$, where $A(t)$ is an $n \times n$ matrix-valued T-periodic function, that is, satisfying $A(t+\mathsf{T}) \equiv A(t)$. Such problems occur in physical problems with "parametric forcing," for example, when an object is being shaken. Such systems also occur after "linearization" of nonlinear systems, as we will see in Chapter 18.

Suppose we have a fundamental matrix $Z(t)$ for system (5.103). Even though the coefficients in the ODE system are periodic with period T, it may or may not be true that $Z(t)$ or even one of its columns is periodic with period T. It may even happen that a column of $Z(t)$, that is, a vector solution $\mathbf{x}(t)$, is periodic with period 2T but not period T. This is called **period doubling**.

Our work will be easier if we use $X(t) = Z(t)\big(Z(0)\big)^{-1}$, the **principal fundamental matrix at** $t = 0$, instead of $Z(t)$. Because $X(0) = I$, the unique solution of the IVP

$$\left\{\begin{matrix} \dot{\mathbf{x}} = A(t)\mathbf{x} \\ \mathbf{x}(0) = \mathbf{x}_0 \end{matrix}\right\} \tag{5.104}$$

is $\mathbf{x}(t) = X(t)\mathbf{x}_0$.

Suppose (5.103) *does* have a nonzero solution that is T-periodic. Then there is an initial condition vector $\mathbf{x}_0 \neq \mathbf{0}$ such that

$$\mathbf{x}(t + \mathsf{T}) = \mathbf{x}(t)$$

for all t. In particular, it follows that, at $t = 0$, we must have

$$\mathbf{x}(\mathsf{T}) = \mathbf{x}(0), \tag{5.105}$$

that is,

$$X(\mathsf{T})\mathbf{x}_0 = \mathbf{x}_0. \tag{5.106}$$

Pictorially, (5.105) says that $\mathbf{x}(t)$ returns to where it started after T units of time.

Equation (5.106) implies that at $t = \mathsf{T}$, the principal fundamental matrix, $X(\mathsf{T})$, should have $\mu = 1$ as one of its eigenvalues, with corresponding eigenvector being the initial condition, $\mathbf{x}_0$, that produces a T-periodic solution.

Theorem 5.22

System (5.103) has a T-periodic solution if, and only if, $X(\mathsf{T})$ has $\mu = 1$ as one of its eigenvalues.

Why? We've already derived why (5.103) has a T-periodic solution only if $X(\mathsf{T})$ has $\mu = 1$ as one of its eigenvalues.

On the other hand, we should explain why $X(\mathsf{T})$ having $\mu = 1$ as one of its eigenvalues guarantees that (5.103) has a T-periodic solution: if $X(\mathsf{T})\mathbf{x}_0 = \mathbf{x}_0$, then let $\mathbf{x}(t) \triangleq X(t)\mathbf{x}_0$

and let $\mathbf{z}(t) \triangleq X(t+\mathsf{T})\mathbf{x}_0$. We will see in the following why uniqueness of solutions (see Theorem 5.1) in Section 5.1 of IVP (5.104) implies that $\mathbf{x}(t) \equiv \mathbf{z}(t) \equiv \mathbf{x}(t+\mathsf{T})$ so that $\mathbf{x}(t)$ is a T-periodic solution.

First, we know that both $\mathbf{x}(t)$ and $\mathbf{z}(t)$ are solutions of ODE system (5.103), the latter because the chain rule gives us

$$\dot{\mathbf{z}}(t) = \frac{d}{dt}[X(t+\mathsf{T})]\mathbf{x}_0 = \dot{X}(t+\mathsf{T}) \cdot \frac{d}{dt}[t+\mathsf{T}]\mathbf{x}_0 = \dot{X}(t+\mathsf{T})\mathbf{x}_0 = A(t+\mathsf{T})X(t+\mathsf{T})\mathbf{x}_0$$

$$= A(t)\,(X(t+\mathsf{T})\mathbf{x}_0) = A(t)\mathbf{z}(t),$$

after using Theorem 5.5 in Section 5.2, that is, $\dot{X}(t) = A(t)X(t)$.

Second, we have

$$\mathbf{z}(0) = X(0+\mathsf{T})\mathbf{x}_0 = \mathbf{x}_0,$$

so $\mathbf{x}(t)$ and $\mathbf{z}(t)$ satisfy the same initial condition. By uniqueness of solutions of the IVP (5.104), $\mathbf{x}(t) \equiv \mathbf{z}(t)$ and thus $\mathbf{x}(t)$ is a T-periodic solution. □

Corollary 5.2

If $X(t)$ has an eigenvalue $\mu = -1$, then (5.103) has a solution that is periodic with period 2T and is not T-periodic.

Why? Note that $A(t)$ being T-periodic implies that $A(t)$ is also 2T-periodic. The rest of the explanation is similar to that for Theorem 5.22. □

5.8.1 The Stroboscopic, or "Return," Map

Lemma 5.6

$X(t+\mathsf{T}) \equiv X(t)X(\mathsf{T})$.

Why? We postpone the explanation, which is similar to that for Theorem 5.22, to later in this section so as to not interrupt the flow of ideas.

For all initial conditions $\mathbf{x}(0) = \mathbf{x}_0$, we have

$$\mathbf{x}(\mathsf{T}) = X(\mathsf{T})\mathbf{x}(0) = X(\mathsf{T})\mathbf{x}_0,$$

so Lemma 5.6 yields

$$\mathbf{x}(2\mathsf{T}) = X(\mathsf{T}+\mathsf{T})\mathbf{x}(0) = X(\mathsf{T})\mathbf{x}(\mathsf{T}) = \big(X(\mathsf{T})\big)^2\mathbf{x}_0,$$

$$\mathbf{x}(3\mathsf{T}) = \big(X(\mathsf{T})\big)^3\mathbf{x}_0, \ldots.$$

This tells us that for a linear homogeneous periodic system of ODEs, if we look at the periodic returns of $\mathbf{x}(t)$, that is, the sequence $\mathbf{x}(0), \mathbf{x}(\mathsf{T}), \mathbf{x}(2\mathsf{T}), \mathbf{x}(3\mathsf{T}), \ldots$, then we are dealing

with the solution of

$$\mathbf{x}_{k+1} = X(\mathrm{T})\mathbf{x}_k, \tag{5.107}$$

a system of linear constant coefficient homogeneous difference equations LCCHSΔ! Because the sequence $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots$ can be thought of as a sequence of photographs of the state of a physical system, we refer to $X(\mathrm{T})$ as the **stroboscopic map**. For example, think of "Milk Drop Coronet," 1957, by Harold Edgerton, one photograph in a movie of the impact of a drop falling into milk. See http://www.vam.ac.uk/vastatic/microsites/photography/photographer.php?photographerid=ph019&row=4.

As for any other LCCHSΔ, the eigenvalues and eigenvectors of the matrix of coefficients play important roles in the solution. The matrix $X(\mathrm{T})$ is called a **monodromy matrix** and its eigenvalues, $\mu_1, \mu_2, \ldots, \mu_n$, are called the **characteristic multipliers** for system (5.103). Let's explain why they are called "multipliers."

Suppose $\mathbf{v}^{(j)}$ is an eigenvector of $X(\mathrm{T})$ corresponding to an eigenvalue, μ_j. We have

$$X(\mathrm{T})\mathbf{v}^{(j)} = \mu_j \mathbf{v}^{(j)}. \tag{5.108}$$

It follows that

$$X(\mathrm{T})^k \mathbf{v}^{(j)} = \mu_j^k \mathbf{v}^{(j)}.$$

If the initial condition for a solution of the original ODE system (5.103) is $\mathbf{x}(0) = \mathbf{v}^{(j)}$, then we have for $k = 1, 2, ...$

$$\mathbf{x}(k\mathrm{T}) = \mu_j^k \mathbf{x}(0),$$

that is, the system returns to an evermore *multiplied* version of where it started.

Theorem 5.23

If $Z(t)$ is any fundamental matrix for a linear homogeneous T-periodic system (5.103), then the eigenvalues of $Z(\mathrm{T})$ are the characteristic multipliers of system (5.103).

5.8.2 Floquet Representation

Theorem 5.24

(Floquet) For system (5.103), there exists a constant matrix C, possibly complex, and a T-periodic matrix $P(t)$ satisfying

$$X(t) = P(t)e^{tC} \quad \text{and} \quad P(0) = I. \tag{5.109}$$

Before explaining this theorem, first let's explain Lemma 5.6: to explain why $X(t + \mathrm{T}) \equiv X(t)X(\mathrm{T})$, we use a "uniqueness of solutions" explanation similar to what we gave for

Theorem 5.22. Let $Y(t) \triangleq X(t)X(\mathsf{T})$. First, we will explain why $Y(t)$ satisfies the matrix differential equation $\dot{Y}(t) = A(t)Y(t)$: we have

$$\dot{Y}(t) = \left(\dot{X}(t)\right) X(\mathsf{T}) = (A(t)X(t))\, X(\mathsf{T}) = A(t)\,(X(t)X(\mathsf{T}) = A(t)Y(t).$$

Second, let $U(t) \triangleq X(t + \mathsf{T})$. We will explain why $U(t)$ satisfies the matrix differential equation $\dot{U}(t) = A(t)U(t)$: by the chain rule and periodicity of $A(t)$,

$$\dot{U}(t) = \frac{d}{dt}[X(t + \mathsf{T})] = \dot{X}(t + \mathsf{T}) \cdot \frac{d}{dt}[t + \mathsf{T}] = \dot{X}(t + \mathsf{T}) = A(t + \mathsf{T})X(t + \mathsf{T}) = A(t)U(t).$$

Now,

$$Y(0) = X(0)X(\mathsf{T}) = IX(\mathsf{T}) = X(\mathsf{T}) \text{ and } U(0) = X(0 + \mathsf{T}) = X(\mathsf{T}).$$

Because $Y(t)$ and $U(t)$ satisfy both the same ODE and the same initial condition, it follows that $Y(t) \equiv U(t)$, that is, $X(t)X(\mathsf{T}) \equiv X(t + \mathsf{T})$. □

We now give a *partial* explanation why the Floquet representation theorem is true in the sense of explaining how to find C and $P(t)$; we will explain later what other details we're omitting. *If* we could find C, then by periodicity of $P(t)$, T would need to satisfy

$$X(\mathsf{T}) = P(\mathsf{T})e^{\mathsf{T}C} = P(0)e^{\mathsf{T}C} = Ie^{\mathsf{T}C};$$

hence, we need to choose C so that

$$e^{\mathsf{T}C} = X(\mathsf{T}). \tag{5.110}$$

Later we will discuss solving (5.110) for C; for now, let's suppose that we have found such a C. Then we would use (5.109) to find $P(t)$:

$$P(t) \triangleq X(t)\left(e^{tC}\right)^{-1} = X(t)e^{-tC}.$$

We need to explain why $P(t)$ is periodic with period T and satisfies $P(0) = I$. Using Lemmas 5.6 and 5.1 in Section 5.2 and Theorem 5.8 in Section 5.2, we calculate that

$$P(t + \mathsf{T}) = X(t + \mathsf{T})\left(e^{-(t+\mathsf{T})C}\right) = X(t)X(\mathsf{T})\left(e^{-\mathsf{T}C}e^{-tC}\right)$$

$$= X(t)\left(X(\mathsf{T})e^{-\mathsf{T}C}\right)e^{-tC} = X(t)Ie^{-tC} = P(t),$$

as we desired. Also, $P(0) = X(0) \cdot e^{O} = I \cdot I = I$, as we desired.

As for solving (5.110) for C, we will only look at the special case where $X(\mathsf{T})$ is diagonalizable and all of its eigenvalues are positive, real numbers. Write

$$X(\mathsf{T}) = QDQ^{-1},$$

where the diagonal matrix $D = diag(\mu_1, \ldots, \mu_n)$ and the eigenvalues of D equal those of $X(\mathrm{T})$, that is, $\mu_1, \ldots, \mu_n$ are the eigenvalues of $X(\mathrm{T})$. If we try to find C in the form $C = SES^{-1}$, where E is diagonal, then we want to solve

$$QDQ^{-1} = X(\mathrm{T}) = e^{\mathrm{T}C} = e^{\mathrm{T}SES^{-1}} = S\left(e^{\mathrm{T}E}\right)S^{-1}.$$

But, we can do that. Let $S = Q$ and

$$E = diag\left(\frac{1}{\mathrm{T}}\ln(\mu_1), \ldots, \frac{1}{\mathrm{T}}\ln(\mu_n)\right). \square$$

Asides: If an eigenvalue of a monodromy matrix is a negative real number or is complex, then we need greater care in constructing the matrix E because we don't yet have a concept of the natural logarithm of such a number. Also, if a monodromy matrix is not diagonalizable, then we would could use a more general result called the "Jordan normal form" to construct a Floquet representation of solutions.

Example 5.38

Find a Floquet representation for solutions and characteristic multipliers of

$$\begin{Bmatrix} \dot{x}_1 = & \left(-1 + \frac{\sin t}{2+\cos t}\right)x_1 & \\ \dot{x}_2 = & & -x_2 \quad +2(\sin t)x_1 \end{Bmatrix}.$$

Method: The system is periodic with period $\mathrm{T} = 2\pi$. The first equation in the system is solvable using the method of separation of variables:

$$\ln|x_1| = c + \int\left(-1 + \frac{\sin t}{2+\cos t}\right)dt = -t - \ln|2+\cos t|,$$

where c is an arbitrary constant, yields

$$x_1(t) = c_1 e^{-t}(2+\cos t)^{-1}.$$

The initial value is

$$x_1(0) = \frac{1}{3}c_1,$$

so the general solution of the first ODE in the system can be written as

$$x_1(t) = e^{-t}\cdot\frac{3}{2+\cos t}\cdot x_1(0).$$

Substitute that into the second ODE in the system, rearrange terms to put it into the standard form of a first-order linear ODE, and multiply through by the integrating factor of e^t to get

$$\frac{d}{dt}\left[\,e^t x_2\,\right] = e^t(\dot{x}_2 + x_2) = e^t\cdot 2\sin t\cdot x_1 = e^t e^{-t}\cdot 2\sin t\cdot\frac{3}{2+\cos t}\cdot x_1(0) = \left(\frac{6\sin t}{2+\cos t}\right)x_1(0).$$

Indefinite integration with respect to t of both sides yields

$$e^t x_2(t) = c_2 - 6\ln(2+\cos t)x_1(0),$$

so

$$x_2(t) = e^{-t}\Big(c_2 - 6\ln(2+\cos t)x_1(0)\Big).$$

The initial value is

$$x_2(0) = c_2 - (6\ln 3)x_1(0),$$

so the solutions of the second ODE can be written in the form

$$x_2(t) = e^{-t}\Big(6\ln 3 - 6\ln(2+\cos t)\Big)x_1(0) + e^{-t}x_2(0).$$

To find a fundamental matrix, first summarize the general solution by writing it in matrix times vector form:

$$\mathbf{x}(t) = \begin{bmatrix} e^{-t}\,\dfrac{3}{2+\cos t}\,x_1(0) \\ \\ 6\ln\Big(\dfrac{3}{2+\cos t}\Big)x_1(0) + e^{-t}x_2(0) \end{bmatrix} = \begin{bmatrix} e^{-t}\,\dfrac{3}{2+\cos t} & 0 \\ \\ e^{-t}\,6\ln\Big(\dfrac{3}{2+\cos t}\Big) & e^{-t} \end{bmatrix}\mathbf{x}(0).$$

So, a fundamental matrix is given by

$$X(t) = e^{-t}\begin{bmatrix} \dfrac{3}{2+\cos t} & 0 \\ \\ 6\ln\Big(\dfrac{3}{2+\cos t}\Big) & 1 \end{bmatrix}.$$

In particular, substitute in $t = 2\pi$ to see that in this problem $X(\mathrm{T}) = X(2\pi) = e^{-2\pi}\,I$, so the characteristic multipliers, being the eigenvalues of $X(\mathrm{T})$, are $\mu = e^{-2\pi},\ e^{-2\pi}$. We can take $Q = I, D = e^{-2\pi}I$ in the calculation of the Floquet representation; hence, $S = Q = I$,

$$E = diag\left(\frac{1}{2\pi}\ln(e^{-2\pi}), \frac{1}{2\pi}\ln(e^{-2\pi})\right) = -I,$$

and $C = -I$. The Floquet representation here is $X(t) = P(t)e^{tC} = P(t)\left(e^{-t}I\right)$, so in this example,

$$P(t) = e^{t}X(t) = \begin{bmatrix} \frac{3}{2+\cos t} & 0 \\ \\ 6\ln(\frac{3}{2+\cos t}) & 1 \end{bmatrix}.$$

To summarize, a Floquet representation is given by

$$X(t) = P(t)e^{tC} = \left(\begin{bmatrix} \dfrac{3}{2+\cos t} & 0 \\ \\ 6\ln\Big(\dfrac{3}{2+\cos t}\Big) & 1 \end{bmatrix}\right)\left(e^{-t}I\right). \bigcirc$$

5.8.3 Stability

The following definitions for periodic systems are the same as for LCCHS given in Definition 5.7 in Section 5.3 and are similar to the definitions for LCCHSΔ given in Definition 5.10 in Section 5.7.

Definition 5.12

Linear homogeneous periodic system (5.103) is

(a) **Asymptotically stable** if all its solutions have $\lim_{t\to\infty} \mathbf{x}(t) = \mathbf{0}$
(b) **Neutrally stable** if it is not asymptotically stable but all its solutions are bounded on $[0, \infty)$, that is, for $j = 1, \ldots, n$, each component $x_j(t)$ of $\mathbf{x}(t)$ there exists an M_j such that for all $t \geq 0$, we have $|x_j(t)| \leq M_j$
(c) **Unstable** if it is neither asymptotically stable nor neutrally stable, that is, there is at least one solution that is not bounded on $[0, \infty)$

Akin to the stability results for LCCHS in Theorem 5.11, we have

Theorem 5.25

Linear homogeneous periodic system (5.103) is

(a) Asymptotically stable if all characteristic multipliers μ satisfy $|\mu| < 1$
(b) Neutrally stable if all characteristic multipliers μ satisfy $|\mu| \leq 1$ and no such μ is both deficient and has modulus 1
(c) Unstable if there is a characteristic multiplier whose modulus is greater than 1 or is both deficient and has modulus equal to 1

Why? To be very brief, an explanation for this result follows from the Floquet representation of solutions and Theorem 5.21 for LCCHSΔ. □

It is very tempting to think that we can determine stability of a linear homogeneous periodic system (5.103) by looking at the eigenvalues of the matrix $A(t)$. Unfortunately, this is not possible! In the following example, the matrix $A(t)$ has eigenvalues that have negative real part for all t, yet the system is not even neutrally stable, let alone asymptotically stable. The results of Theorem 5.11 in Section 5.3 for LCCHS don't carry over to a linear homogeneous periodic system (5.103).

Example 5.39

The system (Markus and Yamabe, 1960)

$$\dot{\mathbf{x}} = \begin{bmatrix} -1 + \frac{3}{2}\cos^2(t) & 1 - \frac{3}{2}\cos t \sin t \\ -1 - \frac{3}{2}\cos t \sin t & -1 + \frac{3}{2}\sin^2(t) \end{bmatrix} \mathbf{x}$$

has a solution $\mathbf{x}(t) = e^{t/2}\begin{bmatrix} -\cos t \\ \sin t \end{bmatrix}$. The system is neither asymptotically stable nor neutrally stable, even though the eigenvalues of $A(t)$ are $\lambda(t) = \dfrac{-1 \pm i\sqrt{7}}{4}$, which have negative real part. ○

For an ODE, the time constant indicates how long it takes for a solution to decay to $\frac{1}{e}$ of its initial value. Suppose a linear homogeneous periodic system (5.103) is asymptotically stable. The time constant τ for that system of ODEs can be *defined* by

$$\tau = \frac{1}{r_{min}},$$

where $0 < r_{min} = -\max\{\ln(|\mu|) : \mu \text{ is a characteristic multiplier}\}$ is the slowest decay rate. Because each solution $\mathbf{x}(t)$ may include many different decaying functions, "weighted" by time-periodic vectors, we *can't* guarantee that $\mathbf{x}(\tau) = \frac{1}{e}\mathbf{x}(0)$. Nevertheless, for physical intuition, it is still useful to think of the time constant as being about how long it takes for the solution to decay in a standard way.

5.8.4 Hill's Equation

Assume $q(t)$ is T-periodic. The second-order scalar ODE

$$\ddot{y} + (\lambda + q(t))y = 0 \tag{5.111}$$

is called **Hill's equation**. We will see later why λ is called an "eigenvalue parameter."

For each value of λ, we can choose two solutions $y_1(t;\lambda), y_2(t;\lambda)$ that satisfy the IVPs

$$\left\{\begin{matrix} \ddot{y}_1 + (\lambda + q(t))y_1 = 0 \\ y_1(0;\lambda) = 1,\ \ \dot{y}_1(0;\lambda) = 0 \end{matrix}\right\}, \quad \left\{\begin{matrix} \ddot{y}_2 + (\lambda + q(t))y_2 = 0 \\ y_2(0;\lambda) = 0,\ \ \dot{y}_2(0;\lambda) = 1 \end{matrix}\right\},$$

just as we did in explaining why Theorem 5.3 in Section 5.2 was true. [The notation $y(t;\lambda)$ indicates that y is a function of t that also depends upon the parameter λ that is in ODE (5.111)]. We form the matrix

$$X(t;\lambda) \triangleq \begin{bmatrix} y_1(t;\lambda) & y_2(t;\lambda) \\ \dot{y}_1(t;\lambda) & \dot{y}_2(t;\lambda) \end{bmatrix}$$

and it is the principal fundamental matrix for ODE system

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ -\lambda - q(t) & 0 \end{bmatrix}\mathbf{x}. \tag{5.112}$$

Example 5.40

Find the principal fundamental matrix for (5.112) in the special case that $\lambda > 0$ and $q(t) \equiv 0$.

Method: The second-order ODE is the undamped harmonic oscillator model $\ddot{y} + \lambda y = 0$. It has solutions $y_1(t;\lambda) = \cos(\sqrt{\lambda}\ t$ and $y_2(t;\lambda) = \frac{1}{\sqrt{\lambda}}\sin(\sqrt{\lambda}\ t)$, so the principal fundamental matrix is given by

$$X(t;\lambda) \triangleq \begin{bmatrix} \cos(\sqrt{\lambda}\ t) & \frac{1}{\sqrt{\lambda}}\sin(\sqrt{\lambda}\ t) \\ -\sqrt{\lambda}\sin(\sqrt{\lambda}\ t) & \cos(\sqrt{\lambda}\ t) \end{bmatrix}. \bigcirc$$

Returning to the general situation, the monodromy matrix is $X(\mathsf{T};\lambda)$. The existence of a T-periodic solution of Hill's equation is equivalent to

$$\begin{aligned}
0 = |\, X(\mathsf{T};\lambda) - 1\cdot I \,| &= \begin{vmatrix} y_1(\mathsf{T};\lambda) - 1 & y_2(\mathsf{T};\lambda) \\ \dot{y}_1(\mathsf{T};\lambda) & \dot{y}_2(\mathsf{T};\lambda) - 1 \end{vmatrix} \\
&= \big(y_1(\mathsf{T};\lambda) - 1\big)\big(\dot{y}_2(\mathsf{T};\lambda) - 1\big) - y_2(\mathsf{T};\lambda)\dot{y}_1(\mathsf{T};\lambda) \\
&= \big(y_1(\mathsf{T};\lambda)\dot{y}_2(\mathsf{T};\lambda) - y_2(\mathsf{T};\lambda)\dot{y}_1(\mathsf{T};\lambda)\big) + 1 - \dot{y}_2(\mathsf{T};\lambda) - y_1(\mathsf{T};\lambda) \\
&= (1) + 1 - \dot{y}_2(\mathsf{T};\lambda) - y_1(\mathsf{T};\lambda) \triangleq \mathsf{H}(\lambda).
\end{aligned}$$

Why? Because Abel's Theorem 3.12 in Section 3.3 and $p(t) \equiv 0$ imply that the Wronskian determinant satisfies

$$\begin{aligned}
|X(\mathsf{T};\lambda)| &= y_1(\mathsf{T};\lambda)\dot{y}_2(\mathsf{T};\lambda) - y_2(\mathsf{T};\lambda)\dot{y}_1(\mathsf{T};\lambda) = W(y_1, y_2)(\mathsf{T}) \\
&= \exp\left(-\int_0^{\mathsf{T}} 0 d\tau\right) W(y_1, y_2)(0) = e^0\left(y_1(0;\lambda)\dot{y}_2(0;\lambda) - y_2(0;\lambda)\dot{y}_1(0;\lambda)\right) \\
&= 1\cdot|X(0;\lambda)| = 1.
\end{aligned}$$

The function $\mathsf{H}(\lambda) \triangleq 2 - \dot{y}_2(\mathsf{T};\lambda) - y_1(\mathsf{T};\lambda)$ is known as the **Hill's discriminant**. Using a modern ODE-IVP numerical "solvers" such as MATLAB's ode45, which we will discuss briefly in Chapter 8, we can find approximate numerical solutions, λ, of the "characteristic equation":

$$\mathsf{H}(\lambda) = 0.$$

We call these values of λ eigenvalues, analogous to finding the eigenvalues of a matrix by solving the characteristic equation $\mathcal{P}(\lambda) = 0$.

Hill's equation originally came up in a model of the three-body celestial dynamics problem of predicting the orbit of the moon under the gravitational influence of both the earth and the sun. A special case of Hill's equation is the Mathieu equation, which models vibration problems including those for machinery and oil rigs on the ocean. Hill's equation also can model vibrations of a crystal lattice.

5.8.5 Periodic Solution of a Nonhomogeneous ODE System

Here we will study the existence of a T-periodic solution of a nonhomogeneous system:

$$\dot{\mathbf{x}} = A(t)\mathbf{x}(t) + \mathbf{f}(t), \tag{5.113}$$

where the matrix $A(t)$ and the vector $\mathbf{f}(t)$ are both periodic with period T. Using the variation of parameters formula (5.41) in Section 5.4, the solution of (5.113) is

$$\mathbf{x}(t) = X(t)\left(\mathbf{x}_0 + \int_0^t (X(s))^{-1}\,\mathbf{f}(s)ds\right).$$

Just as for a linear homogeneous solution, the $\mathbf{x}(t)$ is a T-periodic solution of (5.113) if, and only if,

$$\mathbf{x}(0) = \mathbf{x}(\mathrm{T}),$$

that is,

$$\mathbf{x}_0 = \mathbf{x}(0) = \mathbf{x}(\mathrm{T}) = X(\mathrm{T})\mathbf{x}_0 + X(\mathrm{T})\int_0^{\mathrm{T}} (X(s))^{-1}\,\mathbf{f}(s)ds,$$

that is,

$$(I - X(\mathrm{T}))\,\mathbf{x}_0 = X(\mathrm{T})\int_0^{\mathrm{T}} (X(s))^{-1}\,\mathbf{f}(s)ds. \tag{5.114}$$

So, (5.113) has a T-periodic solution $\mathbf{x}(t)$ if, and only if, (5.114) has a solution $\mathbf{x}_0$.

Theorem 5.26

(Noncritical systems) If the corresponding linear homogeneous system of ODEs (5.103), that is, $\dot{\mathbf{x}} = A(t)\mathbf{x}$, has no T-periodic solution, then the nonhomogeneous system (5.113) is guaranteed to have a T-periodic solution.

Why? By Theorem 5.22, if $\dot{\mathbf{x}} = A(t)\mathbf{x}$ has no T-periodic solution, then $\mu = 1$ is not a characteristic multiplier for the corresponding homogeneous system (5.103), that is, $\mu = 1$ is not an eigenvalue of the monodromy matrix $X(\mathrm{T})$ for (5.103), so

$$0 \neq |\,X(\mathrm{T}) - I\,| = (-1)^n\,|\,I - X(\mathrm{T})|\,.$$

It follows that $(I - X(\mathrm{T}))$ is invertible, so we *can* solve (5.114):

$$\mathbf{x}_0 = \big(I - X(\mathrm{T})\big)^{-1}\left(X(\mathrm{T})\int_0^{\mathrm{T}} (X(s))^{-1}\,\mathbf{f}(s)ds\right).$$

So, in the noncritical case, a T-periodic solution is given by

$$\mathbf{x}(t) = X(t)\left((I - X(\mathsf{T}))^{-1} X(\mathsf{T}) \left(\int_0^{\mathsf{T}} (X(s))^{-1} \mathbf{f}(s)ds\right) + \int_0^t (X(s))^{-1} \mathbf{f}(s)ds\right). \square \qquad (5.115)$$

Theorem 5.13 in Section 5.5, the nonresonant case of sinusoidal forcing, is a special case of Theorem 5.26. If A is a constant matrix, then A is T-periodic for every T. The LCCHS $\dot{\mathbf{x}} = A\mathbf{x}$ is nonresonant for sinusoidal forcing if $\pm i\omega$ are not eigenvalues of A, in which case the nonhomogeneous system

$$\dot{\mathbf{x}} = A\mathbf{x} + (\cos \omega t)\mathbf{k},$$

where $A, \mathbf{k}$ are constant, has a T-periodic solution:

$$\mathbf{x}(t) = -\mathcal{R}e\left(e^{i\omega t}(A - i\omega I)^{-1}\mathbf{k}\right).$$

Example 5.41

Define the square wave function $f(t)$ as defined in (4.30) in Section 4.5, that is,

$$f(t) = \begin{Bmatrix} 1, & 0 < t < \frac{\mathsf{T}}{2} \\ -1, & \frac{\mathsf{T}}{2} < t < \mathsf{T} \end{Bmatrix}.$$

For ODE $\dot{y} + 2y = f(t)$, (a) find a periodic solution, and (b) find the steady-state solution.

Method:

(a) Using Laplace transforms as in Example 4.34 in Section 4.5 or using the method of variation of parameters in (5.44) in Section 5.4, we find that the solutions of the ODE are given by

$$y(t) = e^{-2t}y_0 + e^{-2t}\int_0^t e^{2u}f(u)du,$$

where $y(0) = y_0$. A periodic solution $y(t)$ must have $y_0 = y(\mathsf{T})$, that is,

$$y_0 = e^{-2\mathsf{T}}y_0 + e^{-2\mathsf{T}}\int_0^{\mathsf{T}} e^{2u}f(u)du.$$

Take the term that involves y_0 to the left-hand side, factor out y_0, and then divide through by $(1 - e^{-2\mathsf{T}})$ to get

$$y_0 = y_0^\star = \frac{e^{-2\mathsf{T}}}{1 - e^{-2\mathsf{T}}}\int_0^{\mathsf{T}} e^{2u}f(u)du.$$

In Example 4.34 in Section 4.5, implicitly we found that

$$e^{-2\mathsf{T}}\int_0^{\mathsf{T}} e^{2u}f(u)du = \frac{1}{2}\left(-1 - e^{-2\mathsf{T}} + 2e^{-\mathsf{T}}\right) = -\frac{1}{2}\left(1 - e^{-\mathsf{T}}\right)^2.$$

The ODE has exactly one periodic solution:

$$y_P(t) = e^{-2t}\left(\frac{e^{-2\mathsf{T}}}{1-e^{-2\mathsf{T}}}\cdot\left(-\frac{1}{2}\left(1-e^{-\mathsf{T}}\right)^2\right)+\int_0^t e^{2u}f(u)\,du\right)$$

$$= e^{-2t}\left(-\frac{1}{2}\,\frac{e^{-2\mathsf{T}}}{(1-e^{-\mathsf{T}})(1+e^{\mathsf{T}})}\cdot\left(\left(1-e^{-\mathsf{T}}\right)^2\right)+\int_0^t e^{2u}f(u)\,du\right)$$

$$= e^{-2t}\left(-\frac{1}{2}\,\frac{e^{-2\mathsf{T}}\left(1-e^{-\mathsf{T}}\right)}{(1+e^{\mathsf{T}})}+\int_0^t e^{2u}f(u)\,du\right).$$

(b) All solutions of the ODE can be written in the form

$$y(t) = e^{-2t}y_0 + e^{-2t}\int_0^t e^{2u}f(u)\,du = e^{-2t}(y_0 - y_0^\star + y_0^\star) + e^{-2t}\int_0^t e^{2u}f(u)\,du$$

$$= e^{-2t}\,(y_0 - y_0^\star) + y_P(t).$$

The transient solution is $e^{-2t}\,(y_0 - y_0^\star)$ and the steady-state solution is $y_S(t) = y_P(t)$. ○

By the way, the solution graphed in Figure 4.19 is not the periodic solution because it has $y(0) = 0 \neq y_0^\star$. Nevertheless, we can see a steady-state solution hiding in Figure 4.19.

Learn More About It

A standard, useful reference to linear periodic homogeneous ODEs is *Hill's Equation*, Wilhelm Magnus and Stanley Winkler, John Wiley & Sons, © 1966 (or Dover Publications, © 1979). Example 5.38 and Problems 5.8.6.9 and 5.8.6.10 were inspired by an example in *NonLinear Ordinary Differential Equations*, by R. Grimshaw, Blackwell Scientific Publications, © 1990. A version of Problem 5.8.6.7 is on p. 82 of *A Second Course in Elementary Differential Equations*, by Paul Waltman, Academic Press, Inc., © 1986. Problem 5.8.6.11 is on pp. 166–167 in *Differential Equations, Classical to Controlled*, Dahlard L. Lukes, Academic Press, Inc., © 1982. Hill's equation has its origins in "On the part of the motion of lunar perigee which is a function of the mean motions of the sun and moon," by G. W. Hill, *Acta Math.* **8** (1886), 1–36.

5.8.6 Problems

1. (a) For the system $(\star)$ $\dot{\mathbf{x}} = \begin{bmatrix} 1+\sin t & 0 \\ e^{\cos t} & 0 \end{bmatrix}\mathbf{x}$, explain why $Z(t) = \begin{bmatrix} e^{t-\cos t} & 0 \\ e^t & 1 \end{bmatrix}$ is a fundamental matrix.
   (b) Find the corresponding Floquet representation for $(\star)$.
2. For the scalar ODE $\dot{y} = (\alpha + \sin t)y$, where α is an unspecified constant, (a) find the Floquet representation, and (b) discuss stability for the three cases $\alpha < 0$, $\alpha = 0, \alpha > 0$.

3. What does the Floquet theorem say about the scalar linear T-periodic ODE $\dot{y} = p(t)y$? What does the method of integrating factor say about that ODE, as well as the corresponding scalar linear nonhomogeneous ODE T-periodic ODE $\dot{y} = p(t)y + f(t)$? Also, if $\int_0^{\mathsf{T}} p(t)\,dt < 0$, what can you say about the solutions $y(t)$?
4. In Example 5.39, we gave one solution of a linear periodic homogeneous system $(\star)$.
   (a) Think of that solution as the first column of a fundamental matrix $Z(t)$. Use that solution to find an eigenvector of $Z(t)$ and to explain why e^{π} is an eigenvalue of $Z(2\pi)$. Then, use Theorem 5.23 to explain why e^{π} is a characteristic multiplier for system $(\star)$.
   (b) Explain why $|B| = \gamma_1\gamma_2$ if B is a 2×2 matrix with eigenvalues γ_1, γ_2. [This fact is totally independent of the context of studying system $(\star)$, that is, this is just a fact about eigenvalues of matrices.]
   (c) Use your results from parts (a) and (b), along with the results of Abel's theorem for systems, that is, Theorem 5.7 in Section 5.2, and Theorem 5.23. To explain why $e^{-2\pi}$ is a second multiplier for system $(\star)$. The trace of $A(t)$ is $-\frac{1}{2}$ for system $(\star)$.
5. Consider a T-periodic Hill's equation (5.111), that is, $(\star)$ $\ddot{y}+(\lambda+q(t))y = 0$. Assume that the corresponding solutions $y_1(t;\lambda), y_2(t;\lambda)$ have

$$|\dot{y}_2(\mathsf{T};\lambda) + y_1(\mathsf{T};\lambda)| < 2.$$

   Explain why all solutions of $(\star)$ is neutrally stable.
6. Consider a T-periodic Hill's equation (5.111), that is, $(\star)$ $\ddot{y}+(\lambda+q(t))y = 0$. Assume that the corresponding solutions $y_1(t;\lambda), y_2(t;\lambda)$ have

$$|\dot{y}_2(\mathsf{T};\lambda) + y_1(\mathsf{T};\lambda)| > 2.$$

   Explain why $(\star)$ has a solution satisfying $\lim_{t\to\infty} |y(t)| = \infty$.
7. Suppose $X(t)$ is the principal fundamental matrix at $t = 0$ for a T-periodic system of ODEs $(\star)$ $\dot{\mathbf{x}} = A(t)\mathbf{x}$. Suppose also that $A(-t) \equiv -A(t)$, that is, $A(t)$ is an *odd* function. Recall from Problem 5.2.5.27 that it follows that $X(-t) \equiv X(t)$. Explain why *all* solutions of $(\star)$ are T-periodic by using the steps in the following:
   (a) First, explain why *if* we can explain why $X(\mathsf{T}) = I$, then we can conclude that all solutions of $(\star)$ are T-periodic.
   (b) Use the result of Problem 5.2.5.27 to explain why $X(-\frac{\mathsf{T}}{2} + \mathsf{T}) = X(\frac{\mathsf{T}}{2} - \mathsf{T})$.
   (c) Use Lemma 5.6 to explain why $X(-\frac{\mathsf{T}}{2} + \mathsf{T}) = X(-\frac{\mathsf{T}}{2})X(\mathsf{T})$.
   (d) Use the results of parts (c) and (b) to explain why $X(-\frac{\mathsf{T}}{2})X(\mathsf{T}) = X(-\frac{\mathsf{T}}{2})$.
   (e) Use the result of part (e) and the invertibility of $X(t)$ for all t to explain why $X(\mathsf{T}) = I$.
8. Define a 2π-periodic function

$$b(t) \triangleq \begin{Bmatrix} h_1, & 0 < t < \pi \\ h_2, & \pi < t < 2\pi \end{Bmatrix}.$$

For the scalar ODE $\ddot{y} + b(t)\dot{y} + y = 0$,

(a) Find the characteristic multipliers for the equivalent 2π-periodic system.

(b) Explain why the system is asymptotically stable if, and only if, $h_1 + h_2 > 0$.

(c) Interpret physically the result of part (c).

In problems 9 and 10, find the Floquet representation for solutions of the given system. One can use separation of variables to solve the ODE involving only x_1, substitute that solution into the ODE involving $\dot{x}_2$, and solve that ODE using an integrating factor. After that you'll be able to find a fundamental matrix, find the monodromy matrix [Hint: What's the period of the coefficients in the system of ODEs?], and continue from there.

9. $\begin{Bmatrix} \dot{x}_1 = & \left(-1 - \frac{\cos t}{2+\sin t}\right) x_1 & \\ \dot{x}_2 = & 2(\cos t)x_1 & -x_2 \end{Bmatrix}$

10. $\begin{Bmatrix} \dot{x}_1 = & \left(-1 + \frac{\sin t}{2+\cos t}\right) x_1 & \\ \dot{x}_2 = & -(\sin t)x_1 & -x_2 \end{Bmatrix}$

11. Suppose A_0 and Ω are real, constant $n \times n$ matrices, suppose that $e^{t\Omega}$ is T-periodic, and suppose that $A(t) \triangleq e^{t\Omega}A_0e^{-t\Omega}$. Explain why $\dot{\mathbf{x}} = A(t)\mathbf{x}$ has Floquet representation $X(t) = P(t)e^{tC}$ where $P(t) = e^{t\Omega}$ and $C = A_0 - \Omega$. [Hint: Use (5.29) in Section 5.2, that is, that B being a constant $n \times n$ matrix implies $Be^{tB} = e^{tB}B$.]

12. Why can no characteristic multiplier be zero?

13. Find the characteristic multipliers for Meissner's equation $\ddot{y} + (\delta + \epsilon b(t))y = 0$, where

$$b(t) \triangleq \begin{Bmatrix} 1, & 0 < t < \pi \\ -1, & \pi < t < 2\pi \end{Bmatrix}.$$

Note: Because the ODE has a coefficient that is not continuous, to solve the ODE, you must consider two separate ODEs, after substituting in ± 1 values for $b(t)$:

$$(a)\ \ddot{y} + (\delta + \epsilon)y = 0,\ 0 < t < \pi$$
$$(b)\ \ddot{y} + (\delta - \epsilon)y = 0,\ \pi < t < 2\pi.$$

Let $y_1(t;\delta,\epsilon)$ solve (a) with initial data $y_1(0;\delta,\epsilon) = 1, \dot{y}_1(0;\delta,\epsilon) = 0$ and then define $A = y_1(\pi;\delta,\epsilon), B = \dot{y}_1(\pi;\delta,\epsilon)$. Then solve (b) on the interval $\pi < t < 2\pi$ with initial data $y_1(\pi;\delta,\epsilon) = A, \dot{y}_1(\pi;\delta,\epsilon) = B$. In this way, you will find the first column of a fundamental matrix. To find the second column, let $y_2(t;\delta,\epsilon)$ solve (a) with initial data $y_2(0;\delta,\epsilon) = 0, \dot{y}_2(0;\delta,\epsilon) = 1$, etc. [This ODE is similar to the Kronig–Penney model in the subject of quantum mechanics.]

14. Find a condition of the form $0 = \int_0^{2\pi} g(s)f(s)ds$ that guarantees that ODE $\ddot{y} + y = f(t)$ has a 2π-periodic solution. Use formula (5.114) for the equivalent system in $\mathbb{R}^2$.

15. Suppose that $A(t)$ and $f(t)$ are T-periodic and that the system $\dot{\mathbf{y}} = -A(t)^T\mathbf{y}$ has a T-periodic solution $\mathbf{y}(t)$. If the system $\dot{\mathbf{y}} = A(t)\mathbf{y} + \mathbf{f}(t)$ has a T-periodic solution $\mathbf{x}(t)$, explain why $0 = \int_0^{\mathrm{T}} \mathbf{y}^T(s)\mathbf{f}(s)ds$ must be true.

In problems 16 and 17, determine if the system $\dot{\mathbf{x}} = A(t)\mathbf{x}$ is asymptotically stable, neutrally stable, or unstable.

16. The system of Problem 5.8.6.9
17. The system of Problem 5.8.6.10
18. Suppose k is a positive constant. Consider the system of ODEs $(\star)$ $\dot{\mathbf{x}} = A(t)\mathbf{x}$, where $A = \begin{bmatrix} k\cos 2kt & k - k\sin 2kt \\ -k + k\sin 2kt & -2k + k\cos 2kt \end{bmatrix}$ is periodic with period π/k.
    (a) Verify that $\mathbf{x}^{(1)}(t) \triangleq e^{-kt}\begin{bmatrix} \sin kt \\ \cos kt \end{bmatrix}$ is a solution of $(\star)$ and thus gives us that one characteristic multiplier is $\mu_1 = e^{-\pi}$.
    (b) Use Abel's Theorem 5.7 in Section 5.2 to explain why the product of the two characteristic multipliers is $\mu_1\mu_2 = e^{-2\pi}$, even though we do not know a second basic solution of $(\star)$.
    (c) Use parts (a) and (b) to explain why $(\star)$ is asymptotically stable.

Key Terms

adjoint system: Problem 5.2.5.28
admittance: before Example 5.35 in Section 5.7
asymptotically stable: Definitions 5.7 in Section 5.3, (5.11) in Section 5.7, (5.12) in Section 5.8
basic solution: Definition 5.2 in Section 5.2, Definition 5.10 in Section 5.7
Casorati determinant: before Theorem 5.19 in Section 5.7
Cayley–Hamilton Theorem: Theorem 5.15 in Section 5.6
characteristic multipliers: after (5.108) in Section 5.8
closed vector subspace: after Lemma 5.4 in Section 5.6
companion form: Definition 5.6 in Section 5.2
compartment models, **compartments**: after (5.12) in Section 5.1
complete set of basic solutions: Definition 5.2 in Section 5.2, Definition 5.10 in Section 5.7
completely controllable: Definition 5.8 in Section 5.6
completely observable: Definition 5.9 in Section 5.6
complexification: (5.58) in Section 5.5
control, **control function**: after (5.64) in Section 5.6
drive: after (5.66) in Section 5.6
Floquet representation: Theorem 5.24 in Section 5.8
fundamental matrix (of solutions): Definition 5.3 in Section 5.2
gene: before (5.88) in Section 5.7
general solution: Definition 5.2 in Section 5.2, Definition 5.10 in Section 5.7
generalized eigenvector: after (5.35) in Section 5.3
Hill's discriminant: before (5.113) in Section 5.8
Hill's equation: (5.111) in Section 5.8
impedance: before Example 5.35 in Section 5.7
ladder network: Examples 5.35 in Section 5.7, (5.36) in Section 5.7
LCCHS: after (5.4) in Section 5.1
linear system: (5.3) in Section 5.1
monodromy matrix: before (5.108) in Section 5.8

multivariable control system: before (5.80) in Section 5.6
neutrally stable: Definitions 5.7 in Section 5.3, (5.11) in Section 5.7, (5.12) in Section 5.8
nonresonance assumption: after (5.51) in Section 5.5
optimal control: before (5.80) in Section 5.6
period doubling: after (5.103) in Section 5.8
piecewise continuous: before (5.70) in Section 5.6
principal fundamental matrix at $t = 0$: before (5.104) in Section 5.8
recessive (gene): before (5.88) in Section 5.7
sex-linked gene: before (5.88) in Section 5.7
single input control system: (5.64) in Section 5.6
sinusoidal forcing: before Theorem 5.13 in Section 5.5
solution: Definition 5.1 in Section 5.1
state of the system: after (5.64) in Section 5.6
stroboscopic map: before (5.108) in Section 5.8
systems of second order equations: before Example 5.5 in Section 5.1
trace: before Theorem 5.7 in Section 5.2
unstable: Definitions 5.7 in Section 5.3, (5.11) in Section 5.7, (5.12) in Section 5.8
Wronskian: Definition 5.4 in Section 5.2
$z-$**transforms**: Example 5.37 in Section 5.7

MATLAB® Commands

ode45: after Example 5.40 in Section 5.8
roots: Problem 5.2.5.33

References

Killough, G.G. and Eckerman, K.F. A conversational eigenanalysis program for solving differential equation. Midyear topical meeting of the Health Physics Society, Pasco, WA, February 5, 1984, Technical Report CONF-840202-21 of the Oak Ridge National Lab., TN.

Riggs, D.S. Quantitative aspects of iodine metabolism in man. *Pharmacological Reviews* 4, 285–369, 1952.

6

Geometry, Calculus, and Other Tools

6.1 Dot Product, Cross Product, Lines, and Planes

In this chapter, we study many geometric and differential calculus results that have many useful and powerful applications to problems of engineering and science.

6.1.1 Dot Product and Cross Product

We have been using the dot product, which is familiar also from its physical applications. For example, the **work**, W, is given by

$$W = \mathbf{F} \bullet \mathbf{u} \triangleq ||\mathbf{F}||\,||\mathbf{u}||\cos\theta,$$

where
$\mathbf{F}$ is a constant force
$\mathbf{u}$ is the displacement
θ is the angle between the vectors $\mathbf{F}$ and $\mathbf{u}$ satisfying $0 \le \theta \le \pi$.

Algebraically, for two vectors $\mathbf{r}_1, \mathbf{r}_2$ in $\mathbb{R}^3$,

$$\mathbf{r}_1 \bullet \mathbf{r}_2 = \begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} \bullet \begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix} = x_1x_2 + y_1y_2 + z_1z_2.$$

Denote

$$\hat{\boldsymbol{\imath}} \triangleq \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \hat{\boldsymbol{\jmath}} \triangleq \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \hat{\boldsymbol{k}} \triangleq \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{6.1}$$

The cross product between vectors $\mathbf{r}_1 = x_1\hat{\boldsymbol{\imath}} + y_1\hat{\boldsymbol{\jmath}} + z_1\hat{\boldsymbol{k}}$ and $\mathbf{r}_2 = x_2\hat{\boldsymbol{\imath}} + y_2\hat{\boldsymbol{\jmath}} + z_2\hat{\boldsymbol{k}}$ is defined by

$$\mathbf{r}_1 \times \mathbf{r}_2 = \begin{vmatrix} \hat{\boldsymbol{\imath}} & \hat{\boldsymbol{\jmath}} & \hat{\boldsymbol{k}} \\ x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \end{vmatrix} = \begin{vmatrix} y_1 & z_1 \\ y_2 & z_2 \end{vmatrix} \hat{\boldsymbol{\imath}} - \begin{vmatrix} x_1 & z_1 \\ x_2 & z_2 \end{vmatrix} \hat{\boldsymbol{\jmath}} + \begin{vmatrix} x_1 & y_1 \\ x_2 & y_2 \end{vmatrix} \hat{\boldsymbol{k}}$$

$$= (y_1z_2 - z_1y_2)\hat{\boldsymbol{\imath}} + (z_1x_2 - x_1z_2)\hat{\boldsymbol{\jmath}} + (x_1y_2 - y_1x_2)\hat{\boldsymbol{k}}.$$

Theorem 6.1

(Properties of the cross product)

(a) $\mathbf{x} \times \mathbf{y}$ is perpendicular to both $\mathbf{x}$ and $\mathbf{y}$.

(b) $||\mathbf{x} \times \mathbf{y}|| = ||\mathbf{x}||\,||\mathbf{y}|| \sin\theta$, where θ is the angle between the vectors $\mathbf{x}$ and $\mathbf{y}$ such that $0 \leq \theta \leq \pi$.

(c) $\mathbf{x} \times \mathbf{y} = \mathbf{0}$ if, and only if, $\{\mathbf{x}, \mathbf{y}\}$ is linearly dependent, that is, $\mathbf{x}$ and $\mathbf{y}$ are parallel or one of them is the zero vector.

(d) $\mathbf{y} \times \mathbf{x} = -\mathbf{x} \times \mathbf{y}$.

(e) $\hat{\boldsymbol{\imath}} \times \hat{\boldsymbol{\jmath}} = \hat{\mathbf{k}}$, $\hat{\boldsymbol{\jmath}} \times \hat{\mathbf{k}} = \hat{\boldsymbol{\imath}}$, $\hat{\mathbf{k}} \times \hat{\boldsymbol{\imath}} = \hat{\boldsymbol{\jmath}}$.

(f) $\hat{\boldsymbol{\imath}} \times \hat{\boldsymbol{\imath}} = \mathbf{0}$, $\hat{\boldsymbol{\jmath}} \times \hat{\boldsymbol{\jmath}} = \mathbf{0}$, $\hat{\mathbf{k}} \times \hat{\mathbf{k}} = \mathbf{0}$.

Example 6.1

A force $\mathbf{F}$ acting on a lever arm at a position $\mathbf{r}$ relative to an axis applies the **torque** $\boldsymbol{\tau} \triangleq \mathbf{r} \times \mathbf{F}$, as depicted in Figure 6.1.

Example 6.2

A charge q moving with a velocity $\mathbf{v}$ in a magnetic flux density $\mathbf{B}$ experiences the **Lorentz force** $\mathbf{F} = q\,\mathbf{v} \times \mathbf{B}$.

Example 6.3

Find a unit vector normal to both of the vectors $\mathbf{r}_1 = \hat{\boldsymbol{\imath}} + 2\hat{\boldsymbol{\jmath}} + 3\hat{\mathbf{k}}$, and $\mathbf{r}_2 = \hat{\boldsymbol{\jmath}} - \hat{\mathbf{k}}$.

Method: Let $\mathbf{u} = \mathbf{r}_1 \times \mathbf{r}_2 = \begin{vmatrix} \hat{\boldsymbol{\imath}} & \hat{\boldsymbol{\jmath}} & \hat{\mathbf{k}} \\ 1 & 2 & 3 \\ 0 & 1 & -1 \end{vmatrix} = -5\hat{\boldsymbol{\imath}} + \hat{\boldsymbol{\jmath}} + \hat{\mathbf{k}}$. So, $\widehat{\mathbf{u}} \triangleq \frac{1}{||\mathbf{u}||}\,\mathbf{u} = \frac{1}{\sqrt{27}}\,(-5\hat{\boldsymbol{\imath}} + \hat{\boldsymbol{\jmath}} + \hat{\mathbf{k}})$ is such a unit vector, as is also $-\widehat{\mathbf{u}}$. ◯

A $\widehat{\ }$ symbol over a vector signifies that it has unit length.

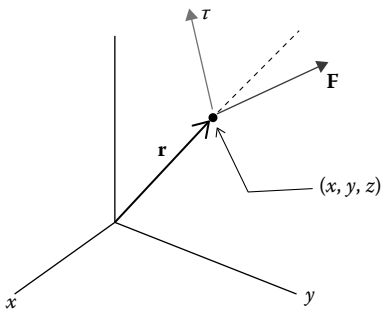


FIGURE 6.1
Torque.

6.1.2 Lines

Definition 6.1

(a) Given a point $P = (x, y, z)$, its **position vector** is

$$\mathbf{r} \triangleq \overrightarrow{OP} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = x\,\hat{\boldsymbol{\imath}} + y\,\hat{\boldsymbol{\jmath}} + z\,\hat{\mathbf{k}}.$$

(b) Given two points $P_0 = (x_0, y_0, z_0)$, and $P_1 = (x_1, y_1, z_1)$, the **vector from** P_0 **to** P_1 is

$$\overrightarrow{P_0P_1} = (x_1 - x_0)\hat{\boldsymbol{\imath}} + (y_1 - y_0)\hat{\boldsymbol{\jmath}} + (z_1 - z_0)\hat{\mathbf{k}}.$$

In particular, when $P_0 = O = (0, 0, 0)$, the origin, and $P_1 = P = (x, y, z)$, we get that $\overrightarrow{OP} = \mathbf{r}$ is the position vector of point P. Note also that $\overrightarrow{P_0P_1} = \mathbf{r}_1 - \mathbf{r}_0$ is the difference of the position vectors of P_0 and P_1.

Remark

Given any two points P_0, and P_1 on a given line L, let $\mathbf{v} = \overrightarrow{P_0P_1} = u\,\hat{\boldsymbol{\imath}} + v\,\hat{\boldsymbol{\jmath}} + w\,\hat{\mathbf{k}}$ and let $\mathbf{r}_0$ be the position vector of P_0. Then the line L consists of all points P whose position vectors $\mathbf{r}$ are of the form

$$\mathbf{r} = \mathbf{r}_0 + t\mathbf{v}, \text{ where } -\infty < t < \infty, \tag{6.2}$$

that is,

$$\left\{ \begin{aligned} x &= x_0 + tu \\ y &= y_0 + tv \\ z &= z_0 + tw \end{aligned} \right\}, \text{ where } -\infty < t < \infty. \tag{6.3}$$

The latter are the **parametric equations** of the line L.

6.1.3 Planes

Example 6.4

Suppose a force $\mathbf{F} = 2\hat{\boldsymbol{\imath}} + \hat{\boldsymbol{\jmath}} - 3\hat{\mathbf{k}}$ is applied at a point P whose position vector is $\mathbf{r} = x\,\hat{\boldsymbol{\imath}} + y\,\hat{\boldsymbol{\jmath}} + z\,\hat{\mathbf{k}}$ and is on a lever arm passing through the point $P_0 = (1, -2, 4)$. Find the set of all points P that maximize the ratio of the magnitude of the torque to the distance from P_0 and describe that set geometrically.

Method: We can write $\mathbf{r} = \mathbf{r}_0 + \mathbf{R}$, where $\mathbf{R}$ is a scalar multiple of the direction vector of the lever arm. Because the torque is $\boldsymbol{\tau} = \mathbf{r} \times \mathbf{F}$, its magnitude is

$$||\boldsymbol{\tau}|| = ||\mathbf{R}||\,||\mathbf{F}||\,\sin\theta,$$

where θ is the angle between the vectors $\mathbf{F}$ and $\mathbf{R}$ with $0 \le \theta \le \pi$.

The ratio

$$\frac{||\boldsymbol{\tau}||}{||\mathbf{R}||} = ||\mathbf{F}||\,\sin\theta$$

is maximized when $\theta = \frac{\pi}{2}$, that is, $\mathbf{F}$ is perpendicular to the lever arm. So, the desired set of points consists of those whose position vectors are $\mathbf{r} = \mathbf{r}_0 + \mathbf{R}$ satisfying

$$0 = \mathbf{R} \bullet \mathbf{F} = (\mathbf{r} - \mathbf{r}_0) \bullet \mathbf{F} = ((x-1)\hat{\boldsymbol{\imath}} + (y-(-2))\hat{\boldsymbol{\jmath}} + (z-4)\hat{\boldsymbol{k}}) \bullet (2\hat{\boldsymbol{\imath}} + \hat{\boldsymbol{\jmath}} - 3\hat{\boldsymbol{k}}),$$

that is,

$$0 = 2(x-1) + (y+2) - 3(z-4), \tag{6.4}$$

that is,

$$2x + y - 3z = -12. \tag{6.5}$$

Both (6.4) and (6.5) are equations of a plane that passes through the point $P_0 = (1, -2, 4)$. ◯

In general, a plane is the set of all points (x, y, z) satisfying a single equation of the form

$$Ax + By + Cz = D$$

or

$$A(x - x_0) + B(y - y_0) + C(z - z_0) = 0 \tag{6.6}$$

where A, B, C, D are scalar constants. In (6.6) the point $P_0 = (x_0, y_0, z_0)$ lies on the plane. The vector $\mathbf{n} = A\,\hat{\boldsymbol{\imath}} + B\,\hat{\boldsymbol{\jmath}} + C\,\hat{\boldsymbol{k}}$ is **normal** to the plane described by (6.6), which can be written as

$$\mathbf{n} \bullet (\mathbf{r} - \mathbf{r}_0) = 0 \tag{6.7}$$

Example 6.5

Find an equation of the plane that contains the points $(x, y, z) = (1, 0, 1)$, $(0, 2, 1)$, and $(2, 1, 0)$.

Method 1: Lying in the plane are the three vectors connecting the points:

$$\mathbf{v}_1 = \overrightarrow{(0,2,1) - (1,0,1)} = -\hat{\boldsymbol{\imath}} + 2\hat{\boldsymbol{\jmath}}$$

$$\mathbf{v}_2 = \overrightarrow{(2,1,0) - (0,2,1)} = 2\hat{\boldsymbol{\imath}} - \hat{\boldsymbol{\jmath}} - \hat{\boldsymbol{k}}$$

$$\mathbf{v}_3 = \overrightarrow{(1,0,1) - (2,1,0)} = -\hat{\boldsymbol{\imath}} - \hat{\boldsymbol{\jmath}} + \hat{\boldsymbol{k}}.$$

No two of these nonzero vectors are parallel, so any pair of them is linearly independent and thus spans the plane. If we take the cross product of any two of them, for example, $\mathbf{v}_1 \times \mathbf{v}_2$, we will get a vector normal to the plane because the cross product is perpendicular to those two. So,

$$\mathbf{n} \triangleq \mathbf{v}_1 \times \mathbf{v}_2 = (-\hat{\boldsymbol{\imath}} + 2\hat{\boldsymbol{\jmath}}) \times (2\hat{\boldsymbol{\imath}} - \hat{\boldsymbol{\jmath}} - \hat{\boldsymbol{k}}) = \cdots = -2\hat{\boldsymbol{\imath}} - \hat{\boldsymbol{\jmath}} - 3\hat{\boldsymbol{k}}$$

is normal to the plane. Using the point $P_0 = (1, 0, 1)$, the plane consists of all of the points that satisfy

$$0 = \mathbf{n} \bullet (\mathbf{r} - \mathbf{r}_0) = (-2\hat{\boldsymbol{\imath}} - \hat{\boldsymbol{\jmath}} - 3\hat{\boldsymbol{k}}) \bullet ((x-1)\hat{\boldsymbol{\imath}} - (y-0)\hat{\boldsymbol{\jmath}} + (z-1)\hat{\boldsymbol{k}}),$$

that is,

$$0 = -2(x-1) - (y-0) - 3(z-1). \bigcirc \tag{6.8}$$

Method 2: The plane is described by an equation of the form $Ax + By + Cz - D = 0$. Substituting in the three given points, we get a homogeneous system of three linear equations, respectively, in three unknowns:

$$\left\{\begin{matrix} A \cdot 1 + B \cdot 0 + C \cdot 1 - D = 0 \\ A \cdot 0 + B \cdot 2 + C \cdot 1 - D = 0 \\ A \cdot 2 + B \cdot 1 + C \cdot 0 - D = 0 \end{matrix}\right\}.$$

On the augmented matrix

$$\left[\begin{array}{cccc|c} 1 & 0 & 1 & -1 & 0 \\ 0 & 2 & 1 & -1 & 0 \\ 2 & 1 & 0 & -1 & 0 \end{array}\right]$$

do the row operations $-2R_1 + R_3 \to R_3$, $R_2 \leftrightarrow R_3$, $-2R_2 + R_3 \to R_3$, $\frac{1}{5}R_3 \to R_3$, $2R_3 + R_2 \to R_2$, $-R_3 + R_1 \to R_1$ to find its RREF

$$\left[\begin{array}{cccc|c} \textcircled{1} & 0 & 0 & -0.4 & 0 \\ 0 & \textcircled{1} & 0 & -0.2 & 0 \\ 0 & 0 & \textcircled{1} & -0.6 & 0 \end{array}\right].$$

There are infinitely many nontrivial solutions with $A = 0.4D, B = 0.2D$, and $C = 0.6C$. For example, for $D = 5$, $A = 2, B = 1$, and $C = 3$, the plane is

$$2x + y + 3z = 5,$$

which agrees with (6.8), found by Method 1. $\bigcirc$

In any kind of mathematical work, it is very good if we can use two different methods—and, of course, arrive at the same solution. This not only checks our conclusion but also enables us to understand the problem from different perspectives. In Example 6.5, Method 1 viewed the problem geometrically, and Method 2 viewed the problem algebraically. Each person may find one method more satisfying than the other and may even find yet a third or fourth method most satisfying.

6.1.4 Problems

1. For the vectors $\mathbf{A} = \hat{\boldsymbol{\imath}} - \hat{\boldsymbol{\jmath}}$ and $\mathbf{B} = 2\hat{\boldsymbol{\jmath}} + \hat{\boldsymbol{k}}$, find (a) $\mathbf{A} \bullet \mathbf{B}$, (b) $\mathbf{A} \times \mathbf{B}$, and (c) the angle between $\mathbf{A}$ and $\mathbf{B}$.

For problems 2–5, find parametric equations of the line satisfying the following given conditions:

No two of these nonzero vectors are parallel, so any pair of them is linearly independent and thus spans the plane. If we take the cross product of any two of them, for example, $\mathbf{v}_1 \times \mathbf{v}_2$, we will get a vector normal to the plane because the cross product is perpendicular to those two. So,

$$\mathbf{n} \triangleq \mathbf{v}_1 \times \mathbf{v}_2 = (-\hat{\boldsymbol{\imath}} + 2\hat{\boldsymbol{\jmath}}) \times (2\hat{\boldsymbol{\imath}} - \hat{\boldsymbol{\jmath}} - \hat{\boldsymbol{k}}) = \cdots = -2\hat{\boldsymbol{\imath}} - \hat{\boldsymbol{\jmath}} - 3\hat{\boldsymbol{k}}$$

is normal to the plane. Using the point $P_0 = (1, 0, 1)$, the plane consists of all of the points that satisfy

$$0 = \mathbf{n} \bullet (\mathbf{r} - \mathbf{r}_0) = (-2\hat{\boldsymbol{\imath}} - \hat{\boldsymbol{\jmath}} - 3\hat{\boldsymbol{k}}) \bullet ((x-1)\hat{\boldsymbol{\imath}} - (y-0)\hat{\boldsymbol{\jmath}} + (z-1)\hat{\boldsymbol{k}}),$$

that is,

$$0 = -2(x-1) - (y-0) - 3(z-1). \bigcirc \tag{6.8}$$

Method 2: The plane is described by an equation of the form $Ax + By + Cz - D = 0$. Substituting in the three given points, we get a homogeneous system of three linear equations, respectively, in three unknowns:

$$\left\{\begin{matrix} A \cdot 1 + B \cdot 0 + C \cdot 1 - D = 0 \\ A \cdot 0 + B \cdot 2 + C \cdot 1 - D = 0 \\ A \cdot 2 + B \cdot 1 + C \cdot 0 - D = 0 \end{matrix}\right\}.$$

On the augmented matrix

$$\left[\begin{array}{cccc|c} 1 & 0 & 1 & -1 & 0 \\ 0 & 2 & 1 & -1 & 0 \\ 2 & 1 & 0 & -1 & 0 \end{array}\right]$$

do the row operations $-2R_1 + R_3 \to R_3$, $R_2 \leftrightarrow R_3$, $-2R_2 + R_3 \to R_3$, $\frac{1}{5}R_3 \to R_3$, $2R_3 + R_2 \to R_2$, $-R_3 + R_1 \to R_1$ to find its RREF

$$\left[\begin{array}{cccc|c} \textcircled{1} & 0 & 0 & -0.4 & 0 \\ 0 & \textcircled{1} & 0 & -0.2 & 0 \\ 0 & 0 & \textcircled{1} & -0.6 & 0 \end{array}\right].$$

There are infinitely many nontrivial solutions with $A = 0.4D, B = 0.2D$, and $C = 0.6C$. For example, for $D = 5$, $A = 2, B = 1$, and $C = 3$, the plane is

$$2x + y + 3z = 5,$$

which agrees with (6.8), found by Method 1. $\bigcirc$

In any kind of mathematical work, it is very good if we can use two different methods—and, of course, arrive at the same solution. This not only checks our conclusion but also enables us to understand the problem from different perspectives. In Example 6.5, Method 1 viewed the problem geometrically, and Method 2 viewed the problem algebraically. Each person may find one method more satisfying than the other and may even find yet a third or fourth method most satisfying.

6.1.4 Problems

1. For the vectors $\mathbf{A} = \hat{\boldsymbol{\imath}} - \hat{\boldsymbol{\jmath}}$ and $\mathbf{B} = 2\hat{\boldsymbol{\jmath}} + \hat{\boldsymbol{k}}$, find (a) $\mathbf{A} \bullet \mathbf{B}$, (b) $\mathbf{A} \times \mathbf{B}$, and (c) the angle between $\mathbf{A}$ and $\mathbf{B}$.

For problems 2–5, find parametric equations of the line satisfying the following given conditions:

2. Passes through points $(1, 3, 5)$ and $(0, -1, 4)$.
3. Passes through points $(0, 3, 5)$ and $(2, 1, -1)$.
4. Passes through the point $(2, 4, 1)$ and is perpendicular to the plane $x + 2y - z = -1$.
5. Passes through the point $(2, 4, 1)$ and is perpendicular to the plane $-x + 2y + 3z = 4$.

For problems 6–12, find an equation of the plane satisfying the given conditions.

6. Passes through the points $(1, 0, 6)$, $(2, 1, 8)$, and $(3, 2, 4)$.
7. Passes through the points $(0, 1, 1)$, $(1, 0, 2)$, and $(2, 3, 0)$.
8. Passes through the points $(1, 3, 5)$, $(2, -1, 4)$, and $(-1, 1, 0)$.
9. Passes through the point $(2, 4, 1)$ and is perpendicular to the line whose parametric equations are $\mathbf{r} = (2 + 3t)\hat{\boldsymbol{\imath}} + (4 - 2t)\hat{\boldsymbol{\jmath}} + (1 + t)\hat{\boldsymbol{k}}, -\infty < t < \infty$.
10. Passes through the point $(2, 4, 1)$ and is perpendicular to the line whose parametric equations are $\mathbf{r} = 3t\,\hat{\boldsymbol{\imath}} + (5 - 2t)\hat{\boldsymbol{\jmath}} + (3 + t)\hat{\boldsymbol{k}}, -\infty < t < \infty$.
11. Passes through the point $(2, 4, 1)$ and is perpendicular to the line that passes through the points $(0, 1, 3)$, and $(-1, 2, 4)$.
12. Contains both of the lines whose parametric equations are $\mathbf{r} = (2+3t)\hat{\boldsymbol{\imath}} + (4-2t)\hat{\boldsymbol{\jmath}} + (1 + t)\hat{\boldsymbol{k}}, -\infty < t < \infty$. and $\mathbf{r} = 2t\,\hat{\boldsymbol{\imath}} + (5 - t)\hat{\boldsymbol{\jmath}} + (-1 + 2t)\hat{\boldsymbol{k}}, -\infty < t < \infty$.
13. Find a unit vector normal to both of the vectors $2\hat{\boldsymbol{\imath}} - \hat{\boldsymbol{\jmath}} + \hat{\boldsymbol{k}}$ and $4\hat{\boldsymbol{\imath}} + \hat{\boldsymbol{k}}$.
14. Suppose the force $\mathbf{F} = \hat{\boldsymbol{\imath}} - 2\hat{\boldsymbol{\jmath}} + 4\hat{\boldsymbol{k}}$ is applied at a point P whose position vector is $\mathbf{r} = x\,\hat{\boldsymbol{\imath}} + y\,\hat{\boldsymbol{\jmath}} + z\,\hat{\boldsymbol{k}}$ and is on a lever arm passing through the point $P_0 = (4, 1, 2)$. Find the set of all points P that maximize the ratio of the magnitude of the torque to the distance from P_0, and describe that set geometrically.
15. For what angle(s) θ between the force vector $\mathbf{F}$ and the displacement vector $\mathbf{u}$ do we get the maximum magnitude of the work, W?
16. A charge q moves with a constant speed of 10 m/s in a magnetic field $\mathbf{B} = -2\hat{\boldsymbol{\imath}} + \hat{\boldsymbol{\jmath}} + 3\hat{\boldsymbol{k}}$, in teslas. What is the smallest charge, in coulombs, that can experience 1 N of Lorentz force, assuming the direction of travel is chosen to maximize that force?
17. A particle whose mass is m and electric charge is $-q$ moves through a constant magnetic flux density $\mathbf{B}_0$. Assuming there are no other forces on the particle, Newton's second law of motion yields the (vector) differential equation $m\ddot{\mathbf{r}} = -q\mathbf{v} \times \mathbf{B}_0$, which along with the relationship between position and velocity yields the system of ordinary differential equations (ODEs):

$$\dot{\mathbf{r}} = \mathbf{v}, \;\; m\dot{\mathbf{v}} = -q\mathbf{v} \times \mathbf{B}_0.$$

Let the particle have $\mathbf{v}(0) = \mathbf{v}_0 \neq \mathbf{0}$ as its initial velocity vector and $\mathbf{r}(0) = \mathbf{r}_0$ as its initial velocity vector.

(a) Explain why the particle's motion remains in the plane $\mathbf{r} = \mathbf{r}(0) + \mathit{Span}\{\mathbf{v}(0), \mathbf{a}(0)\}$, where $\mathbf{a}(0) \triangleq -\frac{q}{m}\,\mathbf{v}_0 \times \mathbf{B}_0$ is the particle's initial acceleration, as long as $\mathbf{a}(0) \neq \mathbf{0}$.

(b) Explain why we can express $\mathbf{r}(t) = \mathbf{r}_0 + g(t)\mathbf{v}(0) + h(t)\mathbf{a}(0)$ for some unknown functions $g(t), h(t)$.

(c) Explain why $g(t)$ and $h(t)$ both satisfy the scalar ODE of undamped harmonic motion, $\ddot{y} = -(||\mathbf{a}_0||^2/||\mathbf{v}_0||^2)\, y$.

(d) Solve for $g(t), h(t)$ and thus find the motion of the particle. This is called **cyclotron motion***.

18. Suppose that $\mathbf{A}$ and $\mathbf{B}$ are orthogonal vectors. Why is $||\mathbf{A} + \mathbf{B}||^2 = ||\mathbf{A}||^2 + ||\mathbf{B}||^2$? [Note that this is the Pythagorean Theorem.]

6.2 Trigonometry, Polar, Cylindrical, and Spherical Coordinates

Any point in the xy-plane can be represented by polar coordinates (r, θ) that satisfy

$$x = r\cos\theta, \quad y = r\sin\theta,$$

where $r \triangleq \sqrt{x^2 + y^2} \geq 0$, as shown in Figure 6.2. The Pythagorean theorem implies the most basic of trigonometric identities:

$$\cos^2\theta + \sin^2\theta \equiv 1. \tag{6.9}$$

The other four basic trigonometric functions are defined by

$$\tan\theta \triangleq \frac{y}{x} = \frac{\sin\theta}{\cos\theta}, \text{ if } x \neq 0.$$

$$\sec\theta \triangleq \frac{r}{x} = \frac{1}{\cos\theta}, \text{ if } x \neq 0.$$

$$\csc\theta \triangleq \frac{r}{y} = \frac{1}{\sin\theta}, \text{ if } y \neq 0.$$

$$\cot\theta \triangleq \frac{x}{y} = \frac{\cos\theta}{\sin\theta}, \text{ if } y \neq 0.$$

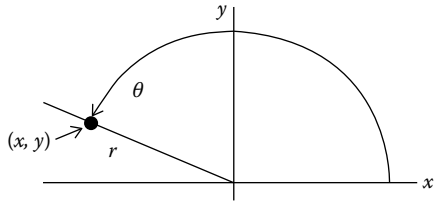


FIGURE 6.2
Polar coordinates.

* From Prof. Paul Diament's EE3401Y Supplementary Class Notes, Columbia University, Spring, 1974.

Identity (6.9) implies the identities

$$\tan^2\theta + 1 \equiv \sec^2\theta, \quad \cot^2\theta + 1 \equiv \csc^2\theta. \tag{6.10}$$

Recall that an inverse function f^{-1} for a function f with domain $\mathcal{D}$ is defined by the relationship

$$y = f(\theta),\ \theta \text{ in } \mathcal{D} \iff \theta = f^{-1}(y),\ y \text{ in } f(\mathcal{D}),$$

assuming f is **one-to-one** on $\mathcal{D}$, that is, $f(\theta_1) \neq f(\theta_2)$ when θ_1, θ_2 are distinct elements of the domain, $\mathcal{D}$. The set $f(\mathcal{D}) = \{f(\theta) :\ \theta \text{ in } \mathcal{D}\}$ is called the **range** or **image** of $\mathcal{D}$. Also, if f has domain $\mathcal{D}$ and has inverse function f^{-1}, then we have the *cancellation properties*

$$f^{-1}(f(\theta)) = \theta, \text{ for all } \theta \text{ in } \mathcal{D},$$

and

$$f(f^{-1}(y)) = y, \text{ for all } y \text{ in } f(\mathcal{D}).$$

The function $f(\theta) \triangleq \sin\theta$ has derivative, $\cos\theta$, that is positive for $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$, so f is strictly decreasing on $\mathcal{D}$. Because of this, $\sin\theta$ is one-to-one on domain $\mathcal{D} \triangleq -\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}$. The inverse function is $\sin^{-1}(y)$, also known as $\arcsin(y)$, which is defined for y in $f(\mathcal{D}) = \{y :\ -1 \leq y \leq 1\}$.

The function $f(\theta) \triangleq \tan\theta$ has derivative, $\sec^2\theta$, that is positive for $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$, so f is strictly increasing on $\mathcal{D}$. Because of this, $\tan\theta$ is one-to-one on domain $\mathcal{D} \triangleq -\frac{\pi}{2} < \theta < \frac{\pi}{2}$. The inverse function is $\tan^{-1}(z)$, also known as $\arctan(z)$, which is defined for z in $f(\mathcal{D}) = \{z :\ -\infty < z < \infty\}$.

The function $f(\theta) \triangleq \cos\theta$ has derivative, $-\sin\theta$, that is negative for $0 < \theta < \pi$, so f is strictly decreasing on $\mathcal{D}$. Because of this, $\cos\theta$ is one-to-one on domain $\mathcal{D} \triangleq 0 \leq \theta \leq \pi$. The inverse function is $\cos^{-1}(x)$, also known as $\arccos(x)$, which is defined for x in $f(\mathcal{D}) = \{x :\ -1 \leq x \leq 1\}$.

As we noted in (3.37) in Section 3.3,

$$\theta = \left\{ \begin{array}{rl} \arctan(y/x), & \text{if } x > 0 \\ \pi + \arctan(y/x), & \text{if } x < 0 \\ \frac{\pi}{2}, & \text{if } x = 0 \text{ and } y > 0 \\ -\frac{\pi}{2}, & \text{if } x = 0 \text{ and } y < 0 \\ \textit{D.N.E.}, & \text{if } (x,y) = (0,0) \end{array} \right\}. \tag{6.11}$$

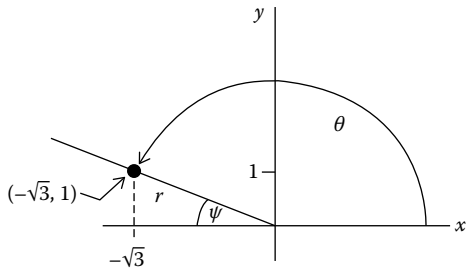


FIGURE 6.3
Example 6.6.

Example 6.6

Find the polar coordinates of the point $(x, y) = (-\sqrt{3}, 1)$.

Method: $r = \sqrt{(-\sqrt{3})^2 + (1)^2} = 2$ and $\tan\theta = \frac{y}{x} = \frac{1}{-\sqrt{3}} = -\frac{1}{\sqrt{3}}$. Unfortunately, $\theta \neq \tan^{-1}(-\frac{1}{\sqrt{3}})$ because (x, y) is in the second quadrant. It helps to draw the picture in Figure 6.3: The reference angle $\psi = \frac{\pi}{6}$ follows from a $30° - 60° - 90°$ triangle having sides of lengths $1, \sqrt{3}, 2$, respectively, opposite those angles. Because $\theta = \pi - \psi$, $\theta = \frac{5\pi}{6}$. To summarize, the polar coordinates of $(-\sqrt{3}, 1)$ are $(r, \theta) = (2, \frac{5\pi}{6})$. ◯

6.2.1 Cylindrical Coordinates

Given a point (x, y, z) in $\mathbb{R}^3$, we can replace the x and y coordinates by their polar coordinates r, θ to get the point's cylindrical coordinates (r, θ, z):

$$x = r\cos\theta, \quad y = r\sin\theta, \quad z = z.$$

Geometrically, to find the cylindrical coordinates of (x, y, z), project (x, y, z) to the point $(x, y, 0)$ in the $z = 0$ plane; after that, find the polar coordinates (r, θ) of (x, y). This is shown in Figure 6.4. The point (x, y, z) has position vector

$$\mathbf{r} = r\cos\theta\,\hat{\boldsymbol{\imath}} + r\sin\theta\,\hat{\boldsymbol{\jmath}} + z\,\hat{\boldsymbol{k}}.$$

So,

$$||\mathbf{r}|| = ||r\cos\theta\,\hat{\boldsymbol{\imath}} + r\sin\theta\,\hat{\boldsymbol{\jmath}} + z\,\hat{\boldsymbol{k}}|| = \sqrt{(r\cos\theta)^2 + (r\sin\theta)^2 + z^2} = \sqrt{r^2 + z^2}.$$

6.2.2 Spherical Coordinates

Given a point (x, y, z) in $\mathbb{R}^3$, whose position vector $\mathbf{r} = x\,\hat{\boldsymbol{\imath}} + y\,\hat{\boldsymbol{\jmath}} + z\,\hat{\boldsymbol{k}}$, define

$$\rho \triangleq ||\mathbf{r}|| = \sqrt{x^2 + y^2 + z^2}.$$

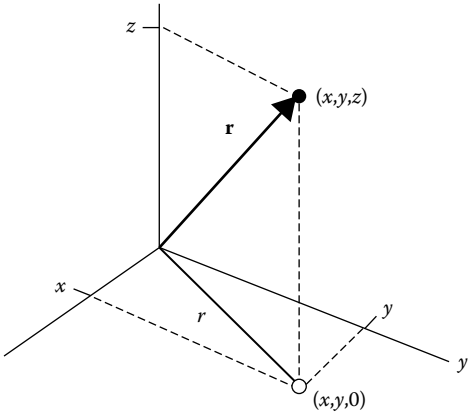


FIGURE 6.4
Cylindrical coordinates.

Define ϕ to be the angle measured downward from the positive z-axis to the vector $\mathbf{r}$, where $0 \leq \phi \leq \pi$. We have $z = \rho \cos\phi$ and $r = \rho \sin\phi$, as shown in Figure 6.5. Geometrically, $(0,0,z)$ is the orthogonal projection of (x,y,z) onto the z-axis, and $(x,y,0)$ is the orthogonal projection of (x,y,z) onto the $z=0$ plane. Using polar coordinates, $x = r\cos\theta, y = r\sin\theta$. Substituting in $r = \rho\sin\phi$, we get the point's spherical coordinates (ρ,ϕ,θ):

$$x = r\,\cos\theta = \rho\,\sin\phi\,\cos\theta, \quad y = r\,\sin\theta = \rho\,\sin\phi\,\sin\theta, \quad z = \rho\,\cos\phi. \tag{6.12}$$

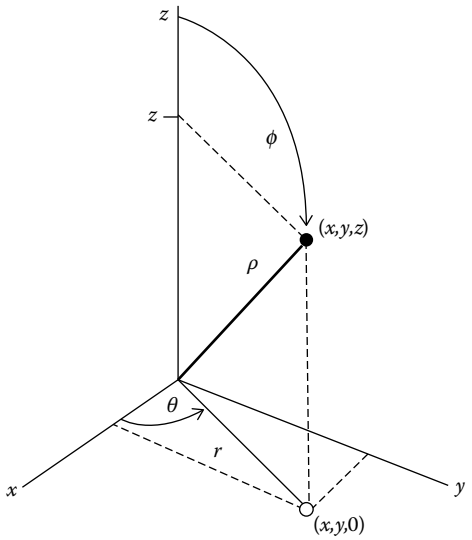


FIGURE 6.5
Spherical coordinates.

Identity (6.9) implies the identities

$$\tan^2\theta + 1 \equiv \sec^2\theta, \quad \cot^2\theta + 1 \equiv \csc^2\theta. \tag{6.10}$$

Recall that an inverse function f^{-1} for a function f with domain $\mathcal{D}$ is defined by the relationship

$$y = f(\theta),\ \theta \text{ in } \mathcal{D} \iff \theta = f^{-1}(y),\ y \text{ in } f(\mathcal{D}),$$

assuming f is **one-to-one** on $\mathcal{D}$, that is, $f(\theta_1) \neq f(\theta_2)$ when θ_1, θ_2 are distinct elements of the domain, $\mathcal{D}$. The set $f(\mathcal{D}) = \{f(\theta) :\ \theta \text{ in } \mathcal{D}\}$ is called the **range** or **image** of $\mathcal{D}$. Also, if f has domain $\mathcal{D}$ and has inverse function f^{-1}, then we have the *cancellation properties*

$$f^{-1}(f(\theta)) = \theta, \text{ for all } \theta \text{ in } \mathcal{D},$$

and

$$f(f^{-1}(y)) = y, \text{ for all } y \text{ in } f(\mathcal{D}).$$

The function $f(\theta) \triangleq \sin\theta$ has derivative, $\cos\theta$, that is positive for $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$, so f is strictly decreasing on $\mathcal{D}$. Because of this, $\sin\theta$ is one-to-one on domain $\mathcal{D} \triangleq -\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}$. The inverse function is $\sin^{-1}(y)$, also known as $\arcsin(y)$, which is defined for y in $f(\mathcal{D}) = \{y :\ -1 \leq y \leq 1\}$.

The function $f(\theta) \triangleq \tan\theta$ has derivative, $\sec^2\theta$, that is positive for $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$, so f is strictly increasing on $\mathcal{D}$. Because of this, $\tan\theta$ is one-to-one on domain $\mathcal{D} \triangleq -\frac{\pi}{2} < \theta < \frac{\pi}{2}$. The inverse function is $\tan^{-1}(z)$, also known as $\arctan(z)$, which is defined for z in $f(\mathcal{D}) = \{z :\ -\infty < z < \infty\}$.

The function $f(\theta) \triangleq \cos\theta$ has derivative, $-\sin\theta$, that is negative for $0 < \theta < \pi$, so f is strictly decreasing on $\mathcal{D}$. Because of this, $\cos\theta$ is one-to-one on domain $\mathcal{D} \triangleq 0 \leq \theta \leq \pi$. The inverse function is $\cos^{-1}(x)$, also known as $\arccos(x)$, which is defined for x in $f(\mathcal{D}) = \{x :\ -1 \leq x \leq 1\}$.

As we noted in (3.37) in Section 3.3,

$$\theta = \left\{ \begin{array}{rl} \arctan(y/x), & \text{if } x > 0 \\ \pi + \arctan(y/x), & \text{if } x < 0 \\ \frac{\pi}{2}, & \text{if } x = 0 \text{ and } y > 0 \\ -\frac{\pi}{2}, & \text{if } x = 0 \text{ and } y < 0 \\ \textit{D.N.E.}, & \text{if } (x,y) = (0,0) \end{array} \right\}. \tag{6.11}$$

Identity (6.9) implies the identities

$$\tan^2\theta + 1 \equiv \sec^2\theta, \quad \cot^2\theta + 1 \equiv \csc^2\theta. \tag{6.10}$$

Recall that an inverse function f^{-1} for a function f with domain $\mathcal{D}$ is defined by the relationship

$$y = f(\theta),\ \theta \text{ in } \mathcal{D} \iff \theta = f^{-1}(y),\ y \text{ in } f(\mathcal{D}),$$

assuming f is **one-to-one** on $\mathcal{D}$, that is, $f(\theta_1) \neq f(\theta_2)$ when θ_1, θ_2 are distinct elements of the domain, $\mathcal{D}$. The set $f(\mathcal{D}) = \{f(\theta) :\ \theta \text{ in } \mathcal{D}\}$ is called the **range** or **image** of $\mathcal{D}$. Also, if f has domain $\mathcal{D}$ and has inverse function f^{-1}, then we have the *cancellation properties*

$$f^{-1}(f(\theta)) = \theta, \text{ for all } \theta \text{ in } \mathcal{D},$$

and

$$f(f^{-1}(y)) = y, \text{ for all } y \text{ in } f(\mathcal{D}).$$

The function $f(\theta) \triangleq \sin\theta$ has derivative, $\cos\theta$, that is positive for $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$, so f is strictly decreasing on $\mathcal{D}$. Because of this, $\sin\theta$ is one-to-one on domain $\mathcal{D} \triangleq -\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}$. The inverse function is $\sin^{-1}(y)$, also known as $\arcsin(y)$, which is defined for y in $f(\mathcal{D}) = \{y :\ -1 \leq y \leq 1\}$.

The function $f(\theta) \triangleq \tan\theta$ has derivative, $\sec^2\theta$, that is positive for $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$, so f is strictly increasing on $\mathcal{D}$. Because of this, $\tan\theta$ is one-to-one on domain $\mathcal{D} \triangleq -\frac{\pi}{2} < \theta < \frac{\pi}{2}$. The inverse function is $\tan^{-1}(z)$, also known as $\arctan(z)$, which is defined for z in $f(\mathcal{D}) = \{z :\ -\infty < z < \infty\}$.

The function $f(\theta) \triangleq \cos\theta$ has derivative, $-\sin\theta$, that is negative for $0 < \theta < \pi$, so f is strictly decreasing on $\mathcal{D}$. Because of this, $\cos\theta$ is one-to-one on domain $\mathcal{D} \triangleq 0 \leq \theta \leq \pi$. The inverse function is $\cos^{-1}(x)$, also known as $\arccos(x)$, which is defined for x in $f(\mathcal{D}) = \{x :\ -1 \leq x \leq 1\}$.

As we noted in (3.37) in Section 3.3,

$$\theta = \left\{ \begin{array}{rl} \arctan(y/x), & \text{if } x > 0 \\ \pi + \arctan(y/x), & \text{if } x < 0 \\ \frac{\pi}{2}, & \text{if } x = 0 \text{ and } y > 0 \\ -\frac{\pi}{2}, & \text{if } x = 0 \text{ and } y < 0 \\ \textit{D.N.E.}, & \text{if } (x,y) = (0,0) \end{array} \right\}. \tag{6.11}$$

Example 6.7

Find the spherical coordinates of the point $(x, y) = (-1, 1, -\sqrt{2})$.

Method: $\rho = \sqrt{x^2 + y^2 + z^2} = \sqrt{(-1)^2 + 1^2 + (-\sqrt{2})^2} = 2$, so

$$-\sqrt{2} = z = \rho \cos\phi = 2\cos\phi$$

implies $\cos\phi = -\frac{1}{\sqrt{2}}$. Because $0 \le \phi \le \pi$, $\phi = \cos^{-1}(-\frac{1}{\sqrt{2}}) = \frac{3\pi}{4}$. It follows that $r = \rho \sin\phi = 2\sin(\frac{3\pi}{4}) = \sqrt{2}$. Our point has $x = -1, y = 1$, so

$$\begin{aligned} -1 &= x = r\cos\theta = \sqrt{2}\cos\theta \text{ and} \\ 1 &= y = r\sin\theta = \sqrt{2}\sin\theta. \end{aligned}$$

It follows that $\theta = \frac{3\pi}{4}$. The spherical coordinates of the point are $(\rho, \phi, \theta) = (2, \frac{3\pi}{4}, \frac{3\pi}{4})$. ◯

6.2.3 Right-Handed Orthogonal Bases for $\mathbb{R}^3$

The **standard orthonormal (o.n) basis** for $\mathbb{R}^3$ is $\{\hat{\boldsymbol{\imath}}, \hat{\boldsymbol{\jmath}}, \hat{\boldsymbol{k}}\}$. It was also denoted $\{\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \mathbf{e}^{(3)}\}$ in Chapter 1, where we noted that they are the columns of the identity matrix, I_3. We will see in this section that in natural ways, cylindrical and spherical coordinates give different o.n. bases for $\mathbb{R}^3$.

Cylindrical coordinates (r, θ, z) are based on the orthogonal projection P given by

$$P\mathbf{r} = P\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ z \end{bmatrix} = z\,\hat{\boldsymbol{k}}.$$

and its orthogonal complementary projection $I - P$ given by

$$(I - P)\mathbf{r} = (I - P)\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x \\ y \\ 0 \end{bmatrix} = x\,\hat{\boldsymbol{\imath}} + y\,\hat{\boldsymbol{\jmath}}.$$

Instead of expressing $(I - P)\mathbf{r}$ in terms of $\hat{\boldsymbol{\imath}}$ and $\hat{\boldsymbol{\jmath}}$, we can find orthogonal basis vectors more appropriate in cylindrical coordinates. As before, it helps to use polar coordinates in the $z = 0$ plane:

$$(I - P)\mathbf{r} = x\,\hat{\boldsymbol{\imath}} + y\,\hat{\boldsymbol{\jmath}} = (r\cos\theta)\hat{\boldsymbol{\imath}} + (r\sin\theta)\hat{\boldsymbol{\jmath}} = r(\cos\theta\,\hat{\boldsymbol{\imath}} + \sin\theta\,\hat{\boldsymbol{\jmath}}),$$

so it makes sense to define

$$\hat{\mathbf{e}}_r \triangleq \cos\theta\,\hat{\boldsymbol{\imath}} + \sin\theta\,\hat{\boldsymbol{\jmath}},$$

which is a unit vector in the direction of the planar vector $x\,\hat{\boldsymbol{\imath}} + y\,\hat{\boldsymbol{\jmath}}$. Note that the vector $\hat{\mathbf{e}}_r$ varies as θ varies.

It makes sense to define a vector

$$\hat{\mathbf{e}}_z \triangleq \hat{\boldsymbol{k}}.$$

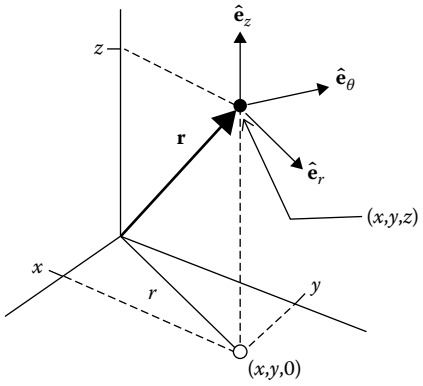


FIGURE 6.6
Right-handed basis $\{\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z\}$ in cylindrical coordinates.

Note that

$$\mathbf{r} = r\,\hat{\mathbf{e}}_r + z\,\hat{\mathbf{e}}_z.$$

So far, in cylindrical coordinates, we have a set of two o.n. vectors, $\{\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_z\}$. To get an o.n. basis for $\mathbb{R}^3$, we need a third unit vector, $\hat{\mathbf{w}}$, that is orthogonal to both $\hat{\mathbf{e}}_z.\hat{\mathbf{e}}_r$. Even better, we would like $\{\hat{\mathbf{e}}_r, \hat{\mathbf{w}}, \hat{\mathbf{e}}_z\}$ to give a **right-handed** o.n. basis, that is, that

$$\hat{\mathbf{e}}_r \times \hat{\mathbf{w}} = \hat{\mathbf{e}}_z, \quad \hat{\mathbf{w}} \times \hat{\mathbf{e}}_z = \hat{\mathbf{e}}_r, \quad \hat{\mathbf{e}}_z \times \hat{\mathbf{r}} = \hat{\mathbf{w}}.$$

The latter gives us an easy way to find

$$\hat{\mathbf{w}} \triangleq \hat{\mathbf{e}}_z \times \hat{\mathbf{r}} = \hat{\boldsymbol{k}} \times (\cos\theta\,\hat{\boldsymbol{\imath}} + \sin\theta\,\hat{\boldsymbol{\jmath}}) = -\sin\theta\,\hat{\boldsymbol{\imath}} + \cos\theta\,\hat{\boldsymbol{\jmath}},$$

using Theorem 6.1(d) and (e) in Section 6.1. We denote

$$\hat{\mathbf{e}}_\theta \triangleq -\sin\theta\,\hat{\boldsymbol{\imath}} + \cos\theta\,\hat{\boldsymbol{\jmath}}. \tag{6.13}$$

Figure 6.6 shows the right-handed o.n. basis at an example of a point in $\mathbb{R}^3$.

$$\{\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z\} = \left\{ \begin{bmatrix} \cos\theta \\ \sin\theta \\ 0 \end{bmatrix}, \begin{bmatrix} -\sin\theta \\ \cos\theta \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\} \tag{6.14}$$

at an example of a point in $\mathbb{R}^3$.

6.2.4 Orthonormal Basis in Spherical Coordinates

In ways analogous to what we did for cylindrical coordinates, we can also find a right-handed o.n. basis in spherical coordinates (ρ, ϕ, θ). One natural vector in that basis is the

unit vector in the direction of the position vector, that is,

$$\hat{\mathbf{e}}_\rho \triangleq \sin\phi\,\cos\theta\,\hat{\boldsymbol{\imath}} + \sin\phi\,\sin\theta\,\hat{\boldsymbol{\jmath}} + \cos\phi\,\hat{\boldsymbol{k}}. \tag{6.15}$$

Note that

$$\mathbf{r} = \rho\,\hat{\mathbf{e}}_\rho.$$

Analogously to what we did in cylindrical coordinates, we define again

$$\hat{\mathbf{e}}_\theta \triangleq -\sin\theta\,\hat{\boldsymbol{\imath}} + \cos\theta\,\hat{\boldsymbol{\jmath}}. \tag{6.16}$$

A third vector $\hat{\mathbf{e}}_\phi$ can be chosen so that $\{\hat{\mathbf{e}}_\rho, \hat{\mathbf{e}}_\phi, \hat{\mathbf{e}}_\theta\}$ is a *right-handed* o.n. basis for $\mathbb{R}^3$: Let

$$\hat{\mathbf{e}}_\phi \triangleq \hat{\mathbf{e}}_\theta \times \hat{\mathbf{e}}_\rho = \cdots = \cos\phi\,\cos\theta\,\hat{\boldsymbol{\imath}} + \cos\phi\,\sin\theta\,\hat{\boldsymbol{\jmath}} - \sin\phi\,\hat{\boldsymbol{k}}. \tag{6.17}$$

Figure 6.7 shows the right-handed o.n. basis at an example of a point in $\mathbb{R}^3$.

$$\{\hat{\mathbf{e}}_\rho, \hat{\mathbf{e}}_\phi, \hat{\mathbf{e}}_\theta\} = \left\{ \begin{bmatrix} \sin\phi\,\cos\theta \\ \sin\phi\,\sin\theta \\ \cos\phi \end{bmatrix}, \begin{bmatrix} \cos\phi\,\cos\theta \\ \cos\phi\,\sin\theta \\ -\sin\phi \end{bmatrix}, \begin{bmatrix} -\sin\theta \\ \cos\theta \\ 0 \end{bmatrix} \right\}. \tag{6.18}$$

We can express the vectors in the o.n. basis in spherical coordinates in terms of the vectors in the o.n. basis in cylindrical coordinates, specifically

$$\hat{\mathbf{e}}_\rho = \sin\phi\,\hat{\mathbf{e}}_r + \cos\phi\,\hat{\mathbf{e}}_z,$$

$$\hat{\mathbf{e}}_\phi = \cos\phi\,\hat{\mathbf{e}}_r - \sin\phi\,\hat{\mathbf{e}}_z,$$

and $\hat{\mathbf{e}}_\theta$ is the same in the two bases.

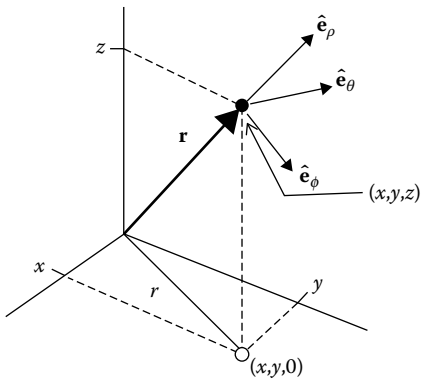


FIGURE 6.7
Right-handed basis $\{\hat{\mathbf{e}}_\rho, \hat{\mathbf{e}}_\phi, \hat{\mathbf{e}}_\theta\}$ in spherical coordinates.

6.2.5 Relationships to the Standard o.n. Basis

Recall from Corollary 2.6 in Section 2.4 that if $\{\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3\}$ is an o.n. basis for $\mathbb{R}^3$ and $\mathbf{x}$ is in $\mathbb{R}^3$, then

$$\mathbf{x} = \langle \mathbf{x}, \mathbf{q}_1 \rangle \mathbf{q}_1 + \langle \mathbf{x}, \mathbf{q}_2 \rangle \mathbf{q}_2 + \langle \mathbf{x}, \mathbf{q}_3 \rangle \mathbf{q}_3.$$

Concerning the relationship between the o.n. bases $\{\hat{\boldsymbol{\imath}}, \hat{\boldsymbol{\jmath}}, \hat{\boldsymbol{k}}\}$ and $\{\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z\}$, we use Corollary 2.6 in Section 2.4 to summarize the information of (6.14) in Table 6.1.

Symmetry of the dot product, that is, $\mathbf{x} \bullet \mathbf{y} = \mathbf{y} \bullet \mathbf{x}$, and Corollary 2.6 in Section 2.4, imply that

$$\hat{\boldsymbol{\imath}} = (\hat{\boldsymbol{\imath}} \bullet \hat{\mathbf{e}}_r)\hat{\mathbf{e}}_r + (\hat{\boldsymbol{\imath}} \bullet \hat{\mathbf{e}}_\theta)\hat{\mathbf{e}}_\theta + (\hat{\boldsymbol{\imath}} \bullet \hat{\mathbf{e}}_z)\hat{\mathbf{e}}_z = \cos\theta\, \hat{\mathbf{e}}_r - \sin\theta\, \hat{\mathbf{e}}_\theta,$$

$$\hat{\boldsymbol{\jmath}} = (\hat{\boldsymbol{\jmath}} \bullet \hat{\mathbf{e}}_r)\hat{\mathbf{e}}_r + (\hat{\boldsymbol{\jmath}} \bullet \hat{\mathbf{e}}_\theta)\hat{\mathbf{e}}_\theta + (\hat{\boldsymbol{\jmath}} \bullet \hat{\mathbf{e}}_z)\hat{\mathbf{e}}_z = \sin\theta\, \hat{\mathbf{e}}_r + \cos\theta\, \hat{\mathbf{e}}_\theta,$$

and

$$\hat{\boldsymbol{k}} = (\hat{\boldsymbol{k}} \bullet \hat{\mathbf{e}}_r)\hat{\mathbf{e}}_r + (\hat{\boldsymbol{k}} \bullet \hat{\mathbf{e}}_\theta)\hat{\mathbf{e}}_\theta + (\hat{\boldsymbol{k}} \bullet \hat{\mathbf{e}}_z)\hat{\mathbf{e}}_z = \hat{\mathbf{e}}_z.$$

Similarly, concerning the relationship between the o.n. bases $\{\hat{\boldsymbol{\imath}}, \hat{\boldsymbol{\jmath}}, \hat{\boldsymbol{k}}\}$ and $\{\hat{\mathbf{e}}_\rho, \hat{\mathbf{e}}_\phi, \hat{\mathbf{e}}_\theta\}$, we use Corollary 2.6 in Section 2.4 to summarize the information of (6.18) in Table 6.2.

Symmetry of the dot product and Corollary 2.6 in Section 2.4, imply that

$$\hat{\boldsymbol{\imath}} = (\hat{\boldsymbol{\imath}} \bullet \hat{\mathbf{e}}_\rho)\hat{\mathbf{e}}_\rho + (\hat{\boldsymbol{\imath}} \bullet \hat{\mathbf{e}}_\phi)\hat{\mathbf{e}}_\phi + (\hat{\boldsymbol{\imath}} \bullet \hat{\mathbf{e}}_\theta)\hat{\mathbf{e}}_\theta = \sin\phi\, \cos\theta\, \hat{\mathbf{e}}_\rho + \cos\phi\, \cos\theta\, \hat{\mathbf{e}}_\phi - \sin\theta\, \hat{\mathbf{e}}_\theta,$$

$$\hat{\boldsymbol{\jmath}} = (\hat{\boldsymbol{\jmath}} \bullet \hat{\mathbf{e}}_\rho)\hat{\mathbf{e}}_\rho + (\hat{\boldsymbol{\jmath}} \bullet \hat{\mathbf{e}}_\phi)\hat{\mathbf{e}}_\phi + (\hat{\boldsymbol{\jmath}} \bullet \hat{\mathbf{e}}_z)\hat{\mathbf{e}}_\theta = \sin\phi\, \sin\theta\, \hat{\mathbf{e}}_\rho + \cos\phi\, \sin\theta\, \hat{\mathbf{e}}_\phi + \cos\theta\, \hat{\mathbf{e}}_\theta,$$

and

$$\hat{\boldsymbol{k}} = (\hat{\boldsymbol{k}} \bullet \hat{\mathbf{e}}_\rho)\hat{\mathbf{e}}_\rho + (\hat{\boldsymbol{k}} \bullet \hat{\mathbf{e}}_\phi)\hat{\mathbf{e}}_\phi + (\hat{\boldsymbol{k}} \bullet \hat{\mathbf{e}}_\theta)\hat{\mathbf{e}}_\theta = \cos\phi\, \hat{\mathbf{e}}_\rho - \sin\phi\, \hat{\mathbf{e}}_\phi.$$

TABLE 6.1

Entries Are Values of $\mathbf{x} \bullet \mathbf{y}$

$\mathbf{x} \setminus \mathbf{y}$	$\hat{\boldsymbol{\imath}}$	$\hat{\boldsymbol{\jmath}}$	$\hat{\boldsymbol{k}}$
$\hat{\mathbf{e}}_r$	$\cos\theta$	$\sin\theta$	0
$\hat{\mathbf{e}}_\theta$	$-\sin\theta$	$\cos\theta$	0
$\hat{\mathbf{e}}_z$	0	0	1

TABLE 6.2

Entries Are Values of $\mathbf{x} \bullet \mathbf{y}$

$\mathbf{x} \setminus \mathbf{y}$	$\hat{\boldsymbol{\imath}}$	$\hat{\boldsymbol{\jmath}}$	$\hat{\boldsymbol{k}}$
$\hat{\mathbf{e}}_\rho$	$\sin\phi\, \cos\theta$	$\sin\phi\, \sin\theta$	$\cos\phi$
$\hat{\mathbf{e}}_\phi$	$\cos\phi\, \cos\theta$	$\cos\phi\, \sin\theta$	$-\sin\phi$
$\hat{\mathbf{e}}_\theta$	$-\sin\theta$	$\cos\theta$	0

6.2.6 Problems

1. Find the exact polar coordinates of the point $(x, y) = (\sqrt{6}, -\sqrt{2})$.
2. Find the exact polar coordinates of the point $(x, y) = (-1, -1)$.
3. What are the exact (r, θ, z) coordinates of the points (a) $(x, y, z) = (-\sqrt{3}, 1, 5)$ and (b) $(x, y, z) = (-1, \sqrt{3}, 4)$?
4. What are the exact (ρ, ϕ, θ) coordinates of the point $(x, y, z) = (\frac{3}{2}, -\frac{\sqrt{3}}{2}, -\sqrt{3})$?
5. What are the exact (ρ, ϕ, θ) coordinates of the point (a) $(x, y, z) = (-\frac{\sqrt{2}}{2}, -\frac{\sqrt{6}}{2}, -\sqrt{2})$, (b) $(x, y, z) = (-\frac{\sqrt{6}}{2}, \frac{\sqrt{2}}{2}, -\sqrt{2})$?
6. What are the exact (ρ, ϕ, θ) coordinates of the point $(x, y, z) = (\frac{\sqrt{3}}{2}, -\frac{3}{2}, -\sqrt{3})$?
7. Find $\hat{\mathbf{e}}_r$, $\hat{\mathbf{e}}_\theta$, and $\hat{\mathbf{e}}_z$ in terms of $\hat{\mathbf{e}}_\rho$, $\hat{\mathbf{e}}_\phi$, and $\hat{\mathbf{e}}_\theta$.

6.3 Curves and Surfaces

In calculus, the first curves we study are graphs, for example, described by $y = f(x)$, and likewise the first surfaces we study are graphs, for example, described by $z = f(x, y)$. We also study curves and surfaces that are parametrized.

A curve $\mathcal{C}$ is **parametrized** by t if the curve's points, P, have position vectors, $\overrightarrow{OP} = \mathbf{r} = \mathbf{r}(t)$, where the function $\mathbf{r}(t)$ is defined *and continuous* on an interval of t values, for example, $a \leq t \leq b$. If the parametrization gives motion along the curve, then the parameter, t, measures time. In any case, an arrow on the curve indicates the direction of "travel" as t increases.

Example 6.8

$$\mathcal{C}: \ \mathbf{r} = \mathbf{r}_0 + t\mathbf{v}, \ 0 \leq t \leq 1, \tag{6.19}$$

is a straight line segment from $\mathbf{r}_0$ to $\mathbf{r}_0 + \mathbf{v}$, as shown in Figure 6.8.

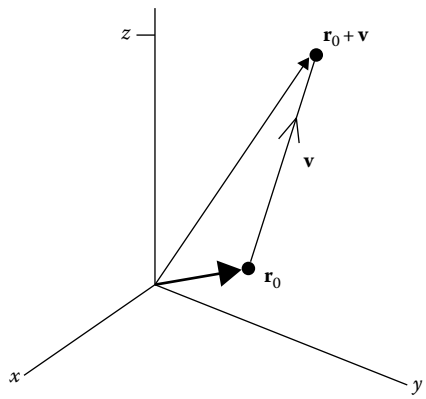


FIGURE 6.8
Line segment from $\mathbf{r}_0$ to $\mathbf{r}_0 + \mathbf{v}$ in Example 6.8.

In particular, if $\mathbf{v} = \mathbf{r}_1 - \mathbf{r}_0$, then the line goes from $\mathbf{r}_0$ to $\mathbf{r}_1$. The line in Example 6.8 lies in the plane $z = 0$ only if both $\mathbf{v} \bullet \hat{\boldsymbol{k}} = \mathbf{r}_0 \bullet \hat{\boldsymbol{k}} = 0$.

Example 6.9

$$\mathcal{C}: \ \mathbf{r} = a\cos t\,\hat{\boldsymbol{\imath}} + a\sin t\,\hat{\boldsymbol{\jmath}},\ 0 \le t \le 2\pi, \tag{6.20}$$

shown in Figure 6.9, is a circle of radius $|a| > 0$ and center at the origin in the $z = 0$ plane.

Usually we will take $a > 0$. The direction of travel on the circle in Example 6.9 is counterclockwise in the xy-plane, even if $a < 0$.

Example 6.10

Parametrize the ellipse shown in Figure 6.10.

Method: Recall that the standard form of an equation for an ellipse, whose center is at the origin, is

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \tag{6.21}$$

where a, b are positive constants. But having a 1 on the right-hand side and the sum of squares on the left-hand side remind us of the Pythagorean identity in trigonometry, $\cos^2\theta + \sin^2\theta \equiv 1$. Looking at (6.21) suggests rewriting the Pythagorean identity as

$$\frac{a^2\cos^2\theta}{a^2} + \frac{b^2\sin^2\theta}{b^2} = 1.$$

This suggests parametrizing the ellipse using $x = a\cos t,\ y = b\sin t$, that is, writing

$$\mathcal{C}: \ \mathbf{r}(t) = a\,\cos t\,\hat{\boldsymbol{\imath}} + b\,\sin t\,\hat{\boldsymbol{\jmath}}, 0 \le t \le 2\pi\,. \bigcirc \tag{6.22}$$

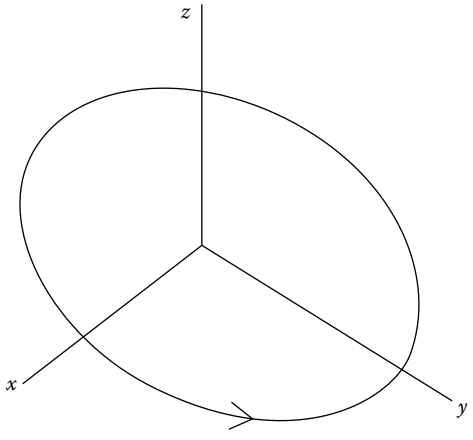


FIGURE 6.9
Circle of radius a in the plane $z = 0$ in Example 6.9.

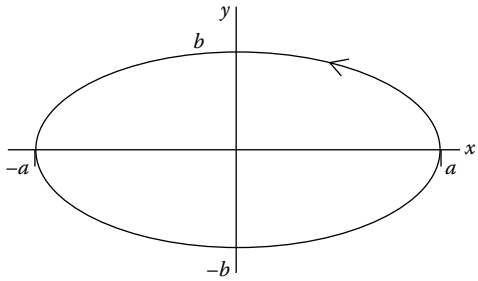


FIGURE 6.10
Ellipse in the plane $z = 0$ in Example 6.10.

Example 6.11

In calculus, we learned about curves written in the form $r = f(\theta)$. This is really just shorthand for the parametrization

$$\mathcal{C}: \ \mathbf{r}(\theta) = f(\theta) \cdot (\cos\theta\, \hat{\boldsymbol{\imath}} + \sin\theta\, \hat{\boldsymbol{\jmath}}), \alpha \leq \theta \leq \beta. \tag{6.23}$$

Example 6.12

Parametrize the circle $x^2 + y^2 = 4x$ shown in Figure 6.11.

Method: In polar coordinates, the equation of the circle is

$$r^2 = 4r\cos\theta. \tag{6.24}$$

For those points on the circle that have $r \neq 0$, we can divide (6.24) by r to get

$$r = 4\cos\theta. \tag{6.25}$$

By Example 6.11, let's try the parametrization

$$\mathbf{r}(\theta) = 4\cos\theta \cdot (\cos\theta\, \hat{\boldsymbol{\imath}} + \sin\theta\, \hat{\boldsymbol{\jmath}}).$$

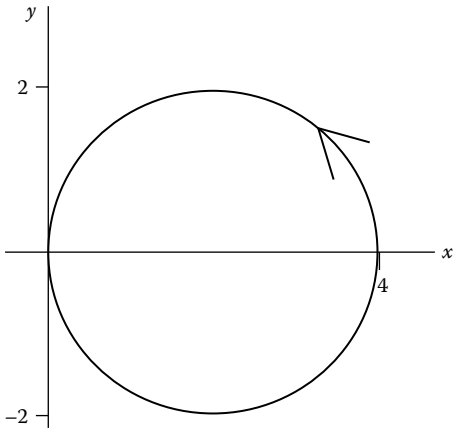


FIGURE 6.11
Example 6.12.

Are we done? No, not yet. The first issue is, "Does this parametrization give the correct direction of travel on the curve?" The answer is, "Yes," as we can see from $\mathbf{r}(0) = 4\hat{\boldsymbol{\imath}}$ and $\mathbf{r}(\frac{\pi}{2}) = \mathbf{0}$. Also, we need to find an interval of θ values that will give the whole curve. While the interval $0 \leq \theta \leq 2\pi$ gives the circle twice, the interval $0 \leq \theta \leq \pi$ gives the circle once. To summarize, our curve can be parametrized by

$$\mathcal{C}: \ \mathbf{r}(\theta) = 4\cos\theta \cdot (\cos\theta\,\hat{\boldsymbol{\imath}} + \sin\theta\,\hat{\boldsymbol{\jmath}}),\ 0 \leq \theta \leq \pi. \bigcirc \tag{6.26}$$

Using the identities $\cos^2\theta = \frac{1}{2}(1 + \cos 2\theta)$ and $\sin\theta\cos\theta = \frac{1}{2}\sin 2\theta$, we can rewrite (6.26) as $\mathcal{C}: \ \mathbf{r}(\theta) = (2 + 2\cos 2\theta)\,\hat{\boldsymbol{\imath}} + 2\sin 2\theta\,\hat{\boldsymbol{\jmath}},\ 0 \leq \theta \leq \pi$. This explains why $0 \leq \theta \leq \pi$ works so well to give the circle once.

One thing in Example 6.12 that may bother us is that in deriving the equation in polar coordinates, $r = 4\cos\theta$, we divided by r and said that was o. k. for $r \neq 0$, but our circle passes through the origin. This illustrates the principle that in the process of finding a parametrization, or, more generally, the solving of any mathematical problem, we may perform a dubious step *as long as* the solution we arrive at can be verified to satisfy all of the requirements of the problem. In other words, a dubious step is allowed if it does not ultimately produce a spurious solution or an incorrect "solution" and does not lose any solution.

Example 6.12 naturally leads to the next concepts.

Definition 6.2

A curve $\mathcal{C}: \ \mathbf{r} = \mathbf{r}(t),\ \alpha \leq t \leq \beta$, is

(a) **Closed** if $\mathbf{r}(\beta) = \mathbf{r}(\alpha)$

(b) **Simple** if $\mathbf{r}(t_1) \neq \mathbf{r}(t_2)$ for all t_1, t_2 with $\alpha \leq t_1 < t_2 \leq \beta$ except possibly $\mathbf{r}(\alpha) = \mathbf{r}(\beta)$.

The curves of Example 6.8, parametrized by (6.19); Example 6.9, parametrized by (6.20); Example 6.10, parametrized by (6.22); and Example 6.11, parametrized by (6.26), are all simple. The latter three are also closed.

A "figure eight" curve, for example, the one shown in Figure 6.12, is not simple, no matter how it is parametrized. Figure 6.12 was drawn using the *Mathematica*™ command:

ParametricPlot[{2Cos[t], Sin[2t]}, {t, 0, 2Pi},PlotStyle → Thick].

6.3.1 Curves and Calculus

Suppose a point P_0 is on a simple curve

$$\mathcal{C}: \ \mathbf{r} = \mathbf{r}(t),\ \alpha \leq t \leq \beta.$$

Suppose at some t_0 in the interval $\alpha < t_0 < \beta$, the position vector is $\overrightarrow{OP_0} = \mathbf{r}(t_0)$. As shown in Figure 6.13, the tangent vector to the curve $\mathcal{C}$ at P_0 is given by

$$\mathbf{T} \triangleq \dot{\mathbf{r}}(t_0), \tag{6.27}$$

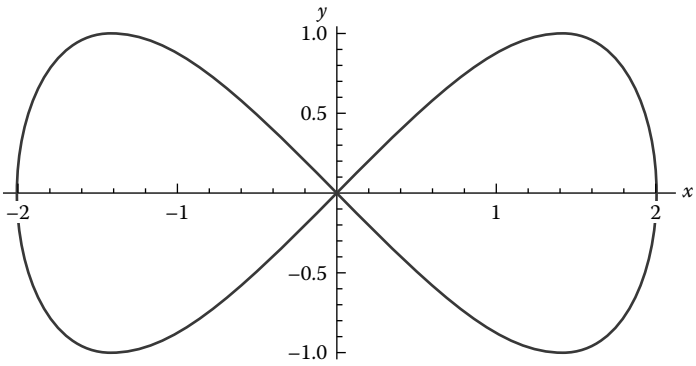


FIGURE 6.12
Figure eight.

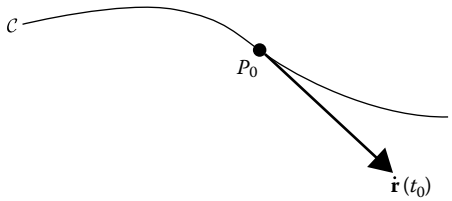


FIGURE 6.13
Tangent vector to curve at a point.

assuming the derivative exists at t_0 *and* is not the zero vector. If

$$\mathbf{r}(t) = x(t)\,\hat{\boldsymbol{\imath}} + y(t)\,\hat{\boldsymbol{\jmath}} + z(t)\,\hat{\boldsymbol{k}},$$

then

$$\dot{\mathbf{r}}(t) = \dot{x}(t)\,\hat{\boldsymbol{\imath}} + \dot{y}(t)\,\hat{\boldsymbol{\jmath}} + \dot{z}(t)\,\hat{\boldsymbol{k}},$$

assuming all three scalar function derivatives exist.

Linear approximation of the three scalar functions $x(t), y(t)$, and $z(t)$ implies linear approximation of points on the curve:

$$\mathbf{r}(t_0 + \triangle t) \approx \mathbf{r}(t_0) + (\triangle t)\mathbf{T}\,.$$

If t measures time, then the tangent vector shows the instantaneous direction (and speed) of travel along the curve.

If the curve is simple and $\dot{\mathbf{r}}(t_0) = \mathbf{0}$, then there is no tangent vector to the curve at the point whose position vector is $\mathbf{r}(t_0)$. If t measures time, then instantaneously the motion stops at $t = t_0$ if $\dot{\mathbf{r}}(t_0) = \mathbf{0}$.

By definition $\mathbf{T} \neq \mathbf{0}$, so the unit tangent vector to $\mathcal{C}$ at P_0 exists and is

$$\widehat{\mathbf{T}} = \frac{1}{||\mathbf{T}||}\,\mathbf{T}\,.$$

A special case of a parametrized curve is given by the solution of a system of ODEs $\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x})$: if the system in $\mathbb{R}^3$ has a solution

$$\mathbf{x}(t) = \begin{bmatrix} x(t) \\ y(t) \\ z(t) \end{bmatrix},$$

or a system in $\mathbb{R}^2$ has a solution

$$\mathbf{x}(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix},$$

then at a point $\mathbf{x}(t_0)$ the solution curve has tangent vector $\mathbf{F}(\mathbf{x}(t_0))$ *at that point*. The collection of all tangent vectors at their respective points forms a **vector field**.

Example 6.13

Consider the ODE system in the plane

$$\dot{\mathbf{x}} \triangleq \begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} y \\ -\omega^2 x \end{bmatrix}. \tag{6.28}$$

Recalling results from Section 3.3 and Section 5.3, you will explain in Problem 6.3.4.30 why the solutions of (6.28) can be written in the form

$$\mathbf{x}(t) = \begin{bmatrix} A\cos(\omega t - \delta) \\ -\omega A \sin(\omega t - \delta) \end{bmatrix}, \tag{6.29}$$

where A, ω, δ are scalar constants and A, ω are positive. The vector field

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} y \\ -\omega^2 x \end{bmatrix}$$

is shown in Figure 6.14a. We can see that the tangent vectors suggest that the solutions, $\mathbf{r}(t) = x(t)\,\hat{\boldsymbol{\imath}} + \dot{x}(t)\,\hat{\boldsymbol{\jmath}}$, travel clockwise around the origin, perhaps along circles, ellipses, spirals, or some similar-looking curves. We know the solutions, as shown in Figure 6.14b, are ellipses because

$$\frac{(x(t))^2}{A^2} + \frac{(\dot{x}(t))^2}{(\omega A)^2} = \left(\frac{x(t)}{A}\right)^2 + \left(\frac{\dot{x}(t)}{\omega A}\right)^2 = \cos^2(\omega t - \delta) + \sin^2(\omega t - \delta) \equiv 1.$$

The solutions are shown superimposed on the vector field in Figure 6.14c.

Example 6.14

Suppose a solution of ODE system (6.28) passes through the points $(x, \dot{x}) = (0, 4)$ and $(-\sqrt{2}, 0)$, for some constant ω. Find the frequency of vibration, ω.

Method: The second data point tells us that for some δ, "time" t_0, and positive A, ω, we get $(-\sqrt{2}, 0) = (A\cos(\omega t_0 - \delta), -\omega A \sin(\omega t_0 - \delta))$. Let $\theta \triangleq \omega t_0 - \delta$, so this becomes $-\sqrt{2} = A\cos\theta$ and $0 = A\sin\theta$. It follows that $A = \sqrt{2}$ and $\theta = (2k+1)\pi$ for any integer k.

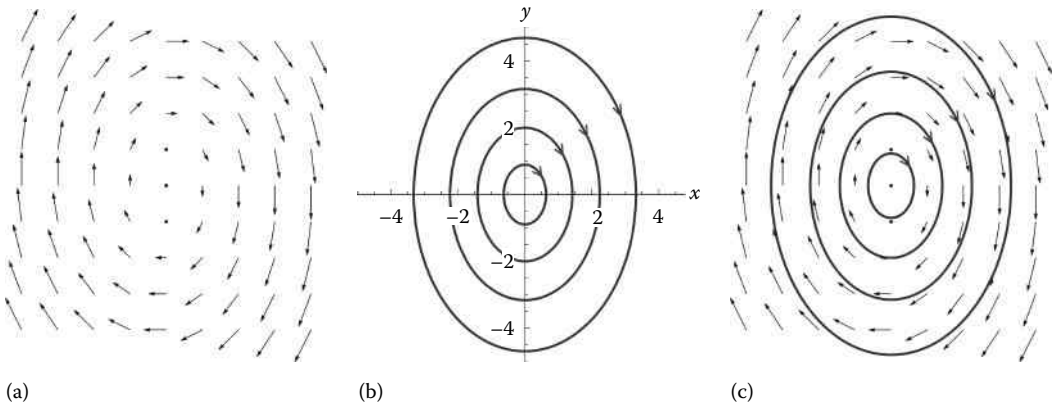


FIGURE 6.14
Solutions of undamped oscillator ODE system are ellipses, for Example 6.13. (a) Oscillator vector field, (b) oscillator solution curves, and (c) oscillator vector field and solutions.

Now, substitute the value of A into the form of the solution and the first data point to get $(0,4) = (\sqrt{2}\cos(\omega t_1 - \delta), -\omega\sqrt{2}\sin(\omega t_1 - \delta))$, for some δ, "time" t_1, and positive ω. Let $\varphi = \omega t_1 - \delta$, so this becomes $0 = \sqrt{2}\cos\varphi$ and $4 = -\omega\sqrt{2}\sin\varphi$. It follows that $4 = \omega\sqrt{2}$ and $\varphi = \left(n - \frac{1}{2}\right)\pi$ for some integer n. The frequency of vibration is $\omega = \frac{4}{\sqrt{2}} = 2\sqrt{2}$. ○

In our work in Example 6.14 to find ω, we didn't actually need to find θ and φ, but finding their values was part of the method.

6.3.2 Zhukovskii Airfoil

One interesting class of curves, called **Zhukovskii airfoils**, gives shapes that look like the cross sections of airplane wings. For example, define

$$x(\xi,\eta) \triangleq \xi \cdot \frac{1+\xi^2+\eta^2}{2(\xi^2+\eta^2)}, \quad y(\xi,\eta) \triangleq \eta \cdot \frac{-1+\xi^2+\eta^2}{2(\xi^2+\eta^2)}, \tag{6.30}$$

where

$$\xi(t) \triangleq -0.07 + \rho_0\cos(t), \quad \eta(t) \triangleq 0.05 + \rho_0\sin(t), \quad \text{and} \quad \rho_0 = \sqrt{(1+0.07)^2 + (0.05)^2}.$$

The curve plotted by *Mathematica* is shown in Figure 6.15a. This curve is simple, closed, and piecewise smooth, even though it has a cusp at the point $(x,y) = (1,0)$. A similar-looking closed, but not simple, curve that is not a Zhukovskii airfoil is shown in Figure 6.15b.

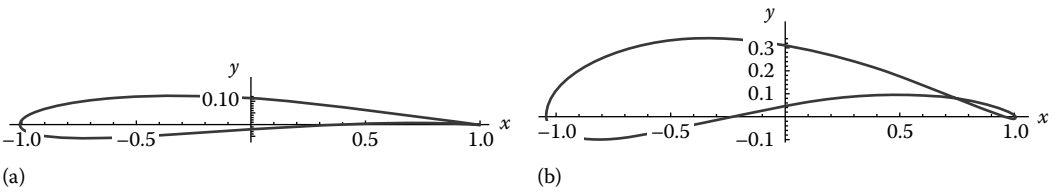


FIGURE 6.15
(a) A Zhukovskii airfoil. (b) Not a Zhukovskii airfoil.

6.3.3 Surfaces

A **parametrized surface** is given by

$$\mathcal{S}: \ \mathbf{r} = \mathbf{r}(u, v), \ (u, v) \text{ in some planar region } \mathcal{D}.$$

The independent variables u, v are called the **parameters** in the parametrization. A special case is when the surface is a graph of a function, for example, $z = f(x, y)$, which we will see can be parametrized using x, y as the parameters.

Example 6.15

Find a parametrization of the surface $\mathcal{S}$ that is the part of the paraboloid $z = 3 - x^2 - y^2$ that lies on or above the xy-plane.

Method: The paraboloid is shown in Figure 6.16. The paraboloid intersects the xy-plane where $z = 0$; hence $0 = 3 - x^2 - y^2$, that is, the circle $x^2 + y^2 = 3$. To be on the part of the paraboloid that lies *above* the $z = 0$ plane, that is, satisfies $z > 0$, we need

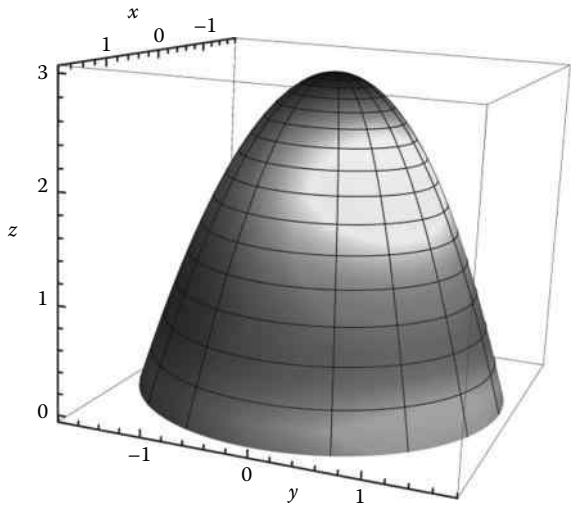


FIGURE 6.16
Paraboloidal surface in Example 6.15.

$x^2 + y^2 = 3 - z < 3$, that is, $x^2 + y^2 < 3$. So, choose the planar region:

$$\mathcal{D} = \{(x, y) : \ x^2 + y^2 \leq 3\}.$$

This suggests using x, y as the parameters. We have

$$\mathcal{S} : \ \mathbf{r} = x\,\hat{\boldsymbol{\imath}} + y\,\hat{\boldsymbol{\jmath}} + (3 - x^2 - y^2)\,\hat{\boldsymbol{k}}, \ (x, y) \text{ in } \mathcal{D}. \bigcirc$$

Example 6.16

Use polar coordinates to find a different parametrization of the surface in Example 6.15.

Method: Because $\mathcal{D}$ is the disk of radius $\sqrt{3}$ and center at the origin in the xy-plane, it makes sense to use polar coordinates: Let

$$\mathcal{D} = \{(r, \theta) : \ 0 \leq r \leq \sqrt{3}, \ 0 \leq \theta \leq 2\pi\},$$

$x = r\cos\theta$, $y = r\sin\theta$. So, another parametrization for $\mathcal{S}$ is

$$\mathcal{S} : \ \mathbf{r} = r\cos\theta\,\hat{\boldsymbol{\imath}} + r\sin\theta\,\hat{\boldsymbol{\jmath}} + (3 - r^2)\,\hat{\boldsymbol{k}}, \ (r, \theta) \text{ in } \mathcal{D}. \bigcirc$$

Here are some famous surfaces expressed in cylindrical coordinates:

$$\begin{aligned} \textit{Sphere} : &\ \ r^2 + z^2 = a^2, \ \text{ some positive constant } a. \\ \textit{Cylinder} : &\ \ r = a, \ \text{ some positive constant } a. \\ \textit{Cone} : &\ \ z = a + b\,r, \ \text{ some constants } a, b. \\ \textit{Paraboloid} : &\ \ z = a + b\,r^2, \ \text{ some constants } a, b. \end{aligned}$$

Example 6.17

Find the curve(s) of intersection of the sphere $x^2 + y^2 + z^2 = 5$ and the cone $z = 7 - 4\sqrt{x^2 + y^2}$.

Method: These two surfaces are $r^2 + z^2 = 5$ and $z = 7 - 4r$. They intersect when $5 = r^2 + z^2 = r^2 + (7 - 4r)^2 = 17r^2 - 56r + 49$, that is,

$$0 = 17r^2 - 56r + 44.$$

The quadratic formula gives

$$r_\pm = \frac{56 \pm \sqrt{56^2 - 4 \cdot 17 \cdot 44}}{2 \cdot 17} = \frac{56 \pm \sqrt{144}}{34} = \left\{ \begin{array}{ll} 2, & \text{if } + \\ \frac{22}{17}, & \text{if } - \end{array} \right\}.$$

Because both $r_\pm$ are less than $\sqrt{5}$, neither is a spurious solution. So, there are two circles of intersection points:

$$x^2 + y^2 = 4, \ z = -1 \text{ and}$$

$$x^2 + y^2 = \left(\frac{22}{17}\right)^2, \ z = 7 - 4 \cdot \frac{22}{17} = \frac{31}{17}. \bigcirc$$

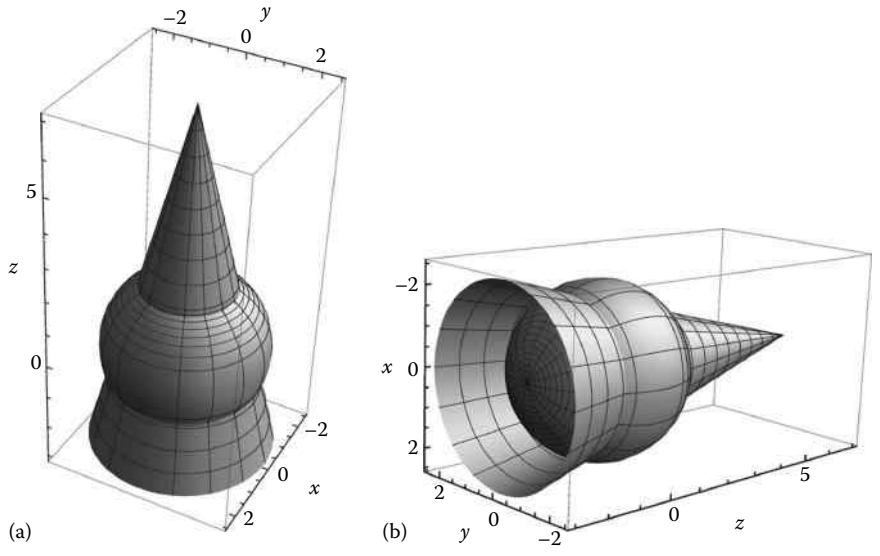


FIGURE 6.17
Intersection of sphere and cone.

Two views of the intersection of the sphere and the cone are given in Figure 6.17.

Example 6.18

A sphere of radius a and center at the origin can be described by the equation $\rho = a$. Find two different parametrizations of the part of the sphere that lies on or above the xy-plane.

Method:

(a) Because the problem involves a sphere, it makes sense to try to work in spherical coordinates, given in (6.12) in Section 6.2:

$$x = \rho \sin\phi \cos\theta, \quad y = \rho \sin\phi \sin\theta, \quad z = \rho \cos\phi.$$

On the sphere, $\rho \equiv a$, so

$$\mathbf{r} = a \sin\phi \cos\theta\, \hat{\boldsymbol{\imath}} + a \sin\phi \sin\theta\, \hat{\boldsymbol{\jmath}} + a \cos\phi\, \hat{\mathbf{k}}.$$

Our parameters are ϕ, θ. Are we done? Not yet: A parametrization consists of a formula for the position vector *and* the description of the domain $\mathcal{D}$ in which those parameters lie.

The part of the Earth that lies on or above the xy-plane consists of the northern hemisphere and the equator. In (x, y, z) coordinates, we need $z \geq 0$. Because $z = a\cos\phi$, that requires $\cos\phi \geq 0$, that is, $0 \leq \phi \leq \frac{\pi}{2}$. The sphere has circular symmetry about the z-axis, so $0 \leq \theta \leq 2\pi$. To summarize, the northern hemisphere and the equator can be parametrized by

$$\mathcal{S}: \ \mathbf{r} = a\sin\phi \cos\theta\, \hat{\boldsymbol{\imath}} + a\sin\phi \sin\theta\, \hat{\boldsymbol{\jmath}} + a\cos\phi\, \hat{\mathbf{k}},$$

$$\text{for } (\phi, \theta) \text{ in } \mathcal{D} = \left\{(\phi, \theta): \ 0 \leq \theta \leq 2\pi, 0 \leq \phi \leq \frac{\pi}{2}\right\}.$$

(b) Cylindrical coordinates are also relatively convenient for working with spheres. Our sphere is $r^2 + z^2 = a^2$; hence, $z = \pm\sqrt{a^2 - r^2}$. As in part (a), we require $z \geq 0$. So, the part of the sphere that lies on or above the xy-plane is the graph of $z = +\sqrt{a^2 - r^2}$. This gives a value of z only for $r \leq a$. In fact, our part of the sphere lies above the disk $0 \leq r \leq a$ in the xy-plane.

Similar to our work in Example 6.16, which also involved a surface that was a graph, we get a parametrization

$$\mathcal{S}: \ \mathbf{r} = r\cos\theta\,\hat{\boldsymbol{\imath}} + r\sin\theta\,\hat{\boldsymbol{\jmath}} + \sqrt{a^2 - r^2}\,\hat{\boldsymbol{k}},$$
$$\text{for } (r,\theta) \text{ in } \mathcal{D} = \{(r,\theta): \ 0 \leq \theta \leq 2\pi, 0 \leq r \leq a\}. \bigcirc$$

Here are some famous surfaces expressed in spherical coordinates:

$$\textit{Sphere}: \quad \rho = a, \ \text{ some positive constant } a$$
$$\textit{Sphere}: \quad \rho = 2a\cos\phi, \ \text{ some positive constant } a$$
$$\textit{The plane } z = 0: \quad \phi = \frac{\pi}{2}$$
$$\textit{Halfplane}: \quad \theta = \alpha, \ \text{ some constant } \alpha$$
$$\textit{Double cusped horn}: \quad \rho = 2a\sin\phi, \ \text{ some positive constant } a$$

Example 6.19

Explain why $\rho = 2a\cos\phi$, where a is positive constant, gives a sphere. Find its radius and center.

Method: Multiply both sides of the defining equation by ρ to get $x^2 + y^2 + z^2 = \rho^2 = 2a(\rho\,\cos\phi) = 2a\,z$. Completing the square gives

$$0 = x^2 + y^2 + z^2 - 2a\,z = x^2 + y^2 + (z^2 - 2a\,z + a^2 - a^2) = x^2 + y^2 + (z^2 - a)^2 - a^2,$$

that is,

$$x^2 + y^2 + (z^2 - a)^2 = a^2.$$

This is an equation of a sphere of radius a and center at $(x, y, z) = (0, 0, a)$. $\bigcirc$

The equation of a cusped horn looks like that for a sphere, but having $\sin\phi$ instead of $\cos\phi$ makes a big difference! Multiply both sides of the defining equation by ρ to get $r^2 + z^2 = \rho^2 = 2a(\rho\,\sin\phi) = 2a\,r$. Completing the square gives $z^2 = a^2 - (r - a)^2$, that is,

$$z = \pm\sqrt{a^2 - (r - a)^2}.$$

An example is drawn in Figure 6.18.

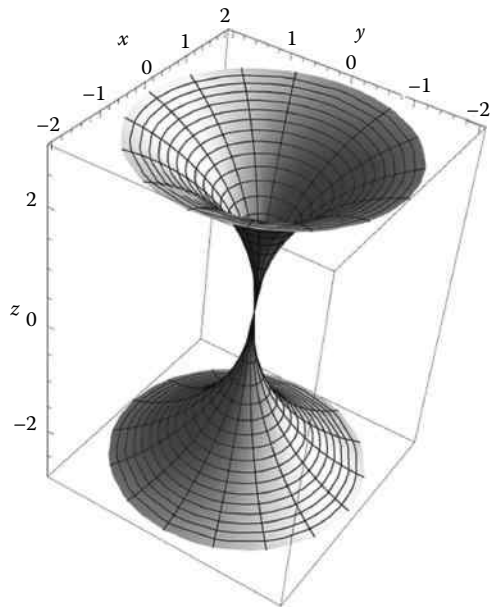


FIGURE 6.18
Double cusped horn.

Definition 6.3

A surface $\mathcal{S}:\ \mathbf{r} = \mathbf{r}(u, v),\ (u, v)$ in some planar region $\mathcal{D}$, is **simple** if

$$\mathbf{r}(u_1, v_1) \neq \mathbf{r}(u_2, v_2)$$

when (u_1, v_1) and (u_2, v_2) are different points in $\mathcal{D}$, except possibly for a curve of points in $\mathcal{S}$.

For example, the side of a cylinder is a simple surface.

6.3.4 Problems

For problems 1–3, find a parametrization of the curve and sketch it.

1. The line from $(1, 2, 4)$ to $(-1, 0, 3)$.
2. The circle $x^2 + y^2 = -6y$.
3. The circle $2x^2 + 2y^2 = 16y$.

In problems 4–10, sketch the curve and put an arrow on it to indicate the direction of travel.

4. $\mathcal{C}: \mathbf{r} = t^2\,\hat{\boldsymbol{\imath}} + t^4\,\hat{\boldsymbol{k}},\ 1 \leq t \leq 2.$
5. $\mathcal{C}: \mathbf{r} = 6\cos 2t\,\hat{\boldsymbol{\imath}} + 2\sin 2t\,\hat{\boldsymbol{k}},\ 0 \leq t \leq \pi.$

6. $\mathcal{C}: \mathbf{r} = 2\cos 2t\,\hat{\boldsymbol{\imath}} + \sin t\,\hat{\boldsymbol{k}},\ 0 \le t \le \frac{\pi}{2}$, using technology, for example, *Mathematica* or a graphing calculator.
7. $r = \sin 3\theta$.
8. $r = 1 - \cos\theta$.
9. $r = 2 - \sin\theta$.
10. The intersection of the surfaces $r = 4$ and $2x + z = 9$.

For problems 11–13, find a parametrization of the curve and state whether your parametrization gives a simple and/or closed curve.

11. The ellipse $\frac{x^2}{9} + \frac{y^2}{4} = 1$.
12. Ellipse $\frac{x^2}{9} + y^2 = 4$.
13. Ellipse $x^2 + 9y^2 = 12x$.
14. What, if anything, is the difference(s) between the curves given by (6.20) and the curve $\mathcal{C}_2:\ \mathbf{r} = a\cos(2\pi - t)\hat{\boldsymbol{\imath}} + a\sin(2\pi - t)\hat{\boldsymbol{\jmath}},\ 0 \le t \le 2\pi$?
15. What, if anything, is the difference(s) between the curves given by (6.26) and the curve $\mathcal{C}_2:\ \mathbf{r}(\theta) = 4\cos\theta \cdot (\cos\theta\,\hat{\boldsymbol{\imath}} + \sin\theta\,\hat{\boldsymbol{\jmath}}),\ 0 \le \theta \le 2\pi$?
16. A fly's position vector is $\mathbf{r}(t) = -\cos 2t\,\hat{\boldsymbol{\imath}} + \sin 2t\,\hat{\boldsymbol{\jmath}} + t\,\hat{\boldsymbol{k}}$.
(a) Sketch the flight path of the fly, and (b) find the unit tangent vector as a function of time, t. (c) Describe in words the flight path of the fly.
17. Find parametric equations of the tangent line to the planar curve parametrized by $\mathbf{r}(t) = \sin t\,\hat{\boldsymbol{\imath}} + (\frac{6}{\pi}t - 1)\hat{\boldsymbol{\jmath}}$ at the point $(\frac{\sqrt{3}}{2}, 1)$.
18. Find parametric equations of the tangent line at the point $(2, 1, 5)$ on the curve found by intersecting the surface $z = x^2y - 2x + 5$ with the plane $y = 1$.
19. Find the exact points(s) on the planar curve parametrized by $\mathbf{r}(t) = \sin t\,\hat{\boldsymbol{\imath}} + (t + 2\cos t)\hat{\boldsymbol{\jmath}}$ at which the tangent line is parallel to the y-axis.

For problems 20–24, describe in words the surface in $\mathbb{R}^3$, say what kind of geometrical object it is, and sketch it.

20. $x^2 + z^2 = 4$.
21. $\rho = 4\cos\phi$.
22. $\phi = \frac{\pi}{12}$.
23. $r = a\cos\theta$, where a is an unspecified constant.
24. $z = 5 - 3r$.

For problems 25–29, find two different parametrizations of each surface and describe what kind of geometrical object it is.

25. $x^2 + y^2 + z^2 = -z$.
26. $x^2 + y^2 + z^2 = 4x$.
27. $x^2 + y^2 = z$.
28. $\sqrt{3}\,(x^2 + y^2) = z$.

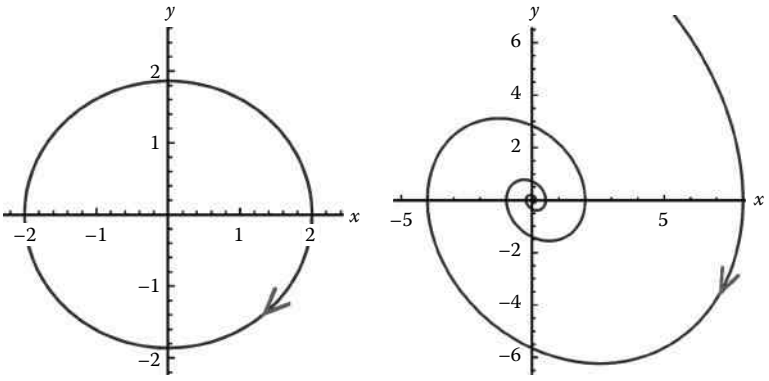


FIGURE 6.19
Figures useful for (a) Problem 6.3.4.31 and (b) Problem 6.3.4.32.

29. $\sqrt{4x^2 + 2y^2} - 2z = 0.$
30. Explain why the solutions of ODE system (6.28) can be written in the form (6.29).
31. Suppose that for some positive constant ω, a solution of the ODE system

$$\dot{\mathbf{x}} \triangleq \begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} y \\ -\omega^2 x \end{bmatrix}$$

passes through the points $(2, 0)$ $(0, -\sqrt{3})$, as shown in Figure 6.19a. Find the exact value of ω, if possible.

32. [Small project] Suppose that for some positive constants α, ω, a solution of the damped oscillator ODE system

$$\dot{\mathbf{x}} \triangleq \begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} y \\ -\omega^2 x - 2\alpha y \end{bmatrix}$$

passes through the points $(x, \dot{x}) = (8, 0)$, $(0, -4\sqrt{2})$, and $(2, 0)$, as shown in Figure 6.19b. Find the values of α, ω exactly, if possible, or with at least six significant digits of accuracy. [Caution: This problem is much more subtle and complicated than Problem 6.3.4.31.]

33. Use cylindrical coordinates to find a parametrization of the surface $r = f(\theta, z)$, where θ and z vary over all possible physical points on the infinite cylinder $r = 1$.
34. Use spherical coordinates to find a parametrization of the surface $\rho = f(\phi, \theta)$, where ϕ and θ vary over all possible physical points on the unit sphere.
35. Sketch the solid $\mathcal{V} = \{(\rho, \phi, \theta) : 0 \le \theta \le 2\pi,\ 0 \le \phi \le \frac{\pi}{2}, 0 \le \rho \le 2\cos\phi\}$ and describe it in words. [Hint: To begin, multiply both sides of $\rho = 2\cos\phi$ by ρ to get $x^2 + y^2 + z^2 = \rho^2 = 2\rho\ \cos\phi$.]

6.4 Partial Derivatives

If $f = f(x, y)$, then its partial derivatives are defined by

$$\frac{\partial}{\partial x}\left[f\right] = \frac{\partial f}{\partial x}(x,y) \triangleq \lim_{\triangle x \to 0} \frac{f(x+\triangle x, y) - f(x,y)}{\triangle x}, \tag{6.31}$$

assuming the limit exists, and

$$\frac{\partial}{\partial y}\left[f\right] = \frac{\partial f}{\partial y}(x,y) \triangleq \lim_{\triangle y \to 0} \frac{f(x, y+\triangle y) - f(x,y)}{\triangle y}, \tag{6.32}$$

assuming the limit exists.

When taking the partial derivative with respect to x, treat y as if it were constant, for example,

$$\frac{\partial}{\partial x}\left[\, y \, \sin 2x - y^2 \,\right] = y \, \frac{\partial}{\partial x}\left[\, \sin 2x \,\right] - \frac{\partial}{\partial x}\left[\, y^2 \,\right] = y \, 2 \cos 2x - 0 = 2y \, \cos 2x.$$

Similarly, when taking the partial derivative with respect to y, treat x as if it were constant, for example, the chain rule in Calculus I implies

$$\frac{\partial}{\partial y}\left[\, \sin xy \,\right] = \cos xy \cdot \frac{\partial}{\partial y}\left[\, xy \,\right] = (\cos xy) \cdot x = x \, \cos xy.$$

Higher order derivatives are defined by

$$\frac{\partial^2 f}{\partial x^2} \triangleq \frac{\partial}{\partial x}\left[\frac{\partial f}{\partial x}\right], \quad \frac{\partial^2 f}{\partial x \partial y} \triangleq \frac{\partial}{\partial x}\left[\frac{\partial f}{\partial y}\right],$$

$$\frac{\partial^2 f}{\partial y \partial x} \triangleq \frac{\partial}{\partial y}\left[\frac{\partial f}{\partial x}\right], \quad \frac{\partial^2 f}{\partial y^2} \triangleq \frac{\partial}{\partial y}\left[\frac{\partial f}{\partial y}\right],$$

assuming the limits exist individually.

Theorem 6.2

(Clairaut's theorem) If $\frac{\partial^2 f}{\partial x \partial y}(x,y)$ and $\frac{\partial^2 f}{\partial y \partial x}(x,y)$ both exist and are continuous at (x_0, y_0), then they are equal, that is,

$$\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) = \frac{\partial^2 f}{\partial y \partial x}(x_0, y_0).$$

6.4.1 Linear Approximation

If $f = f(x, y)$, then the generalization of linear approximation is

$$f(x_0 + \triangle x, y_0 + \triangle y) \approx f(x_0, y_0) + \triangle x \frac{\partial f}{\partial x}(x_0, y_0) + \triangle y \frac{\partial f}{\partial y}(x_0, y_0) \triangleq f(x_0, y_0) + L(\triangle x, \triangle y). \tag{6.33}$$

Example 6.20

As we will derive in Section 10.4, the speed of sound is approximately

$$v_s = \sqrt{\frac{\gamma P}{\varrho}},$$

where
P is the atmospheric pressure
ϱ is the atmospheric gas mass density
γ is the adiabatic gas parameter in the relationship $PV^\gamma =$ constant.

For example, for $\gamma = 1.4$ and "standard" atmospheric conditions at sea level, that is, temperature $= 273.15$ K, $P = P_0 =$ (1 atmosphere) $= 101,323\,\text{N/m}^2$, and mass density $\varrho = \varrho_0 = 1.293\,\text{kg/m}^3$, the speed of sound is approximately $v_{s,0} = 331.22$ m/s.

Predict the approximate effect of a 2% decrease in the atmospheric pressure and a 3% increase in the mass density.

Method: Consider v_s to be a function of (P, ϱ), that is,

$$v_s(P, \varrho) = \sqrt{\frac{\gamma P}{\varrho}} = \gamma^{1/2} P^{1/2} \varrho^{-1/2}.$$

With $\triangle P = -0.02P$ and $\triangle \varrho = +0.03\varrho$, linear approximation at $(P, \varrho) = (P_0, \varrho_0)$ gives

$$\begin{aligned} v_s &\approx v_{s,0} + \triangle P \frac{\partial v_s}{\partial P}(P_0, \varrho_0) + \triangle \varrho \frac{\partial v_s}{\partial \varrho}(P_0, \varrho_0) \\ &= v_{s,0} + (-0.02P_0)\gamma^{1/2} \cdot \frac{1}{2} P_0^{-1/2} \varrho_0^{-1/2} + (0.03\varrho_0)\gamma^{1/2} P_0^{1/2} \cdot \left(-\frac{1}{2}\right) \varrho_0^{-3/2}, \end{aligned}$$

so

$$\begin{aligned} v_s &\approx v_{s,0} - 0.02 \cdot \frac{1}{2} \cdot \gamma^{1/2} \cdot P_0^{1/2} \varrho_0^{-1/2} + 0.03 \cdot \left(-\frac{1}{2}\right) \gamma^{1/2} \cdot P_0^{1/2} \varrho_0^{-1/2} \\ &= v_{s,0} - 0.01 \cdot v_{s,0} - 0.015 \cdot v_{s,0} = 0.975 v_{s,0} \approx 322.94 \text{ m/s}. \end{aligned}$$

So, a 2% decrease in the atmospheric pressure and a 3% increase in the mass density have the effect of decreasing the speed of sound by about 2.5%, thus adding the individual effects of about 1% and about 1.5%. ◯

TABLE 6.3

Kinematic Viscosity of Steam $\mu/\varrho = \nu(P,T)$, in $10^7\ \mathrm{m}^2/\mathrm{s}$

P\T	400	450	500	550	600	650	700
2.5	29.5	34.6	40.2	46.0	52.1	58.6	65.4
5	14.5	17.0	19.8	22.8	25.9	29.2	32.7
7.5	9.3	11.2	13.1	15.1	17.2	19.4	21.8
10	6.82	8.21	9.67	11.23	12.82	14.52	16.30
12.5	5.26	6.44	7.66	8.91	10.21	11.60	13.01
15	4.21	5.26	6.28	7.36	8.47	9.62	10.85
17.5	3.44	4.40	5.33	6.25	7.23	8.23	9.27

Example 6.21

Table 6.3 gives the kinematic viscosity of steam, $\nu = \nu(P,T)$, in $10^7\ \mathrm{m}^2/\mathrm{s}$, where the pressure, P, is measured in units of 10^6 Pa (pascals) and the temperature, T, is measured in °C. Use estimates of partial derivatives to approximate $\nu(16, 590)$.

Method: Linear approximation, with $\triangle P = 1$ and $\triangle T = -10$, gives

$$\nu(16,590) \approx \nu(15,600) + (1)\frac{\partial \nu}{\partial P}(15,600) + (-10)\frac{\partial \nu}{\partial T}(15,600).$$

Tu use this, we need to estimate the two partial derivatives at $(P,T) = (15,600)$. The simplest estimates are

$$\frac{\partial \nu}{\partial P}(15,600) \approx \frac{\nu(17.5,600) - \nu(15,600)}{17.5 - 15} = \frac{7.23 - 8.47}{2.5} \approx -0.496$$

and

$$\frac{\partial \nu}{\partial T}(15,600) \approx \frac{\nu(15,600) - \nu(15,550)}{600 - 550} = \frac{8.47 - 7.36}{50} \approx 0.0222.$$

So,

$$\nu(16,590) \approx 8.47 + (1)(-0.496) + (-10)(0.0222) \approx 7.752. \bigcirc$$

Using the dot product in $\mathbb{R}^2$, we can write part of (6.33) as

$$L(\triangle x, \triangle y) \triangleq \triangle x\, \frac{\partial f}{\partial x}(x_0,y_0) + \triangle y\, \frac{\partial f}{\partial y}(x_0,y_0) = \begin{bmatrix} \frac{\partial f}{\partial x}(x_0,y_0) \\ \\ \frac{\partial f}{\partial y}(x_0,y_0) \end{bmatrix} \bullet \begin{bmatrix} \triangle x \\ \\ \triangle y \end{bmatrix}.$$

This motivates defining the **gradient vector**:

$$\boldsymbol{grad}\, f = \nabla f \triangleq \frac{\partial f}{\partial x}\,\hat{\boldsymbol{\imath}} + \frac{\partial f}{\partial y}\,\hat{\boldsymbol{\jmath}}.$$

Using this definition of $\boldsymbol{grad}\, f$,

$$L(\triangle x, \triangle y) = \nabla f(x_0,y_0) \bullet (\triangle x\, \hat{\boldsymbol{\imath}} + \triangle y\, \hat{\boldsymbol{\jmath}}).$$

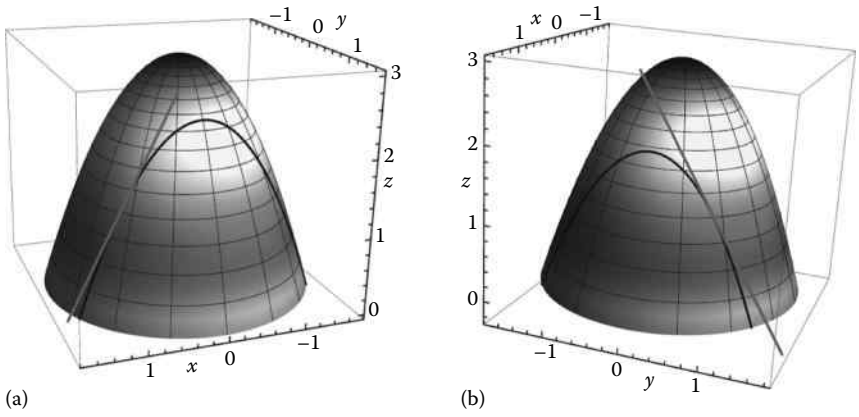


FIGURE 6.20
Partial derivatives with respect to (a) x and (b) y.

Because of the linear approximation formula (6.33), we define

$$(D_{\widehat{\mathbf{u}}}f)(x_0, y_0) \triangleq \nabla f(x_0, y_0) \bullet \widehat{\mathbf{u}}. \tag{6.34}$$

This is called the **directional derivative** of f in the direction of the unit vector $\widehat{\mathbf{u}}$.

Assuming $\nabla f(x_0, y_0) \neq \mathbf{0}$, the direction in which f increases the most is the unit vector $\widehat{\mathbf{u}}$ in the direction of $\nabla f(x_0, y_0)$, because $\mathbf{v} \bullet \widehat{\mathbf{u}} = ||\mathbf{v}||\ ||\widehat{\mathbf{u}}||\cos\theta$ is greatest when $\widehat{\mathbf{u}}$ is in the direction of $\mathbf{v}$.

Here are geometric interpretations of the partial derivatives and the directional derivative: Consider a surface $z = f(x, y)$. Intersecting the surface with the plane $y = y_0$ gives the curve $z = f(x, y_0)$ as x varies, and $\frac{\partial f}{\partial x}(x_0, y_0)$ is the slope of the tangent line to that curve at the point $(x, y, z) = \big(x_0, y_0, f(x_0, y_0)\big)$. This tangent line lies in the plane $y = y_0$. The curve and the tangent line are shown in Figure 6.20a.

Similarly, intersecting the same surface $z = f(x, y)$ with the plane $x = x_0$ gives the curve $z = f(x_0, y)$ as y varies, and $\frac{\partial f}{\partial y}(x_0, y_0)$ is the slope of the tangent line to that curve at the point $(x, y, z) = \big(x_0, y_0, f(x_0, y_0)\big)$. This tangent line lies in the plane $x = x_0$. The curve and the tangent line are shown in Figure 6.20b.

Finally, if $\widehat{\mathbf{u}} = a\,\hat{\boldsymbol{\imath}} + b\,\hat{\boldsymbol{\jmath}}$ is a unit vector, consider the parametric curve

$$\mathcal{C}:\ \mathbf{r}(t) = (x_0 + at)\hat{\boldsymbol{\imath}} + (y_0 + bt)\hat{\boldsymbol{\jmath}} + f(x_0 + at, y_0 + bt)\hat{\mathbf{k}},\ -\delta < t < \delta.$$

It also lies on the surface $z = f(x, y)$. Using the linear approximation in (6.33) we get

$$f(x_0 + at, y_0 + bt) \approx f(x_0, y_0) + at\frac{\partial f}{\partial x}(x_0, y_0) + bt\frac{\partial f}{\partial y}(x_0, y_0).$$

So, the tangent vector to $\mathcal{C}$ at the point $(x_0, y_0, f(x_0, y_0))$ is

$$\frac{d\mathbf{r}}{dt} = a\,\hat{\boldsymbol{\imath}} + b\,\hat{\boldsymbol{\jmath}} + \left(a\frac{\partial f}{\partial x}(x_0, y_0) + b\frac{\partial f}{\partial y}(x_0, y_0)\right)\hat{\mathbf{k}} = \widehat{\mathbf{u}} + (D_{\widehat{\mathbf{u}}}f)(x_0, y_0).$$

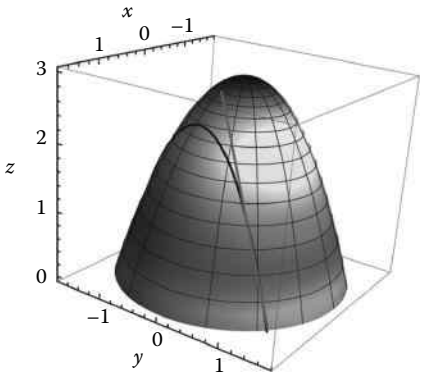


FIGURE 6.21
Directional derivative.

In some sense, $(D_{\widehat{\mathbf{u}}}f)(x_0, y_0)$ is the "slope" of the tangent vector to this curve, at the point $(x_0, y_0, f(x_0, y_0))$, because the individual partial derivatives $\frac{\partial f}{\partial x}(x_0, y_0)$ and $\frac{\partial f}{\partial y}(x_0, y_0)$ are special cases of $(D_{\widehat{\mathbf{u}}}f)(x_0, y_0)$ for $b = 0$ and $a = 0$, respectively. The geometric interpretation of the directional derivative is depicted in Figure 6.21.

6.4.2 Multivariable Chain Rules

Recall that if $f = f(x)$ and $x = x(t)$, then the single-variable chain rule says that

$$\frac{d}{dt}\big[f(x(t))\big] = \frac{df}{dx}(x(t)) \cdot \frac{dx}{dt}(t) = \left(\frac{df}{dx}(x(t))\right)\dot{x}(t),$$

where the $\dot{}$ signifies the derivative with respect to t. In a straightforward way, this generalizes to the multivariable chain rule

$$\frac{d}{dt}\Big[\, f\big(x(t), y(t)\big)\,\Big] = \frac{\partial f}{\partial x}\big(x(t), y(t)\big) \cdot \frac{dx}{dt}(t) + \frac{\partial f}{\partial y}\big(x(t), y(t)\big) \cdot \frac{dy}{dt}(t),$$

that is,

$$\frac{d}{dt}\big[f\big(x(t), y(t)\big)\big] = \left(\frac{\partial f}{\partial y}\big(x(t), y(t)\big)\right)\dot{x}(t) + \left(\frac{\partial f}{\partial y}\big(x(t), y(t)\big)\right)\dot{y}(t)\,. \tag{6.35}$$

This follows from the calculation

$$\begin{aligned}
&\frac{d}{dt}\big[f\big(x(t), y(t)\big)\big] = \\
&\quad = \lim_{\triangle t \to 0} \frac{f\big(x(t+\triangle t), y(t+\triangle t)\big) - f\big(x(t), y(t+\triangle t)\big) + f\big(x(t), y(t+\triangle t)\big) - f\big(x(t), y(t)\big)}{\triangle t} \\
&\quad = \lim_{\triangle t \to 0} \left(\frac{f\big(x(t)+\triangle x, y(t)+\triangle y\big) - f\big(x(t), y(t)+\triangle y\big)}{\triangle t} + \frac{f\big(x(t), y(t)+\triangle y\big) - f\big(x(t), y(t)\big)}{\triangle t}\right),
\end{aligned}$$

where $\triangle x \triangleq x(t+\triangle t) - x(t)$, $\triangle y \triangleq y(t+\triangle t) - y(t)$. So,

$$\frac{d}{dt}\left[f\big(x(t),y(t)\big)\right] = \lim_{\triangle t\to 0}\frac{f\big(x(t)+\triangle x, y(t)+\triangle y\big) - f\big(x(t),y(t)+\triangle y\big)}{\triangle x}\cdot\frac{\triangle x}{\triangle t}$$
$$+\lim_{\triangle t\to 0}\frac{f\big(x(t),y(t)+\triangle y\big) - f\big(x(t),y(t)\big)}{\triangle y}\cdot\frac{\triangle y}{\triangle t}.$$

Implicitly we're assuming that $x(t), y(t)$ are differentiable, so as $\triangle t \to 0$, so do $\triangle x \to 0$ and $\triangle y \to 0$. So,

$$\frac{d}{dt}\left[f\big(x(t),y(t)\big)\right] = \lim_{\substack{\triangle x\to 0\\ \triangle y\to 0}}\frac{f\big(x(t)+\triangle x, y(t)+\triangle y\big) - f\big(x(t),y(t)+\triangle y\big)}{\triangle x}\cdot\lim_{\triangle t\to 0}\frac{\triangle x}{\triangle t}$$
$$+\lim_{\triangle y\to 0}\frac{f\big(x(t),y(t)+\triangle y\big) - f\big(x(t),y(t)\big)}{\triangle y}\cdot\lim_{\triangle t\to 0}\frac{\triangle y}{\triangle t}$$
$$= \left(\frac{\partial f}{\partial y}\big(x(t),y(t)\big)\right)\dot{x}(t) + \left(\frac{\partial f}{\partial y}\big(x(t),y(t)\big)\right)\dot{y}(t),$$

that is, (6.35) is true.

The aforementioned calculations are similar to those we did to explain (3.16) in Section 3.2, which is actually a special case of (6.35).

The multivariable chain rule (6.35) can be rewritten as

$$\frac{d}{dt}\left[f\big(x(t),y(t)\big)\right] = \nabla f\big(x(t),y(t)\big)\bullet\dot{\mathbf{r}}(t), \tag{6.36}$$

where $\mathbf{r}(t) = x(t)\hat{\boldsymbol{\imath}} + y(t)\hat{\boldsymbol{\jmath}}$.

Suppose, instead, that x, y are functions not of a single independent variable but instead two independent variables, u, v, that is,

$$x = x(u,v), \quad y = y(u,v).$$

We have multivariable chain rules

$$\frac{\partial}{\partial u}\left[f\big(x(u,v),y(u,v)\big)\right] = \nabla f\big(x(u,v),y(u,v)\big)\bullet\frac{\partial\mathbf{r}}{\partial u} = \frac{\partial f}{\partial x}\cdot\frac{\partial x}{\partial u} + \frac{\partial f}{\partial y}\cdot\frac{\partial y}{\partial u} \tag{6.37}$$

and

$$\frac{\partial}{\partial v}\left[\big(x(u,v),y(u,v)\big)\right] = \nabla f\big(x(u,v),y(u,v)\big)\bullet\frac{\partial\mathbf{r}}{\partial v} = \frac{\partial f}{\partial x}\cdot\frac{\partial x}{\partial v} + \frac{\partial f}{\partial y}\cdot\frac{\partial y}{\partial v}. \tag{6.38}$$

Example 6.22

Suppose $f = f(x, y)$ and x, y are expressed in polar coordinates. Find

$$\frac{\partial f}{\partial r}, \frac{\partial f}{\partial \theta}, \text{ and } \frac{\partial^2 f}{\partial \theta^2}.$$

Method: With r, θ playing the roles of u, v, we get

$$\begin{aligned}\frac{\partial f}{\partial r} &= \frac{\partial}{\partial r}\left[f(x,y)\right] = \frac{\partial}{\partial r}\left[f(r\cos\theta, r\sin\theta)\right] \\ &= \frac{\partial f}{\partial x}\cdot\frac{\partial x}{\partial r} + \frac{\partial f}{\partial y}\cdot\frac{\partial y}{\partial r} = \frac{\partial f}{\partial x}\cdot\frac{\partial}{\partial r}[r\cos\theta] + \frac{\partial f}{\partial y}\cdot\frac{\partial}{\partial r}[r\sin\theta] \\ &= \cos\theta\,\frac{\partial f}{\partial x} + \sin\theta\,\frac{\partial f}{\partial y}.\end{aligned}$$

Similarly,

$$\frac{\partial f}{\partial \theta} = \frac{\partial}{\partial \theta}\Big[f(r\cos\theta, r\sin\theta)\Big] = \frac{\partial f}{\partial x}\cdot\frac{\partial}{\partial \theta}[r\cos\theta] + \frac{\partial f}{\partial y}\cdot\frac{\partial}{\partial \theta}[r\sin\theta];$$

hence,

$$\frac{\partial f}{\partial \theta} = -r\sin\theta\,\frac{\partial f}{\partial x} + r\cos\theta\,\frac{\partial f}{\partial y}. \tag{6.39}$$

Finally,

$$\frac{\partial^2 f}{\partial \theta^2} = \frac{\partial}{\partial \theta}\left[\frac{\partial f}{\partial \theta}(r\cos\theta, r\sin\theta)\right] = \frac{\partial}{\partial \theta}\left[-r\sin\theta\,\frac{\partial f}{\partial x} + r\cos\theta\,\frac{\partial f}{\partial y}\right].$$

Use the same product rule as for functions of a single variable, along with the multivariable chain rule (6.37), to get

$$\begin{aligned}\frac{\partial^2 f}{\partial \theta^2} &= -r\cos\theta\,\frac{\partial f}{\partial x} - r\sin\theta\cdot\left(\frac{\partial^2 f}{\partial x^2}\cdot\frac{\partial}{\partial \theta}[r\cos\theta] + \frac{\partial^2 f}{\partial y\partial x}\cdot\frac{\partial}{\partial \theta}[r\sin\theta]\right) \\ &\quad - r\sin\theta\,\frac{\partial f}{\partial y} + r\cos\theta\cdot\left(\frac{\partial^2 f}{\partial x\partial y}\cdot\frac{\partial}{\partial \theta}[r\cos\theta] + \frac{\partial^2 f}{\partial^2 y}\cdot\frac{\partial}{\partial \theta}[r\sin\theta]\right) \\ &= -r\cos\theta\,\frac{\partial f}{\partial x} + r^2\sin^2\theta\,\frac{\partial^2 f}{\partial x^2} - r^2\sin\theta\,\cos\theta\,\frac{\partial^2 f}{\partial y\partial x} \\ &\quad - r\sin\theta\,\frac{\partial f}{\partial y} - r^2\sin\theta\,\cos\theta\,\frac{\partial^2 f}{\partial x\partial y} + r^2\cos^2\theta\,\frac{\partial^2 f}{\partial^2 y}.\end{aligned}$$

If Clairaut's theorem, that is, Theorem 6.2, applies, then this simplifies to

$$\frac{\partial^2 f}{\partial \theta^2} = -r\cos\theta\,\frac{\partial f}{\partial x} - r\sin\theta\,\frac{\partial f}{\partial y} + r^2\sin^2\theta\,\frac{\partial^2 f}{\partial x^2} - 2r^2\sin\theta\,\cos\theta\,\frac{\partial^2 f}{\partial y\partial x} + r^2\cos^2\theta\,\frac{\partial^2 f}{\partial^2 y}. \bigcirc$$

6.4.3 Gradient Vector in $\mathbb{R}^3$

If $F = F(x, y, z)$ then the **gradient vector** is defined by

$$\textbf{\textit{grad}}\, F = \nabla F \triangleq \frac{\partial F}{\partial x}\,\hat{\boldsymbol{\imath}} + \frac{\partial F}{\partial y}\,\hat{\boldsymbol{\jmath}} + \frac{\partial F}{\partial z}\,\hat{\boldsymbol{k}}.$$

Linear approximation is given by

$$\begin{aligned} &F(x_0 + \triangle x, y_0 + \triangle y, z_0 + \triangle z) \\ &\quad \approx F(x_0, y_0, z_0) + \triangle x \frac{\partial F}{\partial x}(x_0, y_0, z_0) + \triangle y \frac{\partial F}{\partial y}(x_0, y_0, z_0) + \triangle z \frac{\partial F}{\partial z}(x_0, y_0, z_0), \end{aligned}$$

that is,

$$F(x_0 + \triangle x, y_0 + \triangle y, z_0 + \triangle z) \approx F(x_0, y_0, z_0) + L(\triangle x, \triangle y, \triangle z),$$

where

$$L(\triangle x, \triangle y, \triangle z) = \nabla F(x_0, y_0, z_0) \bullet (\triangle x\,\hat{\boldsymbol{\imath}} + \triangle y\,\hat{\boldsymbol{\jmath}} + \triangle z\,\hat{\boldsymbol{k}}).$$

Similarly, the directional derivative is defined by

$$(D_{\widehat{\mathbf{u}}} F)(x_0, y_0, z_0) \triangleq \nabla F(x_0, y_0, z_0) \bullet \widehat{\mathbf{u}},$$

for any unit vector $\widehat{\mathbf{u}}$ in $\mathbb{R}^3$, and the multivariable chain rules are given by

$$\frac{dF}{dt} = \frac{\partial F}{\partial x} \cdot \frac{dx}{dt} + \frac{\partial F}{\partial y} \cdot \frac{dy}{dt} + \frac{\partial F}{\partial z} \cdot \frac{dz}{dt} = \nabla F|_{at\ \mathbf{r}(t)} \bullet \frac{d\mathbf{r}}{dt} \tag{6.40}$$

and, for example,

$$\frac{\partial F}{\partial u} = \frac{\partial F}{\partial x} \cdot \frac{\partial x}{\partial u} + \frac{\partial F}{\partial y} \cdot \frac{\partial y}{\partial u} + \frac{\partial F}{\partial z} \cdot \frac{\partial z}{\partial u} = \nabla F|_{at\ \mathbf{r}(u,v)} \bullet \frac{\partial \mathbf{r}}{\partial u}\,. \tag{6.41}$$

If x, y, z are functions of three variables u, v, w, then, for example,

$$\frac{\partial F}{\partial w} = \frac{\partial F}{\partial x} \cdot \frac{\partial x}{\partial w} + \frac{\partial F}{\partial y} \cdot \frac{\partial y}{\partial w} + \frac{\partial F}{\partial z} \cdot \frac{\partial z}{\partial w} = \nabla F|_{at\ \mathbf{r}(u,v,w)} \bullet \frac{\partial \mathbf{r}}{\partial w}, \tag{6.42}$$

where $\mathbf{r} = x\,\hat{\boldsymbol{\imath}} + y\,\hat{\boldsymbol{\jmath}} + z\,\hat{\boldsymbol{k}}$.

6.4.4 Scalar Potential Functions

Definition 6.4

Given a vector field $\mathbf{F}(x,y,z) = F_x(x,y,z)\hat{\boldsymbol{\imath}} + F_y(x,y,z)\hat{\boldsymbol{\jmath}} + F_z(x,y,z)\hat{\mathbf{k}}$, it can be of physical importance to find a scalar **potential function** $f = f(x,y,z)$ such that

$$\mathbf{F} = \nabla f,$$

if it is possible to do so. If there is such a function f, we say that $\mathbf{F}$ is **exact**.

If there is a potential function, then

$$F_x\,\hat{\boldsymbol{\imath}} + F_y\,\hat{\boldsymbol{\jmath}} + F_z\,\hat{\mathbf{k}} = \nabla f = \frac{\partial f}{\partial x}\,\hat{\boldsymbol{\imath}} + \frac{\partial f}{\partial y}\,\hat{\boldsymbol{\jmath}} + \frac{\partial f}{\partial z}\,\hat{\mathbf{k}},$$

so

$$F_x = \frac{\partial f}{\partial x}, \quad F_y = \frac{\partial f}{\partial y}, \quad F_z = \frac{\partial f}{\partial z}.$$

If, in addition, all of the first partial derivatives $F_x(x,y,z)$, $F_x(x,y,z)$, and $F_x(x,y,z)$ are defined and continuous, then Clairaut's theorem (Theorem 6.2) implies that

$$\frac{\partial F_y}{\partial x} \triangleq \frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x} \triangleq \frac{\partial F_x}{\partial y}.$$

This and similar calculations give the **exactness criterion** for a vector field: If $\mathbf{F}$ is exact then

$$\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y} = 0, \quad \frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z} = 0, \quad \text{and} \quad \frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x} = 0. \tag{6.43}$$

If $\mathbf{F}$ is a planar vector field with no dependence on z and, also, $F_z \equiv 0$, then the exactness criterion reduces to

$$\frac{\partial F_y}{\partial x}(x,y) - \frac{\partial F_x}{\partial y}(x,y) = 0. \tag{6.44}$$

If we call $M = M(x,y) = F_x(x,y)$ and $N = N(x,y) = F_y(x,y)$, we see that the exactness criterion (6.44) is the same as the exactness criterion (3.17) in Section 3.2 for a first-order ODE! In Section 3.2, we saw how to use anti-partial differentiation to *calculate* a potential function $\phi(x,y)$ used to solve an exact first-order ODE. The same method works in this section to find a scalar potential function in $\mathbb{R}^3$.

Example 6.23

If possible, find a potential function for the vector field

$$\mathbf{F} = \mathbf{F}(\mathbf{r}) = (z + y\cos(xy))\hat{\boldsymbol{\imath}} + (2 - y + x\cos(xy))\hat{\boldsymbol{\jmath}} + (x + z)\hat{\boldsymbol{k}}.$$

Method: First, for safety, let's check the exactness criterion—if it fails, we shouldn't try to produce a nonexistent potential function. Using the single-variable product and chain rules, we calculate

$$\begin{aligned}\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y} &= \frac{\partial}{\partial x}\left[\,2 - y + x\cos(xy)\,\right] - \frac{\partial}{\partial y}\left[\,z + y\cos(xy)\,\right] \\ &= (0 - 0 + 1\cdot\cos(xy) + x\cdot(-\sin(xy)\cdot y)) \\ &\quad - (0 + 1\cdot\cos(xy) + y\cdot(-\sin(xy)\cdot x)) = 0;\end{aligned}$$

$$\frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z} = \frac{\partial}{\partial y}\left[\,x + z\,\right] - \frac{\partial}{\partial z}\left[\,2 - y + x\cos(xy)\,\right] = (0 - 0 + 0) - (0 - 0 + 0) = 0;$$

and

$$\frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x} = \frac{\partial}{\partial z}\left[\,z + y\cos(xy)\right] - \frac{\partial}{\partial x}\left[\,x + z\,\right] = (1 + 0) - (1 + 0) = 0.$$

So, there should exist a scalar potential function.

Because we need to have $\frac{\partial f}{\partial x} = F_x$,

$$f = \int \frac{\partial f}{\partial x}\,\partial x = \int F_x\,\partial x = \int (z + y\cos(xy))\partial x = xz + \int y\cos(w)\partial x,$$

where we substitute $w = xy$. Because $\frac{\partial w}{\partial x} = y$, we get

$$f = xz + \int \cos(w)\partial w = xz + \sin(w) + g(y,z) = xz + \sin(xy) + g(y,z); \tag{6.45}$$

$g(y,z)$ can be an arbitrary function of *only* y, z, because $\frac{\partial}{\partial x}\left[\,g(y,z)\,\right] \equiv 0$.

Substitute (6.45) into $F_y = \frac{\partial f}{\partial y}$ to get

$$2 - y + \cancel{x\cos(xy)} = F_y = \frac{\partial f}{\partial y} = \frac{\partial}{\partial y}\left[xz + \sin(xy) + g(y,z)\right] = 0 + \cancel{x\cos(xy)} + \frac{\partial g}{\partial y};$$

hence, we need

$$2 - y = \frac{\partial g}{\partial y}.$$

It follows that

$$g(y,z) = \int \frac{\partial g}{\partial y}\,\partial y = \int (2 - y)\,\partial y = 2y - \frac{y^2}{2} + h(z),$$

where $h(z)$ can be an arbitrary function of *only* z. Substitute this into (6.45) to get

$$f = xz + \sin(xy) + 2y - \frac{y^2}{2} + h(z). \tag{6.46}$$

Substitute (6.46) into $F_z = \frac{\partial f}{\partial z}$ to get

$$\not{x} + z = F_z = \frac{\partial f}{\partial z} = \frac{\partial}{\partial z}\left[xz + \sin(xy) + 2y - \frac{y^2}{2} + h(z) \right] = \not{x} + 0 + 0 - 0 + \frac{dh}{dz};$$

hence $z = h'(z)$. Because $h = h(z)$ is a function of z alone, its partial and ordinary derivatives with respect to z are equal.

So, $h(z) = \frac{z^2}{2} + c$, where c is an arbitrary constant. To summarize,

$$f = xz + \sin(xy) + 2y - \frac{y^2}{2} + \frac{z^2}{2} + c \tag{6.47}$$

is a scalar potential function for the given vector field, for any constant c. ◯

Theorem 6.3

(Existence of a potential function) Suppose a vector field $\mathbf{F} = F_x\,\hat{\boldsymbol{\imath}} + F_y\,\hat{\boldsymbol{\jmath}} + F_z\,\hat{\boldsymbol{k}}$ satisfies the exactness criterion (6.43) for all (x, y, z) in some open region $\mathcal{V}$ in $\mathbb{R}^3$. If, in addition, F_x, F_y, and F_z are all continuous in $\mathcal{V}$, then $\mathbf{F}$ is exact in $\mathcal{V}$, that is, there is a function f such that $\mathbf{F}(x, y, z) \equiv \boldsymbol{\nabla} f(x, y, z)$ for all (x, y, z) in $\mathcal{V}$.

6.4.5 Problems

1. If $f(x, y) = \sin(xy^2 + 3)$, find $\frac{\partial^2 f}{\partial x \partial y}$.
2. If $f(x, y) = xe^{y^2} + \cos(2x - 3y + \pi)$, find $\frac{\partial^2 f}{\partial x \partial y}$.
3. The elliptic cone $z^2 = 4x^2 + y^2$ intersects the plane $y = -2$ in a hyperbola. Find parametric equations of the tangent line to this hyperbola at the point $(-1, -2, 2\sqrt{2})$.
4. Table 6.4 gives the rated horsepower capacity, H, of a roller chain as a function of the sprocket speed, ω, in revolutions per minute, and the ANSI chain number N. Use the table to estimate the rated horsepower capacity for a sprocket speed of 175 revolutions per minute and ANSI chain number 100, *if* such a chain existed. Be clear about which values from the table you used, how you found approximate partial derivative values, and how you used them.
5. Find the directional derivative of $f(x, y) = e^{-x^2+y}$ at the point $(1, 2)$ in the direction of the vector $-\hat{\boldsymbol{\imath}} + 2\,\hat{\boldsymbol{\jmath}}$.
6. The temperature at a point (x, y, z) is given by $T(x, y, z) = 5 - \frac{6}{\sqrt{x^2+4y^2+9z^2}}$, in °C.
   (a) In what unit vector direction is the temperature increasing the most at the point

TABLE 6.4

Horsepower Data

$\omega \backslash N$	80	120	156
150	7.75	25.1	56.3
200	10.0	32.5	72.9
300	14.5	46.8	105

$P = (-1, -2, 2)$? (b) What is the maximum value of the directional derivative there, assuming x, y, z are measured in meters?

7. Define $f(x, y, z) = \frac{1}{\sqrt{x^2+y^2+z^2}}$. Find (a) $\nabla f(x, y, z)$ and (b) the directional derivative of f in the radial direction.
8. Consider the surface $z = 3 - x^2 - y^2$. (a) Find an equation, or the parametric equations, for the curve obtained by fixing $y = 0.8$ and letting x vary. (b) Use your work for part (a) to find the parametric equations of the tangent line to that curve at the point $(x, y, z) = (1, 0.8, 1.36)$. (c) Use the pictures associated with parts (a) and (b) to give the pictorial interpretation of the partial derivative with respect to x of the function $f(x, y) = 3 - x^2 - y^2$ at the point $(x, y) = (1, 0.8)$.
9. Consider the surface $z = 3 - x^2 - y^2$. (a) Find equations for, or the parametric equations for, the curve obtained by fixing $x = 1$ and letting y vary. (b) Use your work for part (a) to find the parametric equations of the tangent line to that curve at the point $(x, y, z) = (1, 0.8, 1.36)$. (c) Use the pictures associated with parts (a) and (b) to give the pictorial interpretation of the partial derivative with respect to y of the function $f(x, y) = 3 - x^2 - y^2$ at the point $(x, y) = (1, 0.8)$.
10. Consider the surface $z = 3 - x^2 - y^2$. (a) Find an equation, or the parametric equations, for the curve obtained by letting $x = 1 + at$ and $y = 0.8 + bt$ as t varies. (b) Use your work for part (a) to find the parametric equations of the tangent line to that curve at the point $(x, y, z) = (1, 0.8, 1.36)$. (c) Use the pictures associated with parts (a) and (b) to give the pictorial interpretation of the directional derivative of the function $f(x, y) = 3 - x^2 - y^2$ at the point $(x, y) = (1, 0.8)$ in the direction of the vector $\mathbf{u} = 0.6\,\hat{\boldsymbol{\imath}} + 0.8\,\hat{\boldsymbol{\jmath}}$.
11. If $x = -3 + 3t$, $y = 4 - t^2$, and $z = f(x, y) = x^{2y}$, evaluate $\frac{dz}{dt}(2)$.
12. If $x = x(t)$, $y = y(t)$, and $f = f(x, y)$, find $\frac{d}{dt}\left[(f(x(t), y(t))^2\right]$ in terms of f, $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, $\frac{dx}{dt}$, $\frac{dy}{dt}$ and t.
13. If $x = ue^{-v}$, $y = ve^{u}$, $f(x, y) = \frac{x}{y} - x^2 y$ and $g(u, v) = f(x(u, v), y(u, v))$, evaluate $\frac{\partial g}{\partial u}(-2, 1)$ and $\frac{\partial g}{\partial v}(-2, 1)$.
14. Suppose $\mathbf{r}(t)$ is of constant magnitude. [Physically, this corresponds to motion on a sphere whose center is at the origin.] Why is $\mathbf{r}(t) \bullet \dot{\mathbf{r}}(t) \equiv 0$? What does this say about the tangent vector to a point on a circle?
15. If $z = f(x, y)$, $x = r\cos\theta$, and $y = r\sin\theta$, explain why

$$\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2 = \left(\frac{\partial z}{\partial r}\right)^2 + \left(\frac{1}{r} \cdot \frac{\partial z}{\partial \theta}\right)^2$$

is an identity, at least where everything exists.

16. If $z = f(x, y)$, $x = u^2 + v^2$, and $y = u^2 - v^2$, explain why

$$\left(\frac{\partial z}{\partial x}\right)^2 - \left(\frac{\partial z}{\partial y}\right)^2 = \frac{1}{4uv}\left(\frac{\partial z}{\partial u}\right)\left(\frac{\partial z}{\partial v}\right)$$

is an identity, at least where everything exists.

17. Use linear approximation to estimate $\ln\left((1.02)^2 + (2.99)^3\right)$.
18. Determine a scalar potential function for the vector field

$$\mathbf{F} \triangleq (3x^2y + y^2z)\hat{\boldsymbol{\imath}} + (x^3 + 2xyz + z)\,\hat{\boldsymbol{\jmath}} + (xy^2 + y + z)\hat{\boldsymbol{k}}.$$

    You may assume that the given vector field is exact.
19. Determine a scalar potential function for the vector field

$$\mathbf{F} \triangleq (yz - 2x)\hat{\boldsymbol{\imath}} + (xz + \cos z)\,\hat{\boldsymbol{\jmath}} + (xy - y\sin z)\hat{\boldsymbol{k}}.$$

    You may assume that the given vector field is exact.
20. Let $\mathbf{F} \triangleq (1 - y - z)\,\hat{\boldsymbol{\imath}} - x\,\hat{\boldsymbol{\jmath}} - (x + e^{-2z})\hat{\boldsymbol{k}}$. Explain why $\mathbf{F}$ is exact and find a potential function for it.
21. Determine a scalar potential function for the vector field

$$\mathbf{F} \triangleq \cos y\,\hat{\boldsymbol{\imath}} + (-x\sin y + z)\,\hat{\boldsymbol{\jmath}} + (y - 1)\hat{\boldsymbol{k}}.$$

22. Explain why (6.37) is true, using a derivation similar to the derivation of (6.35).
23. Let $\rho = ||\mathbf{r}||$ and $\mathbf{v} = \dot{\mathbf{r}}$. Fill in the blanks in the following with integers to make the equation an identity. Explain how you arrived at your conclusions and write the identity.

$$\frac{d}{dt}\left[\rho^{-3}\mathbf{r}\right] = \rho^{-5}\left(\_\_\,\mathbf{r} \times (\mathbf{v} \times \mathbf{r}) + \_\_\,\rho^2\mathbf{v}\right).$$

    [Hint: You may use the result about vector triple products that will be mentioned in Problem 6.8.4.18.]
24. In this problem, you will establish **Leibniz's rule**:

$$\frac{d}{dt}\left[\int_a^t f(t,s)\,ds\right] = f(t,t) + \int_a^t \frac{\partial f}{\partial t}(t,s)\,ds,$$

    assuming a is a constant. How? Begin by using the definition of ordinary differentiation,

$$\frac{d}{dt}\left[\int_a^t f(t,s)\,ds\right] \triangleq \lim_{\triangle t \to 0} \frac{1}{\triangle t}\left(\int_a^{t+\triangle t} f(t+\triangle t, s)\,ds - \int_a^t f(t,s)\,ds\right).$$

    After that, inside the big parentheses, subtract and add $\int_a^{t+\triangle t} f(t,s)\,ds$.

6.5 Tangent Plane and Normal Vector

Suppose a point $(x_0, y_0, f(x_0, y_0))$ is on the surface $z = f(x, y)$. At that point, the tangent plane to the surface consists of all of the points (x, y, z) for which the change

$$\triangle z \triangleq z - f(x_0, y_0)$$

agrees with the linear approximation of change in f corresponding to simultaneous changes $\triangle x = x - x_0$ and $\triangle y = y - y_0$. Denote $z_0 = f(x_0, y_0)$. Rewriting (6.33) in Section 6.4 gives

$$z - z_0 = \triangle z = L(\triangle x, \triangle y) = \left(\frac{\partial f}{\partial x}(x_0, y_0)\right)\triangle x + \left(\frac{\partial f}{\partial y}(x_0, y_0)\right)\triangle y$$

$$= \left(\frac{\partial f}{\partial x}(x_0, y_0)\right)(x - x_0) + \left(\frac{\partial f}{\partial y}(x_0, y_0)\right)(y - y_0).$$

So, the plane is described by the equation:

$$-\left(\frac{\partial f}{\partial x}(x_0, y_0)\right)(x - x_0) - \left(\frac{\partial f}{\partial y}(x_0, y_0)\right)(y - y_0) + (z - z_0) = 0. \tag{6.48}$$

Example 6.24

Find an equation of the tangent plane to the surface $z = x^2 + 3y^2$ at the point $(2, -1, 7)$, and find a normal vector for that plane.

Method: $z = f(x, y) \triangleq x^2 + 3y^2$, so

$$\frac{\partial f}{\partial x}(x_0, y_0) = 2x_0 = 2 \cdot 2 = 4, \quad \frac{\partial f}{\partial y}(x_0, y_0) = 6y_0 = 6 \cdot (-1) = -6.$$

An equation of the tangent plane is given by $-4(x - 2) - (-6)(y - (-1)) + z - 7 = 0$, that is,

$$-4(x - 2) + 6(y + 1) + z - 7 = 0.$$

Because a plane $A(x - x_0) + B(y - y_0) + C(z - z_0) = 0$ has a normal vector $\mathbf{n} = A\,\hat{\boldsymbol{\imath}} + B\,\hat{\boldsymbol{\jmath}} + C\,\hat{\mathbf{k}}$, in this problem $\mathbf{n} = -4\,\hat{\boldsymbol{\imath}} + 6\,\hat{\boldsymbol{\jmath}} + \hat{\mathbf{k}}$ is a normal vector to the plane. ○

Suppose a point $P_0 = (x_0, y_0, z_0)$ has position vector $\mathbf{r}_0$ on a parametrized surface:

$$\mathcal{S}: \ \mathbf{r} = \mathbf{r}(u, v) = x(u, v)\,\hat{\boldsymbol{\imath}} + y(u, v)\,\hat{\boldsymbol{\jmath}} + z(u, v)\,\hat{\mathbf{k}}, \ (u, v) \text{ in } \mathcal{D}.$$

Assuming the surface is simple, as we defined at the end of Section 6.3, there is a single choice of (u_0, v_0) in $\mathcal{D}$ for which $\mathbf{r}_0 = \mathbf{r}(u_0, v_0)$. We can define two parametrized curves by

$$\mathcal{C}: \ \mathbf{r} = \mathbf{r}(u, v_0), \ u \text{ in interval } I_u,$$

where I_u is chosen small enough that (u, v_0) is in $\mathcal{D}$ for all u in I_u, and

$$\mathcal{C}: \ \mathbf{r} = \mathbf{r}(u_0, v), \ v \text{ in interval } I_v,$$

where I_v is chosen small enough that (u_0, v) is in $\mathcal{D}$ for all v in I_v.

At $\mathbf{r}_0$ there are correspondingly two tangent vectors $\mathbf{T}_u, \mathbf{T}_v$ defined by

$$\mathbf{T}_u \triangleq \frac{\partial \mathbf{r}}{\partial u}(u_0, v_0), \quad \mathbf{T}_v \triangleq \frac{\partial \mathbf{r}}{\partial v}(u_0, v_0).$$

If $\mathbf{T}_u \neq \mathbf{0}$, $\mathbf{T}_v \neq \mathbf{0}$, and $\{\mathbf{T}_u, \mathbf{T}_v\}$ is linearly independent, then the vector

$$\mathbf{n} = \mathbf{T}_u \times \mathbf{T}_v = \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v},$$

evaluated at $\mathbf{r}_0$, is a nonzero vector normal to the plane:

$$\mathbf{r}_0 + \text{span}\{\mathbf{T}_u, \mathbf{T}_v\}. \tag{6.49}$$

If either $\mathbf{T}_u = \mathbf{0}$ or $\mathbf{T}_v = \mathbf{0}$, or more generally $\{\mathbf{T}_u, \mathbf{T}_v\}$ is linearly dependent, then the set given by (6.49) is a point or a line, not a plane.

In the special case that $\mathcal{S}$ is a surface $z = f(x, y)$, that is,

$$\mathcal{S}: \ \mathbf{r} = \mathbf{r}(x, y) = x\,\hat{\boldsymbol{\imath}} + y\,\hat{\boldsymbol{\jmath}} + f(x, y)\,\hat{\mathbf{k}}, \ (x, y) \text{ in } \mathcal{D},$$

then at a point $(x_0, y_0, f(x_0, y_0))$, we get

$$\mathbf{T}_x \triangleq \frac{\partial \mathbf{r}}{\partial x}(x_0, y_0) = \hat{\boldsymbol{\imath}} + \frac{\partial f}{\partial x}(x_0, y_0)\,\hat{\mathbf{k}},$$

$$\mathbf{T}_y \triangleq \frac{\partial \mathbf{r}}{\partial x}(x_0, y_0) = \hat{\boldsymbol{\jmath}} + \frac{\partial f}{\partial y}(x_0, y_0)\,\hat{\mathbf{k}};$$

hence

$$\mathbf{n} = \mathbf{T}_x \times \mathbf{T}_y = \left(\hat{\boldsymbol{\imath}} + \frac{\partial f}{\partial x}(x_0, y_0)\,\hat{\mathbf{k}}\right) \times \left(\hat{\boldsymbol{\jmath}} + \frac{\partial f}{\partial y}(x_0, y_0)\,\hat{\mathbf{k}}\right) = -\frac{\partial f}{\partial x}(x_0, y_0)\,\hat{\boldsymbol{\imath}} - \frac{\partial f}{\partial y}(x_0, y_0)\,\hat{\boldsymbol{\jmath}} + \hat{\mathbf{k}},$$

which agrees with (6.48). This is very good! One way to understand a general result is by seeing what it means in a special case and, of course, seeing that in the special case the general result agrees with a previously known result.

Besides the special case of a surface $z = f(x, y)$ and the general case of a parametrized surface $\mathcal{S}: \ \mathbf{r} = \mathbf{r}(u, v)$, there is the case of a level set, that is, $\mathcal{S} = \{(x, y, z) : F(x, y, z) = k\}$, where k is a constant. On a level set $\mathcal{S}$, at a point whose position vector is $\mathbf{r}_0$,

$$\mathbf{n} = \nabla F|_{\mathbf{r}_0}$$

is a normal vector, as you will establish in Problem 6.5.1.10.

Example 6.25

Find an equation of the tangent plane to the sphere $\rho = a$ at the point P_0 whose position vector is $\mathbf{r}_0$ and find a unit vector normal to that plane.

Method: On the sphere,

$$\mathbf{r} = a\,\sin\phi\,\cos\theta\,\hat{\boldsymbol{\imath}} + a\,\sin\phi\,\sin\theta\,\hat{\boldsymbol{\jmath}} + a\,\cos\phi\,\hat{\boldsymbol{k}} = a\hat{\mathbf{e}}_\rho,$$

so

$$\frac{\partial \mathbf{r}}{\partial \phi} = a\,\cos\phi\,\cos\theta\,\hat{\boldsymbol{\imath}} + a\,\cos\phi\,\sin\theta\,\hat{\boldsymbol{\jmath}} - a\,\sin\phi\,\hat{\boldsymbol{k}} = a\,\hat{\mathbf{e}}_\phi,$$

as in (6.17) in Section 6.2, and

$$\frac{\partial \mathbf{r}}{\partial \theta} = -a\,\sin\phi\,\sin\theta\,\hat{\boldsymbol{\imath}} + a\,\sin\phi\,\cos\theta\,\hat{\boldsymbol{\jmath}} = a\,\sin\phi\,\hat{\mathbf{e}}_\theta,$$

as in (6.16) in Section 6.2.

It follows that a normal vector to the tangent plane is given by

$$\mathbf{n} = \frac{\partial \mathbf{r}}{\partial \phi} \times \frac{\partial \mathbf{r}}{\partial \theta} = (a\hat{\mathbf{e}}_\phi) \times (a\,\sin\phi\,\hat{\mathbf{e}}_\theta) = a^2 \sin\phi\,\hat{\mathbf{e}}_\phi \times \hat{\mathbf{e}}_\theta = a^2 \sin\phi\,\hat{\mathbf{e}}_\rho,$$

unless $\phi_0 = 0$ or $\phi_0 = \pi$, that is, unless the point is on the sphere's North Pole or South Pole.

By the way, Problem 6.4.5.14 gives an elegant way to see why the radial vector is a normal vector at a point on a sphere.

It follows that for $0 \neq \phi_0 \neq \pi$, the unit normal vector is

$$\widehat{\mathbf{n}} = \frac{1}{||\mathbf{n}||}\,\mathbf{n} = \hat{\mathbf{e}}_\rho(\phi_0, \theta_0). \tag{6.50}$$

This agrees with our intuition that at points on a sphere, a radial vector is normal to the surface. In fact, (6.50) is also correct for $\phi_0 = 0$ and $\phi_0 = \pi$, that is, if the point is on the North or South Pole.

In (6.50) we saw that the radial vector depends upon the position.

An equation of the tangent plane at a point on a sphere is

$$0 = \widehat{\mathbf{n}} \bullet (\mathbf{r} - \mathbf{r}_0) = \hat{\mathbf{e}}_\rho(\phi_0, \theta_0) \bullet (\mathbf{r} - a\,\hat{\mathbf{e}}_\rho) = \hat{\mathbf{e}}_\rho(\phi_0, \theta_0) \bullet \mathbf{r} - a\hat{\mathbf{e}}_\rho(\phi_0, \theta_0) \bullet \hat{\mathbf{e}}_\rho(\phi_0, \theta_0).$$

But,

$$\hat{\mathbf{e}}_\rho(\phi_0, \theta_0) = \frac{1}{a} \cdot a\,\hat{\mathbf{e}}_\rho(\phi_0, \theta_0) = \frac{1}{a}\,\mathbf{r}_0,$$

so an equation of the tangent plane is

$$0 = (\hat{\mathbf{e}}_\rho(\phi_0, \theta_0) \bullet \mathbf{r}) - a = \left(\frac{1}{a}\,\mathbf{r}_0 \bullet \mathbf{r}\right) - a,$$

that is, $\frac{1}{a}\,\mathbf{r}_0 \bullet \mathbf{r} = a$, that is,

$$\mathbf{r}_0 \bullet \mathbf{r} = x_0 x + y_0 y + z_0 z = a^2. \bigcirc \tag{6.51}$$

As a quick reality check, is the point whose position vector is $\mathbf{r}_0$ on the tangent plane given by (6.51)? It should be! In fact, $\mathbf{r} = \mathbf{r}_0 = x_0\,\hat{\boldsymbol{\imath}} + y_0\,\hat{\boldsymbol{\jmath}} + z_0\,\hat{\boldsymbol{k}}$ does satisfy

$$a^2 = \mathbf{r}_0 \bullet \mathbf{r} = x_0 x + y_0 y + z_0 z = x_0 x_0 + y_0 y_0 + z_0 z_0 = x_0^2 + y_0^2 + z_0^2.$$

Example 6.26

Find an equation of the tangent plane at the point $(-\frac{15}{\sqrt{17}}, \frac{15}{\sqrt{17}}, 2)$ on a power plant's cooling tower that is part of the hyperboloid of one sheet

$$\mathcal{S}: \ \mathbf{r} = \mathbf{r}(t,s) = 5\cosh t\, \cos s\, \hat{\boldsymbol{\imath}} + 3\cosh t\, \sin s\, \hat{\boldsymbol{\jmath}} + 2\sinh t\, \hat{\boldsymbol{k}}, \ -\infty < t < \infty,\ 0 \le s \le 2\pi,$$

also known as

$$\frac{x^2}{5^2} + \frac{y^2}{3^2} - \frac{z^2}{2^2} = 1.$$

Method: We calculate*

$$\begin{aligned}\mathbf{n} = \frac{\partial \mathbf{r}}{\partial t} \times \frac{\partial \mathbf{r}}{\partial s} &= (5\sinh t\, \cos s\, \hat{\boldsymbol{\imath}} + 3\sinh t\, \sin s\, \hat{\boldsymbol{\jmath}} + 2\cosh t\, \hat{\boldsymbol{k}}) \\ &\quad \times (-5\cosh t\, \sin s\, \hat{\boldsymbol{\imath}} + 3\cosh t\, \cos s\, \hat{\boldsymbol{\jmath}}) \\ &= -6\cosh^2 t\, \cos s\, \hat{\boldsymbol{\imath}} - 10\cosh^2 t \sin s\, \hat{\boldsymbol{\jmath}} + 15\, \cosh t\, \sinh t\, \hat{\boldsymbol{k}}.\end{aligned}$$

To simplify the algebra, divide through by $\cosh t$, because it is never zero. That gives a normal vector

$$\mathbf{n}_1 = -6\cosh t\, \cos s\, \hat{\boldsymbol{\imath}} - 10\cosh t \sin s\, \hat{\boldsymbol{\jmath}} + 15\, \sinh t\, \hat{\boldsymbol{k}}. \tag{6.52}$$

In order to proceed, we must find the parameter values (t_0, s_0) for which

$$-\frac{15}{\sqrt{17}}\, \hat{\boldsymbol{\imath}} + \frac{15}{\sqrt{17}}\, \hat{\boldsymbol{\jmath}} + 2\hat{\boldsymbol{k}} = \mathbf{r}(t_0, s_0) = 5\cosh t_0\, \cos s_0\, \hat{\boldsymbol{\imath}} + 3\cosh t_0\, \sin s_0\, \hat{\boldsymbol{\jmath}} + 2\sinh t_0\, \hat{\boldsymbol{k}},$$

that is,

$$-\frac{15}{\sqrt{17}} = 5\, \cosh t_0\, \cos s_0, \qquad \frac{15}{\sqrt{17}} = 3\cosh t_0\, \sin s_0, \quad \text{and} \quad 2 = 2\sinh t_0.$$

The latter implies† $t_0 = \operatorname{arcsinh}(1) = \ln(1 + \sqrt{2})$.

Because of the identity $\cosh^2 t - \sinh^2 t \equiv 1$ and the fact that $\sinh t_0 = 1$,

$$\cosh t_0 = \sqrt{1 + \sinh^2 t_0} = \sqrt{1 + 1^2} = \sqrt{2}.$$

Because $-\frac{15}{\sqrt{17}} = 5\, \cosh t_0\, \cos s_0$, we get $-\frac{15}{\sqrt{17}} = 5\sqrt{2}\cos s_0$; hence,

$$\cos s_0 = -\frac{3}{\sqrt{34}}.$$

Similarly, $\frac{15}{\sqrt{17}} = 3\cosh t_0\, \sin s_0$ implies

$$\sin s_0 = \frac{5}{\sqrt{34}}.$$

---

* Note that $\cosh t \triangleq \frac{1}{2}(e^t + e^{-t})$ and $\sinh t \triangleq \frac{1}{2}(e^t - e^{-t})$ satisfy $\cosh^2 t - \sinh^2 t = 1$, $\frac{d}{dt}[\cosh t] = \sinh t$, and $\frac{d}{dt}[\sinh t] = \cosh t$.

† $1 = \sinh t_0 = \frac{1}{2}(e^{t_0} - e^{-t_0}) \iff 2 = e^{t_0} - e^{-t_0} \iff 2e^{t_0} = (e^{t_0})^2 - 1 \iff 0 = (e^{t_0})^2 - 2e^{t_0} - 1 \iff 0 < e^{t_0} = \frac{2+\sqrt{8}}{2} = 1 + \sqrt{2}$.

Because $(\cos s_0, \sin s_0)$ is in the second quadrant, it follows that $\frac{\pi}{2} < s_0 < \pi$ and thus $s_0 = \cos^{-1}(-\frac{3}{\sqrt{34}})$.

To find a vector $\widehat{\mathbf{n}}$ that is normal to the surface at the point, we actually only need to know that $\cosh t_0 = \sqrt{2}, \sinh t_0 = 1, \cos s_0 = -\frac{3}{\sqrt{34}}, \sin s_0 = \frac{5}{\sqrt{34}}$, so (6.52) gives

$$\mathbf{n}_1 = \cdots = \frac{18}{\sqrt{17}}\,\hat{\boldsymbol{\imath}} - \frac{50}{\sqrt{17}}\,\hat{\boldsymbol{\jmath}} + 15\,\hat{\boldsymbol{k}}. \tag{6.53}$$

At this point, we might as well multiply the unit normal vector by $\sqrt{17}$ to get a simpler-looking normal vector, $\mathbf{n} = 18\,\hat{\boldsymbol{\imath}} - 50\,\hat{\boldsymbol{\jmath}} + 15\sqrt{17}\,\hat{\boldsymbol{k}}$.

An equation of the tangent plane at the point P_0 is

$$18\left(x + \frac{15}{\sqrt{17}}\right) - 50\left(y - \frac{15}{\sqrt{17}}\right) + 15\sqrt{17}\,(z-2) = 0,$$

or

$$18x - 50y + 15\sqrt{17}\,z = 30\sqrt{17}. \bigcirc$$

We saw that the normal vector to the tangent plane of a surface $z = f(x, y)$ points in the direction of greatest increase of the function f. This has a natural interpretation in terms of **level curves** or, more generally, **level sets**,

$$\{(x, y) :\ k = f(x, y)\},$$

or $k = f(x, y)$ for short, where k is a scalar constant. The level sets are also called "contours" in a **contour plot**. At a point (x_0, y_0) on a level set $k = f(x, y)$, the gradient vector

$$\nabla f(x_0, y_0) = \frac{\partial f}{\partial x}(x_0, y_0)\,\hat{\boldsymbol{\imath}} + \frac{\partial f}{\partial y}(x_0, y_0)\,\hat{\boldsymbol{\jmath}}$$

points in the direction of greatest increase of f and hence is *normal to the level set* at the point (x_0, y_0), as shown in Figure 6.22.

A nice example of this is the subject of **isobars**, that is, "lines" of constant atmospheric pressure on a map. Not only does the normal to an isobar give the direction of greatest increase or decrease of atmospheric pressure, the closeness of successive isobars indicates

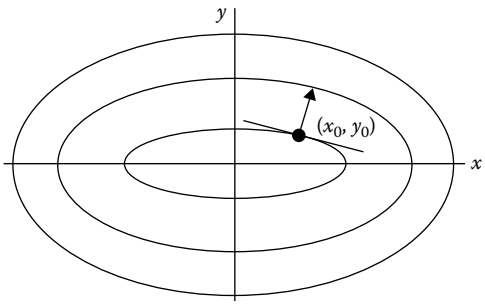


FIGURE 6.22
Normal to a level curve.

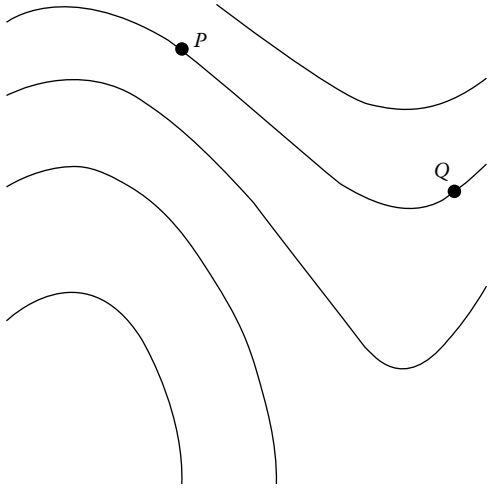


FIGURE 6.23
Isobars.

strength of the winds. For example, we expect stronger winds at point P than at point Q for the isobars shown in the fictional map in Figure 6.23.

As to the direction of the wind, it would be normal to the isobars *if* the Earth were not rotating. But the Earth's rotation and the resulting "Coriolis acceleration" actually cause the wind direction to be almost parallel to the isobars, at least at altitudes above 1000 m, where ground effects are insignificant! This is called **geostrophic wind** and will be explored in Problem 6.8.4.6.

Similarly, a set of points (x, y, z) that satisfy an equation of the form $k = F(x, y, z)$, for some scalar constant k, might be a part of a surface. Suppose $P_0 = (x_0, y_0, z_0)$ is in that set, that is, $k = F(x_0, y_0, z_0)$. By linear approximation,

$$F(x,y,z) \approx F(x_0,y_0,z_0) + (x-x_0)\frac{\partial F}{\partial x}(x_0,y_0,z_0) + (y-y_0)\frac{\partial F}{\partial y}(x_0,y_0,z_0) + (z-z_0)\frac{\partial F}{\partial z}(x_0,y_0,z_0),$$

so the direction of greatest increase of F is

$$\nabla F(x_0,y_0,z_0) = \frac{\partial F}{\partial x}(x_0,y_0,z_0)\,\hat{\boldsymbol{\imath}} + \frac{\partial F}{\partial y}(x_0,y_0,z_0)\,\hat{\boldsymbol{\jmath}} + \frac{\partial F}{\partial z}(x_0,y_0,z_0)\,\hat{\boldsymbol{k}},$$

as long as $\nabla F(x_0, y_0, z_0) \neq \mathbf{0}$.

Example 6.27

(Example 6.26 again) Find a vector normal to the plane tangent to the surface

$$\frac{x^2}{5^2} + \frac{y^2}{3^2} - \frac{z^2}{2^2} = 1.$$

at the point $(-\frac{15}{\sqrt{17}}, \frac{15}{\sqrt{17}}, 2)$.

Method: Define $f(x,y,z) \triangleq \frac{x^2}{5^2} + \frac{y^2}{3^2} - \frac{z^2}{2^2}$, so $P_0 \triangleq (-\frac{15}{\sqrt{17}}, \frac{15}{\sqrt{17}}, 2)$ is on the level set $1 = f(x,y,z)$. We calculate

$$\nabla f = \frac{2x}{5^2}\,\hat{\boldsymbol{\imath}} + \frac{2y}{3^2}\,\hat{\boldsymbol{\jmath}} - \frac{2z}{2^2}\,\hat{\boldsymbol{k}}.$$

So

$$\mathbf{n} = \nabla f(-\frac{15}{\sqrt{17}}, \frac{15}{\sqrt{17}}, 2) = \frac{2}{25}\cdot\left(-\frac{15}{\sqrt{17}}\right)\hat{\boldsymbol{\imath}} + \frac{2}{9}\cdot\left(\frac{15}{\sqrt{17}}\right)\hat{\boldsymbol{\jmath}} - \frac{2}{4}\cdot 2\hat{\boldsymbol{k}}$$

$$= -\frac{6}{5\sqrt{17}}\,\hat{\boldsymbol{\imath}} + \frac{10}{3\sqrt{17}}\,\hat{\boldsymbol{\jmath}} - \hat{\boldsymbol{k}},$$

which agrees with (6.53) except for a factor of -1. ◯

6.5.1 Problems

In problems 1–4, find an equation of the tangent plane to the surface at the given point.

1. $3x^2 - y^2 + xz = 7$ at $(1,2,8)$
2. $z = e^{xy} - 2e^x + e^{2y}$ at $(2,1,0)$
3. $z = x^2 - 3y^2 + xy$ at $(1,-1,-3)$
4. $\mathbf{r} = u\hat{\boldsymbol{\imath}} + v^3\,\hat{\boldsymbol{\jmath}} + ue^{-uv^2}\,\hat{\boldsymbol{k}}$ at the point $(x,y,z) = (-1,-8,-e^4)$
5. Find an upward-pointing unit vector that is perpendicular to the plane containing the vectors $\hat{\boldsymbol{\imath}} + \hat{\boldsymbol{\jmath}} - 2\hat{\boldsymbol{k}}$ and $3\hat{\boldsymbol{\imath}} - 2\hat{\boldsymbol{\jmath}} + \hat{\boldsymbol{k}}$.
6. A helicoid is parametrized by $\mathbf{r}(u,v) = u\cos v\,\hat{\boldsymbol{\imath}} + u\sin v\,\hat{\boldsymbol{\jmath}} + v\,\hat{\boldsymbol{k}}$. Find an equation of the tangent plane to the helicoid at the point $(x,y,z) = (0,\pi,\frac{\pi}{2})$.
7. If a surface is given by $z = g(x,y)$ near a point $P_0 = (x_0,y_0,z_0)$, define $f(x,y,z) \triangleq z - g(x,y)$. (a) Use ∇f to find a vector normal to the surface at P_0. (b) For the particular situation of Example 6.26, does this method produce a normal vector that agrees with (6.53)?
8. By expressing $\frac{x^2}{5^2} + \frac{y^2}{3^2} - \frac{z^2}{2^2} = 1$ as a surface in the form $z = f(x,y)$ near the point $P_0 = (-\frac{15}{\sqrt{17}}, \frac{15}{\sqrt{17}}, 2)$, find yet another way to find a vector normal to the tangent plane to the surface there.
9. Two surfaces $\phi(x,y,z) = 0$ and $\psi(x,y,z) = 0$ are defined to be **orthogonal** if at *every* point at which they intersect their normal vectors are orthogonal. Are the surfaces $x^2 - y^2 + z^2 = 4$ and $z = \frac{1}{xy^2}$ orthogonal? [Time-saving hint: Do *not* begin by trying to find all of their points of intersection.]
10. Suppose a point has position vector $\mathbf{r}_0$ on a level set $\mathcal{S} = \{(x,y,z) : F(x,y,z) = k\}$, where k is a constant. Explain why a nonzero vector $\mathbf{n} = \nabla F|_{\mathbf{r}_0}$ is a normal vector by considering all curves $\mathcal{C} : \mathbf{r} = \mathbf{r}(t)$ that lie on $\mathcal{S}$ and using the multivariable chain rule (6.40) in Section 6.4.

In problems 11–13, sketch the level curves of the function and find the normal vector to a point on several of the level curves.

11. $f(x,y) = 2x^2 + y^2$

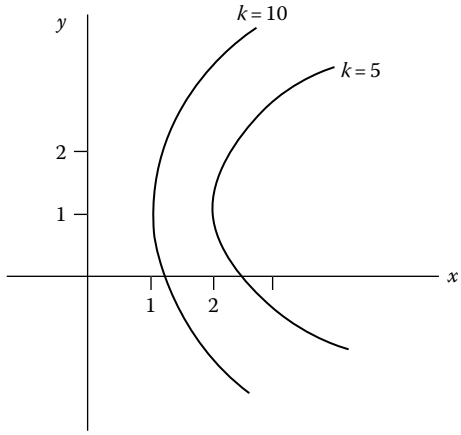


FIGURE 6.24
Level curves of $f = f(x, y)$.

12. $f(x, y) = \frac{2y}{x^2+y^2}$,
13. $f(x, y) = \frac{x^2+y^2}{4x}$
14. Sketched in Figure 6.24 are parts of two level curves for a function $f(x, y)$, for levels $k = 5$ and $k = 10$. (a) Find the unit vector in which f is increasing the most at the point $(x, y) = (2, 1)$, and (b) use that unit vector to find a point (x, y) where $f(x, y) \approx 6$. Be accurate and check that your conclusion would make sense in the picture.
15. Define $f(x, y) = x^2+2y^2$. (a) At the point $(x, y) = (2, 1)$, in what unit vector direction is f increasing the most? (b) For what value of k does the level curve $f(x, y) = k$ pass through the point $(2, 1)$? Draw that level curve and the unit normal vector to it at that point.

6.6 Area, Volume, and Linear Transformations

Recall Theorem 6.1(b) in Section 6.1, that is,

$$||\mathbf{A} \times \mathbf{B}|| = ||\mathbf{A}||\,||\mathbf{B}||\,\sin\theta, \tag{6.54}$$

where θ is the angle between the vectors $\mathbf{A}$ and $\mathbf{B}$ with $0 \leq \theta \leq \pi$. The parallelogram determined by $\mathbf{A}$ and $\mathbf{B}$ has

$$\text{Area} = (\text{base})\ (\text{height}) = ||\mathbf{A}||\ (||\mathbf{B}||\,\sin\theta) = ||\mathbf{A} \times \mathbf{B}||,$$

as shown in Figure 6.25.

Example 6.28

Find the area of the triangle with vertices $(4, 0, 0)$, $(0, 2, 0)$, $(0, 0, \frac{4}{3})$.

Method: Denote by P, Q, R the three given points. The area of the triangle shown in Figure 6.26 is one-half of the area of the parallelogram determined by vectors

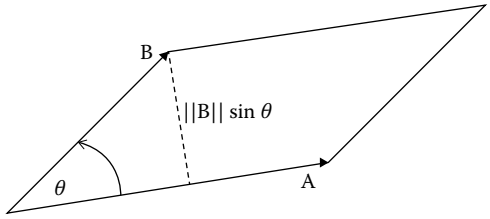


FIGURE 6.25
Parallelogram.

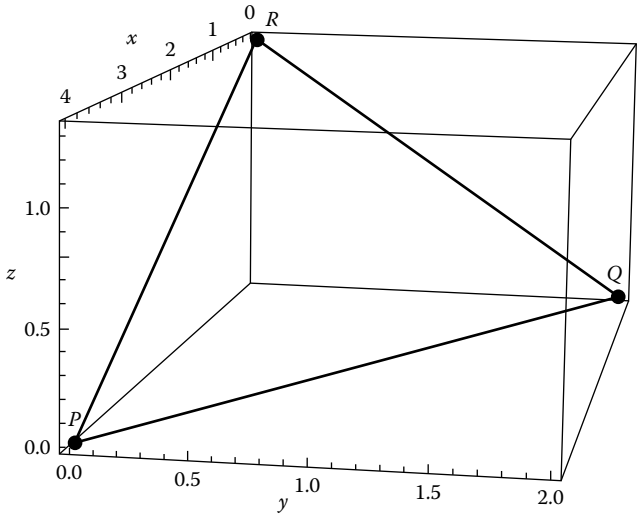


FIGURE 6.26
Triangle with vertices $(4,0,0)$, $(0,2,0)$, $(0,0,\frac{4}{3})$.

$\mathbf{A} \triangleq \overrightarrow{PQ} = -4\,\hat{\boldsymbol{\imath}} + 2\,\hat{\boldsymbol{\jmath}}$ and $\mathbf{B} \triangleq \overrightarrow{PR} = -4\,\hat{\boldsymbol{\imath}} + \frac{4}{3}\,\hat{\boldsymbol{k}}$. Because $||\mathbf{A}|| = \sqrt{4^2 + 2^2} = \sqrt{20}$, $||\mathbf{B}|| = \sqrt{4^2 + \left(\frac{4}{3}\right)^2} = \frac{\sqrt{160}}{3}$, and

$$\cos\theta = \frac{\mathbf{A} \bullet \mathbf{B}}{||\mathbf{A}||\,||\mathbf{B}||} = \frac{16}{\sqrt{20} \cdot \frac{\sqrt{160}}{3}} = \frac{6}{5\sqrt{2}}$$

implies

$$\sin\theta = \sqrt{1 - \cos^2\theta} = \frac{\sqrt{7}}{5},$$

so the area is

$$\text{Area} = \frac{1}{2}\,\text{base} \times \text{height} = \frac{1}{2} \cdot \sqrt{20} \cdot \frac{\sqrt{160}}{3} \cdot \frac{\sqrt{7}}{5} = \frac{4\sqrt{14}}{3}. \bigcirc$$

If **A** and **B** happen to lie in the *xy*-plane, that is,

$$\mathbf{A} = a\,\hat{\boldsymbol{\imath}} + b\,\hat{\boldsymbol{\jmath}} \text{ and } \mathbf{B} = c\,\hat{\boldsymbol{\imath}} + d\,\hat{\boldsymbol{\jmath}},$$

then

$$\mathbf{A} \times \mathbf{B} = (ad - bc)\hat{\mathbf{k}}.$$

So, in this special case,

$$\text{Area} = ||\mathbf{A} \times \mathbf{B}|| = ||(ad - bc)\hat{\mathbf{k}}|| = |ad - bc| = \left| det \begin{bmatrix} a & c \\ b & d \end{bmatrix} \right|,$$

and we see that the area of the parallelogram is given by the absolute value of a determinant.

Theorem 6.4

If **A**, **B** are vectors in the *xy*-plane, then the parallelogram they determine has

$$\text{Area} = \left| det \left[\mathbf{A} \,\vdots\, \mathbf{B} \right] \right|. \tag{6.55}$$

Now look at what happens in $\mathbb{R}^3$: first, we will study the special case where $\mathbf{A} = a\,\hat{\boldsymbol{\imath}} + b\,\hat{\boldsymbol{\jmath}}$ and $\mathbf{B} = c\,\hat{\boldsymbol{\imath}} + d\,\hat{\boldsymbol{\jmath}}$ lie in the *xy*-plane and **C** is parallel to the *z*-axis, that is, $\mathbf{C} = w\,\hat{\mathbf{k}}$ for some scalar *w*. As shown in Figure 6.27, the parallelepiped determined by the vectors **A**, **B**, and **C** has

$$\text{Volume} = (\text{Base area})\ (\text{height}) = |ad - bc|\ |w| = \left| det \begin{bmatrix} a & c & 0 \\ b & d & 0 \\ 0 & 0 & w \end{bmatrix} \right| = \left| det \left[\mathbf{A} \,\vdots\, \mathbf{B} \,\vdots\, \mathbf{C} \right] \right|.$$

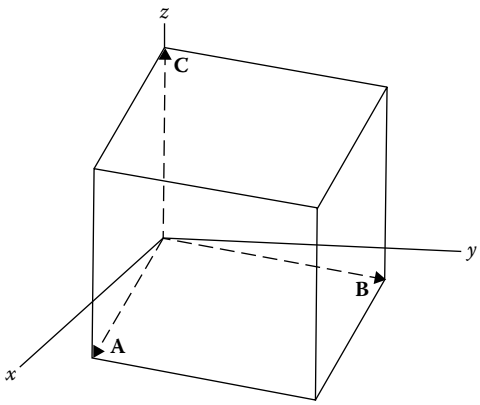


FIGURE 6.27
Volume of parallelepiped #1.

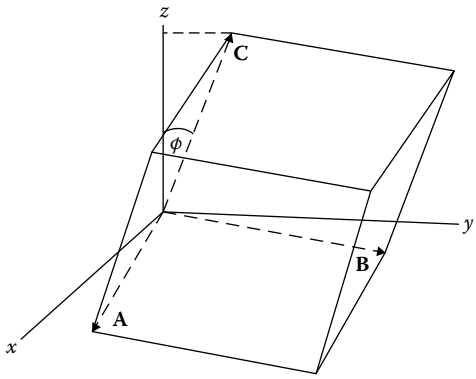


FIGURE 6.28
Volume of parallelepiped #2.

So, at least in this special case, the volume of a parallelepiped is given by the absolute value of a determinant.

In fact, **C** need not be a multiple of $\hat{\boldsymbol{k}}$. Assuming again that **A** and **B** lie in the xy-plane, the volume of the parallelepiped determined by the vectors **A**, **B**, and **C** is given by

$$\text{Volume} = (\text{Base area})\ (\text{height}) = (ad - bc)\ (\text{length of the projection of } \mathbf{C} \text{ on the } z\text{−axis}), \tag{6.56}$$

as shown in Figure 6.28. But, $\mathbf{A} \times \mathbf{B} = (ad - bc)\hat{\boldsymbol{k}}$, and

$$\text{the length of the projection of } \mathbf{C} \text{ on the } z\text{−axis} = ||\mathbf{C}||\ |\cos\phi| = |\mathbf{C} \bullet \hat{\boldsymbol{k}}|, \tag{6.57}$$

where ϕ is the angle from the positive *z-axis* to the vector **C**, as in spherical coordinates. Putting together the information in (6.56) and (6.57), we see that in this special case, the volume of the parallelepiped is given by

$$\text{Volume} = |\,(\mathbf{A} \times \mathbf{B}) \bullet \mathbf{C}\,|.$$

By the way, $(\mathbf{A} \times \mathbf{B}) \bullet \mathbf{C}$ is called a **scalar triple product**.

Denoting

$$\mathbf{C} = \begin{bmatrix} u \\ v \\ w \end{bmatrix},$$

we observe that

$$det \begin{bmatrix} a & c & u \\ b & d & v \\ 0 & 0 & w \end{bmatrix} = (ad - bc)w,$$

hence, the parallelepiped determined by **A**, **B**, and **C** has volume given by

$$\text{Volume} = \text{(Base area) (height)} = |(ad - bc)|\,|w| = \left| det \begin{bmatrix} a & c & u \\ b & d & v \\ 0 & 0 & w \end{bmatrix} \right| .$$

It follows that

$$\text{Volume} = \left| det \left[\mathbf{A} \,\vdots\, \mathbf{B} \,\vdots\, \mathbf{C} \right] \right| . \tag{6.58}$$

Finally, to be completely general, drop the assumption that **A** and **B** lie in the xy-plane. Suppose for the moment that we can find a real, orthogonal matrix Q that rotates both **A** and **B** into the xy-plane. Recall that we saw in Section 2.4, specifically Problems 2.4.3.12 and 2.4.3.11, that multiplication by an orthogonal matrix preserves lengths of vectors, that is, $||Q\mathbf{A}|| = ||\mathbf{A}||$, and likewise preserves angles between vectors, because $(Q\mathbf{A})\bullet(Q\mathbf{B}) = \mathbf{A}\bullet\mathbf{B}$.

After rotation, the vectors $Q\mathbf{A}$ and $Q\mathbf{B}$ lie in the xy-plane, so the volume of the parallelepiped determined by $Q\mathbf{A}$, $Q\mathbf{B}$, and $Q\mathbf{C}$ is, by (6.58),

$$\text{Volume} = \left| det \left[Q\mathbf{A} \,\vdots\, Q\mathbf{B} \,\vdots\, Q\mathbf{C} \right] \right| .$$

But, by Theorem 1.9 in Section 1.2, $Q\left[\mathbf{A} \,\vdots\, \mathbf{B} \,\vdots\, \mathbf{C} \right] = \left[Q\mathbf{A} \,\vdots\, Q\mathbf{B} \,\vdots\, Q\mathbf{C} \right]$, so the volume of the parallelepiped is given by

$$\text{Volume} = \left| det(Q \left[\mathbf{A} \,\vdots\, \mathbf{B} \,\vdots\, \mathbf{C} \right]) \right| = \left| det(Q)\, det \left[\mathbf{A} \,\vdots\, \mathbf{B} \,\vdots\, \mathbf{C} \right] \right| = |det(Q)| \left| det \left[\mathbf{A} \,\vdots\, \mathbf{B} \,\vdots\, \mathbf{C} \right] \right| .$$

But all orthogonal matrices have determinant equal to ± 1, so (6.58), that is, Volume $= \left| det \left[\mathbf{A} \,\vdots\, \mathbf{B} \,\vdots\, \mathbf{C} \right] \right|$, holds in the most general case, assuming we can find a real, orthogonal matrix Q that rotates both **A** and **B** into the xy-plane.

Now we will explain why such a real, orthogonal matrix Q can be found. The fact that we can find it, and how we find it, are interesting for their own sakes! We will consider two cases: (i) when $\mathbf{n} \triangleq \mathbf{A} \times \mathbf{B} \neq \mathbf{0}$ and (ii) if $\mathbf{A} \times \mathbf{B} = \mathbf{0}$.

In the first case, define $\widehat{\mathbf{n}} = \frac{1}{||\mathbf{n}||}\,\mathbf{n}$. We will explain why we can find a real, orthogonal matrix Q such that $Q\widehat{\mathbf{n}} = \hat{\boldsymbol{k}}$: think of $\widehat{\mathbf{n}}$ as the unit normal to a point P_0 on the sphere $\rho = 1$, specifically $\widehat{\mathbf{n}} = \overrightarrow{OP_0}$. In spherical coordinates,

$$\widehat{\mathbf{n}} = \overrightarrow{OP_0} = \sin\phi_0 \cos\theta_0\, \hat{\boldsymbol{\imath}} + \sin\phi_0 \sin\theta_0\, \hat{\boldsymbol{\jmath}} + \cos\phi_0\, \hat{\boldsymbol{k}}.$$

The matrix

$$Q_0 \triangleq \begin{bmatrix} \cos\theta_0 & \sin\theta_0 & 0 \\ -\sin\theta_0 & \cos\theta_0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

rotates vectors around the z-axis clockwise by an angle θ_0. Specifically, we have

$$Q_0\,\hat{\mathbf{n}} = Q_0 \begin{bmatrix} \sin\phi_0\cos\theta_0 \\ \sin\phi_0\sin\theta_0 \\ \cos\phi_0 \end{bmatrix} = \begin{bmatrix} \sin\phi_0 \\ 0 \\ \cos\phi_0 \end{bmatrix}.$$

So, Q_0 rotates $\hat{\mathbf{n}}$ to lie in the xz-plane. To rotate $Q_0\hat{\mathbf{n}}$ to be $\hat{\boldsymbol{k}}$, we can use the matrix

$$Q_1 \triangleq \begin{bmatrix} \cos\phi_0 & 0 & -\sin\phi_0 \\ 0 & 1 & 0 \\ \sin\phi_0 & 0 & \cos\phi_0 \end{bmatrix}.$$

So, we have arranged to get

$$Q_1(Q_0\,\hat{\mathbf{n}}) = Q_1 \begin{bmatrix} \sin\phi_0 \\ 0 \\ \cos\phi_0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

You should check that the matrix

$$Q \triangleq Q_1Q_0 = \begin{bmatrix} \cos\phi_0 & 0 & -\sin\phi_0 \\ 0 & 1 & 0 \\ \sin\phi_0 & 0 & \cos\phi_0 \end{bmatrix} \begin{bmatrix} \cos\theta_0 & \sin\theta_0 & 0 \\ -\sin\theta_0 & \cos\theta_0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \cos\phi_0\cos\theta_0 & \cos\phi_0\sin\theta_0 & -\sin\phi_0 \\ -\sin\theta_0 & \cos\theta_0 & 0 \\ \sin\phi_0\cos\theta_0 & \sin\phi_0\sin\theta_0 & \cos\phi_0 \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{e}}_\phi^T \\ \hat{\mathbf{e}}_\theta^T \\ \hat{\mathbf{e}}_\rho^T \end{bmatrix}$$

is a real, orthogonal matrix. By construction, it rotates $\mathbf{n}$ to $\hat{\boldsymbol{k}}$, as we desired. Further, by Problem 6.6.5.14, $\hat{\boldsymbol{k}} = Q\mathbf{n}$ is orthogonal to both $Q\mathbf{A}$ and $Q\mathbf{B}$. Thus, $Q\mathbf{A}$ and $Q\mathbf{B}$ lie in the xy-plane, as we desired.

In the second case, that is, when $\mathbf{n} \triangleq \mathbf{A} \times \mathbf{B} = \mathbf{0}$, then $\mathbf{A}$ and $\mathbf{B}$ are parallel, so it is easy to find a real, orthogonal matrix that rotates $\mathbf{A}$, and thus also $\mathbf{B}$, into the x-axis and thus into the xy-plane.

Thus, we have established the next result.

Theorem 6.5

If $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are vectors in the $\mathbb{R}^3$ then the parallelepiped they determine has volume given by (6.58), that is,

$$\text{Volume} = \left| det \left[\mathbf{A} \,\vdots\, \mathbf{B} \,\vdots\, \mathbf{C} \right] \right|.$$

6.6.1 Linear Transformations

Suppose A is a real, 2×2 matrix $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$. Take a point (x, y) in the plane and multiply its position vector, $\begin{bmatrix} x \\ y \end{bmatrix}$, on the left by the matrix A. This gives the vector

$$A \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a_{11}x + a_{12}y \\ a_{21}x + a_{22}y \end{bmatrix}, \tag{6.59}$$

which is the position vector of a new point, $(a_{11}x + a_{12}y, a_{21}x + a_{22}y)$, in the plane.

If S is a set of points in the plane, then

$$A(S) \triangleq \{A\mathbf{x} : \ \mathbf{x} \text{ in } S\}$$

is a set of points* in the plane. We call $A(S)$ the **image** of S under the transformation $\mathbf{x} \mapsto A\mathbf{x}$ given by (6.59).

A real, 2×2 matrix A thus defines a **transformation** or **map** of the plane. In fact, (6.59) defines a **linear transformation** of the plane because

$$A(\alpha \mathbf{x}_1 + \beta \mathbf{x}_2) = \alpha A(\mathbf{x}_1) + \beta (A\mathbf{x}_2)$$

for all scalars α, β, and all vectors $\mathbf{x}_1, \mathbf{x}_2$ in $\mathbb{R}^2$.

Example 6.29

Determine the images of the square $S = \{(x, y) : \ 0 \le x \le 1, 0 \le y \le 1\}$ under the transformations defined by the matrices:

$$(a) \begin{bmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}, \quad (b) \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, \quad (c) \begin{bmatrix} \cos\frac{\pi}{6} & -\sin\frac{\pi}{6} \\ \sin\frac{\pi}{6} & \cos\frac{\pi}{6} \end{bmatrix}.$$

Also, for each of (a), (b), and (c), sketch the **mapping**, that is, the set S and its image $A(S)$.

* At this point, we are acting as if there is no difference between a point and its position vector, so that we don't have to use an excess of words.

Method: The **unit square** $S \triangleq \{(x,y) : 0 \le x \le 1, 0 \le y \le 1\}$ is the parallelogram determined by the two vectors:

$$\mathbf{e}^{(1)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{e}^{(2)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

To see what a matrix A does to S, it helps to see what A does to each of $\mathbf{e}^{(1)}, \mathbf{e}^{(2)}$: because

$$A(S) = \{A(c_1\mathbf{e}^{(1)} + c_2\mathbf{e}^{(2)}) : 0 \le c_1 \le 1, 0 \le c_2 \le 1\}, \tag{6.60}$$

we see that $A(S)$ is the parallelogram determined by the vectors $A\mathbf{e}^{(1)}, A\mathbf{e}^{(2)}$.

(a) Because

$$\begin{bmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} = 2\,\mathbf{e}^{(1)} \quad \text{and} \quad \begin{bmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{1}{2} \end{bmatrix} = \frac{1}{2}\,\mathbf{e}^{(2)},$$

it follows that

$$\begin{bmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} (S) = \left\{ (x,y) : 0 \le x \le 2, 0 \le y \le \frac{1}{2} \right\},$$

which is a rectangle. We can think of this rectangle as having been produced by stretching the unit square in the x direction and compressing it in the y direction. The mapping is shown in Figure 6.29. The images of the points B, C, D are B', C', D'.

(b) Because

$$\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix},$$

$\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} (S)$ is the parallelogram determined by the vectors $\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix}$. The mapping is shown in Figure 6.30. The images of the points B, C, D are B', C', D'.

(c) Because

$$\begin{bmatrix} \cos\frac{\pi}{6} & -\sin\frac{\pi}{6} \\ \sin\frac{\pi}{6} & \cos\frac{\pi}{6} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \cos\frac{\pi}{6} \\ \sin\frac{\pi}{6} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \cos\frac{\pi}{6} & -\sin\frac{\pi}{6} \\ \sin\frac{\pi}{6} & \cos\frac{\pi}{6} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -\sin\frac{\pi}{6} \\ \cos\frac{\pi}{6} \end{bmatrix},$$

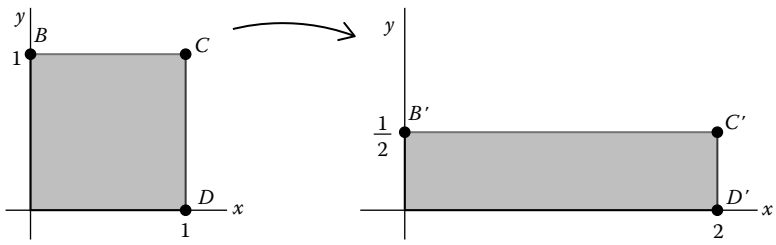


FIGURE 6.29
Example 6.29(a).

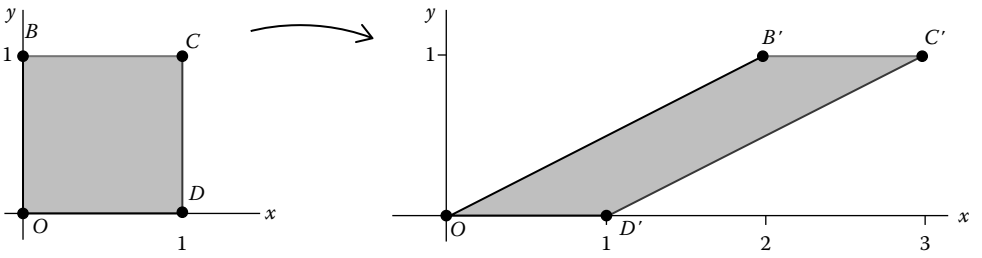


FIGURE 6.30
Example 6.29(b).

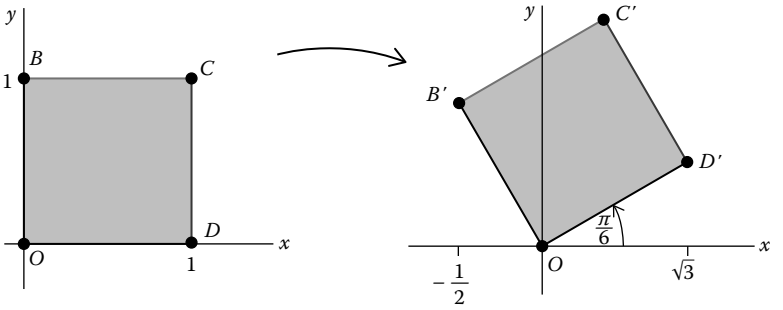


FIGURE 6.31
Example 6.29(c).

$$\begin{bmatrix} \cos\frac{\pi}{6} & -\sin\frac{\pi}{6} \\ \sin\frac{\pi}{6} & \cos\frac{\pi}{6} \end{bmatrix} (S)$$

is the square obtained by the counterclockwise rotation of the unit square by an angle of $\frac{\pi}{6}$. The mapping is shown in Figure 6.31. The images of the points B, C, D are B', C', D'. ○

Two by two matrices of the form

$$\begin{bmatrix} k & 0 \\ 0 & h \end{bmatrix},$$

where h, k are positive constants, define **stretching and/or compressing** transformations. Matrices of the form

$$\begin{bmatrix} 1 & 0 \\ h & 1 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}$$

define **shear** transformations. Matrices of the form

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

define **rotation** transformations. Matrices of the form

$$\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \text{or} \quad \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

define **reflection** transformations.

6.6.2 Linear Transformations, Area, and Volume

Because of the formula for area given in (6.55), linear transformations defined by a real, 2×2 matrix A have the property that

$$\text{Area of } A(\mathcal{D}) = |\, det(A)\,| \text{ (Area of } \mathcal{D})$$

for any 2D region $\mathcal{D}$. Why? Because of (6.60), $A(S)$ is the parallelogram determined by the vectors $A\mathbf{e}^{(1)}, A\mathbf{e}^{(2)}$, so the area formula for a parallelogram implies

$$\text{Area of } A(S) = |\, det\left(\left[\, A\mathbf{e}^{(1)} \,\vdots\, A\mathbf{e}^{(2)} \,\right]\right)| = |\, det\left(A\left[\, \mathbf{e}^{(1)} \,\vdots\, \mathbf{e}^{(2)} \,\right]\right)| = |\, det(AI_2)\,| = |\, det(A)\,|.$$

If a rectangle R is of the form $R = \{(x_0, y_0) + (x, y) : \ 0 \le x \le a, 0 \le y \le b\}$, then a similar calculation explains why

$$\begin{aligned} \text{Area of } A(R) &= |\, det\left(\left[\, aA\mathbf{e}^{(1)} \,\vdots\, bA\mathbf{e}^{(2)} \,\right]\right)| = |\, det\left(A\left[\, a\mathbf{e}^{(1)} \,\vdots\, b\mathbf{e}^{(2)} \,\right]\right)| \\ &= |\, det(A)\,|\,|\, det\left[\, a\mathbf{e}^{(1)} \,\vdots\, b\mathbf{e}^{(2)} \,\right]| = |\, det(A)\,|\, ab. \end{aligned}$$

(The translation by (x_0, y_0) does not affect the area.)

Example 6.30

By what factors do the transformations defined by the given 2×2 matrices multiply area?

$$\text{(a)} \begin{bmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}, \quad \text{(b)} \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, \quad \text{(c)} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \quad \text{(d)} \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}.$$

Method:

(a) $|det(A)| = 1$, so area is unaffected, that is, is multiplied by a factor of 1.
(b) $|det(A)| = 1$, so area is unaffected, that is, is multiplied by a factor of 1.
(c) $|det(A)| = 1$, so area is unaffected, that is, is multiplied by a factor of 1.
(d) $|det(A)| = 2$, so area is multiplied by a factor of 2. ◯

Figure 6.32 shows the effect of the linear transformation defined by the matrix in part (d) of Example 6.30. That transformation shears in both ways and stretches.

Because all orthogonal matrices have determinant equal to ± 1, orthogonal matrices define area preserving linear transformations. One example of a 2×2 orthogonal matrix is the rotation matrix in part (c) of Example 6.30.

In $\mathbb{R}^3$ we get results similar to those in $\mathbb{R}^2$: Any 3×3 matrix A defines a linear transformation of $\mathbb{R}^3$ by $\mathbf{x} \mapsto A\mathbf{x}$ and

$$\text{Volume of } A(\mathcal{V}) = |det(A)| \text{ (Volume of } \mathcal{V})$$

for any 3D region $\mathcal{V}$.

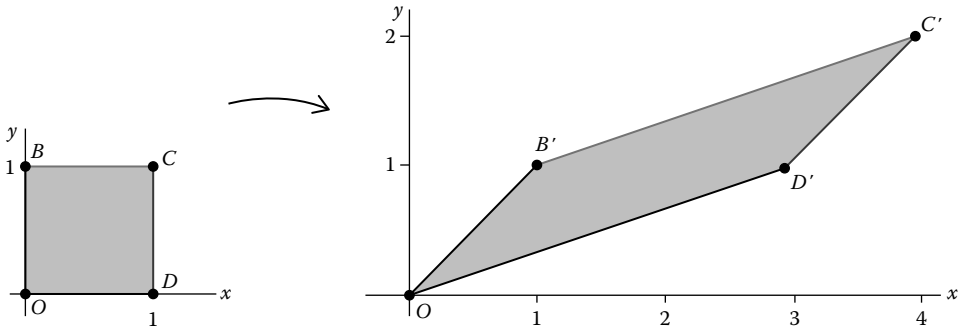


FIGURE 6.32
Example 6.30(d).

Stretching and/or compressing matrices have the form

$$\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix},$$

where a_{11}, a_{22}, and a_{33} are positive constants.

Shear transformations on $\mathbb{R}^3$ are upper or lower triangular matrices, just as for shear transformations on $\mathbb{R}^2$:

$$\begin{bmatrix} 1 & a_{12} & a_{13} \\ 0 & 1 & a_{23} \\ 0 & 0 & 1 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 1 & 0 & 0 \\ a_{21} & 1 & 0 \\ a_{31} & a_{32} & 1 \end{bmatrix}.$$

The matrices

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

are reflections, and the matrices of the forms

$$\begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} \cos\phi & 0 & -\sin\phi \\ 0 & 1 & 0 \\ \sin\phi & 0 & \cos\phi \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\psi & -\sin\psi \\ 0 & \sin\psi & \cos\psi \end{bmatrix}$$

are rotation matrices, as are products of such matrices. An example of such a product, $Q = Q_1 Q_0$, was used in the derivation of volume formula (6.58). In that derivation, Q_0 was a rotation by angle $-\theta_0$ around the z-axis, and Q_1 was a rotation by angle ϕ_0 around the y-axis.

Similar to what happens for 2×2 matrices, multiplication by a real, 3×3 matrix A transforms volume by a factor of $|det(A)|$. In particular, multiplication by a rotation or reflection matrix does not change volume, as we would expect.

In general, a product $A_nA_{n-1}\cdots A_2A_1$ defines a linear transformation that consists of *first* doing the transformation defined by matrix A_1, and then after that doing the transformation defined by matrix A_2, . . ., and finishing by doing the transformation defined by matrix A_n.

If we perform a succession of linear transformations corresponding to a matrix

$$A_nA_{n-1}\cdots A_2A_1,$$

then the effect on volume is to multiply it by the factor

$$|\,det(A_nA_{n-1}\cdots A_2A_1)\,| = \prod_{i=1}^{n} |\,det(A_i)\,|.$$

6.6.3 Change of Variables, Area, and Volume

Definition 6.5

If $x = x(u,v)$ and $y = y(u,v)$, then the map is a **change of variables** from the uv-plane to the xy-plane, as long as the **Jacobian matrix**

$$\left[\frac{\partial(x,y)}{\partial(u,v)}\right] \triangleq \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}$$

is invertible.

For example, if A is a real, 2×2 matrix, then

$$\begin{bmatrix} x \\ y \end{bmatrix} \triangleq A \begin{bmatrix} u \\ v \end{bmatrix}$$

defines a change of variables as long as A is invertible, that is, as long as $|det(A)| \neq 0$. In fact, this change of variables multiplies area by a factor of $|det(A)|$: The rectangle

$$\mathcal{D} \triangleq \{(u,v) :\ u_0 \leq u \leq u_0 + \triangle u, v_0 \leq v \leq v_0 + \triangle v\} \tag{6.61}$$

has Area$(\mathcal{D}) = \triangle u\ \triangle v$, and its image under the change of variables (Figure 6.33),

$$A(\mathcal{D}) = \{A \begin{bmatrix} u \\ v \end{bmatrix} :\ \begin{bmatrix} u \\ v \end{bmatrix} \text{ in } \mathcal{D}\},$$

has Area$(A(\mathcal{D})) = |det(A)|$ Area$(\mathcal{D})$.

Similarly, a change of variables

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x(u,v) \\ y(u,v) \end{bmatrix}$$

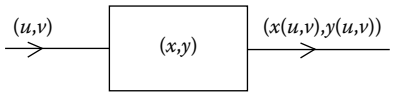


FIGURE 6.33
Change of variables in $\mathbb{R}^2$.

multiplies the area of $\mathcal{D}$, given in (6.61), by approximately

$$\left| det\left(\left[\frac{\partial(x,y)}{\partial(u,v)} \right] \right) \right|. \tag{6.62}$$

Why? Because the image of $\mathcal{D}$ under the change of variables is approximately

$$\left\{ \begin{bmatrix} x_0 + \left(\frac{\partial x}{\partial u}(u_0,v_0)\right)(u-u_0) + \left(\frac{\partial x}{\partial v}(u_0,v_0)\right)(v-v_0) \\ \\ y_0 + \left(\frac{\partial y}{\partial u}(u_0,v_0)\right)(u-u_0) + \left(\frac{\partial y}{\partial v}(u_0,v_0)\right)(v-v_0) \end{bmatrix} : u_0 \le u \le u_0 + \triangle u, v_0 \le v \le v_0 + \triangle v \right\},$$

where $x_0 \triangleq x(u_0,v_0)$ and $y_0 \triangleq y(u_0,v_0)$.

As the area of $\mathcal{D}$ gets closer and closer to zero, the approximation in (6.62) gets better and better.

The object in (6.62) is the absolute value of the determinant of the Jacobian matrix, or "the absolute value of the Jacobian determinant."

Example 6.31

Polar coordinates can be thought of as a change of variables from the $r\theta$-plane to the xy-plane:

$$\begin{bmatrix} x \\ y \end{bmatrix} \triangleq \begin{bmatrix} r\cos\theta \\ r\sin\theta \end{bmatrix}.$$

What effect does this change of variables have on area?

Method: We calculate

$$\left| det\left(\left[\frac{\partial(x,y)}{\partial(r,\theta)} \right] \right) \right| = \left| det\left(\begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{bmatrix} \right) \right| = \left| det\left(\begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix} \right) \right| = \cdots = |r| = r.$$

If an "infinitesimal" element of area in the xy-plane is given by $dx\,dy$, then the corresponding element of area in the $r\theta$-plane is

$$r\,dr\,d\theta = \left| det\left(\left[\frac{\partial(x,y)}{\partial(r,\theta)} \right] \right) \right| dr\,d\theta,$$

as illustrated in Figure 6.34. ◯

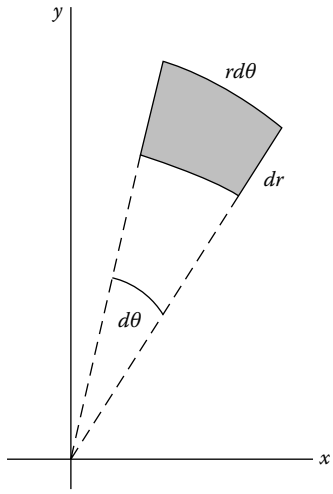


FIGURE 6.34
Example 6.30(d): Element of area in polar coordinates.

Similarly a change of variables on $\mathbb{R}^3$ is a map:

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} \mapsto \begin{bmatrix} x(u,v,w) \\ y(u,v,w) \\ z(u,v,w) \end{bmatrix}. \tag{6.63}$$

Because volume in $\mathbb{R}^3$ is also given by a determinant, specifically (6.58), linear approximation implies that the (possibly nonlinear) transformation (6.63) multiplies volume by approximately

$$\left| det\left(\left[\frac{\partial(x,y,z)}{\partial(u,v,w)} \right] \right) \right| \triangleq \left| det \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} & \frac{\partial y}{\partial w} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} & \frac{\partial z}{\partial w} \end{bmatrix} \right|.$$

Again, this is the absolute value of the determinant of the Jacobian matrix, or "the absolute value of the Jacobian determinant."

Example 6.32

Spherical coordinates can be thought of as a change of variables from $\rho\phi\theta$-space to xyz-space:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} \triangleq \begin{bmatrix} \rho \sin\phi \cos\theta \\ \rho \sin\phi \sin\theta \\ \rho \cos\phi \end{bmatrix}.$$

What effect does this change of variables have on volume?

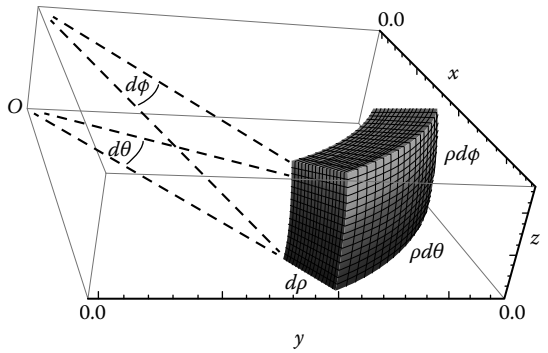


FIGURE 6.35
Example 6.32. Element of volume in spherical coordinates.

Method: We calculate

$$\left|det\left(\left[\frac{\partial(x,y,z)}{\partial(\rho,\phi,\theta)}\right]\right)\right| = \left|det\left(\begin{bmatrix} \frac{\partial x}{\partial\rho} & \frac{\partial x}{\partial\phi} & \frac{\partial x}{\partial\theta} \\ \frac{\partial y}{\partial\rho} & \frac{\partial x}{\partial\phi} & \frac{\partial y}{\partial\theta} \\ \frac{\partial z}{\partial\rho} & \frac{\partial z}{\partial\phi} & \frac{\partial z}{\partial\theta} \end{bmatrix}\right)\right|$$

$$= \left|det\left(\begin{bmatrix} \sin\phi\,\cos\theta & \rho\,\cos\phi\,\cos\theta & -\rho\,\sin\phi\,\sin\theta \\ \sin\phi\,\sin\theta & \rho\,\cos\phi\,\sin\theta & \rho\,\cos\phi\,\cos\theta \\ \cos\phi & -\rho\sin\phi & 0 \end{bmatrix}\right)\right|$$

$$= \cdots = |\rho^2\sin\phi| = \rho^2\sin\phi.$$

Note that $\sin\phi \geq 0$ because $0 \leq \phi \leq \pi$ in spherical coordinates.

If an "infinitesimal" element of volume in xyz space is given by $dx\,dy\,dz$, then the corresponding element of volume in $\rho\phi\theta$-space is

$$\rho^2\sin\phi\,d\rho\,d\phi\,d\theta = \left|det\left(\left[\frac{\partial(x,y,z)}{\partial(\rho,\phi,\theta)}\right]\right)\right| d\rho\,d\phi\,d\theta.$$

This is illustrated in Figure 6.35. ○

6.6.4 Element of Surface Area

If

$$\mathcal{S}:\ \mathbf{r} = \mathbf{r}(u,v) = x(u,v)\,\hat{\boldsymbol{\imath}} + y(u,v)\,\hat{\boldsymbol{\jmath}} + z(u,v)\,\hat{\boldsymbol{k}},\ (u,v) \text{ in } \mathcal{D},$$

is a parametrized surface, then linear approximation gives

$$\mathbf{r}(u,v) \approx \mathbf{r}_0 + (u-u_0)\frac{\partial\mathbf{r}}{\partial u}(u_0,v_0) + (v-v_0)\frac{\partial\mathbf{r}}{\partial v}(u_0,v_0),$$

where $\mathbf{r}_0 \triangleq \mathbf{r}(u_0, v_0)$. So, the image of a rectangle $\mathcal{D}$ given by (6.61) is approximately the parallelogram given by

$$\{\mathbf{r}_0 + (u-u_0)\frac{\partial \mathbf{r}}{\partial u}(u_0,v_0) + (v-v_0)\frac{\partial \mathbf{r}}{\partial v}(u_0,v_0) : u_0 \le u \le u_0 + \Delta u, v_0 \le v \le v_0 + \Delta v\}.$$

This parallelogram has

$$\text{Area} = \left|\left|\left(\triangle u\,\frac{\partial \mathbf{r}}{\partial u}(u_0,v_0)\right) \times \left(\triangle v\,\frac{\partial \mathbf{r}}{\partial v}(u_0,v_0)\right)\right|\right| = \left|\left|\frac{\partial \mathbf{r}}{\partial u}(u_0,v_0) \times \frac{\partial \mathbf{r}}{\partial v}(u_0,v_0)\right|\right| \; |\triangle u|\,|\triangle v|,$$

by the original formula for area given in (6.54). So, an **element of surface area** is

$$dS \triangleq \left|\left|\frac{\partial \mathbf{r}}{\partial u}(u_0,v_0) \times \frac{\partial \mathbf{r}}{\partial v}(u_0,v_0)\right|\right| du\,dv. \tag{6.64}$$

A special case of the surface area formula occurs when the surface is "flat," for example, when it lies in the xy-plane and is thus given by

$$\mathcal{S}: \; \mathbf{r} = \mathbf{r}(u,v) = x(u,v)\,\hat{\boldsymbol{\imath}} + y(u,v)\,\hat{\boldsymbol{\jmath}}, \; (u,v) \text{ in } \mathcal{D}. \tag{6.65}$$

In this case,

$$dS = \left|\left|\left(\frac{\partial x}{\partial u}\,\hat{\boldsymbol{\imath}} + \frac{\partial y}{\partial u}\,\hat{\boldsymbol{\jmath}}\right) \times \left(\frac{\partial x}{\partial v}\,\hat{\boldsymbol{\imath}} + \frac{\partial y}{\partial v}\,\hat{\boldsymbol{\jmath}}\right)\right|\right| du\,dv = \cdots = \left|\frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}\right| du\,dv$$

$$= \left|det\left(\left[\,\frac{\partial(x,y)}{\partial(u,v)}\,\right]\right)\right| du\,dv.$$

This agrees with (6.62). So, a change of variables multiplies area by the same factor, $\left|\left|\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v}\right|\right|$, that appears in the formula for surface area in the special case of a flat surface. This makes a lot of sense because a change of variables in $\mathbb{R}^2$ is really a parametrization of a flat surface.

Learn More About It

For more about linear transformations, including homogeneous coordinates, translation, and perspective, see *Linear Algebra*, 3rd edn., by David C. Lay, Addison-Wesley, © 2003, Section 2.7.

6.6.5 Problems

In problems 1 and 2, find the area of the parallelogram satisfying the given conditions.

1. Parallelogram has vertices $(0,0,0), (1,0,-2), (3,3,-2), (2,3,0)$
2. Parallelogram is determined by the vectors $3\hat{\boldsymbol{\imath}} + \hat{\boldsymbol{\jmath}}$ and $2\hat{\boldsymbol{\jmath}} + \hat{\boldsymbol{k}}$

In problems 3 and 4, the given vectors determine a parallelepiped. Find its volume and sketch it.

3. $\left\{ \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 4 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix} \right\}$
4. $\left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 3 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} \right\}$

In problems 5 and 6, find the volume of the parallelepiped with the given vertices.

5. $(0,0,0), (1,1,0), (0,2,0), (1,3,0), (1,2,3), (2,3,3), (1,4,3), (2,5,3)$
6. $(0,0,0), (1,0,1), (0,2,0), (1,2,1), (1,2,3), (2,2,4), (1,4,3), (2,4,4)$

For problems 7–10, sketch the effect of the linear transformation defined by the given matrix, and describe that effect in words. [Note: For a linear transformation defined by a product of matrices, the matrix furthest on the right acts first, then the one on its left, then the one left of that, etc.]

7. $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$
8. $\begin{bmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{bmatrix}$
9. $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{bmatrix}$
10. $\begin{bmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$
11. Find a 2×2 matrix that gives the linear transformation sketched in Figure 6.36.

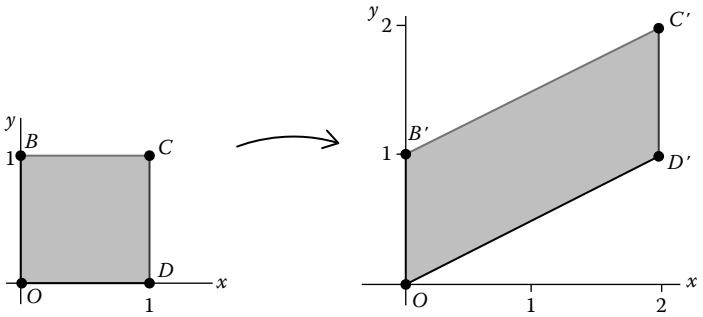


FIGURE 6.36
Problem 6.6.5.11.

In problems 12 and 13, by what factor is area or volume multiplied as a result of the linear transformation defined by the given matrix?

12. $\begin{bmatrix} 30 & 29 \\ 31 & 30 \end{bmatrix}$

13. $\begin{bmatrix} 1 & 2 & 4 \\ -1 & 0 & 5 \\ 1 & 0 & 3 \end{bmatrix}$

14. Explain why $Q\mathbf{n}$ is orthogonal to both $Q\mathbf{A}$ and $Q\mathbf{B}$ if Q is an orthogonal matrix and $\mathbf{n} \triangleq \mathbf{A} \times \mathbf{B}$.

15. (a) Write down three different examples of 3×3 elementary matrices.
    (b) Find the effect on volume for each of the matrices you wrote in part (a).
    (c) Formulate a conjecture about the effect of elementary matrices on volume.

16. In crystallography, the "primitive cell" of a lattice $\mathcal{L}$ is the parallelepiped, three of whose edges are the vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$, which are assumed to be linearly independent. The $\mathbf{a}, \mathbf{b}, \mathbf{c}$ are called the "translation vectors." The **lattice** $\mathcal{L}$ is defined to be the position vectors of an infinite set of points such that $\mathcal{L} = n_1\mathbf{a} + n_2\mathbf{b} + n_3\mathbf{c} + \mathcal{L}$, for all integers n_1, n_2, n_3. In other words, the lattice looks the same under translations by multiples of the vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$.

    The "reciprocal (or dual) lattice," $\mathcal{L}^\star$, is defined to be the set of all vectors $\mathbf{k}$ for which $e^{i2\pi\mathbf{k}\bullet\mathbf{r}} = 1$ for all vectors $\mathbf{r}$ in $\mathcal{L}$.

    (a) Find the volume, V, of the primitive cell and

    (b) Explain why the vectors $\mathbf{A}, \mathbf{B}, \mathbf{C}$ give a primitive cell of the reciprocal lattice, $\mathcal{L}^\star$, where

$$\mathbf{A} \triangleq \frac{1}{\mathbf{a} \bullet (\mathbf{b} \times \mathbf{c})} \mathbf{b} \times \mathbf{c}, \quad \mathbf{B} \triangleq \frac{1}{\mathbf{a} \bullet (\mathbf{b} \times \mathbf{c})} \mathbf{c} \times \mathbf{a}, \quad \mathbf{C} \triangleq \frac{1}{\mathbf{a} \bullet (\mathbf{b} \times \mathbf{c})} \mathbf{a} \times \mathbf{b}.$$

    [Hint: You may use the result about scalar triple products mentioned in Problem 6.8.4.15.]

    (c) Explain why $\left[\, \mathbf{a} \;\vdots\; \mathbf{b} \;\vdots\; \mathbf{c} \,\right]^{-1} = \left[\, \mathbf{A} \;\vdots\; \mathbf{B} \;\vdots\; \mathbf{C} \,\right]^T$.

    Each of the vectors $\mathbf{A}, \mathbf{B}, \mathbf{C}$ is normal to two of the faces of the primitive cell parallelepiped.

6.7 Differential Operators and Curvilinear Coordinates

The foundation of calculus in $\mathbb{R}^2$ or $\mathbb{R}^3$ is the *gradient operator*. In $\mathbb{R}^2$ this is defined by

$$\textbf{\textit{grad}}(f) = \nabla f \triangleq \frac{\partial f}{\partial x}\,\hat{\boldsymbol{\imath}} + \frac{\partial f}{\partial y}\,\hat{\boldsymbol{\jmath}},$$

and in $\mathbb{R}^3$ the gradient operator is defined by

$$\textbf{\textit{grad}}(f) = \nabla f \triangleq \frac{\partial f}{\partial x}\,\hat{\boldsymbol{\imath}} + \frac{\partial f}{\partial y}\,\hat{\boldsymbol{\jmath}} + \frac{\partial f}{\partial z}\,\hat{\mathbf{k}}.$$

Mathematical physics also uses two operators on vector fields: the **divergence**, defined by

$$div(\mathbf{F}) = \nabla \bullet \mathbf{F} \triangleq \left(\hat{\imath}\,\frac{\partial}{\partial x} + \hat{\jmath}\,\frac{\partial}{\partial y} + \hat{k}\,\frac{\partial}{\partial z}\right) \bullet \left(F_x\,\hat{\imath} + F_y\,\hat{\jmath} + F_z\,\hat{k}\right) = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z},$$

and the **curl** operator, defined by

$$\boldsymbol{curl}(\mathbf{F}) = \nabla \times \mathbf{F} \triangleq \left(\hat{\imath}\,\frac{\partial}{\partial x} + \hat{\jmath}\,\frac{\partial}{\partial y} + \hat{k}\,\frac{\partial}{\partial z}\right) \left(F_x\,\hat{\imath} + F_y\,\hat{\jmath} + F_z\,\hat{k}\right),$$

hence,

$$\boldsymbol{curl}(\mathbf{F}) = \nabla \times \mathbf{F} = \left(\frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z}\right)\hat{\imath} + \left(\frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x}\right)\hat{\jmath} + \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right)\hat{k}$$

$$= \det \begin{bmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_x & F_y & F_z \end{bmatrix}.$$

There are many physical applications of these operators. Here are a few: **Fourier's law of heat conduction** states that if u is the temperature, then the rate of heat flux is $\mathbf{q} = -[\kappa]\nabla u$. In general, $[\kappa]$ could be a 3×3 matrix, which allows for heat to flow more intensely in some directions than in others. We may call $[\kappa]$ a "tensor." A special case is when $[\kappa]$ is replaced by a scalar κ.

Similarly, if c is the concentration of a substance, then **Fick's law of diffusion** states that the rate of flow of the substance is $\mathbf{J} = -[A]\nabla c$ for some matrix $[A]$ or scalar A. So, heat flow is analogous to flow of a substance.

Suppose a fluid whose mass density is $\varrho = \varrho(x,y,z)$ flows with velocity vector $\mathbf{v} = u\,\hat{\imath} + v\,\hat{\jmath} + w\,\hat{k}$. Then $\varrho\,\mathbf{v}$ is the rate of mass flow. The requirement that

$$\nabla \bullet (\varrho\,\mathbf{v}) \equiv 0 \tag{6.66}$$

is a physical conservation law that puts a constraint on the velocity vector field. If ϱ is constant, then (6.66) would require

$$0 = \varrho(\nabla \bullet \mathbf{v}).$$

The fluid is called **incompressible** if $\boxed{\nabla \bullet \mathbf{v} \equiv 0}$.

Continuing with the situation of fluid flow,

$$\boldsymbol{\omega} \triangleq \nabla \times \mathbf{v}$$

is called the **vorticity** of the fluid. We say the fluid is **irrotational** if

$$\nabla \times \mathbf{v} \equiv \mathbf{0}.$$

The **Laplacian operator**, denoted by ∇^2 or Δ, is defined by

$$\nabla^2 f \triangleq (\nabla \bullet \nabla)[f] = \nabla \bullet (\nabla f) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}$$

for a scalar function f. If $\mathbf{F} = F_x\,\hat{\boldsymbol{i}} + F_y\,\hat{\boldsymbol{j}} + F_z\,\hat{\boldsymbol{k}}$ is a vector field then we define the Laplacian on it component by component, that is,

$$\nabla^2[F_x\,\hat{\boldsymbol{i}} + F_y\,\hat{\boldsymbol{j}} + F_z\,\hat{\boldsymbol{k}}] \triangleq (\nabla^2 F_x)\,\hat{\boldsymbol{i}} + (\nabla^2 F_y)\,\hat{\boldsymbol{j}} + (\nabla^2 F_z)\,\hat{\boldsymbol{k}}.$$

6.7.1 Properties of the Operators *grad*, div, and *curl*

Theorem 6.6

For any scalar functions f and g, vector fields $\mathbf{F}$ and $\mathbf{G}$ for which the derivatives exist, and constant scalar α,

(1) $\nabla(f+g) = \nabla f + \nabla g$, $\nabla \bullet (\mathbf{F}+\mathbf{G}) = \nabla \bullet \mathbf{F} + \nabla \bullet \mathbf{G}$, $\nabla \times (\mathbf{F}+\mathbf{G}) = \nabla \times \mathbf{F} + \nabla \times \mathbf{G}$.
(2) $\nabla(\alpha f) = \alpha \nabla f$, $\nabla \bullet (\alpha \mathbf{F}) = \alpha \nabla \bullet \mathbf{F}$, $\nabla \times (\alpha \mathbf{F}) = \alpha \nabla \times \mathbf{F}$.
(3) $\nabla(fg) = g\nabla f + f\nabla g$.
(4) $\nabla \bullet (g\mathbf{F}) = g\,\nabla \bullet \mathbf{F} + (\nabla g) \bullet \mathbf{F}$.
(5) $\nabla \times (g\mathbf{F}) = g\,\nabla \times \mathbf{F} + (\nabla g) \times \mathbf{F}$.
(6) $\nabla \bullet (\mathbf{F} \times \mathbf{G}) = \mathbf{G} \bullet (\nabla \times \mathbf{F}) - \mathbf{F} \bullet (\nabla \times \mathbf{G})$.
(7) $\nabla \times (\mathbf{F} \times \mathbf{G}) = (\mathbf{G} \bullet \nabla)\mathbf{F} - (\mathbf{F} \bullet \nabla)\mathbf{G} - (\nabla \bullet \mathbf{F})\mathbf{G} + (\nabla \bullet \mathbf{G})\mathbf{F}$.
(8) $\nabla \bullet (\nabla \times \mathbf{F}) = 0$, if $\nabla \times \mathbf{F}$ is continuously differentiable.
(9) $\nabla \times (\nabla f) = \mathbf{0}$, if ∇f is continuously differentiable.
(10) $\nabla \times (\nabla \times \mathbf{F}) = \nabla(\nabla \bullet \mathbf{F}) - \nabla^2 \mathbf{F}$.

The next result will turn out to be useful in studying partial differential equations.

Corollary 6.1

$$\boxed{(\nabla f) \bullet (\nabla g) = \nabla \bullet (g \nabla f) - g\nabla^2 f}\ .$$

Why? This follows from Theorem 6.6(4) with $\mathbf{F} \triangleq \nabla f$. □

Recall from Definition 6.4 that a vector field $\mathbf{F}$ has a *potential function* f if $\mathbf{F} = \nabla f$ continuously on an open region, in which case we say that $\mathbf{F}$ is *exact*.

Corollary 6.2

If $\mathbf{F}$ is exact and continuously differentiable, then $\boldsymbol{curl}(\mathbf{F}) = \mathbf{0}$, that is, $\mathbf{F}$ is irrotational.

6.7.2 Curvilinear Coordinates

Suppose there is a change of variables from coordinates (u, v, w) to (x, y, z), that is,

$$x = x(u, v, w), \quad y = y(u, v, w), \quad z = z(u, v, w),$$

and $\mathbf{r} = \mathbf{r}(u, v, w) = x(u, v, w)\,\hat{\boldsymbol{\imath}} + y(u, v, w)\,\hat{\boldsymbol{\jmath}} + z(u, v, w)\,\hat{\boldsymbol{k}}$ is continuously differentiable. We already know from Section 6.6 that under this change of variables, volume is multiplied by

$$\left|\det\left(\frac{\partial(x, y, z)}{\partial(u, v, w)}\right)\right|.$$

Let's find the differential operators ***grad***, *div*, ***curl***, and *Laplacian* in curvilinear coordinates.

Definition 6.6

Suppose $\mathcal{C}_1 :\ \mathbf{r}_1 = \mathbf{r}_1(t)$ and $\mathcal{C}_2 :\ \mathbf{r}_2 = \mathbf{r}_2(\tau)$ are two curves that intersect at a point $\mathbf{r}_0 = x_0\,\hat{\boldsymbol{\imath}} + y_0\,\hat{\boldsymbol{\jmath}} + z_0\,\hat{\boldsymbol{k}}$. We say the two curves **intersect orthogonally** at $\mathbf{r}_0$ if the corresponding tangent vectors, that is, $\frac{d\mathbf{r}_1}{dt}$ and $\frac{d\mathbf{r}_2}{d\tau}$, are nonzero and orthogonal there.

The two curves intersect at $\mathbf{r}_0$ if $\mathbf{r}_1(t_0) = \mathbf{r}_0 = \mathbf{r}_2(\tau_0)$ for some t_0, τ_0, but nothing is said about comparing the values of t_0 and τ_0.

Definition 6.7

(u, v, w) are **curvilinear coordinates** if at all (u_0, v_0, w_0) the three curves

$$\mathcal{C}_1 :\ \mathbf{r}_1(u) = \mathbf{r}(u, v_0, w_0), \quad \mathcal{C}_2 :\ \mathbf{r}_2(v) = \mathbf{r}(u_0, v, w_0), \quad \mathcal{C}_3 :\ \mathbf{r}_3(w) = \mathbf{r}(u_0, v_0, w)$$

are pairwise orthogonal there, that is, the set of three tangent vectors

$$\{\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3\}$$

is an orthogonal basis for $\mathbb{R}^3$, where

$$\mathbf{T}_1 \triangleq \frac{\partial \mathbf{r}}{\partial u}(u_0, v_0, w_0), \quad \mathbf{T}_2 \triangleq \frac{\partial \mathbf{r}}{\partial v}(u_0, v_0, w_0), \quad \mathbf{T}_3 \triangleq \frac{\partial \mathbf{r}}{\partial w}(u_0, v_0, w_0).$$

Example 6.33

(a) Explain why cylindrical coordinates are curvilinear coordinates except at the origin, (b) find the gradient operator in cylindrical coordinates, and (c) find the Laplacian operator in cylindrical coordinates.

Method:

(a) $x = r\cos\theta$, $y = r\sin\theta$, $z = z$ express (x, y, z) in terms of cylindrical coordinates (r, θ, z), that is,

$$\mathbf{r} = \mathbf{r}(r, \theta, z) = (r\cos\theta)\hat{\boldsymbol{\imath}} + (r\sin\theta)\hat{\boldsymbol{\jmath}} + z\,\hat{\mathbf{k}}.$$

The three curves are

$$\mathcal{C}_1 : \ \mathbf{r}_1(r) = \mathbf{r}(r, \theta_0, z_0) = r\cos\theta_0\,\hat{\boldsymbol{\imath}} + r\sin\theta_0\,\hat{\boldsymbol{\jmath}} + z_0\,\hat{\mathbf{k}},$$

$$\mathcal{C}_2 : \ \mathbf{r}_2(\theta) = \mathbf{r}(r_0, \theta, z_0) = r_0\cos\theta\,\hat{\boldsymbol{\imath}} + r_0\sin\theta\,\hat{\boldsymbol{\jmath}} + z_0\,\hat{\mathbf{k}},$$

$$\mathcal{C}_3 : \ \mathbf{r}_3(z) = \mathbf{r}(r_0, \theta_0, z) = r_0\cos\theta_0\,\hat{\boldsymbol{\imath}} + r_0\sin\theta_0\,\hat{\boldsymbol{\jmath}} + z\,\hat{\mathbf{k}},$$

and the corresponding three tangent vectors are, respectively,

$$\mathbf{T}_1 = \frac{d\mathbf{r}_1}{dr}(r_0) = \cos\theta_0\,\hat{\boldsymbol{\imath}} + \sin\theta_0\,\hat{\boldsymbol{\jmath}} = \mathbf{e}_{r_0},$$

$$\mathbf{T}_2 = \frac{d\mathbf{r}_2}{d\theta}(\theta_0) = -r_0\sin\theta_0\,\hat{\boldsymbol{\imath}} + r_0\cos\theta_0\,\hat{\boldsymbol{\jmath}} = r_0\mathbf{e}_{\theta_0},$$

and

$$\mathbf{T}_3 = \frac{d\mathbf{r}_3}{dz}(z_0) = \hat{\mathbf{k}} = \mathbf{e}_z.$$

We already know from Section 6.2 that $\{\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \mathbf{e}_z\}$ is an o.n. basis for $\mathbb{R}^3$; hence $\{\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3\}$ is an orthogonal basis for $\mathbb{R}^3$ *except* at $r_0 = 0$, as we were asked to explain.

(b) By (6.42) in Section 6.4, the multivariable chain rule, if $f = f(x, y, z) = f(r\cos\theta, r\sin\theta, z)$, then

$$\frac{\partial f}{\partial r} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial r} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial r} = \cos\theta\frac{\partial f}{\partial x} + \sin\theta\frac{\partial f}{\partial y},$$

$$\frac{\partial f}{\partial \theta} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial \theta} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial \theta} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial \theta} = -r\sin\theta\frac{\partial f}{\partial x} + r\cos\theta\frac{\partial f}{\partial y},$$

$$\frac{\partial f}{\partial z} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial z} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial z} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial z} = \frac{\partial f}{\partial z}.$$

These three results can be rewritten as one matrix-vector equation:

$$\begin{bmatrix} \frac{\partial f}{\partial r} \\ \frac{\partial f}{\partial \theta} \\ \frac{\partial f}{\partial z} \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -r\sin\theta & r\cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial z} \end{bmatrix}.$$

It follows that

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial z} \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -r\sin\theta & r\cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial f}{\partial r} \\ \frac{\partial f}{\partial \theta} \\ \frac{\partial f}{\partial z} \end{bmatrix} \tag{6.67}$$

that is,

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \\ \frac{\partial f}{\partial y} \\ \\ \frac{\partial f}{\partial z} \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{e}}_r^T \\ \\ r\hat{\mathbf{e}}_\theta^T \\ \\ \hat{\mathbf{e}}_z^T \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial f}{\partial r} \\ \\ \frac{\partial f}{\partial \theta} \\ \\ \frac{\partial f}{\partial z} \end{bmatrix} \tag{6.68}$$

Because $\{\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z\}$ is an o.n. basis for $\mathbb{R}^3$, the matrix

$$\begin{bmatrix} \hat{\mathbf{e}}_r^T \\ \\ \hat{\mathbf{e}}_\theta^T \\ \\ \hat{\mathbf{e}}_z^T \end{bmatrix}^{-1} = \left[\hat{\mathbf{e}}_r \;\vdots\; \hat{\mathbf{e}}_\theta \;\vdots\; \hat{\mathbf{e}}_z\right] \text{ exists; hence, } \begin{bmatrix} \hat{\mathbf{e}}_r^T \\ \\ r\hat{\mathbf{e}}_\theta^T \\ \\ \hat{\mathbf{e}}_z^T \end{bmatrix}^{-1} = \left[\hat{\mathbf{e}}_r \;\vdots\; \frac{1}{r}\,\hat{\mathbf{e}}_\theta \;\vdots\; \hat{\mathbf{e}}_z\right] \text{ exists.}$$

So,

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \\ \frac{\partial f}{\partial y} \\ \\ \frac{\partial f}{\partial z} \end{bmatrix} = \left[\hat{\mathbf{e}}_r \;\vdots\; \frac{1}{r}\,\hat{\mathbf{e}}_\theta \;\vdots\; \hat{\mathbf{e}}_z\right] \begin{bmatrix} \frac{\partial f}{\partial r} \\ \\ \frac{\partial f}{\partial \theta} \\ \\ \frac{\partial f}{\partial z} \end{bmatrix} = \frac{\partial f}{\partial r}\,\hat{\mathbf{e}}_r + \frac{1}{r}\,\frac{\partial f}{\partial \theta}\,\hat{\mathbf{e}}_\theta + \frac{\partial f}{\partial z}\,\hat{\mathbf{e}}_z. \tag{6.69}$$

Sometimes, we may write (6.69) as

$$\nabla = \hat{\mathbf{e}}_r \frac{\partial}{\partial r} + \hat{\mathbf{e}}_\theta \left(\frac{1}{r}\,\frac{\partial}{\partial \theta}\right) + \hat{\mathbf{e}}_z \frac{\partial}{\partial z}. \tag{6.70}$$

(c) In rectangular coordinates, the Laplacian operator is

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.$$

Equation (6.69) can be rewritten as

$$\begin{bmatrix} \frac{\partial f}{\partial x} \\ \\ \frac{\partial f}{\partial y} \\ \\ \frac{\partial f}{\partial z} \end{bmatrix} = \left[\hat{\mathbf{e}}_r \;\vdots\; \frac{1}{r}\,\hat{\mathbf{e}}_\theta \;\vdots\; \hat{\mathbf{e}}_z\right] \begin{bmatrix} \frac{\partial f}{\partial r} \\ \\ \frac{\partial f}{\partial \theta} \\ \\ \frac{\partial f}{\partial z} \end{bmatrix} = \begin{bmatrix} \cos\theta & -\frac{1}{r}\sin\theta & 0 \\ \\ \sin\theta & \frac{1}{r}\cos\theta & 0 \\ \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{\partial f}{\partial r} \\ \\ \frac{\partial f}{\partial \theta} \\ \\ \frac{\partial f}{\partial z} \end{bmatrix},$$

so

$$\frac{\partial f}{\partial x} = \cos\theta\,\frac{\partial f}{\partial r} - \frac{\sin\theta}{r}\,\frac{\partial f}{\partial \theta}, \frac{\partial f}{\partial y} = \sin\theta\,\frac{\partial f}{\partial r} + \frac{\cos\theta}{r}\,\frac{\partial f}{\partial \theta}, \quad \frac{\partial f}{\partial z} = \frac{\partial f}{\partial z}.$$

It follows that

$$\frac{\partial}{\partial x} = \cos\theta\,\frac{\partial}{\partial r} - \frac{\sin\theta}{r}\,\frac{\partial}{\partial \theta} \quad \text{and} \quad \frac{\partial}{\partial y} = \sin\theta\,\frac{\partial}{\partial r} + \frac{\cos\theta}{r}\,\frac{\partial}{\partial \theta}. \tag{6.71}$$

So,

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x}\left[\frac{\partial f}{\partial x}\right] = \left(\cos\theta\,\frac{\partial}{\partial r} - \frac{\sin\theta}{r}\,\frac{\partial}{\partial \theta}\right)\left[\cos\theta\,\frac{\partial f}{\partial r} - \frac{\sin\theta}{r}\,\frac{\partial f}{\partial \theta}\right]$$

$$= \cos\theta\,\frac{\partial}{\partial r}\left[\cos\theta\,\frac{\partial f}{\partial r} - \frac{\sin\theta}{r}\,\frac{\partial f}{\partial \theta}\right] - \frac{\sin\theta}{r}\,\frac{\partial}{\partial \theta}\left[\cos\theta\,\frac{\partial f}{\partial r} - \frac{\sin\theta}{r}\,\frac{\partial f}{\partial \theta}\right]$$

$$= \cos\theta\left(\cos\theta\,\frac{\partial^2 f}{\partial r^2} + \left(\frac{\sin\theta}{r^2}\,\frac{\partial f}{\partial \theta} - \frac{\sin\theta}{r}\,\frac{\partial^2 f}{\partial r \partial \theta}\right)\right)$$

$$- \frac{\sin\theta}{r}\left(\left(-\sin\theta\,\frac{\partial f}{\partial r} + \cos\theta\,\frac{\partial^2 f}{\partial \theta \partial r}\right) + \left(-\frac{\cos\theta}{r}\,\frac{\partial f}{\partial \theta} - \frac{\sin\theta}{r}\,\frac{\partial^2 f}{\partial \theta^2}\right)\right)$$

$$= \cos^2\theta\,\frac{\partial^2 f}{\partial r^2} - \frac{2\sin\theta\,\cos\theta}{r}\,\frac{\partial^2 f}{\partial r \partial \theta} + \frac{2\sin\theta\,\cos\theta}{r^2}\,\frac{\partial f}{\partial \theta}$$

$$+ \frac{\sin^2\theta}{r^2}\,\frac{\partial^2 f}{\partial \theta^2} + \frac{\sin^2\theta}{r}\,\frac{\partial f}{\partial r}.$$

after using Clairaut's theorem (Theorem 6.2) in Section 6.4 to get $\frac{\partial^2 f}{\partial r \partial \theta} = \frac{\partial^2 f}{\partial \theta \partial r}$. Similarly,

$$\frac{\partial^2 f}{\partial y^2} = \frac{\partial}{\partial y}\left[\frac{\partial f}{\partial y}\right] = \left(\sin\theta\,\frac{\partial}{\partial r} + \frac{\cos\theta}{r}\,\frac{\partial}{\partial \theta}\right)\left[\sin\theta\,\frac{\partial f}{\partial r} + \frac{\cos\theta}{r}\,\frac{\partial f}{\partial \theta}\right]$$

$$= \sin\theta\left(\sin\theta\,\frac{\partial^2 f}{\partial r^2} + \left(-\frac{\cos\theta}{r^2}\,\frac{\partial f}{\partial \theta} + \frac{\cos\theta}{r}\,\frac{\partial^2 f}{\partial r \partial \theta}\right)\right)$$

$$+ \frac{\cos\theta}{r}\left(\left(\cos\theta\,\frac{\partial f}{\partial r} + \sin\theta\,\frac{\partial^2 f}{\partial \theta \partial r}\right) + \left(-\frac{\sin\theta}{r}\,\frac{\partial f}{\partial \theta} + \frac{\cos\theta}{r}\,\frac{\partial^2 f}{\partial \theta^2}\right)\right)$$

$$= \sin^2\theta\,\frac{\partial^2 f}{\partial r^2} + \frac{2\sin\theta\,\cos\theta}{r}\,\frac{\partial^2 f}{\partial r \partial \theta} - \frac{2\sin\theta\,\cos\theta}{r^2}\,\frac{\partial f}{\partial \theta}$$

$$+ \frac{\cos^2\theta}{r^2}\,\frac{\partial^2 f}{\partial \theta^2} + \frac{\cos^2\theta}{r}\,\frac{\partial f}{\partial r}.$$

So,

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}$$

$$= \cos^2\theta\,\frac{\partial^2 f}{\partial r^2} - \cancel{\frac{2\sin\theta\,\cos\theta}{r}\,\frac{\partial^2 f}{\partial r \partial \theta}} + \cancel{\frac{2\sin\theta\,\cos\theta}{r^2}\,\frac{\partial f}{\partial \theta}} + \frac{\sin^2\theta}{r^2}\,\frac{\partial^2 f}{\partial \theta^2} + \frac{\sin^2\theta}{r}\,\frac{\partial f}{\partial r}$$

$$+ \sin^2\theta\,\frac{\partial^2 f}{\partial r^2} + \cancel{\frac{2\sin\theta\,\cos\theta}{r}\,\frac{\partial^2 f}{\partial r \partial \theta}} - \cancel{\frac{2\sin\theta\,\cos\theta}{r^2}\,\frac{\partial f}{\partial \theta}} + \frac{\cos^2\theta}{r^2}\,\frac{\partial^2 f}{\partial \theta^2}$$

$$+ \frac{\cos^2\theta}{r}\,\frac{\partial f}{\partial r} + \frac{\partial^2 f}{\partial z^2}.$$

So, the Laplacian in cylindrical coordinates is

$$\nabla^2 f = \frac{\partial^2 f}{\partial r^2} + \frac{1}{r}\frac{\partial f}{\partial r} + \frac{1}{r^2}\frac{\partial^2 f}{\partial \theta^2} + \frac{\partial^2 f}{\partial z^2}, \tag{6.72}$$

which can be written as

$$\nabla^2 f = \frac{1}{r}\frac{\partial}{\partial r}\left[r\frac{\partial f}{\partial r}\right] + \frac{1}{r^2}\frac{\partial^2 f}{\partial \theta^2} + \frac{\partial^2 f}{\partial z^2}. \tag{6.73}$$

Sometimes, we may write (6.72) as

$$\nabla^2 = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2} + \frac{\partial^2}{\partial z^2}\cdot \bigcirc \tag{6.74}$$

6.7.3 Differential Operators in Curvilinear Coordinates

Suppose (u, v, w) are general curvilinear coordinates. We will derive the gradient operator in those coordinates, state the results for the other differential operators, and give sketches of their derivations. After that, we will state the results in cylindrical and spherical coordinates, which are commonly used in science and engineering.

Recall that (u, v, w) are curvilinear coordinates for position vectors $\mathbf{r}(u, v, w)$ if at every point $\mathbf{r}(u_0, v_0, w_0)$ the three tangent vectors,

$$\mathbf{T}_u = \frac{\partial \mathbf{r}}{\partial u}(u_0, v_0, w_0), \quad \mathbf{T}_v = \frac{\partial \mathbf{r}}{\partial v}(u_0, v_0, w_0), \quad \mathbf{T}_w = \frac{\partial \mathbf{r}}{\partial w}(u_0, v_0, w_0),$$

are nonzero and give an orthogonal basis for $\mathbb{R}^3$. Define the **scale factors** or "length factors"

$$h_u = ||\mathbf{T}_u|| = \left|\left|\frac{\partial \mathbf{r}}{\partial u}\right|\right|, \quad h_v = ||\mathbf{T}_v|| = \left|\left|\frac{\partial \mathbf{r}}{\partial v}\right|\right|, \quad h_w = ||\mathbf{T}_w|| = \left|\left|\frac{\partial \mathbf{r}}{\partial w}\right|\right|.$$

Then

$$\hat{\mathbf{e}}_u \triangleq \frac{1}{h_u}\mathbf{T}_u, \quad \hat{\mathbf{e}}_v \triangleq \frac{1}{h_v}\mathbf{T}_v, \quad \hat{\mathbf{e}}_w \triangleq \frac{1}{h_w}\mathbf{T}_w, \tag{6.75}$$

are unit vectors, so

$$\{\hat{\mathbf{e}}_u, \hat{\mathbf{e}}_v, \hat{\mathbf{e}}_w\} \tag{6.76}$$

is an o.n. basis for $\mathbb{R}^3$.

As we saw in Section 6.6, an element of volume is given by

$$\begin{aligned} dV &= \left|\det\left(\frac{\partial(x,y,z)}{\partial(u,v,w)}\right) du\, dv\, dw\right| = \left|\det\left[\frac{\partial \mathbf{r}}{\partial u}du \;\vdots\; \frac{\partial \mathbf{r}}{\partial v}dv \;\vdots\; \frac{\partial \mathbf{r}}{\partial w}dw\right]\right| \\ &= \left|\det\left[h_u du\hat{\mathbf{e}}_u \;\vdots\; h_v dv\hat{\mathbf{e}}_v \;\vdots\; h_w dw\hat{\mathbf{e}}_w\right]\right| \\ &= \left|\det(\, diag(h_u du, h_v dv, h_w dw))\right| \cdot \left|\det\left[\hat{\mathbf{e}}_u \;\vdots\; \hat{\mathbf{e}}_v \;\vdots\; \hat{\mathbf{e}}_w\right]\right|. \end{aligned}$$

So,

$$dV = h_u\, h_v\, h_w\, du\, dv\, dw\, |\pm 1| = h_u\, h_v\, h_w\, du\, dv\, dw, \tag{6.77}$$

because the determinant of a real, orthogonal matrix, such as $\left[\hat{\mathbf{e}}_u \,\vdots\, \hat{\mathbf{e}}_v \,\vdots\, \hat{\mathbf{e}}_w\right]$, is ± 1, by Problem 2.4.4.16.

We see from (6.77) why we refer to h_u, h_v, h_w as length factors.

By (6.42) in Section 6.4, the multivariable chain rule, we have

$$\begin{aligned}
\frac{\partial f}{\partial u} &= \frac{\partial f}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial u} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial u} \\
\frac{\partial f}{\partial v} &= \frac{\partial f}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial v} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial v} \\
\frac{\partial f}{\partial w} &= \frac{\partial f}{\partial x}\frac{\partial x}{\partial w} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial w} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial w}.
\end{aligned}$$

These three results can be rewritten as one matrix-vector equation:

$$\begin{bmatrix} \frac{\partial f}{\partial u} \\ \frac{\partial f}{\partial v} \\ \frac{\partial f}{\partial w} \end{bmatrix} = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} & \frac{\partial z}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} & \frac{\partial z}{\partial v} \\ \frac{\partial x}{\partial w} & \frac{\partial y}{\partial w} & \frac{\partial z}{\partial w} \end{bmatrix} \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial z} \end{bmatrix}.$$

It follows that

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial z} \end{bmatrix} = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} & \frac{\partial z}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} & \frac{\partial z}{\partial v} \\ \frac{\partial x}{\partial w} & \frac{\partial y}{\partial w} & \frac{\partial z}{\partial w} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial f}{\partial u} \\ \frac{\partial f}{\partial v} \\ \frac{\partial f}{\partial w} \end{bmatrix}, \tag{6.78}$$

that is,

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial z} \end{bmatrix} = \begin{bmatrix} h_u \hat{\mathbf{e}}_u^T \\ h_v \hat{\mathbf{e}}_v^T \\ h_w \hat{\mathbf{e}}_w^T \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial f}{\partial u} \\ \frac{\partial f}{\partial v} \\ \frac{\partial f}{\partial w} \end{bmatrix}. \tag{6.79}$$

Because $\{\hat{\mathbf{e}}_u, \hat{\mathbf{e}}_v, \hat{\mathbf{e}}_w\}$ is an o.n. basis for $\mathbb{R}^3$,

$$\begin{bmatrix} h_u \hat{\mathbf{e}}_u^T \\ h_v \hat{\mathbf{e}}_v^T \\ h_w \hat{\mathbf{e}}_w^T \end{bmatrix}^{-1} = \left[\frac{1}{h_u} \hat{\mathbf{e}}_u \;\vdots\; \frac{1}{h_v} \hat{\mathbf{e}}_v \;\vdots\; \frac{1}{h_w} \hat{\mathbf{e}}_w \right]$$

exists. Similarly to (6.68), and using Lemma 1.2 in Section 1.2, that is, (1.19) in Section 1.2, we get that

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial z} \end{bmatrix} = \left[\frac{1}{h_u} \hat{\mathbf{e}}_u \;\vdots\; \frac{1}{h_v} \hat{\mathbf{e}}_v \;\vdots\; \frac{1}{h_w} \hat{\mathbf{e}}_w \right] \begin{bmatrix} \frac{\partial f}{\partial u} \\ \frac{\partial f}{\partial v} \\ \frac{\partial f}{\partial w} \end{bmatrix}, \tag{6.80}$$

that is,

$$\nabla f = \frac{1}{h_u} \frac{\partial f}{\partial u} \hat{\mathbf{e}}_u + \frac{1}{h_v} \frac{\partial f}{\partial v} \hat{\mathbf{e}}_v + \frac{1}{h_w} \frac{\partial f}{\partial w} \hat{\mathbf{e}}_w$$

gives the gradient operator in (u, v, w) coordinates. Sometimes, we may write (6.80) as

$$\nabla = \frac{1}{h_u} \hat{\mathbf{e}}_u \frac{\partial}{\partial u} + \frac{1}{h_v} \hat{\mathbf{e}}_v \frac{\partial}{\partial v} + \frac{1}{h_w} \hat{\mathbf{e}}_w \frac{\partial}{\partial w}. \tag{6.81}$$

So it was not a coincidence that the vectors $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z$ appeared in our results for cylindrical coordinates in Example 6.33.

It is common to also assume that

$$\{\hat{\mathbf{e}}_u, \hat{\mathbf{e}}_v, \hat{\mathbf{e}}_w\}$$

is a *right-handed* coordinate frame, that is,

$$\hat{\mathbf{e}}_u \times \hat{\mathbf{e}}_v = \hat{\mathbf{e}}_w, \quad \hat{\mathbf{e}}_v \times \hat{\mathbf{e}}_w = \hat{\mathbf{e}}_u, \quad \hat{\mathbf{e}}_w \times \hat{\mathbf{e}}_u = \hat{\mathbf{e}}_v.$$

In fact, this assumption plays a role in deriving the complete results.

Theorem 6.7

Suppose (u, v, w) are curvilinear coordinates that give a right-handed coordinate frame $\{\hat{\mathbf{e}}_u, \hat{\mathbf{e}}_v, \hat{\mathbf{e}}_w\}$ for $\mathbb{R}^3$ defined by (6.75). By (6.80), ***grad*** is given by

$$\nabla f = \frac{1}{h_u} \frac{\partial f}{\partial u} \hat{\mathbf{e}}_u + \frac{1}{h_v} \frac{\partial f}{\partial v} \hat{\mathbf{e}}_v + \frac{1}{h_w} \frac{\partial f}{\partial w} \hat{\mathbf{e}}_w,$$

and the *div*, ***curl***, and Laplacian operators are given by

$$\boldsymbol{\nabla} \bullet \left(F_u \hat{\mathbf{e}}_u + F_v \hat{\mathbf{e}}_v + F_w \hat{\mathbf{e}}_w\right) = \frac{1}{h_u h_v h_w}\left(\frac{\partial}{\partial u}\left[h_v h_w F_u\right] + \frac{\partial}{\partial v}\left[h_w h_u F_v\right] + \frac{\partial}{\partial w}\left[h_u h_v F_w\right]\right), \tag{6.82}$$

$$\begin{aligned}\boldsymbol{\nabla} \times \left(F_u \hat{\mathbf{e}}_u + F_v \hat{\mathbf{e}}_v + F_w \hat{\mathbf{e}}_w\right) &= \frac{1}{h_v h_w}\left(\frac{\partial}{\partial v}\left[h_w F_w\right] - \frac{\partial}{\partial w}\left[h_v F_v\right]\right)\hat{\mathbf{e}}_u \\ &+ \frac{1}{h_w h_u}\left(\frac{\partial}{\partial w}\left[h_u F_u\right] - \frac{\partial}{\partial u}\left[h_w F_w\right]\right)\hat{\mathbf{e}}_v + \frac{1}{h_u h_v}\left(\frac{\partial}{\partial u}\left[h_v F_v\right] - \frac{\partial}{\partial v}\left[h_u F_u\right]\right)\hat{\mathbf{e}}_w,\end{aligned} \tag{6.83}$$

and

$$\nabla^2 f = \frac{1}{h_u h_v h_w}\left(\frac{\partial}{\partial u}\left[\frac{h_v h_w}{h_u}\frac{\partial f}{\partial u}\right] + \frac{\partial}{\partial v}\left[\frac{h_w h_u}{h_v}\frac{\partial f}{\partial v}\right] + \frac{\partial}{\partial w}\left[\frac{h_u h_v}{h_w}\frac{\partial f}{\partial w}\right]\right). \tag{6.84}$$

Why? We've already derived (6.80) by relatively straightforward calculations. But deriving the other results uses some tricks.

Using (6.80), for the function $f(u, v, w) = u$, we calculate

$$\boldsymbol{\nabla} u = \frac{1}{h_u}\hat{\mathbf{e}}_u \frac{\partial u}{\partial u} + \frac{1}{h_v}\hat{\mathbf{e}}_v \frac{\partial u}{\partial v} + \frac{1}{h_w}\hat{\mathbf{e}}_w \frac{\partial u}{\partial w} = 1 \cdot \frac{1}{h_u}\hat{\mathbf{e}}_u + 0 \cdot \frac{1}{h_v}\hat{\mathbf{e}}_v + 0 \cdot \frac{1}{h_w}\hat{\mathbf{e}}_w = \frac{1}{h_u}\hat{\mathbf{e}}_u,$$

so it follows from Theorem 6.6(9) that

$$\mathbf{0} = \boldsymbol{\nabla} \times (\boldsymbol{\nabla} u) = \boldsymbol{\nabla} \times \left(\frac{1}{h_u}\hat{\mathbf{e}}_u\right).$$

Similarly,

$$\boldsymbol{\nabla} \times \left(\frac{1}{h_v}\hat{\mathbf{e}}_v\right) = \boldsymbol{\nabla} \times \left(\frac{1}{h_w}\hat{\mathbf{e}}_w\right) = \mathbf{0}. \tag{6.85}$$

Using (6.85) and Theorem 6.6(6), we get

$$\boldsymbol{\nabla} \bullet \left(\frac{1}{h_v}\hat{\mathbf{e}}_v \times \frac{1}{h_w}\hat{\mathbf{e}}_w\right) = 0.$$

But $\hat{\mathbf{e}}_v \times \hat{\mathbf{e}}_w = \hat{\mathbf{e}}_u$, so (6.85) implies $\boldsymbol{\nabla} \bullet \left(\frac{1}{h_v h_w}\hat{\mathbf{e}}_u\right) = 0$, and similarly, $\boldsymbol{\nabla} \bullet \left(\frac{1}{h_w h_u}\hat{\mathbf{e}}_v\right) = 0$, and $\boldsymbol{\nabla} \times \left(\frac{1}{h_u h_v}\hat{\mathbf{e}}_w\right) = 0$. The rest of the explanations will be in the homework problems.

6.7.4 Summary of Operators in Cylindrical Coordinates

Because $h_r = 1, h_\theta = r, h_z = 1$,

$$\nabla f = \frac{\partial f}{\partial r}\,\hat{\mathbf{e}}_r + \frac{1}{r}\,\frac{\partial f}{\partial \theta}\,\hat{\mathbf{e}}_\theta + \frac{\partial f}{\partial z}\,\hat{\mathbf{e}}_z,$$

$$\nabla \bullet \left(F_r\,\hat{\mathbf{e}}_r + F_\theta\,\hat{\mathbf{e}}_\theta + F_z\,\hat{\mathbf{e}}_z\right) = \frac{1}{r}\frac{\partial}{\partial r}\Big[\,rF_r\,\Big] + \frac{1}{r}\frac{\partial}{\partial \theta}\Big[\,F_\theta\,\Big] + \frac{\partial}{\partial z}\Big[\,F_z\,\Big],$$

$$\nabla \times \left(F_r\,\hat{\mathbf{e}}_r + F_\theta\,\hat{\mathbf{e}}_\theta + F_z\,\hat{\mathbf{e}}_z\right)$$

$$= \left(\frac{1}{r}\frac{\partial}{\partial \theta}\,[\,F_z\,] - \frac{\partial}{\partial z}\,[\,F_\theta\,]\right)\hat{\mathbf{e}}_r + \left(\frac{\partial}{\partial z}\,[\,F_r\,] - \frac{\partial}{\partial r}\,[\,F_z\,]\right)\hat{\mathbf{e}}_\theta + \left(\frac{1}{r}\frac{\partial}{\partial r}\,[\,rF_\theta\,] - \frac{1}{r}\frac{\partial}{\partial \theta}\,[\,F_r\,]\right)\hat{\mathbf{e}}_z,$$

and

$$\nabla^2 f = \frac{1}{r}\frac{\partial}{\partial r}\Big[\,r\,\frac{\partial f}{\partial r}\,\Big] + \frac{1}{r^2}\,\frac{\partial^2 f}{\partial \theta^2} + \frac{\partial^2 f}{\partial z^2}.$$

6.7.5 Summary of Operators in Spherical Coordinates

Because $h_\rho = 1, h_\phi = \rho, h_\theta = \rho \sin\phi$,

$$\nabla f = \frac{\partial f}{\partial \rho}\,\hat{\mathbf{e}}_\rho + \frac{1}{\rho}\,\frac{\partial f}{\partial \phi}\,\hat{\mathbf{e}}_\phi + \frac{1}{\rho\,\sin\phi}\,\frac{\partial f}{\partial \theta}\,\hat{\mathbf{e}}_\theta,$$

$$\nabla \bullet \left(F_\rho\,\hat{\mathbf{e}}_\rho + F_\phi\,\hat{\mathbf{e}}_\phi + F_\theta\,\hat{\mathbf{e}}_\theta\right) = \frac{1}{\rho^2}\,\frac{\partial}{\partial \rho}\Big[\,\rho^2 F_\rho\,\Big] + \frac{1}{\rho\sin\phi}\,\frac{\partial}{\partial \phi}\Big[\,\sin\phi\,F_\phi\,\Big] + \frac{1}{\rho\sin\phi}\,\frac{\partial}{\partial \theta}\Big[\,F_\theta\Big],$$

$$\nabla \times \left(F_\rho\,\hat{\mathbf{e}}_\rho + F_\phi\,\hat{\mathbf{e}}_\phi + F_\theta\,\hat{\mathbf{e}}_\theta\right) = \left(\frac{1}{\rho\sin\phi}\,\frac{\partial}{\partial \phi}\Big[\,\sin\phi\,F_\theta\Big] - \frac{1}{\rho\sin\phi}\,\frac{\partial}{\partial \theta}\Big[\,F_\phi\Big]\right)\hat{\mathbf{e}}_\rho$$

$$+ \left(\frac{1}{\rho\sin\phi}\,\frac{\partial}{\partial \theta}\Big[\,F_\rho\Big] - \frac{1}{\rho}\,\frac{\partial}{\partial \rho}\Big[\,\rho F_\theta\Big]\right)\hat{\mathbf{e}}_\phi + \left(\frac{1}{\rho}\,\frac{\partial}{\partial \rho}\Big[\,\rho F_\phi\Big] - \frac{1}{\rho}\,\frac{\partial}{\partial \phi}\Big[\,F_\rho\Big]\right)\hat{\mathbf{e}}_\theta,$$

and

$$\nabla^2 f = \frac{1}{\rho^2}\,\frac{\partial}{\partial \rho}\Big[\,\rho^2\,\frac{\partial f}{\partial \rho}\,\Big] + \frac{1}{\rho^2\sin\phi}\,\frac{\partial}{\partial \phi}\Big[\,\sin\phi\,\frac{\partial f}{\partial \phi}\,\Big] + \frac{1}{\rho^2\sin^2\phi}\,\frac{\partial^2 f}{\partial \theta^2}.$$

Learn More About It

For complete derivations of the differential operators in curvilinear coordinates and further observations, see "Orthogonal curvilinear coordinates," by T. A. S. Jackson, *Math. Gazette*, (**50**), 1966, 28–30. Problem 6.7.6.30 is from *The Classical Electromagnetic Field*, by Leonard Eyges, Addison-Wesley, © 1972, p.18.

6.7.6 Problems

1. Evaluate ***curl*** $\left(x^2y\,\hat{\boldsymbol{\imath}} + x^2z\,\hat{\boldsymbol{\jmath}} + x^3\,\hat{\boldsymbol{k}}\right)$.

In problems 2–5, find the divergence and curl of the vector fields.

2. $\mathbf{F} \triangleq xy^2z\,\hat{\boldsymbol{\imath}} + 2x^2yz\,\hat{\boldsymbol{k}}$
3. $\mathbf{F} \triangleq \cos(xy^2)\hat{\boldsymbol{\imath}} + \sin(y+z^2)\,\hat{\boldsymbol{\jmath}} + \ln|x-z|\,\hat{\boldsymbol{k}}$
4. $\mathbf{F} \triangleq \left(3r^2 + \ln(r) + z\right)\hat{\mathbf{e}}_r + \left(r^3 + 2 - \cos\theta\right)\hat{\mathbf{e}}_\theta - \left(z^2 - 2r^2\right)\mathbf{e}_z$
5. $\mathbf{F} \triangleq \rho^{-2}(z\,\hat{\boldsymbol{\imath}} + x\,\hat{\boldsymbol{\jmath}} + y\,\hat{\boldsymbol{k}})$

In each of the problems 6 and 7, a "velocity" potential $\phi = \phi(r,\theta)$ is given for 2D irrotational fluid flow past a cylinder (Milne-Thomson, 1968). Find $\mathbf{v} = \nabla\phi$, the velocity of the fluid. Assume a and α are positive constants.

6. $\phi = U\left(r + \frac{a^2}{r}\right)\cos\theta$
7. $\phi = U\left(r + \frac{a^2}{r}\right)\cos(\theta - \alpha)$

   By the way, the potential in Problem 6.7.6.6 corresponds to "parallel flow" far from the cylinder because the velocity vector "at infinity" is given by $V_\infty = U\hat{\boldsymbol{\imath}}$. Similarly, the potential in Problem 6.7.6.7 corresponds has velocity at infinity given by $V_\infty = U(\cos\alpha\,\hat{\boldsymbol{\imath}} + \sin\alpha\,\hat{\boldsymbol{\jmath}})$. This explains why the potential in Problem 6.7.6.7 is referred to as "at an angle of attack α."

In problems 8–10, find the divergence and curl of the vector fields in both (a) rectangular coordinates, and (b) cylindrical coordinates. Table 6.1 in Section 6.2 may be useful. [Note: To find the curl, for example, in rectangular coordinates of a vector field given in cylindrical coordinates, we can find the curl in cylindrical coordinates and convert that into rectangular coordinates.]

8. $\mathbf{F} \triangleq \left(r - \frac{2}{r} + z\right)\hat{\mathbf{e}}_r + \left(\frac{1}{2}r^2 - 2r\right)\hat{\mathbf{e}}_\theta + r\,\mathbf{e}_z$
9. $\mathbf{F} \triangleq x(x^2+y^2)\hat{\boldsymbol{\imath}} + y(x^2+y^2-2z)\,\hat{\boldsymbol{\jmath}} + (z^2 - 3x^2 - 3y^2)\hat{\boldsymbol{k}}$
10. $\mathbf{F} \triangleq (2z + x^2 - y^2)\hat{\boldsymbol{\imath}} - (x^2+y^2)\,\hat{\boldsymbol{\jmath}} + z\,\hat{\boldsymbol{k}}$
11. For the vector field given as follows, find the divergence and curl in both (a) rectangular coordinates, and (b) spherical coordinates. Table 6.2 in Section 6.2 may be useful. [Note: To find the curl, for example, in spherical coordinates of a vector field given in rectangular coordinates, we can find the curl in rectangular coordinates and convert that into spherical coordinates.]

$$\mathbf{F} \triangleq x(x^2+y^2+z^2)\hat{\boldsymbol{\imath}} + (y(x^2+y^2+z^2) - 2(x^2+y^2))\,\hat{\boldsymbol{\jmath}} + \left(z^2 - z(x^2+y^2+z^2)\right)\hat{\boldsymbol{k}}.$$

12. Suppose $\mathbf{F} = \rho^{-4}\mathbf{r}$, where $\mathbf{r} = x\hat{\boldsymbol{\imath}} + y\,\hat{\boldsymbol{\jmath}} + z\hat{\boldsymbol{k}}$ and $\rho = ||\mathbf{r}||$. Find the divergence of $\mathbf{F}$ in as simple a form as possible.
13. If $\mathbf{F} = \rho^{-2}\mathbf{r}$, where $\rho = ||\mathbf{r}||$, find $\nabla \bullet \mathbf{F}$ in as simple a form as possible.

14. If $\mathbf{r} = x\hat{\boldsymbol{\imath}} + y\hat{\boldsymbol{\jmath}} + z\hat{\boldsymbol{k}}$ and $\mathbf{F} \triangleq ||\mathbf{r}||\mathbf{r}$, evaluate ***curl***$(\mathbf{F})$.
15. Suppose that a surface is described in cylindrical coordinates by the equation $z = 2r$. Find an equation of the plane tangent to the surface at the point $(x, y, z) = (-\sqrt{3}, 1, 4)$. [Hint: The equation of the tangent plane may be written in rectangular coordinates.]
16. Explain why the formulas of Theorem 6.6(4) and (5) are correct.
17. Explain why the formula of Theorem 6.6(6) is correct.
18. Explain why the formula of Theorem 6.6(7) is correct.
19. (a) Explain why the formula of Theorem 6.6(9) is correct.
    (b) Explain why $\boldsymbol{\nabla} \times (f\boldsymbol{\nabla} g) = (\boldsymbol{\nabla} f) \times (\boldsymbol{\nabla} g)$ is an identity.
20. Determine a scalar potential function for the vector field
$$\mathbf{F} \triangleq -\frac{\cos z}{r^2}\,\hat{\mathbf{e}}_r + \frac{\cos\theta}{r}\,\hat{\mathbf{e}}_\theta - \left(1 + \frac{\sin z}{r}\right)\mathbf{e}_z.$$
21. In the absence of currents, magnetostatics gives magnetic field
$$\mathbf{H} = -H_0\left(1 - 2\alpha\left(\frac{a}{\rho}\right)^3\right)\cos\phi\,\hat{\mathbf{e}}_\rho + H_0\left(1 + \alpha\left(\frac{a}{\rho}\right)^3\right)\sin\phi\,\mathbf{e}_\phi$$
outside of a metal sphere $\rho = a$, where H_0 is a constant. Find a potential function u satisfying $\mathbf{H} = -\boldsymbol{\nabla} u$. [By the way, $\alpha = \frac{\mu_0 - \mu}{2\mu_0 + \mu}$, where μ_0 is the magnetic permeability* of a vacuum and μ is the magnetic permeability of the metal.]
22. A 2D electric dipole can be idealized by two infinite line charges running parallel to the xy-plane. A good approximation of the electric field is
$$\mathbf{E} = -\frac{p_\ell}{2\pi\epsilon}\cdot\frac{1}{r^2}\left(\sin\theta\,\hat{\mathbf{e}}_r - \cos\theta\,\hat{\mathbf{e}}_\theta\right),$$
where p_ℓ is the magnitude of the dipole moment. Find a potential function u satisfying $\mathbf{E} = -\boldsymbol{\nabla} u$.

In problems 23–26, assume A is an unspecified constant. Determine a scalar potential function for the given vector field.

23. $\mathbf{F} \triangleq (2Ar + 1)(2\pi\theta - \theta^2)\hat{\mathbf{e}}_r + 2(Ar + 1)(\pi - \theta)\,\hat{\mathbf{e}}_\theta + \cos 3z\,\mathbf{e}_z$
24. $\mathbf{F} \triangleq \left(2r - \frac{\cos z}{r^2}\right)\hat{\mathbf{e}}_r + \frac{\cos\theta}{r}\,\hat{\mathbf{e}}_\theta - \left(1 + \frac{\sin z}{r}\right)\mathbf{e}_z$
25. $\mathbf{F} \triangleq \left(\frac{z^2}{2} + A^2\cos\theta + 2Ar\sin\theta\right)\hat{\mathbf{e}}_r + (-A^2\sin\theta + Ar\cos\theta)\hat{\mathbf{e}}_\theta + rz\mathbf{e}_z$
26. $\mathbf{F} \triangleq (-\frac{A}{\rho} + 6\rho\sin^2\phi\,\cos 2\theta)\hat{\mathbf{e}}_\rho + 6\rho\,\sin\phi\cos\phi\cos 2\theta\,\mathbf{e}_\phi - 6\rho\,\sin\phi\,\sin 2\theta\,\hat{\mathbf{e}}_\theta$
27. In particle physics, the "Yukawa potential" is defined by $V(\mathbf{r}) \triangleq -k\,\frac{e^{-\mu\rho}}{\rho}$, where k and μ are scalar constants. Find (a) the corresponding force field $\mathbf{F} \triangleq -\boldsymbol{\nabla} V(\mathbf{r})$ and (b) the Laplacian $\nabla^2 V(\mathbf{r})$.

* From URL http://www.scienceworld.Wolfram.com, the permeability of free space is $\mu_0 \approx 1.2566 \times 10^{-6}$ webers/(amps · m)

28. Assume A, B are unspecified constants. Determine a scalar potential function for the vector field

$$\mathbf{F} \triangleq -\frac{B}{\rho^2}(\cos\phi + \sin\theta)\hat{\mathbf{e}}_\rho - \left(\frac{A}{\rho} + \frac{B}{\rho^2}\right)\sin\phi\,\mathbf{e}_\phi - \left(\frac{A}{\rho} - \frac{B}{\rho^2}\right)\frac{\cos\theta}{\sin\phi}\,\hat{\mathbf{e}}_\theta.$$

29. Finish the explanation for the formula for the divergence operator in right-handed curvilinear coordinates, that is, (6.82) in Theorem 6.7: First, rewrite

$$\begin{aligned}\boldsymbol{\nabla} \bullet \mathbf{F} &\triangleq \boldsymbol{\nabla} \bullet \left(F_u\hat{\mathbf{e}}_u + F_v\hat{\mathbf{e}}_v + F_w\hat{\mathbf{e}}_w\right)\\ &= \boldsymbol{\nabla} \bullet \left(h_v h_w F_u \cdot \left(\frac{1}{h_v h_w}\,\hat{\mathbf{e}}_u\right)\right)\\ &\quad + \boldsymbol{\nabla} \bullet \left(h_w h_u F_v \cdot \left(\frac{1}{h_w h_u}\,\hat{\mathbf{e}}_v\right)\right) + \boldsymbol{\nabla} \bullet \left(h_u h_v F_w \cdot \left(\frac{1}{h_u h_v}\,\hat{\mathbf{e}}_w\right)\right).\end{aligned}$$

For each of these three terms, use the product rule in Theorem 6.6(4) and the "trick" results we gave, for example, $\boldsymbol{\nabla} \bullet \left(\frac{1}{h_v h_w}\,\hat{\mathbf{e}}_u\right) = \mathbf{0}$.

30. Define a "modified Coulomb's law" by $\mathbf{F}(\mathbf{r}) \triangleq K\frac{e^{-\rho/\alpha}}{\rho}\left(\frac{1}{\rho} + \frac{1}{\alpha}\right)\hat{\mathbf{e}}_\rho$, where K and α are scalar constants. (a) Calculate $\boldsymbol{\nabla} \times \mathbf{F}$. (b) Decide if $\mathbf{F}$ is exact; if it is, find a potential function for it. [Hint: Look at the answer for Problem 6.7.6.27(a).] (c) Calculate $\boldsymbol{\nabla} \bullet \mathbf{F}$.
31. Finish the explanation for the formula for the Laplacian operator in right-handed curvilinear coordinates, that is, (6.84) in Theorem 6.7, by using $\nabla^2 = \boldsymbol{\nabla} \bullet (\boldsymbol{\nabla} f)$ and the formulas for the gradient and the divergence in (6.80) and (6.82), respectively.
32. Finish the explanation for the formula for the curl operator in right-handed curvilinear coordinates, that is, (6.83) in Theorem 6.7: begin similarly to Problem 6.7.6.29 by rewriting

$$\begin{aligned}\boldsymbol{\nabla} \times \mathbf{F} &\triangleq \boldsymbol{\nabla} \times \left(F_u\hat{\mathbf{e}}_u + F_v\hat{\mathbf{e}}_v + F_w\hat{\mathbf{e}}_w\right)\\ &= \boldsymbol{\nabla} \times \left(h_u F_u \cdot \left(\frac{1}{h_u}\,\hat{\mathbf{e}}_u\right)\right) + \boldsymbol{\nabla} \times \left(h_v F_v \cdot \left(\frac{1}{h_v}\,\hat{\mathbf{e}}_v\right)\right)\\ &\quad + \boldsymbol{\nabla} \times \left(h_w F_w \cdot \left(\frac{1}{h_w}\,\hat{\mathbf{e}}_w\right).\right)\end{aligned}$$

33. Suppose $\mathbf{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$, $A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$, and $f(x, y, z) \triangleq \mathbf{x}^T A\mathbf{x}$. For unspecified A and $\mathbf{x}$, find (a) a formula for $\boldsymbol{\nabla} f$ in as simple a form as possible, for example, in terms of multiplication of a matrix or sum of matrices times a vector and

(b) a simpler version of that formula in the special case that A is symmetric. [Hint: It may help to multiply out to get a more explicit expression for $f(x, y, z)$.]

34. Suppose we make a change of variables from (x, y, z) coordinates to (u, v, w) coordinates by

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = A \begin{bmatrix} u \\ v \\ w \end{bmatrix},$$

where $A = [a_{ij}]$ is a constant 3×3 matrix, and suppose $f = f(x, y, z)$. Use the chain rule to find a simplified expression for $\nabla_{u,v,w} f \triangleq \begin{bmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} & \frac{\partial f}{\partial w} \end{bmatrix}^T$, the gradient operator in (u, v, w) coordinates, in terms of $\nabla_{x,y,z} f \triangleq \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} & \frac{\partial f}{\partial z} \end{bmatrix}^T$. Your final conclusion should be in terms of the matrix A.

6.8 Rotating Coordinate Frames

Here we will discuss motion of a particle moving in a coordinate frame that is rotating and/or in linear motion relative to a fixed coordinate frame. The results are applicable to motion on a rotating planet and are applicable to mechanical systems in general.

6.8.1 ODEs Describing Rotation

First, let's consider the case of a particle rotating around the z-axis with an angular speed ω, measured in radians per unit of time. The axis of revolution here is $\hat{\boldsymbol{k}}$. The **angular velocity** is defined to be $\boldsymbol{\omega} \triangleq \omega \hat{\boldsymbol{k}} = (\textit{angular speed}) \cdot (\textit{axis of revolution})$.

Denote the initial position vector of the particle by

$$\mathbf{r}_0 = \overrightarrow{OP_0} = x_0\,\hat{\boldsymbol{\imath}} + y_0\,\hat{\boldsymbol{\jmath}} + z_0\,\hat{\boldsymbol{k}} = \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix}.$$

The motion of the particle is described by its position vector

$$\mathbf{r}(t) = \begin{bmatrix} \cos(\omega t)\,x_0 - \sin(\omega t)\,y_0 \\ \sin(\omega t)\,x_0 + \cos(\omega t)\,y_0 \\ z_0 \end{bmatrix} = \begin{bmatrix} \cos(\omega t) & -\sin(\omega t) & 0 \\ \sin(\omega t) & \cos(\omega t) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix},$$

as shown in Figure 6.37.

There is another way to write this: denoting the initial position vector of the particle by

$$\mathbf{r}_0 = x_0\,\hat{\boldsymbol{\imath}} + y_0\,\hat{\boldsymbol{\jmath}} + z_0\,\hat{\boldsymbol{k}} = r_0\,\cos\theta_0\,\hat{\boldsymbol{\imath}} + r_0\,\sin\theta_0\,\hat{\boldsymbol{\jmath}} + z_0\,\hat{\boldsymbol{k}},$$

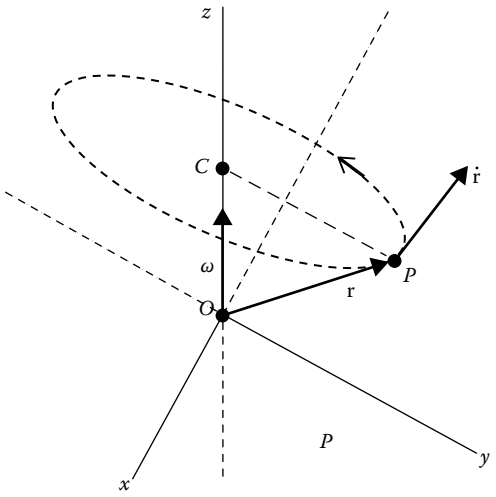


FIGURE 6.37
Particle in circular motion around the z-axis.

then the motion of the particle is described by its position vector

$$\mathbf{r}(t) = r_0 \cos(\omega t + \theta_0)\,\hat{\boldsymbol{\imath}} + r_0 \sin(\omega t + \theta_0)\,\hat{\boldsymbol{\jmath}} + z_0\,\hat{\mathbf{k}}.$$

The latter is in a form convenient for finding a tangent vector, specifically the velocity vector

$$\mathbf{T}(t) = \dot{\mathbf{r}}(t) = -\omega r_0 \sin(\omega t + \theta_0)\,\hat{\boldsymbol{\imath}} + \omega r_0 \cos(\omega t + \theta_0)\,\hat{\boldsymbol{\jmath}} = (\omega\hat{\mathbf{k}}) \times \mathbf{r}(t) = \boldsymbol{\omega} \times \mathbf{r}(t).$$

Theorem 6.8

Any rotation of a particle is described by the system of ODEs

$$\dot{\mathbf{r}}(t) = \boldsymbol{\omega} \times \mathbf{r}(t). \tag{6.86}$$

Why? Refer to Figure 6.38. First, a particle rotating around an axis of revolution $\boldsymbol{\omega}$ describes a circle lying in a plane that has $\boldsymbol{\omega}$ as a normal vector, so the tangent vector, $\dot{\mathbf{r}}(t)$, must be orthogonal to $\boldsymbol{\omega}$. Also, because the point P is on a circle, the tangent vector $\dot{\mathbf{r}}(t)$ must be orthogonal to the radial vector that points from C, the center of the circle, to P. But

$$\mathbf{r} \triangleq \overrightarrow{OP} = \overrightarrow{CP} + (\mathbf{r} \bullet \widehat{\boldsymbol{\omega}})\widehat{\boldsymbol{\omega}} \tag{6.87}$$

is an orthogonal decomposition, where $\widehat{\boldsymbol{\omega}}$ is the unit vector in the direction of $\boldsymbol{\omega}$. For example, if $\boldsymbol{\omega} = \omega\hat{\mathbf{k}}$ then $\widehat{\boldsymbol{\omega}} = \hat{\mathbf{k}}$ and (6.87) says

$$\mathbf{r} \triangleq \overrightarrow{OP} = \overrightarrow{CP} + (\mathbf{r} \bullet \hat{\mathbf{k}})\hat{\mathbf{k}} = (x\,\hat{\boldsymbol{\imath}} + y\,\hat{\boldsymbol{\jmath}}) + z\hat{\mathbf{k}}.$$

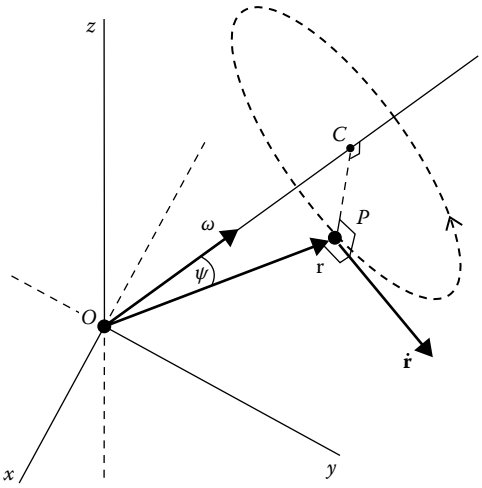


FIGURE 6.38
Particle in circular motion around an axis along $\boldsymbol{\omega}$.

So, we see that, in general, (6.87) is an orthogonal decomposition of $\mathbf{r}$ into its radial vector, $\overrightarrow{CP}$, plus the projection of $\mathbf{r}$ onto the axis of revolution. Now, the tangent vector, $\dot{\mathbf{r}}(t)$, is orthogonal to both $\boldsymbol{\omega}$ and $\overrightarrow{CP}$, so $\dot{\mathbf{r}}(t)$ is orthogonal to $\mathbf{r}$.

Also, because $\dot{\mathbf{r}}(t)$ is orthogonal to the plane spanned by $\boldsymbol{\omega}$ and $\overrightarrow{CP}$, $\dot{\mathbf{r}}(t)$ is orthogonal to the plane spanned by $\boldsymbol{\omega}$ and $\mathbf{r}$. It follows from the properties of the cross product that

$$\dot{\mathbf{r}}(t) = \alpha(\boldsymbol{\omega} \times \mathbf{r}), \tag{6.88}$$

for some scalar α.

But, the linear speed of a particle moving in a circle is (*the radius of the circle*) $\times$ (*angular speed*) $= ||\overrightarrow{CP}||\,\omega$, while (6.88) says the linear speed should be

$$||\dot{\mathbf{r}}(t)|| = |\alpha|\;||\boldsymbol{\omega} \times \mathbf{r}|| = |\alpha|\;||\boldsymbol{\omega}||\;||\mathbf{r}||\;\sin\psi,$$

where ψ is the angle between $\boldsymbol{\omega}$ and $\mathbf{r}$. But $||\overrightarrow{CP}|| = ||\mathbf{r}||\;\sin\psi$, so the linear speed is

$$\omega \cdot ||\overrightarrow{CP}|| = ||\dot{\mathbf{r}}(t)|| = |\alpha|\;||\boldsymbol{\omega}||\;||\overrightarrow{CP}|| = \omega \cdot |\alpha| \cdot ||\overrightarrow{CP}||.$$

It follows that $|\alpha| = 1$. We may define "counterclockwise" rotation about $\boldsymbol{\omega}$ to be the case when $\alpha = 1$. Replacing $\boldsymbol{\omega}$ by $-\boldsymbol{\omega}$ is equivalent to replacing $\alpha = 1$ by $\alpha = -1$. So, rotation is described by the ODE system (6.86). □

If $\boldsymbol{\omega} = \omega_x\,\hat{\boldsymbol{\imath}} + \omega_y\,\hat{\boldsymbol{\jmath}} + \omega_z\,\hat{\mathbf{k}}$ and, as usual, $\mathbf{r} = x\,\hat{\boldsymbol{\imath}} + y\,\hat{\boldsymbol{\jmath}} + z\,\hat{\mathbf{k}}$, then (6.86) can be rewritten as

$$\frac{d}{dt}\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix} \triangleq \Omega \begin{bmatrix} x \\ y \\ z \end{bmatrix}.$$

The solutions of (6.86) can be written as

$$\begin{bmatrix} x(t) \\ y(t) \\ z(t) \end{bmatrix} = e^{t\Omega} \begin{bmatrix} x(0) \\ y(0) \\ z(0) \end{bmatrix}. \tag{6.89}$$

Denoting $\nu_x = \omega_x/\omega, \nu_y = \omega_y/\omega, \nu_z = \omega_z/\omega$, which are the components of the unit vector $\widehat{\boldsymbol{\omega}}$, the fundamental matrix is given by

$$\begin{aligned} & e^{t\Omega} \\ &= \begin{bmatrix} (\nu_y^2 + \nu_z^2)\cos\omega t + \nu_x^2 & \nu_x\nu_y(1-\cos\omega t) - \nu_z \sin\omega t & \nu_z\nu_x(1-\cos\omega t) + \nu_y \sin\omega t \\ \nu_x\nu_y(1-\cos\omega t) + \nu_z \sin\omega t & (\nu_z^2 + \nu_x^2)\cos\omega t + \nu_y^2 & \nu_y\nu_z(1-\cos\omega t) - \nu_x \sin\omega t \\ \nu_z\nu_x(1-\cos\omega t) - \nu_y \sin\omega t & \nu_y\nu_z(1-\cos\omega t) + \nu_x \sin\omega t & (\nu_x^2 + \nu_y^2)\cos\omega t + \nu_z^2 \end{bmatrix}. \end{aligned} \tag{6.90}$$

As an aside, we have derived that the motion of a rotating particle is described by the system of ODEs

$$\dot{\mathbf{r}}(t) = \boldsymbol{\omega} \times \mathbf{r}(t),$$

where ω is the constant angular velocity vector. In that derivation we could have allowed $\boldsymbol{\omega}$ to be time dependent, that is,

$$\dot{\mathbf{r}}(t) = \boldsymbol{\omega}(t) \times \mathbf{r}(t),$$

but then the solutions would *not* be given by (6.89).

6.8.2 Velocity and Acceleration

The usual o.n. basis for $\mathbb{R}^3$ is $\{\hat{\boldsymbol{\imath}}, \hat{\boldsymbol{\jmath}}, \hat{\boldsymbol{k}}\}$. It is a right-handed basis because of Theorem 6.1(e) in Section 6.1.

Suppose $\{\mathbf{e}_\ell = \mathbf{e}_\ell(t) : \ \ell = 1,2,3\}$ is the o.n. set of vectors in $\mathbb{R}^3$ that are the rotations of the initial vectors $\{\hat{\boldsymbol{\imath}}, \hat{\boldsymbol{\jmath}}, \hat{\boldsymbol{k}}\}$ with *constant* angular velocity $\boldsymbol{\omega}$. So,

$$\dot{\mathbf{e}}_\ell = \boldsymbol{\omega} \times \mathbf{e}_\ell, \ \ell = 1,2,3.$$

For example, if $\boldsymbol{\omega}(t) = \omega\,\hat{\boldsymbol{k}}$ for some constant angular speed ω, then

$$\mathbf{e}_1(t) = \begin{bmatrix} \cos(\omega t) \\ \sin(\omega t) \\ 0 \end{bmatrix}, \quad \mathbf{e}_2(t) = \begin{bmatrix} -\sin(\omega t) \\ \cos(\omega t) \\ 0 \end{bmatrix}, \quad \mathbf{e}_3(t) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

In general,

$$\left[\, \mathbf{e}_1(t) \ \vdots \ \mathbf{e}_2(t) \ \vdots \ \mathbf{e}_3(t) \,\right] = e^{t\Omega} \left[\, \hat{\boldsymbol{\imath}} \ \vdots \ \hat{\boldsymbol{\jmath}} \ \vdots \ \hat{\boldsymbol{k}} \,\right] = e^{t\Omega} I_3 = e^{t\Omega}.$$

We call

$$\{\mathbf{e}_1(t), \mathbf{e}_2(t), \mathbf{e}_3(t)\}$$

a **rotating coordinate frame**. In Problem 6.8.4.14, you will check that this set of three vectors is a right-handed basis for $\mathbb{R}^3$.

Suppose

$$\mathbf{r} = \mathbf{r}(t) \triangleq x_1(t)\mathbf{e}_1(t) + x_2(t)\mathbf{e}_2(t) + x_3(t)\mathbf{e}_3(t)$$

expresses the position of a particle relative to the fixed coordinate frame $\{\hat{\boldsymbol{\imath}}, \hat{\boldsymbol{\jmath}}, \hat{\boldsymbol{k}}\}$. Here, $(x_1(t), x_2(t), x_3(t))$ are the coordinates of the particle relative to the rotating coordinate frame. By the product rule,

$$\begin{aligned}\dot{\mathbf{r}}(t) &= \dot{x}_1(t)\mathbf{e}_1(t) + \dot{x}_2(t)\mathbf{e}_2(t) + \dot{x}_3(t)\mathbf{e}_3(t) + x_1(t)\dot{\mathbf{e}}_1(t) + x_2(t)\dot{\mathbf{e}}_2(t) + x_3(t)\dot{\mathbf{e}}_3(t).\\ &= \dot{x}_1(t)\mathbf{e}_1(t) + \dot{x}_2(t)\mathbf{e}_2(t) + \dot{x}_3(t)\mathbf{e}_3(t) + x_1(t)\boldsymbol{\omega} \times \mathbf{e}_1(t) + x_2(t)\boldsymbol{\omega} \times \mathbf{e}_2(t) + x_3(t)\boldsymbol{\omega} \times \mathbf{e}_3(t)\\ &= \dot{x}_1(t)\mathbf{e}_1(t) + \dot{x}_2(t)\mathbf{e}_2(t) + \dot{x}_3(t)\mathbf{e}_3(t) + \boldsymbol{\omega} \times (x_1(t)\mathbf{e}_1(t) + x_2(t)\mathbf{e}_2(t) + x_3(t)\mathbf{e}_3(t)).\end{aligned}$$

It follows that

$$\dot{\mathbf{r}}(t) = \dot{x}_1(t)\mathbf{e}_1(t) + \dot{x}_2(t)\mathbf{e}_2(t) + \dot{x}_3(t)\mathbf{e}_3(t) + \boldsymbol{\omega} \times \mathbf{r}(t). \tag{6.91}$$

Define the *velocity of the particle relative to the rotating coordinate frame* by

$$\mathbf{v}_{rel}(t) \triangleq \dot{x}_1(t)\mathbf{e}_1(t) + \dot{x}_2(t)\mathbf{e}_2(t) + \dot{x}_3(t)\mathbf{e}_3(t).$$

Then (6.91) can be rewritten as

$$\mathbf{v}(t) \triangleq \dot{\mathbf{r}}(t) = \mathbf{v}_{rel}(t) + \boldsymbol{\omega} \times \mathbf{r}(t). \tag{6.92}$$

This gives the velocity relative to the fixed coordinate frame $\{\hat{\boldsymbol{\imath}}, \hat{\boldsymbol{\jmath}}, \hat{\boldsymbol{k}}\}$.

Take another derivative with respect to time to get

$$\ddot{\mathbf{r}}(t) = \dot{\mathbf{v}}(t) = \frac{d}{dt}[\,\mathbf{v}_{rel}(t) + \boldsymbol{\omega} \times \mathbf{r}(t)\,] = \frac{d}{dt}[\,\mathbf{v}_{rel}(t)\,] + \boldsymbol{\omega} \times \dot{\mathbf{r}}(t); \tag{6.93}$$

recall that we are assuming that the angular velocity vector, $\boldsymbol{\omega}$, is constant. To further analyze (6.93), we calculate

$$\begin{aligned}\frac{d}{dt}[\,\mathbf{v}_{rel}(t)\,] &= \frac{d}{dt}[\,\dot{x}_1(t)\mathbf{e}_1(t) + \dot{x}_2(t)\mathbf{e}_2(t) + \dot{x}_3(t)\mathbf{e}_3(t)\,]\\ &= \ddot{x}_1(t)\mathbf{e}_1(t) + \ddot{x}_2(t)\mathbf{e}_2(t) + \ddot{x}_3(t)\mathbf{e}_3(t) + \dot{x}_1(t)\dot{\mathbf{e}}_1(t) + \dot{x}_2(t)\dot{\mathbf{e}}_2(t) + \dot{x}_3(t)\dot{\mathbf{e}}_3(t)\\ &= \ddot{x}_1(t)\mathbf{e}_1(t) + \ddot{x}_2(t)\mathbf{e}_2(t) + \ddot{x}_3(t)\mathbf{e}_3(t) + \dot{x}_1(t)\boldsymbol{\omega} \times \mathbf{e}_1(t) + \dot{x}_2(t)\boldsymbol{\omega} \times \mathbf{e}_2(t) + \dot{x}_3(t)\boldsymbol{\omega} \times \mathbf{e}_3(t),\end{aligned}$$

that is,

$$\frac{d}{dt}[\,\mathbf{v}_{rel}(t)\,] = \mathbf{a}_{rel}(t) + \boldsymbol{\omega} \times \big(\dot{x}_1(t)\mathbf{e}_1(t) + \dot{x}_2(t)\mathbf{e}_2(t) + \dot{x}_3(t)\mathbf{e}_3(t)\big). \tag{6.94}$$

We define the *acceleration of the particle relative to the rotating coordinate frame* by

$$\mathbf{a}_{rel}(t) \triangleq \ddot{x}_1(t)\mathbf{e}_1(t) + \ddot{x}_2(t)\mathbf{e}_2(t) + \ddot{x}_3(t)\mathbf{e}_3(t).$$

Substituting the definition of $\mathbf{v}_{rel}(t)$ into (6.94) gives

$$\frac{d}{dt}[\,\mathbf{v}_{rel}(t)\,] = \mathbf{a}_{rel}(t) + \boldsymbol{\omega} \times \mathbf{v}_{rel}(t).$$

Substituting this and (6.92) into (6.93) gives

$$\ddot{\mathbf{r}}(t) = \dot{\mathbf{v}}(t) = \mathbf{a}_{rel}(t) + \boldsymbol{\omega} \times \mathbf{v}_{rel}(t) + \boldsymbol{\omega} \times \big(\mathbf{v}_{rel}(t) + \boldsymbol{\omega} \times \mathbf{r}(t)\big),$$

that is,

$$a(t) \triangleq \ddot{\mathbf{r}}(t) = \mathbf{a}_{rel}(t) + 2\boldsymbol{\omega} \times \mathbf{v}_{rel}(t) + \boldsymbol{\omega} \times (\boldsymbol{\omega} \times \mathbf{r}(t)). \tag{6.95}$$

The term $2\boldsymbol{\omega} \times \mathbf{v}_{rel}(t)$ is called the **Coriolis acceleration**.

While $\boldsymbol{\omega} \times \boldsymbol{\omega} = \mathbf{0}$, if $\boldsymbol{\omega} \times \mathbf{r}(t) \neq \mathbf{0}$, then the term $\boldsymbol{\omega} \times (\boldsymbol{\omega} \times \mathbf{r}(t)) \neq \mathbf{0}$ because $\boldsymbol{\omega} \times \mathbf{r}(t)$ is perpendicular to $\boldsymbol{\omega}$.

Example 6.34

A baseball pitcher standing at latitude 50°N of the Earth's equator throws a ball in the due north direction at a speed of 90.00 mph. Find the initial Coriolis acceleration of the ball.

Method: Figure 6.39a shows a point at latitude 50°N of the Earth's equator. There, a unit vector pointing due north resolves to be $-\cos 50^\circ\, \hat{\boldsymbol{j}} + \sin 50^\circ\, \hat{\boldsymbol{k}}$. The initial velocity vector of a ball thrown due north resolves to be

$$\mathbf{v}_{rel}(0) = -v_{rel}(0)\, \cos 50^\circ\, \hat{\boldsymbol{j}} + v_{rel}(0)\, \sin 50^\circ\, \hat{\boldsymbol{k}}.$$

Because 60 miles/h equals 88 ft/s, the initial speed of the ball is

$$v_{rel}(0) = \Big(132\, \frac{\cancel{\text{ft}}}{\text{s}}\Big) \cdot \frac{0.304800610\,\text{m}}{1\,\cancel{\text{ft}}} \approx 40.23368052\,\text{m/s},$$

in KMS units. So, rounding all final conclusions to the four significant digits in the given initial speed of the ball,

$$\mathbf{v}_{rel}(0) \approx (-25.86\, \hat{\boldsymbol{j}} + 30.82\, \hat{\boldsymbol{k}})\,\text{m/s}.$$

The Earth does one complete rotation, that is, 2π radians, in about 86164.1 s, so its angular rotational speed is $\omega \approx \frac{2\pi}{86164.1\text{ s}} \approx 7.29212 \times 10^{-5}$/s and its angular velocity vector is $\boldsymbol{\omega} = \omega\, \hat{\boldsymbol{k}}$. The initial Coriolis acceleration is

$$2\boldsymbol{\omega} \times \mathbf{v}_{rel}(0) \approx 2(7.29212 \times 10^{-5}/\text{s})\hat{\boldsymbol{k}} \times (-25.86\, \hat{\boldsymbol{j}} + 30.82\, \hat{\boldsymbol{k}})\,\text{m/s} \approx 0.003771\,\text{m/s}^2\, \hat{\boldsymbol{i}}.\ \bigcirc$$

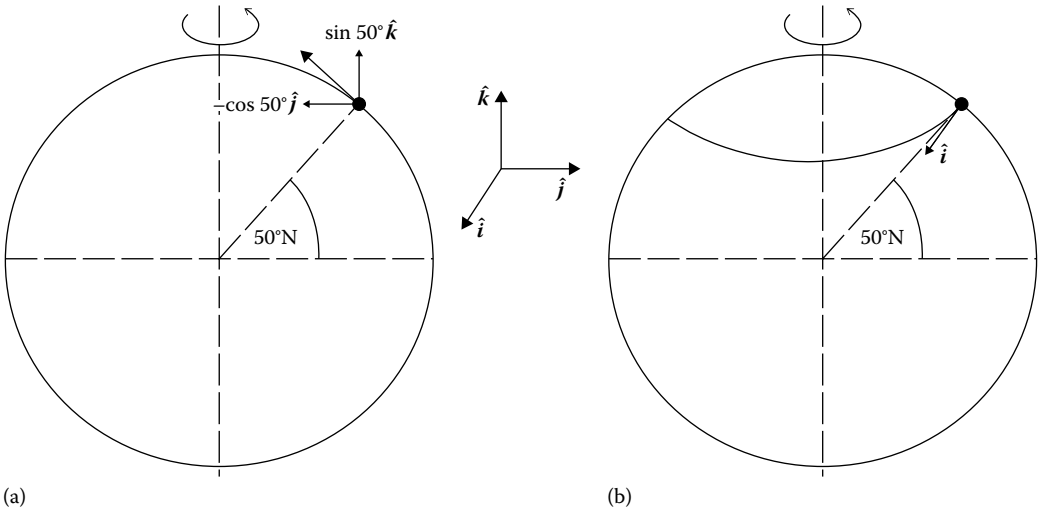


FIGURE 6.39
Example 6.34 (a) Ball thrown due north and (b) ball thrown due west.

Example 6.35

Find the Coriolis acceleration of the ball in Example 6.34 if instead it is thrown due west.

Method: Figure 6.39b shows a point at latitude 50°N of the Earth's equator and a vector pointing due west from that point, that is, in the direction of $\hat{\boldsymbol{\imath}}$. Reusing some of the information from Example 6.9, the initial Coriolis acceleration is

$$2\boldsymbol{\omega} \times \mathbf{v}_{rel}(0) \approx 2(7.29212 \times 10^{-5}/\text{s})\hat{\mathbf{k}} \times 40.23368052\hat{\boldsymbol{\imath}}\,\text{m/s} \approx 0.005868\,\text{m/s}^2\,\hat{\boldsymbol{\jmath}}. \bigcirc$$

So, on the planet Earth, the Coriolis acceleration is so small that it does not explain why a "curveball" curves.

Remark

If the angular velocity vector, $\boldsymbol{\omega}$, is not constant, then, instead of (6.95), we get

$$\ddot{\mathbf{r}}(t) = \mathbf{a}_{rel}(t) + 2\boldsymbol{\omega}(t) \times \mathbf{v}_{rel}(t) + \boldsymbol{\omega}(t) \times (\boldsymbol{\omega}(t) \times \mathbf{r}(t)) + \dot{\boldsymbol{\omega}}(t) \times \mathbf{r}(t), \tag{6.96}$$

where $\dot{\boldsymbol{\omega}}(t)$ is called the *angular acceleration vector*.

6.8.3 Velocity and Acceleration in a Rotating Frame Whose Origin Is Moving

Suppose there is a third, *fixed*, rectangular coordinate frame $\{\hat{I}, \hat{J}, \hat{K}\}$. By "fixed" we mean constant in time, so $\frac{d\hat{I}}{dt} = \frac{d\hat{J}}{dt} = \frac{d\hat{K}}{dt} = \mathbf{0}$. Let O be the center of this third, fixed rectangular frame and let o be the center of both the rotating coordinate frame, $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, and the **intermediate frame**, $\{\hat{\boldsymbol{\imath}}, \hat{\boldsymbol{\jmath}}, \hat{\mathbf{k}}\}$. Suppose o is in motion relative to O, with velocity $\mathbf{V}(t)$ and acceleration, $\mathbf{A}(t)$. A particle in the rotating frame at a point P has coordinates (X, Y, Z)

relative to O, the center of this third, fixed rectangular frame, that is,

$$\mathbf{R}(t) = \overrightarrow{OP} = X(t)\,\hat{\boldsymbol{I}} + Y(t)\,\hat{\boldsymbol{J}} + Z(t)\,\hat{\boldsymbol{K}}.$$

So, relative to the third frame, a particle in the rotating coordinate frame has velocity

$$\dot{\mathbf{R}}(t) = \mathbf{V}(t) + \dot{\mathbf{r}}(t) = \mathbf{V}(t) + \mathbf{v}_{rel}(t) + \boldsymbol{\omega} \times \mathbf{r}(t)$$

and acceleration

$$\ddot{\mathbf{R}}(t) = \mathbf{A}(t) + \ddot{\mathbf{r}}(t) = \mathbf{A}(t) + \mathbf{a}_{rel}(t) + 2\boldsymbol{\omega} \times \mathbf{v}_{rel}(t) + \boldsymbol{\omega} \times (\boldsymbol{\omega} \times \mathbf{r}(t)). \tag{6.97}$$

Remark

Again, if the angular velocity vector, $\boldsymbol{\omega}$, is not constant, then, instead of (6.97), the particle's acceleration is

$$\ddot{\mathbf{R}}(t) = \mathbf{A}(t) + \ddot{\mathbf{r}}(t) = \mathbf{A}(t) + \mathbf{a}_{rel}(t) + 2\boldsymbol{\omega} \times \mathbf{v}_{rel}(t) + \boldsymbol{\omega} \times (\boldsymbol{\omega} \times \mathbf{r}(t)) + \dot{\boldsymbol{\omega}}(t) \times \mathbf{r}(t). \tag{6.98}$$

Learn More About It

Problem 6.8.4.6 is discussed in "The applications of mathematics in meteorology," by B. Haurwitz, *Am. Math. Mon*, (**50**) (1943), 77–84.

6.8.4 Problems

1. At a latitude of 30°N of the equator on the surface of the Earth, find $\boldsymbol{\omega} \times (\boldsymbol{\omega} \times \mathbf{r})$. Give a picture showing your coordinate frame and the location of 30°N.
2. The wind is blowing straight south at 54 km/h at a latitude of 30°N of the equator on the surface of the Earth. (a) Find the Coriolis acceleration vector and (b) the $\boldsymbol{\omega} \times (\boldsymbol{\omega} \times \mathbf{r})$ component of the acceleration. Give a picture showing your coordinate frame, the location of 30°N, and the wind direction. Assume the radius of the Earth is about 6400 km.
3. Suppose that in an inertial reference frame, the velocity of a particle is $\mathbf{v}(t) = \frac{1}{1+(\rho(t))^3}\mathbf{R}$, where $\mathbf{R}$ is a constant vector and ρ is the magnitude of the position vector $\mathbf{r}(t)$, as usual. Find the acceleration of the particle relative to the inertial reference frame in terms of $\mathbf{R}$, ρ, $\mathbf{r}(t)$, and $\mathbf{v}(t)$, the velocity vector relative to the inertial reference frame.
4. A very small insect is on a compact disk rotating at 60 revolutions per second. The insect is traveling along a straight line, according to its point of view, at a constant speed of 0.05 mm/s. What is the Coriolis acceleration *vector* it experiences? Do make clear your coordinates.
5. On the planet Jupiter at latitude 60°N of the equator, find the vector $\boldsymbol{\omega} \times (\boldsymbol{\omega} \times \mathbf{r})$ that is a component of the acceleration. Here are some data about Jupiter: It's period of rotation is about 9.925 h*, and its radius is about 71,000 km.

* According to N.A.S.A. at URL http://nssdc.gsfc.nasa.gov/planetary/factsheet/jupiterfact.html

6. Near the end of Section 6.5, we discussed isobars, for example, shown in Figure 6.23. The closeness of successive isobars shows the magnitude of the horizontal components of the gradient of atmospheric pressure p. But we will see that the direction of the geostrophic wind is almost parallel to the isobars because of the Coriolis acceleration due to the Earth's rotation. For example, Coriolis acceleration effects the "sub tropical jet stream," which is usually at an altitude of between 10,000 and 16,000 m.

   Let $\mathbf{v} = \mathbf{v}(t, \mathbf{r})$ be the wind velocity. Assuming an altitude above 1000 m so that ground effects are insignificant, an approximate model for the velocity is the nonlinear partial differential equation

$$\left(\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \bullet \nabla \mathbf{v}\right) + 2\boldsymbol{\omega} \times \mathbf{v} = -\frac{1}{\varrho_0} \nabla \mathbf{p} - \mathbf{g},$$

   where $\mathbf{g}$ is the acceleration of gravity and we assume that the air density ϱ_0 is approximately constant. The first two terms are the material derivative of velocity and the third term is the Coriolis acceleration.

   For this problem, it helps to choose the coordinate system so that the outward normal to the Earth's surface, at a latitude of ψ north of the equator, is the vector $\hat{\boldsymbol{k}}$; hence, the angular velocity vector is $\boldsymbol{\omega} = \omega \sin \psi \, \hat{\boldsymbol{k}}$ and the acceleration of gravity is $\mathbf{g} = -g\hat{\boldsymbol{k}}$, that is, is in the vertical direction. Notice that the latitude affects the equations; in particular, there is a difference between the northern and southern hemispheres.

   Assuming the wind is steady, that is, $\mathbf{v}(t, \mathbf{r})$ is approximately constant, find two equations for the two horizontal components of the wind, $u = \mathbf{v} \bullet \hat{\boldsymbol{\imath}}$ and $v = \mathbf{v} \bullet \hat{\boldsymbol{\jmath}}$.

   At 39°45′ 32″ N, the latitude of Dayton, Ohio, assume the horizontal gradient of the pressure points due south and has magnitude of 1 millibar *per* 70 km; note that 1 millibar = 100 Pa in pascals, that is, $\mathrm{N/m^2}$. Assume the density of air is about $\varrho_0 = 1.293\,\mathrm{kg/m^3}$. Find the direction and magnitude of the geostrophic wind.

7. An astronaut stands in the crew compartment at the end of a 60 m long cylinder rotating about its lateral midpoint at a speed of ω rads/s. What is the maximum period of rotation that gives the astronaut the sensation of 1 g (gravity) of acceleration?
8. A baseball pitcher standing on the Earth's equator throws a ball in the due north direction at a speed of 90.00 mph. Find the initial Coriolis acceleration of the ball.
9. The velocity at the point $(-1, 2, 3)$, with distances measured in m, from the center of a rotating coordinate frame with angular velocity $\boldsymbol{\omega}$, is measured to be $\mathbf{V} = 12\,\hat{\boldsymbol{\imath}} + 6\,\hat{\boldsymbol{\jmath}}$ m/s. What is $\boldsymbol{\omega}$, if $\omega_x = 3$ rad/s?

In problems 10–13, write down the rotating matrix $e^{t\Omega}$ for the given $\boldsymbol{\omega}$.

10. $\boldsymbol{\omega} = \hat{\boldsymbol{\imath}}$
11. $\boldsymbol{\omega} = \hat{\boldsymbol{\jmath}}$
12. $\boldsymbol{\omega} = \hat{\boldsymbol{\imath}} + \hat{\boldsymbol{\jmath}}$
13. $\boldsymbol{\omega} = \hat{\boldsymbol{\imath}} + \hat{\boldsymbol{\jmath}} + \hat{\boldsymbol{k}}$

14. Suppose a coordinate frame is rotating with constant angular velocity and the frame is given by $\{\mathbf{e}_1(t), \mathbf{e}_2(t), \mathbf{e}_3(t)\}$. Using the explicit representation

$$\left[\, \mathbf{e}_1(t) \;\vdots\; \mathbf{e}_2(t) \;\vdots\; \mathbf{e}_3(t) \,\right] = e^{t\Omega},$$

explain why the vectors in the rotating frame give a right-handed basis for $\mathbb{R}^3$.

15. Explain why scalar triple products satisfy $\mathbf{A} \bullet (\mathbf{B} \times \mathbf{C}) = \mathbf{C} \bullet (\mathbf{A} \times \mathbf{B})$.

16. Given any two vector valued functions $\mathbf{x}(t), \mathbf{z}(t)$, explain why

$$\frac{d}{dt}[\mathbf{x}(t) \bullet \mathbf{z}(t)] \equiv \dot{\mathbf{x}}(t) \bullet \mathbf{z}(t) + \mathbf{x}(t) \bullet \dot{\mathbf{z}}(t).$$

17. Suppose a coordinate frame is rotating with a (possibly nonconstant) angular velocity and the frame is given by $\{\mathbf{e}_1(t), \mathbf{e}_2(t), \mathbf{e}_3(t)\}$, where $\dot{\mathbf{e}}_\ell = \boldsymbol{\omega}(t) \times \mathbf{e}_\ell$, $\ell = 1, 2, 3$. Use the results of Problems 6.8.4.15 and 6.8.4.16 to explain why $\{\mathbf{e}_1(t), \mathbf{e}_2(t), \mathbf{e}_3(t)\}$ is an o.n. basis for $\mathbb{R}^3$. [Hints: Define a scalar function of t by $y(t) \triangleq \mathbf{e}_i(t) \bullet \mathbf{e}_j(t)$, calculate $y(0)$ and $\dot{y}(t)$, and use the uniqueness part of the existence and uniqueness theorem for first-order ODEs.]

18. Define $\mathbf{A} \times (\mathbf{B} \times \mathbf{C})$ to be a **vector triple product**. Explain why vector triple products satisfy $\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = (\mathbf{A} \bullet \mathbf{C})\mathbf{B} - (\mathbf{A} \bullet \mathbf{B})\mathbf{C}$.

19. Given any two vector-valued functions $\mathbf{x}(t), \mathbf{z}(t)$, explain why

$$\frac{d}{dt}[\mathbf{x}(t) \times \mathbf{z}(t)] \equiv \dot{\mathbf{x}}(t) \times \mathbf{z}(t) + \mathbf{x}(t) \times \dot{\mathbf{z}}(t).$$

20. Suppose a coordinate frame is rotating with a (possibly nonconstant) angular velocity and the frame is given by $\{\mathbf{e}_1(t), \mathbf{e}_2(t), \mathbf{e}_3(t)\}$, where $\dot{\mathbf{e}}_\ell = \boldsymbol{\omega}(t) \times \mathbf{e}_\ell$, $\ell = 1, 2, 3$. Use the results of Problems 6.8.4.17 through 6.8.4.19 to explain why $\{\mathbf{e}_1(t), \mathbf{e}_2(t), \mathbf{e}_3(t)\}$ is a right-handed o.n. basis for $\mathbb{R}^3$. [Hints: For example, define a vector function of t by $\mathbf{x}(t) = \mathbf{e}_1(t) \times \mathbf{e}_2(t)$, calculate $\mathbf{x}(0)$ and $\dot{\mathbf{x}}(t)$, and use the uniqueness part of the existence and uniqueness theorem for systems of ODEs.]

Key Terms

angular speed, **angular velocity**: beginning of Section 6.8
change of variables: Definition 6.5 in Section 6.6
closed: Definition 6.2 in Section 6.3
contour plot: after Example 6.26 in Section 6.5
Coriolis acceleration: after (6.95) in Section 6.8
curl: beginning of Section 6.7
curvilinear coordinates: Definition 6.7 in Section 6.7
cyclotron motion: Problem 6.1.4.17
cylindrical coordinates: after Example 6.6 in Section 6.2
directional derivative: (6.34) and before (6.40), both in Section 6.4
divergence: beginning of Section 6.7
element of surface area: (6.64) in Section 6.6
exactness criterion: (6.43) in Section 6.4

Fick's law of diffusion: beginning of Section 6.7
Fourier's Law of Heat Conduction: beginning of Section 6.7
geostrophic wind: before Example 6.27, Problem 6.8.4.6
gradient vector: (6.34) and before (6.40), both in Section 6.4
image: after (6.10) in Section 6.2, after (6.59) in Section 6.6
incompressible: beginning of Section 6.7
intermediate frame: after (6.96) in Section 6.8
intersect orthogonally: Definition 6.6 in Section 6.7
irrotational: beginning of Section 6.7
isobars: after Example 6.26 in Section 6.5
Jacobian matrix: Definition 6.5 in Section 6.6
Laplacian operator: beginning of Section 6.7
lattice: Problem 6.6.5.16
Leibniz's rule: Problem 6.4.5.24
level curves, **level sets**: after Example 6.26 in Section 6.5
linear approximation: (6.33) in Section 6.4
linear transformation: (6.59) in Section 6.6
Lorentz force: Example 6.2 in Section 6.1
map: (6.59) in Section 6.6
mapping: Example 6.29 in Section 6.6
normal: after (6.6) in Section 6.1
one-to-one: after (6.10) in Section 6.2
orthogonal surfaces: Problem 6.5.1.9
parameters: before Example 6.15 in Section 6.3
parametrized: before Example 6.8 in Section 6.3
parametric equations: after (6.3) in Section 6.1
parametrized surface: before Example 6.15 in Section 6.3
position vector: Definition 6.1 in Section 6.1
potential function, **exact**: Definition 6.4 in Section 6.4
range: after (6.10) in Section 6.2
reciprocal or "**dual**" **lattice**: Problem 6.6.5.16
reflection transformation: after Example 6.29 in Section 6.6
right-handed o.n. basis: before (6.13) in Section 6.2
rotating coordinate frame: before (6.91) in Section 6.8
rotation transformation: after Example 6.29 in Section 6.6
scalar triple product: after (6.57) in Section 6.6
scale factors: before (6.75) in Section 6.7
shear transformation: after Example 6.29 in Section 6.6
simple (curve): Definition 6.2 in Section 6.3
simple (surface): Definition 6.3 in Section 6.3
spherical coordinates: Section 6.2.2
standard orthonormal basis for $\mathbb{R}^3$: after Example 6.7 in Section 6.2
stretching and/or compressing transformation: after Example 6.29 in Section 6.6
torque: Example 6.1 in Section 6.1
transformation: (6.59) in Section 6.6
unit square: Example 6.29 in Section 6.6
vector field: before Example 6.13 in Section 6.3
vector from P_0 to P_1: Definition 6.1 in Section 6.1
vector triple product: Problem 6.8.4.18

vorticity: beginning of Section 6.7
work: before (6.1) in Section 6.1
Zhukovskii airfoils: Section 6.3.2

Mathematica Command

ParametricPlot[{2Cos[t], Sin[2t]}, {t, 0, 2Pi}]: before Figure 6.12

Reference

Milne-Thomson, L.M. *Theoretical Hydrodynamics*, Section 6.22. Dover Publications, Mineola, NY, 1968.

7

Integral Theorems, Multiple Integrals, and Applications

7.1 Integrals for a Function of a Single Variable

Suppose $f(x)$ is a function of x on a finite interval $[a,b]$. Recall that the definite integral of f on $[a,b]$ can be defined by the limit of Riemann sums, if the limit exists: If $a = x_0 < x_1 < \ldots < x_n = b$ is a partition of $[a,b]$, x_k^* are sampling points satisfying $x_{k-1} \le x_k^* \le x_k$ for $k = 1,2,\ldots,n$, and $\triangle x_k = x_k - x_{k-1}$ for $k = 1,2,\ldots,n$, then

$$\sum_{k=1}^{n} f(x_k^*)\ \triangle x_k \tag{7.1}$$

is a **Riemann sum**. The $\triangle x_k$'s don't have to be equal. If the limit of the Riemann sums exists* as $n \to \infty$ and $\max\limits_{1\le k\le n} |\triangle x_k| \to 0$, we define the **definite integral** of f from $x = a$ to $x = b$:

$$\int_a^b f(x)dx = \lim_{n\to\infty} \left(\sum_{k=1}^{n} f(x_k^*)\ \triangle x_k \right). \tag{7.2}$$

If the definite integral of f on $[a,b]$ exists and $f(x) \ge 0$ there, then $\int_a^b f(x)dx$ equals the area under the curve $y = f(x)$ and above the interval $a \le x \le b$ on the x-axis, as illustrated in Figure 7.1.

Theorem 7.1

Suppose f is continuous on the finite interval $[a,b]$. Then $\int_a^b f(x)dx$ exists. Further, for each $n = 1,2,\ldots$, we may choose the **regular partition** $a = x_0 < x_1 < \cdots < x_n = b$, where $\triangle x = \frac{b-a}{n}$, $x_k = a + k\ \triangle x$, and sampling at right endpoints $x_k^* = x_k$, and then

* For a limit to *exist*, it must be a finite number. Even for one n, there are infinitely many choices of partitions and sampling points, so saying the "limit of the Riemann sums exists" says that it is the same number no matter what partitions and sampling points the Riemann sums use.

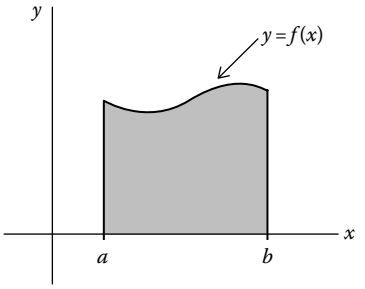


FIGURE 7.1
Definite integral calculates an area.

$$\int_a^b f(x)dx = \lim_{n\to\infty}\left(\sum_{k=1}^{n} f(x_k)\ \triangle x\right).$$

We get the same limit using sampling at left endpoints or sampling at midpoints:

$$\int_a^b f(x)dx = \lim_{n\to\infty}\left(\sum_{k=1}^{n} f(x_{k-1})\ \triangle x\right) = \lim_{n\to\infty}\left(\sum_{k=1}^{n} f\left(\frac{x_{k-1}+x_k}{2}\right)\ \triangle x\right).$$

Theorem 7.2

(Properties of the definite integral) Suppose the definite integrals of f and g on $[a, b]$ exist, c is a constant in $[a, b]$, and α is a constant. Then

(a) $\int_a^b (f(x) + g(x))dx = \int_a^b f(x)dx + \int_a^b g(x)dx$.
(b) $\int_a^b (\alpha f(x))dx = \alpha \int_a^b f(x)dx$.
(c) $\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx$, as illustrated in Figure 7.2.

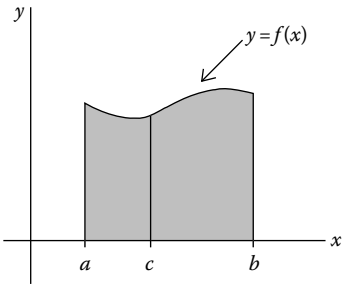


FIGURE 7.2
Definite integral on a union of intervals.

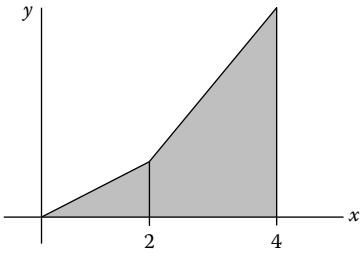


FIGURE 7.3
Definite integral of a piecewise-defined function.

Definition 7.1

If $a < b$ and $\int_a^b f(x)dx$ exists, $\int_b^a f(x)dx \triangleq -\int_a^b f(x)dx$.

Example 7.1

Let $f(x)$ be defined "piecewise" by $f(x) = \begin{Bmatrix} x, & 0 \le x \le 2 \\ 2x, & 2 \le x \le 4 \end{Bmatrix}$. Evaluate $\int_0^4 f(x)dx$.

Method: We get $\int_0^4 f(x)dx = \int_0^2 f(x)dx + \int_2^4 f(x)dx = \int_0^2 x\,dx + \int_2^4 2x\,dx$. Interpreted as an area, we calculate that $\int_0^2 x\,dx$ is the area of a triangle with base of length 2 and height 2, so $\int_0^2 x\,dx = 2$ as seen in Figure 7.3. Similarly, $\int_2^4 x\,dx$ is the area of a trapezoid with base 2 and average height $\frac{2+8}{2} = 5$, so $\int_2^4 x\,dx = 2 \cdot 5 = 10$. Finally,

$$\int_0^4 f(x)dx = \int_0^2 x\,dx + \int_2^4 2x\,dx = 2 + 10 = 12. \bigcirc$$

Example 7.2

Suppose $\int_a^b f(x)dx = 5$, $\int_a^b \big(f(x) + g(x)\big)dx = -4$, and $\int_c^b g(x)dx = 2$. Find $\int_a^c 3g(x)dx$.

Method:

$$\int_a^c 3g(x)dx = 3\int_a^c g(x)dx = 3\left(\int_a^b g(x)dx - \int_c^b g(x)dx\right)$$

$$= 3\left(\int_a^b \big(f(x) + g(x)\big)dx - \int_a^b f(x)dx - \int_c^b g(x)dx\right) = 3(-4 - 5 - 2) = -33. \bigcirc$$

Definition 7.2

$F(x)$ is an **indefinite integral** of $f(x)$ on an open interval I if $\frac{d}{dx}[F(x)] = f(x)$ for all x in I, in which case we denote $F(x) = \int f(x)dx$. Also, we call $F(x)$ an **anti-derivative** for $f(x)$ on I.

Theorem 7.3

If $F_1(x)$ and $F_2(x)$ are anti-derivatives for the same function $f(x)$ on an interval (a, b), then $F_1(x) - F_2(x) \equiv c$ on $[a, b]$ for some constant c.

Theorem 7.4

(The Fundamental Theorem of Calculus) Suppose $f(x)$ is continuous on a finite interval $[a, b]$ and c is in $[a, b]$. Then

(a) $G(x) \triangleq \int_c^x f(t)dt$ satisfies $\frac{d}{dx}[G(x)] = f(x)$, for $a < x < b$.

(b) $\int_a^b f(x)dx = F(b) - F(a) \triangleq \left[F(x)\right]_a^b$, if $F(x)$ is any anti-derivative for $f(x)$ on the interval (a, b).

There is another way to write Theorem 7.4(b):

Theorem 7.5

$[F(x)]_a^b = \int_a^b \left(\frac{d}{dx}[F(x)]\right) dx$, assuming $\frac{d}{dx}[F(x)]$ is continuous on (a, b).

Another notation for Theorem 7.4(b) is

$$\int_a^b f(x)dx = \left[\int f(x)dx\right]_a^b.$$

Theorem 7.6

(**Method of Substitution**) Suppose $w(x)$ is continuously differentiable and monotone on a finite interval $[a, b]$ and $f(x)$ is continuous on the interval $[w(a), w(b)]$. Then

$$\int_a^b f(w(x))\frac{dw}{dx}dx = \int_{w(a)}^{w(b)} f(w)dw.$$

Theorem 7.7

(**Method of Integration by Parts**) Suppose $u(x), v(x)$ are continuously differentiable on a finite interval $[a, b]$. Then

$$\int_a^b u(x)\,\frac{dv}{dx}(x)\,dx = [u(x)v(x)]_a^b - \int_a^b v(x)\,\frac{du}{dx}(x)\,dx.$$

Definition 7.3

The **average value** of a function f on the interval $[a, b]$ is

$$\bar{f} \triangleq \frac{1}{b-a} \int_a^b f(x)\, dx = \frac{\int_a^b f(x)\, dx}{\int_a^b 1\, dx}.$$

In many problems of science and engineering, the average value of a function is a simple quantity used to describe a situation. Even though vastly different functions can have the same average value, nevertheless the average value may be useful information.

7.1.1 Improper Integrals

Definition 7.4

Suppose a is a finite number and $f(x)$ is continuous on the interval $[a, \infty)$. The **improper integral** $\int_a^\infty f(x)dx$

(a) Is **convergent** if $\boxed{\lim_{b\to\infty} \int_a^b f(x)dx = L}$ exists, in which case we say the integral **converges** to L

(b) Is **divergent** if $\lim_{t\to\infty} \int_a^t f(x)dx$ does not exist

Recall that saying a limit "exists" implicitly says that the limit is a finite value. Figure 7.4 depicts the area under the curve $y = x^{-2}$ and over the interval $[1, t]$. For example, in case (a) of Definition 7.4, we can write $\boxed{\int_a^\infty f(x)dx \triangleq \lim_{t\to\infty} \int_a^t f(x)dx}$.

On the other hand, if $\lim_{t\to\infty} \int_a^t f(x)dx = \infty$, that is, the limit is defined but does not exist, then we write nothing more than "the integral is divergent"; we do not claim that we can evaluate it.

Similarly, the convergence or divergence of $\int_{-\infty}^b f(x)dx$ is decided by studying $\lim_{t\to -\infty} \int_t^b f(x)dx$.

Both $\int_a^\infty f(x)dx$ and $\int_{-\infty}^b f(x)dx$ are called improper integrals on **semi-infinite** intervals; $\int_{-\infty}^\infty f(x)dx$ is called an improper integral on the real line.

Definition 7.5

Suppose a is any finite number and $f(x)$ is continuous on the real line, that is, $(-\infty, \infty)$. The **improper integral** $\int_{-\infty}^\infty f(x)dx$

(a) Is **convergent** if both $\int_a^\infty f(x)dx = L_1$ and $\int_{-\infty}^a f(x)dx = L_2$ exist, in which case we say the integral **converges** to $L_1 + L_2$ and write $\boxed{\int_{-\infty}^\infty f(x)dx = \int_{-\infty}^a f(x)dx + \int_a^\infty f(x)dx}$

(b) Is **divergent** if *either* $\int_a^\infty f(x)dx$ or $\int_{-\infty}^a f(x)dx$ is divergent

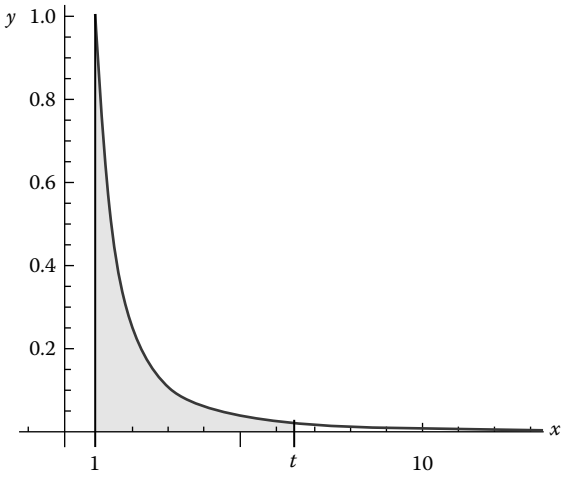


FIGURE 7.4
$\int_1^t x^{-2}dx$.

Example 7.3

Study the convergence of $\int_{-\infty}^{\infty} \frac{2x}{x^2+1}dx$.

Method: Choosing $a = 0$ for convenience,

$$\lim_{t\to\infty} \int_0^t \frac{2x}{x^2+1}dx = \lim_{t\to\infty} \Big[\ln(x^2+1)\Big]_0^t = \lim_{t\to\infty} \Big(\ln(t^2+1) - 0\Big) = \infty,$$

so $\int_{-\infty}^{\infty} \frac{2x}{x^2+1}dx$ diverges. ○

The improper integral in Example 7.3 is divergent *even though*

$$\lim_{t\to\infty} \int_{-R}^{R} \frac{2x}{x^2+1}dx = \lim_{t\to\infty} \Big[\ln(x^2+1)\Big]_{-R}^{R} = \lim_{t\to\infty} \Big[\ln(R^2+1) - \ln\Big((-R)^2+1\Big)\Big] = \lim_{t\to\infty} (0) = 0.$$

7.1.2 Problems

For the given integral, use the method of substitution and the fundamental theorem of calculus to find its value.

1. $\int_{-\pi}^{\frac{\pi}{2}} \cos\left(2t + \frac{\pi}{4}\right) dt$
2. $\int_{-\frac{\pi}{2}}^{0} \sin^3(2x)\cos(2x)\, dx$
3. $\int_0^5 \frac{3x}{1+x^2}\, dx$
4. $\int_0^1 \frac{e^{-x}}{(e^{-x}+1)^2}\, dx$

5. $\int_0^{\frac{\pi}{4}} \cos(\sqrt{x})\, dx$
   [Hint: Try the substitution $w = \sqrt{x}$; after that, you may need to use integration by parts.]

In problems 6–12, for the given improper integral, determine whether it is convergent or divergent. If it is convergent, give its value.

6. $\int_2^{\infty} \frac{1}{x^{1.1}}\, dx$
7. $\int_0^{\infty} \frac{x^2 dx}{(1+x^3)^2}$
8. $\int_0^{\infty} \frac{x\, dx}{(4+x^2)^{3/2}}$
9. $\int_{\frac{4}{\pi}}^{\infty} x^{-2} \cos(x^{-1})\, dx$
10. $\int_1^{\infty} \frac{dx}{\sqrt{1+x^2}}$
    [Hint: $\int \sec\theta d\theta = \ln|\sec\theta + \tan\theta|$ may be useful, after making an inverse trigonometric substitution.]
11. $\int_{-\infty}^{\infty} \frac{dx}{\sqrt{1+x^2}}$
12. $\int_{-\infty}^{\infty} xe^{-x}\, dx$
13. Suppose p is a real constant. For what values of p does $\int_0^{\infty} \frac{1}{x^p}\, dx$ (a) converge and (b) diverge? For those values of p for which the integral converges, state the value to which it converges.
14. (a) Suppose that $\int_n^{n+1} f(x)dx = 1$ for every positive integer n, for a certain function $f(x)$. Say as much as you can about $\int_0^{\infty} f(x)dx$. Also, find a specific example of such a function f.
    (b) Suppose that $\int_n^{n+1} f(x)dx \geq \frac{1}{2}$ for every positive integer n, for a certain function $f(x)$. Say as much as you can about $\int_0^{\infty} f(x)dx$. Also, find a specific example of such a function f.
    (c) Suppose that $\left|\int_n^{n+1} f(x)dx\right| \geq \frac{1}{2}$ for every positive integer n, for a certain function $f(x)$. Say as much as you can about $\int_0^{\infty} f(x)dx$. Also, find a specific example of such a function f.
15. Explain why $f(0) = \int_0^1 xf''(x)\, dx$ for *all* functions $f(x)$ that are twice continuously differentiable on $[0,1]$ and satisfy $f(1) = 0$ and $f'(1) = 0$.
16. Explain why $1 + f(0) = \int_0^1 xf''(x)\, dx$ for *all* functions $f(x)$ that are twice continuously differentiable on $[0,1]$ and satisfy $f(1) = 0$ and $f'(1) = 1$.

7.2 Line Integrals

Suppose $\mathcal{C}$: $\mathbf{r} = \mathbf{r}(t), \alpha \leq t \leq \beta$, is a piecewise smooth parametrized* curve in the xy-plane. We can find its length by a limiting process on piecewise linear curves that approximate $\mathcal{C}$.

* The parameter, t, may or may not be the time variable in a physical problem.

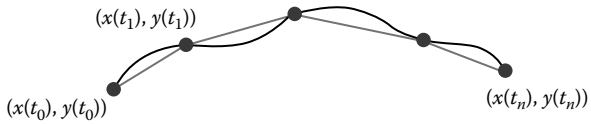


FIGURE 7.5
Piecewise linear approximation of a curve.

Let $\alpha = t_0 < t_1 < \ldots < t_n = \beta$ be a partition and $\triangle t_k = t_k - t_{k-1}$. For $k = 1$, we get

$$\mathbf{r}(t_1) - \mathbf{r}(t_0) = (x(t_1) - x(t_0))\,\hat{\boldsymbol{\imath}} + (y(t_1) - y(t_0))\hat{\boldsymbol{\jmath}} \approx \dot{x}(t_1)\ \triangle t_1\,\hat{\boldsymbol{\imath}} + \dot{y}(t_1)\ \triangle t_1\,\hat{\boldsymbol{\jmath}}.$$

In Figure 7.5, the curve $\mathcal{C}$ is in black, and connecting points on the curve are line segments.

The distance between $\mathbf{r}(t_1)$ and $\mathbf{r}(t_0)$ is

$$||\mathbf{r}(t_1) - \mathbf{r}(t_0)|| \approx \sqrt{(\dot{x}(t_1)^2 + \dot{y}(t_1)^2)(\triangle t_1)^2} = \sqrt{\dot{x}(t_1)^2 + \dot{y}(t_1)^2}\ |\triangle t_1| = ||\dot{\mathbf{r}}(t_1)||\ |\triangle t_1|.$$

An approximation for L, the total arclength of the curve $\mathcal{C}$, is given by

$$L \approx \sum_{k=1}^{n} ||\mathbf{r}(t_k) - \mathbf{r}(t_{k-1})|| = \sum_{k=1}^{n} ||\dot{\mathbf{r}}(t_k))||\ \triangle t_k,$$

after noting that $\triangle t_k > 0$. Take the limit, as both $n \to \infty$ and each $\triangle t_k \to 0$, to get the *exact* total arclength:

$$L = \int_{\alpha}^{\beta} ||\dot{\mathbf{r}}(t)||\ dt. \tag{7.3}$$

Physically, this makes sense because $||\dot{\mathbf{r}}(t)||$ is the speed of travel when t is the time.

Definition 7.6

The **arclength** function is defined by

$$s(t) \triangleq \int_{\alpha}^{t} ||\dot{\mathbf{r}}(u)||\ du.$$

By the fundamental theorem of calculus, specifically Theorem 7.4(a),

$$\frac{ds}{dt} = ||\dot{\mathbf{r}}(t)||. \tag{7.4}$$

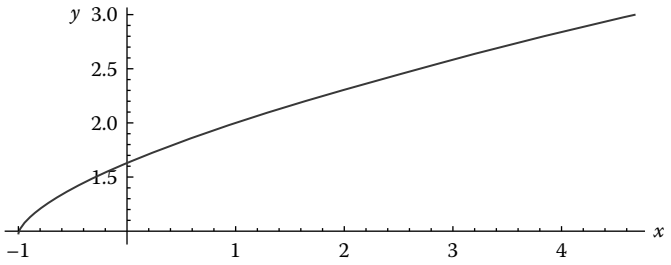


FIGURE 7.6
Example 7.4.

Using this, we have the following:

Definition 7.7

The **element of arclength** is defined by

$$ds \triangleq ||\dot{\mathbf{r}}(t)||\, dt.$$

Example 7.4

Find the total arclength of the curve $\mathcal{C} : \mathbf{r}(t) = (-1+2t^{3/2})\hat{\boldsymbol{\imath}} + (1+t)\hat{\boldsymbol{\jmath}}$, $0 \le t \le 2$, assuming distances are measured in cm.

Method: The curve is shown in Figure 7.6. Measuring t in seconds, $\dot{\mathbf{r}}(t) = (3t^{1/2}\hat{\boldsymbol{\imath}} + \hat{\boldsymbol{\jmath}})$ cm/s, so the speed is $||\dot{\mathbf{r}}(t)|| = \sqrt{(3t^{1/2})^2 + (1)^2}$. The total arclength of this curve is

$$L = \int_0^2 \sqrt{9t+1}\, dt = \int_1^{19} \sqrt{w} \cdot \frac{dw}{9} = \frac{1}{9} \cdot \left[\frac{w^{3/2}}{3/2} \right]_1^{19} = \frac{2}{27}\left(19\sqrt{19} - 1\right) \approx 6.06067 \text{ cm},$$

by using the method of substitution with $w = 9t + 1$. ◯

The curve shown in Figure 7.6 was drawn by the *Mathematica*™ command:

ParametricPlot[{−1 + 2t^(3/2), 1 + t}, {t, 0, 2},PlotStyle → Thick].

Example 7.5

A wire in the shape of the curve $\mathcal{C}$ of Example 7.4 has a linear mass density of $\varrho = f(x, y) \triangleq 0.05 + 0.01y$ in g/cm. Find the total mass of the wire.

Method: Approximate $\mathcal{C}$ by linear pieces, and assume that in each piece the density is approximately constant. The line segment in the kth piece, that is, from $\mathbf{r}(t_{k-1})$ to $\mathbf{r}(t_k)$, has length approximately $||\dot{\mathbf{r}}(t_k)||\, \triangle t_k = \sqrt{9t_k + 1}\, \triangle t_k$ cm. The mass of that piece is *length* × *density*, which is approximately

$$(||\dot{\mathbf{r}}(t_k)||\, \triangle t_k \,\cancel{\text{cm}}) \times (f(x(t_k), y(t_k))\text{g}/\cancel{\text{cm}}) = \sqrt{9t_k + 1}\, \triangle t_k \times (0.05 + 0.01y(t_k)) \text{ g}$$
$$= \sqrt{9t_k + 1} \cdot (0.05 + 0.01(1 + t_k))\, \triangle t_k \text{ g}.$$

Approximate the total mass of the wire by adding up the masses of the pieces:

$$M \approx \sum_{k=1}^{n} \sqrt{9t_k + 1} \cdot (0.05 + 0.01(1 + t_k)) \triangle t_k \text{ g.}$$

After that, take the limit, as both $n \to \infty$ and $\max\limits_{1 \le k \le n} |\triangle t_k| \to 0$, to get

$$\begin{aligned} M &= \int_0^2 \sqrt{9t + 1} \cdot (0.05 + 0.01(1 + t))dt = 0.06 \int_0^2 \sqrt{9t + 1}\, dt + 0.01 \int_0^2 t\sqrt{9t + 1}\, dt \\ &= \frac{0.12}{27}(19\sqrt{19} - 1) + 0.01 \int_1^{19} \frac{w - 1}{9} \sqrt{w} \cdot \frac{dw}{9} \\ &= \frac{0.12}{27}(19\sqrt{19} - 1) + \frac{0.01}{81} \int_1^{19} (w^{3/2} - w^{1/2})dw \end{aligned}$$

by again using the method of substitution with $w = 9t + 1$; hence, $t = \frac{w-1}{9}$. So the total mass of the wire is

$$\begin{aligned} M &= \frac{0.12}{27}(19\sqrt{19} - 1) + \frac{0.01}{81}\Big[\frac{2}{5}w^{5/2} - \frac{2}{3}w^{/2})\Big]_1^{19} \\ &= \frac{0.12}{27}(19\sqrt{19} - 1) + \frac{0.01}{81}\left(\frac{2}{5}(19^2\sqrt{19} - 1) - \frac{2}{3}(19\sqrt{19} - 1)\right) \approx 0.434564 \text{ g.} \bigcirc \end{aligned}$$

Example 7.5 motivates the definition that the total mass of a curve in the plane $\mathcal{C}: \mathbf{r} = \mathbf{r}(t),\ \alpha \le t \le \beta$, whose linear mass density is $\varrho = f(x(t), y(t))$, is

$$M \triangleq \int_\alpha^\beta f(x(t), y(t))||\dot{\mathbf{r}}(t)||\, dt.$$

In general, given a curve $\mathcal{C}: \mathbf{r} = \mathbf{r}(t),\ \alpha \le t \le \beta$, in xyz-space, we calculate its total arclength to be, as in (7.3),

$$L = \int_\alpha^\beta ||\dot{\mathbf{r}}(t)||\, dt$$

and its arclength function to be, as in Definition 7.6,

$$s(t) \triangleq \int_\alpha^t ||\dot{\mathbf{r}}(u)||\, du.$$

We calculate, as in (7.4),

$$\frac{ds}{dt} = ||\dot{\mathbf{r}}(t)||$$

and define, as in Definition 7.7, the element of arclength to be

$$ds \triangleq ||\dot{\mathbf{r}}(t)||dt.$$

A curve in the xy-plane is a special case where $\mathbf{r}(t) = x(t)\hat{\boldsymbol{\imath}} + y(t)\hat{\boldsymbol{\jmath}} + 0 \cdot \hat{\boldsymbol{k}}$.

Using these notations and defining

$$f(\mathbf{r}(t)) = f(x(t), y(t), z(t)),$$

we have

Definition 7.8

The **line integral** of a function is defined by

$$\int_{\mathcal{C}} f(\mathbf{r})\, ds \triangleq \int_{\alpha}^{\beta} f(\mathbf{r}(t))\, ||\dot{\mathbf{r}}(t)||\, dt. \tag{7.5}$$

For example, the total mass of a wire in space is the line integral of the density function $\varrho = f(\mathbf{r}(t))$, that is,

$$M = \int_{\mathcal{C}} \varrho\, ds \triangleq \int_{\alpha}^{\beta} f(\mathbf{r}(t))\, ||\dot{\mathbf{r}}(t)||\, dt.$$

Note that $||\dot{\mathbf{r}}(t)|| = \sqrt{(\dot{x}(t))^2 + (\dot{y}(t))^2 + (\dot{z}(t))^2}$.

Example 7.6

Set up the line integral of a function $f = f(x, y, z)$ over one turn of the helix $\mathcal{C}:\ \mathbf{r}(t) = (a\cos t)\hat{\boldsymbol{\imath}} + (a\sin t)\hat{\boldsymbol{\jmath}} + t\hat{\boldsymbol{k}}$.

The curve shown in Figure 7.7 was drawn by the *Mathematica* command:

ParametricPlot3D[{2Cos[t], 2Sin[t], t}, {t, 0, 2π}],PlotStyle → Thick].

Method: $\dot{\mathbf{r}}(t) = (-a\sin t)\hat{\boldsymbol{\imath}} + (a\cos t)\hat{\boldsymbol{\jmath}} + \hat{\boldsymbol{k}}$, and $0 \le t \le 2\pi$ gives one turn of the helix, so the line integral is

$$\int_{\mathcal{C}} f(\mathbf{r})\, ds = \int_{0}^{2\pi} f(a\cos t, a\sin t, t)\, \sqrt{a^2 + 1}\, dt. \bigcirc$$

Definition 7.9

The average value of a function f on a curve $\mathcal{C}$ is

$$\bar{f} \triangleq \frac{1}{\textit{Length of } \mathcal{C}} \int_{\mathcal{C}} f(\mathbf{r})\, ds = \frac{\int_{\mathcal{C}} f(\mathbf{r})\, ds}{\int_{\mathcal{C}} 1\, ds}.$$

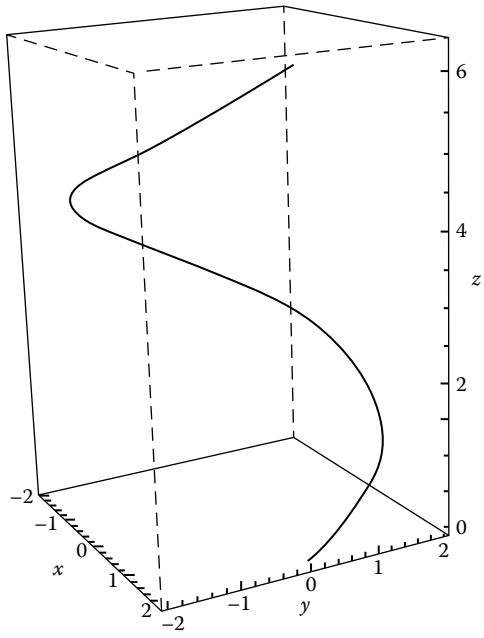


FIGURE 7.7
One turn of a helix with $a = 2$.

Besides the line integrals defined by (7.5) in Definition 7.8, we have the following:

Definition 7.10

$$\int_{\mathcal{C}} f(\mathbf{r})\, dx \triangleq \int_{\alpha}^{\beta} f(\mathbf{r}(t)) \left(\frac{dx}{dt}(t)\right) dt, \tag{7.6}$$

$$\int_{\mathcal{C}} f(\mathbf{r})\, dy \triangleq \int_{\alpha}^{\beta} f(\mathbf{r}(t)) \left(\frac{dy}{dt}(t)\right) dt, \tag{7.7}$$

and

$$\int_{\mathcal{C}} f(\mathbf{r})\, dz \triangleq \int_{\alpha}^{\beta} f(\mathbf{r}(t)) \left(\frac{dz}{dt}(t)\right) dt. \tag{7.8}$$

7.2.1 Line Integrals of Vector-Valued Functions

Suppose $\mathbf{F} = \mathbf{F}(x, y, z)$ is a vector field, that is, a vector-valued function. We define a **line integral of a vector field** by

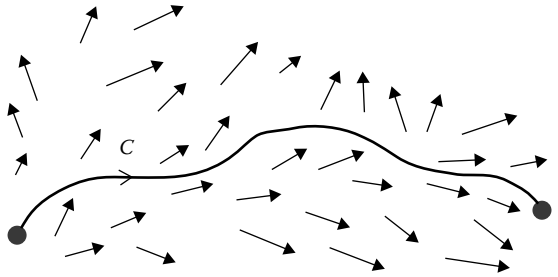


FIGURE 7.8
Curve in a vector field.

$$\int_{\mathcal{C}} \mathbf{F}(\mathbf{r}) \bullet d\mathbf{r} \triangleq \int_{\alpha}^{\beta} \mathbf{F}(\mathbf{r}(t)) \bullet \dot{\mathbf{r}}(t)\, dt. \tag{7.9}$$

For example, if $\mathbf{F}(\mathbf{r})$ is a force field, then the total **work**, W, done by that force on a particle is defined by

$$W \triangleq \int_{\mathcal{C}} \mathbf{F}(\mathbf{r}) \bullet d\mathbf{r}. \tag{7.10}$$

Taking the dot product of $\mathbf{F}$ with $\dot{\mathbf{r}}(t)$, that is, with $\frac{d\mathbf{r}}{dt}$, gives the component of $\mathbf{F}$ in the direction of travel on the curve, that is, in the tangential direction.

Shown in Figure 7.8 is a curve in a vector field.

Another way of writing a line integral of a vector field $\mathbf{F} = P\hat{\boldsymbol{\imath}} + Q\hat{\boldsymbol{\jmath}} + R\hat{\boldsymbol{k}}$ is

$$\int_{\mathcal{C}} \mathbf{F}(\mathbf{r}) \bullet d\mathbf{r} = \int_{\mathcal{C}} (P dx + Q dy + R dz).$$

A simple, closed, parametrized curve in the xy-plane is said to be **positively oriented** if, while standing in the xy-plane with head in the $\hat{\boldsymbol{k}}$ direction, the inside of the curve is on our left as we travel on the curve in the direction of its parametrization. While this definition is admittedly imprecise, it will suffice for our needs.

If $\mathcal{C}$ is a simple, closed, piecewise smooth curve, we notate line integrals on $\mathcal{C}$ by $\oint_{\mathcal{C}}$ instead of $\int_{\mathcal{C}}$. The little arrow pointing in the counterclockwise direction in the notation

$$\oint_{\mathcal{C}}$$

indicates that the parametrized curve $\mathcal{C}$ is positively oriented.

If $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2$ is a **chain**, as illustrated in Figure 7.9, then we define

$$\int_{\mathcal{C}} \ldots \triangleq \int_{\mathcal{C}_1} \ldots + \int_{\mathcal{C}_2} \ldots .$$

There is a similar definition for a chain of any finite number of curves.

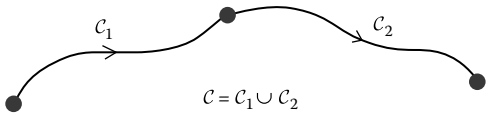


FIGURE 7.9
Chain of two curves.

As a second example, if $\mathbf{v} = \mathbf{v}(x, y, z)$ is a velocity vector field of fluid particles (from the "Eulerian" point of view) and $\mathcal{C}$ is a simple, closed, positively oriented, piecewise smooth curve, then the **circulation** of $\mathbf{v}$ around $\mathcal{C}$ is defined to be

$$\Gamma \triangleq \oint_{\mathcal{C}} \mathbf{v}(\mathbf{r}) \bullet d\mathbf{r}. \tag{7.11}$$

Circulation is related to the concept of **lift** in aerodynamics by

$$Lift \triangleq \varrho_0 \, \Gamma, \tag{7.12}$$

where ϱ_0 is the gas (constant) mass density.

Example 7.7

Find the total circulation around the triangle with "nose up" shown in Figure 7.10, assuming the wind has constant speed U_0.

Method: Let φ be the constant "pitch" angle, so

$$\mathbf{v} = U_0(\cos\varphi\, \hat{\boldsymbol{\imath}} + \sin\varphi\, \hat{\boldsymbol{\jmath}}).$$

Write $\mathcal{C}$ as a chain of three curves as shown in Figure 7.10. Parametrize the curves by

$$\mathcal{C}_1 : \ \mathbf{r}_1 = (1-t)\hat{\boldsymbol{\imath}}, \ 0 \le t \le 1.$$
$$\mathcal{C}_2 : \ \mathbf{r}_2 = t\hat{\boldsymbol{\imath}} + ht\hat{\boldsymbol{\jmath}}, \ 0 \le t \le 1.$$
$$\mathcal{C}_3 : \ \mathbf{r}_3 = (1-t)h\hat{\boldsymbol{\jmath}}, \ 0 \le t \le 1.$$

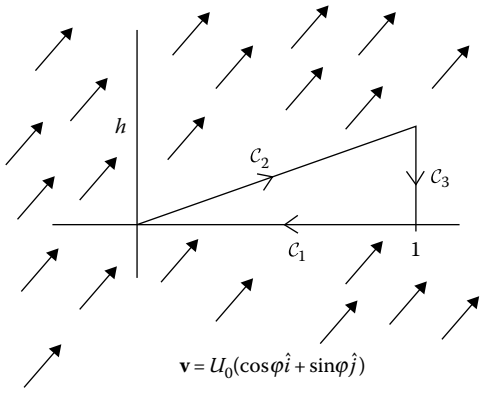


FIGURE 7.10
Triangle nose up in airflow.

We calculate

$$\begin{aligned}\Gamma &= \oint_{\mathcal{C}} \mathbf{v}(\mathbf{r}) \bullet d\mathbf{r} = \int_{\mathcal{C}_1} \mathbf{v}(\mathbf{r}_1) \bullet d\mathbf{r}_1 + \int_{\mathcal{C}_2} \mathbf{v}(\mathbf{r}_2) \bullet d\mathbf{r}_2 + \int_{\mathcal{C}_3} \mathbf{v}(\mathbf{r}_3) \bullet d\mathbf{r}_3 \\ &= \int_0^1 U_0(\cos\varphi \hat{\boldsymbol{\imath}} + \sin\varphi \hat{\boldsymbol{\jmath}}) \bullet (-\hat{\boldsymbol{\imath}}) dt + \int_0^1 U_0(\cos\varphi \hat{\boldsymbol{\imath}} + \sin\varphi \hat{\boldsymbol{\jmath}}) \bullet (\hat{\boldsymbol{\imath}} + h\hat{\boldsymbol{\jmath}}) dt \\ &\quad + \int_0^1 U_0(\cos\varphi \hat{\boldsymbol{\imath}} + \sin\varphi \hat{\boldsymbol{\jmath}}) \bullet (-h\hat{\boldsymbol{\jmath}}) dt \\ &= \int_0^1 U_0(-\cos\varphi) dt + \int_0^1 U_0(\cos\varphi + h\sin\varphi) dt + \int_0^1 U_0(-h\sin\varphi) dt \\ &= -U_0\cos\varphi + U_0(\cos\varphi + h\sin\varphi) + U_0(-h\sin\varphi) = 0.\end{aligned}$$

So, the lift is zero. ◯

7.2.2 Fundamental Theorem of Line Integrals

This result is both a generalization of the fundamental theorem of calculus, specifically in the form of Theorem 7.4, and explained by it!

Theorem 7.8

Suppose $\mathcal{C}$ is a piecewise smooth curve in a domain $\mathcal{D}$ and $\mathbf{F} = \nabla f$ in $\mathcal{D}$. Then

$$\int_{\mathcal{C}} \mathbf{F}(\mathbf{r}) \bullet d\mathbf{r} = f\big(\mathbf{r}(\beta)\big) - f\big(\mathbf{r}(\alpha)\big), \tag{7.13}$$

where $\mathcal{C}$ is parametrized by t in the interval $[\alpha, \beta]$, as depicted in Figure 7.11.

Why?

$$\int_{\mathcal{C}} \mathbf{F}(\mathbf{r}) \bullet d\mathbf{r} \triangleq \int_{\alpha}^{\beta} \mathbf{F}\big(\mathbf{r}(t)\big) \bullet \dot{\mathbf{r}}(t) dt = \int_{\alpha}^{\beta} (\nabla f)\big(\mathbf{r}(t)\big) \bullet \dot{\mathbf{r}}(t) dt.$$

By (6.42) in Section 6.4, the multivariable chain rule, and Theorem 7.4(b), this gives

$$\int_{\mathcal{C}} \mathbf{F}(\mathbf{r}) \bullet d\mathbf{r} = \int_{\alpha}^{\beta} \left(\frac{d}{dt} \big[f(\mathbf{r}(t)) \big] \right) dt = \Big[f(\mathbf{r}(t)) \Big]_{\alpha}^{\beta} = f(\mathbf{r}(\beta)) - f(\mathbf{r}(\alpha)). \; \square$$

A consequence of Theorem 7.8 is that for an exact force field, the work done does not depend on the path taken, that is, the work done is **path independent**.

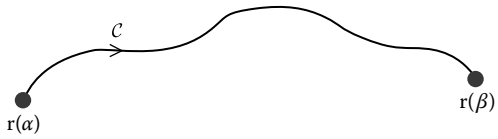


FIGURE 7.11
Fundamental theorem of line integrals.

Example 7.8

Suppose $\mathbf{F} = \frac{k}{||\mathbf{r}||^3}\,\mathbf{r}$, that is, is an inverse square law force field with proportionality constant k, for example, a gravitational or electric force field, due to an object at the origin. If $\mathcal{C}$ is any piecewise smooth curve not passing through the origin, then

$$\int_{\mathcal{C}} \mathbf{F}(\mathbf{r}) \bullet d\mathbf{r} = -\frac{k}{||\mathbf{r}(\beta)||} + \frac{k}{||\mathbf{r}(\alpha)||},$$

where the curve starts at $\mathbf{r}(\alpha)$ and ends at $\mathbf{r}(\beta)$.

Why? $\frac{k}{||\mathbf{r}||^3}\,\mathbf{r} = \nabla\left[-\frac{k}{||\mathbf{r}||}\right]$. ○

The result in Example 7.8 implies that the work done in going from $\mathbf{r}(\beta)$ to $\mathbf{r}(\alpha)$ is the negative of the work done in going from $\mathbf{r}(\alpha)$ to $\mathbf{r}(\beta)$. This is why we call a vector field **conservative** if it is the gradient of a potential function, such as $\mathbf{F} = \nabla\left[-\frac{k}{||\mathbf{r}||}\right]$.

Corollary 7.1

If $\mathcal{C}$ is a simple, closed, piecewise smooth curve in a domain $\mathcal{D}$ in which $\mathbf{F}$ is exact, then $\oint_{\mathcal{C}} \mathbf{F} \bullet d\mathbf{r} = 0$.

Why? By Theorem 7.8, if $\mathcal{C}$ is parametrized by t in the interval $[\alpha, \beta]$ and $\mathbf{F} = \nabla f$ in $\mathcal{D}$, then $\oint_{\mathcal{C}} \mathbf{F} \bullet d\mathbf{r} = f\big(\mathbf{r}(\beta)\big) - f\big(\mathbf{r}(\alpha)\big) = 0$, because $\mathbf{r}(\beta) = \mathbf{r}(\alpha)$. □

Figure 7.12 depicts an example of a simple, closed, piecewise smooth curve.

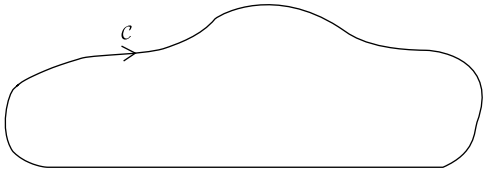


FIGURE 7.12
Simple, closed, piecewise smooth curve.

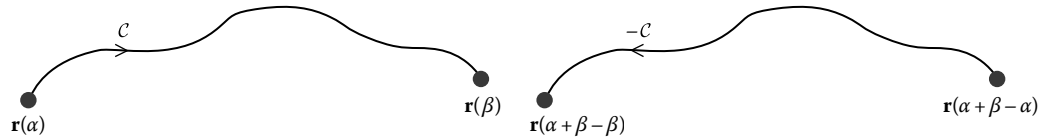


FIGURE 7.13
Opposite curve.

7.2.3 Path Direction

Given a piecewise smooth parametrized curve $\mathcal{C}:\ \mathbf{r}=\mathbf{r}(t),\ \alpha \le t \le \beta$, we can define the **opposite curve** by

$$-\mathcal{C}:\ \mathbf{r}=\mathbf{r}(\alpha+\beta-t),\ \alpha \le t \le \beta. \tag{7.14}$$

The curve $-\mathcal{C}$ consists of the same points as $\mathcal{C}$ but is traversed in the opposite direction, as shown in Figure 7.13.

Theorem 7.9

Suppose $\mathcal{C}$ is a piecewise smooth curve. Then

$$\int_{-\mathcal{C}} f(\mathbf{r})ds = \int_{\mathcal{C}} f(\mathbf{r})ds, \tag{7.15}$$

$$\int_{-\mathcal{C}} f(\mathbf{r})dx = -\int_{\mathcal{C}} f(\mathbf{r})dx, \tag{7.16}$$

and

$$\int_{-\mathcal{C}} \mathbf{F}(\mathbf{r}) \bullet d\mathbf{r} = -\int_{\mathcal{C}} \mathbf{F}(\mathbf{r}) \bullet d\mathbf{r}. \tag{7.17}$$

Why? For example, the substitution $w=\alpha+\beta-t$ gives

$$\int_{-\mathcal{C}} \mathbf{F}(\mathbf{r}) \bullet d\mathbf{r} \triangleq \int_{\alpha}^{\beta} \mathbf{F}(\mathbf{r}(\alpha+\beta-t)) \bullet \dot{\mathbf{r}}(\alpha+\beta-t)dt = \int_{\beta}^{\alpha} \mathbf{F}(\mathbf{r}(w)) \bullet \dot{\mathbf{r}}(w)(-dw)$$

$$= -\int_{\alpha}^{\beta} \mathbf{F}(\mathbf{r}(w)) \bullet \dot{\mathbf{r}}(w)dw = -\int_{\mathcal{C}} \mathbf{F}(\mathbf{r}) \bullet d\mathbf{r}. \ \square$$

Physically, (7.17) makes sense: The work against a force field along the opposite path should be the negative of the work done on the original path.

7.2.4 Other Notations

A parametrized curve $\mathcal{C}:\ \mathbf{r}=\mathbf{r}(t),\ \alpha \le t \le \beta$, has tangent vector

$$\mathbf{T} = \dot{\mathbf{r}}(t)$$

wherever it is differentiable. Where the tangent vector is not zero, the curve has unit tangent vector:

$$\widehat{\mathbf{T}} \triangleq \frac{1}{||\dot{\mathbf{r}}(t)||}\,\dot{\mathbf{r}}(t) = \frac{1}{\frac{ds}{dt}}\,\dot{\mathbf{r}}(t).$$

It follows that

$$\int_{\mathcal{C}} \mathbf{F}(\mathbf{r}) \bullet d\mathbf{r} = \int_{\mathcal{C}} \left(\mathbf{F}(\mathbf{r}) \bullet \widehat{\mathbf{T}}\right) ds, \tag{7.18}$$

because

$$\mathbf{F}(\mathbf{r}) \bullet d\mathbf{r} \triangleq \mathbf{F}(\mathbf{r}) \bullet (\dot{\mathbf{r}})dt = \mathbf{F}(\mathbf{r}) \bullet \left(\frac{ds}{dt}\,\widehat{\mathbf{T}}\right) dt = \left(\mathbf{F}(\mathbf{r}) \bullet \widehat{\mathbf{T}}\right) \left(\frac{ds}{dt}\right) dt \triangleq \left(\mathbf{F}(\mathbf{r}) \bullet \widehat{\mathbf{T}}\right) ds.$$

A third example of a line integral of a vector field is if **E** is an electrostatic field, that is, the field due to some stationary electric charges, measured in newtons/coulomb. Because **E** is the sum of inverse square law vector fields, it is exact in any domain $\mathcal{D}$ that contains none of the charges. Given two points P_1, P_2, neither of which is at any of the charges,

$$\int_{\mathcal{C}} \mathbf{E} \bullet d\mathbf{r}$$

is the same for all piecewise smooth curves $\mathcal{C}$ that remain in such a charge-free domain $\mathcal{D}$, start at P_1 and end at P_2. So, we can define

$$\varphi(P_1, P_2) \triangleq -\int_{P_1}^{P_2} \mathbf{E} \bullet d\mathbf{r} \triangleq -\int_{\mathcal{C}} \mathbf{E} \bullet d\mathbf{r}$$

for any such curve $\mathcal{C}$. Measured in volts, that is, joules/coulomb, $\varphi(P_1, P_2)$ is called the **electrostatic potential difference** between P_1 and P_2. Moving a charge of one coulomb through a potential difference of one volt requires one joule of work.

Suppose $\mathcal{C}$ is a simple, closed, piecewise smooth curve bounding a planar region $\mathcal{D}$. Let **B** be the magnetic flux density and let I be the current through $\mathcal{D}$. Then in a vacuum, Ampère's circuital law states that

$$\oint_{\mathcal{C}} \mathbf{B} \bullet d\mathbf{r} = \mu_0 I.$$

The **magnetic permeability** is $\mu_0 = 4\pi \times 10^{-7}\,\mathrm{m\,kg\,s^{-2}A^{-2}} = \frac{4\pi}{c}$, where c is the speed of light in a vacuum, lengths are measured in meters (m), A stands for ampères, and I is measured in ampères.

Example 7.9

An infinitely long conducting wire of radius a has circular cross sections. Find the magnetic flux density $\mathbf{B}$.

Method: Define the circle

$$\mathcal{C}: \ \mathbf{r} = a\,\hat{\mathbf{e}}_{\mathbf{r}} = a(\cos\theta\,\hat{\boldsymbol{\imath}} + \sin\theta\,\hat{\boldsymbol{\jmath}}),\ 0 \le \theta \le 2\pi;$$

hence, $\frac{d\mathbf{r}}{d\theta} = a\,\hat{\mathbf{e}}_\theta$. Assuming the wire has the z-axis as its longitudinal axis, the magnetic field is $\mathbf{B} = B_0\,\hat{\mathbf{e}}_\theta$, where B_0 is a constant. We calculate

$$\mu_0 I = \oint_{\mathcal{C}} \mathbf{B} \bullet d\mathbf{r} = \int_0^{2\pi} (B_0\,\hat{\mathbf{e}}_\theta) \bullet (a\,\hat{\mathbf{e}}_\theta) d\theta = 2\pi a B_0.$$

So

$$B_0 = \frac{\mu_0 I}{2\pi a},$$

and thus that

$$\mathbf{B} = \left(\frac{\mu_0 I}{2\pi a}\right)\hat{\mathbf{e}}_\theta. \ \bigcirc$$

Example 7.10

Explain why the work done by a force $\mathbf{F}$ on a particle equals the net change in kinetic energy (KE).

Method: Newton's second law of motion says in general that

$$m\,\ddot{\mathbf{r}} = \mathbf{F},$$

so the work done by the force is

$$W = \int_{\mathcal{C}} \mathbf{F} \bullet d\mathbf{r} \triangleq \int_\alpha^\beta \mathbf{F}(\mathbf{r}) \bullet \dot{\mathbf{r}} dt = \int_\alpha^\beta m\ddot{\mathbf{r}} \bullet \dot{\mathbf{r}} dt = \int_\alpha^\beta \frac{1}{2}\, m\, \frac{d}{dt}\Big[\,||\dot{\mathbf{r}}||^2\,\Big] dt = \Big[\frac{1}{2}\, m\, ||\dot{\mathbf{r}}||^2\,\Big]_\alpha^\beta$$
$$= \frac{1}{2}\, m\, ||\dot{\mathbf{r}}(\beta)||^2 - \frac{1}{2}\, m\, ||\dot{\mathbf{r}}(\alpha)||^2.$$

So, the total work done equals the net *change* of the KE. $\bigcirc$

7.2.5 Problems

1. (a) If $\mathcal{C}$ is a straight line segment, parametrize it, calculate $\int_{\mathcal{C}} ds$, and find that it gives the length of the line segment.
   (b) If $\mathcal{C}$ is an arc of a circle of radius a from θ_0 to θ_1, parametrize it, calculate $\int_{\mathcal{C}} ds$, and find that it gives the length of the arc, that is, $a \cdot (\theta_1 - \theta_0)$.
2. A fly's position vector is $\mathbf{r}(t) = -\cos 2t\,\hat{\boldsymbol{\imath}} + \sin 2t\,\hat{\boldsymbol{\jmath}} + t\,\hat{\mathbf{k}}$. Find the exact distance traveled by the fly, that is, the arclength, when it travels from the point $(-1, 0, 0)$ to the point $(-1, 0, 2\pi)$.
3. Find the exact arclength of the curve parametrized by $\mathbf{r}(t) = \frac{1}{2}\,t^2\,\hat{\boldsymbol{\imath}} + \sqrt{6}\,t\,\hat{\boldsymbol{\jmath}} + 3\ln(t)\,\hat{\mathbf{k}}$, $1 \le t \le e$.

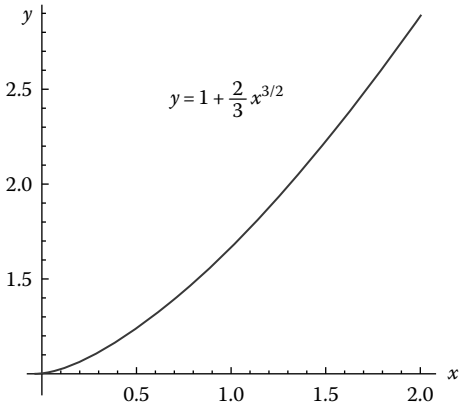


FIGURE 7.14
Problem 7.2.5.6.

4. Find the exact arclength of the curve $y = \ln(\cos x)$ over the interval $0 \le x \le \frac{\pi}{4}$.
5. A helical curve $\mathcal{C}$ is given by $\mathbf{r}(t) = \cos 2t\,\hat{\boldsymbol{\imath}} + \sin 2t\,\hat{\boldsymbol{\jmath}} + t\,\hat{\mathbf{k}}$ for $0 \le t \le \frac{\pi}{4}$. Evaluate $\int_{\mathcal{C}} x\,z\,ds$.
6. Set up, with as much specificity as possible, but do not evaluate, a definite integral that would give the value of $\int_{\mathcal{C}} x\,ds$, where the curve $\mathcal{C}$ is shown in Figure 7.14. Your final conclusion should be in the form $\int\_\_\_\_\_ d\_\_$ with blanks filled in.
7. A wire bent in the shape of the quarter circle $x^2 + y^2 = a^2$ in the first quadrant is made of metal whose density is $\varrho = (1 + x^2 + y^2 - xy)$ g/m. Find the exact center of mass of the wire.
8. A wire bent in the shape of the half circle $x^2 + y^2 = 3$, $y \ge 0$ is made of a material whose density is $\varrho = (1 + \cos^2 x)y$ g/m. Find the total mass of the wire.
9. A wire bent in the shape of the quarter circle $x^2 + y^2 = a^2$ in the first quadrant is made of metal of varying composition whose linearly varying density is $\varrho = 1 + 0.2\theta$ g/m. Find the exact center of mass of the wire.
10. Let $\mathcal{C}$ be the semicircle that runs from $(a, 0)$ to $(-a, 0)$ and $\mathbf{F} = e^x\,\hat{\boldsymbol{\imath}} - y\,\hat{\boldsymbol{\jmath}}$. Find the value of $\int_{\mathcal{C}} \mathbf{F} \bullet d\mathbf{r}$.
11. For the curve shown in Figure 7.15a, evaluate (a) $\int_{\mathcal{C}} xy^2\,ds$, (b) $\int_{\mathcal{C}} xy^2\,dx$, and (c) $\int_{\mathcal{C}} \nabla(xy^2) \bullet d\mathbf{r}$.
12. Find the circulation of the velocity field $\mathbf{v} = \hat{\mathbf{e}}_r - r\,\hat{\mathbf{e}}_\theta$ around the circle $x^2 + y^2 = 4$.
13. Find the circulation of $\mathbf{v} = y\,\hat{\boldsymbol{\imath}} - x\,\hat{\boldsymbol{\jmath}} + xy\,\hat{\mathbf{k}}$ around the half disk $x^2 + y^2 = a^2$, $x \ge 0$.
14. (a) Explain why the circulation around the circle $\mathcal{C} : \mathbf{r} = a(\cos\theta\,\hat{\boldsymbol{\imath}} + \sin\theta\,\hat{\boldsymbol{\jmath}})$, $0 \le \theta \le 2\pi$, of a planar vector field $\mathbf{F}(r, \theta) \triangleq F_r\hat{\mathbf{e}}_{\mathbf{r}} + F_\theta\hat{\mathbf{e}}_\theta$ is $\Gamma = a\int_0^{2\pi} F_\theta(a, \theta)\,d\theta$.
    (b) A streaming motion past a circular cylinder $r = a$ may have fluid velocity

$$\mathbf{v} = U\left(1 - \frac{a^2}{r^2}\right)\cos\theta\,\hat{\mathbf{e}}_{\mathbf{r}} + U\left(\frac{\kappa}{r} + (-1 - \frac{a^2}{r^2})\sin\theta\right)\hat{\mathbf{e}}_\theta,$$

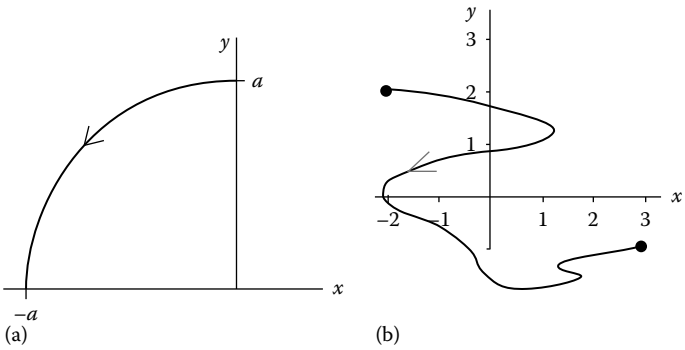


FIGURE 7.15
(a) Problem 7.2.5.11, (b) Problem 7.2.5.18.

where κ is a constant. Explain why the circulation around the cylinder is $2\pi\kappa U$. [By the way, the velocity vector "at infinity" is given by $V_\infty = U\hat{\boldsymbol{\imath}}$.]

15. (a) Find a function $f(x, y, z)$ such that

$$\nabla f = \mathbf{F} \triangleq y^2\,\hat{\boldsymbol{\imath}} + (2xy + e^{3z})\hat{\boldsymbol{\jmath}} + (3ye^{3z} + 1)\hat{\boldsymbol{k}}.$$

    (b) Find the value of $\int_{\mathcal{C}} \mathbf{F} \bullet d\mathbf{r}$, where $\mathcal{C}$ is an unspecified piecewise smooth parametrized curve from the point $(x, y, z) = (1, 0, 1)$ to the point $(0, 1, 1)$.

16. Find the work done by the force field $\mathbf{F} = (1 + 4x^3y^3)\,\hat{\boldsymbol{\imath}} + 3x^4y^2\,\hat{\boldsymbol{\jmath}}$ to move a particle from the point $(x, y) = (-1, 2)$ to the point $(1, 0)$.

17. Find the work done against the gravitational field $-\frac{mMG}{||\mathbf{r}||^3}\,\mathbf{r}$ along an unspecified piecewise smooth parametrized curve from the point $(x, y, z) = (-2, 1, 1)$ to the point $(2, 3, -1)$.

18. Evaluate $\int_{\mathcal{C}} (y\,\hat{\boldsymbol{\imath}} + (x + 2)\,\hat{\boldsymbol{\jmath}}) \bullet d\mathbf{r}$, where $\mathcal{C}$ is shown in Figure 7.15b.

19. Evaluate $\int_{\mathcal{C}} \left(\frac{x}{x^2+y^2}\,\hat{\boldsymbol{\imath}} + \frac{y}{x^2+y^2}\,\hat{\boldsymbol{\jmath}}\right) \bullet d\mathbf{r}$, where $\mathcal{C}$ is shown in Figure 7.16.

20. Evaluate $\int_{\mathcal{C}} \left(-\frac{y}{x^2+y^2}\,\hat{\boldsymbol{\imath}} + \frac{x}{x^2+y^2}\,\hat{\boldsymbol{\jmath}}\right) \bullet d\mathbf{r}$, where $\mathcal{C}$ is shown in Figure 7.16. [Hint: The function $\arctan\left(\frac{y}{x}\right)$ is undefined at $x = 0$, and the function $\arctan\left(\frac{x}{y}\right)$ is undefined at $y = 0$. But there is a trigonometric identity that $\arctan a + \arctan\left(\frac{1}{a}\right) \equiv \frac{\pi}{2}$. So, we can define a function $\phi(x, y) \triangleq \begin{Bmatrix} -\arctan(x/y), & y \neq 0 \\ \arctan(y/x) - \frac{\pi}{2}, & x \neq 0 \end{Bmatrix}$.]

21. The bob of a pendulum, of length ℓ, starts at angular position $\varphi = \alpha$ and travels to $\varphi = 0$, as depicted in Figure 7.17. The bob has weight w, so the force of gravity is always $\mathbf{F} = -\,w\,\hat{\boldsymbol{\jmath}}$. What is the work done by gravity during the travel of the bob, assuming the pendulum arm is of negligible weight? [Hint: What is the relationship between φ and the usual angle θ in polar coordinates?]

22. Use the method of substitution for integration of a function of a single variable to explain why line integration with respect to arclength does not depend on the parametrization of the curve. Specifically, if one curve has two piecewise smooth

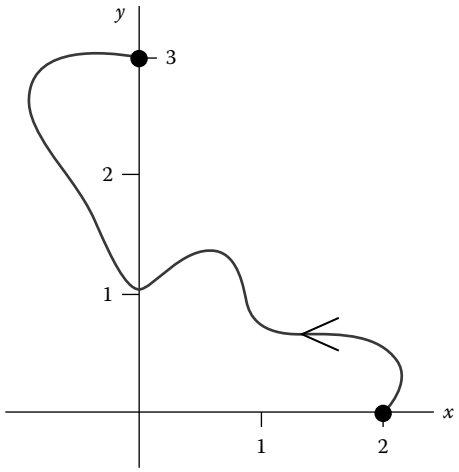


FIGURE 7.16
Problems 7.2.5.19 and 7.2.5.20.

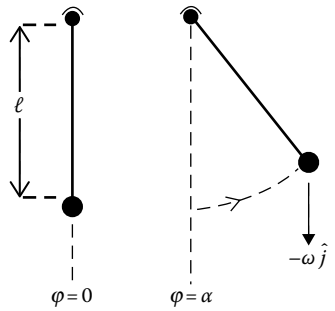


FIGURE 7.17
Problem 7.2.5.21.

parametrizations $\mathcal{C}_1 :\ \mathbf{r} = \mathbf{r}_1(t),\ \alpha_1 \le t \le \beta_1$, $\mathcal{C}_1 :\ \mathbf{r} = \mathbf{r}_2(t),\ \alpha_2 \le t \le \beta_2$ for which there is a piecewise smooth, increasing function $w = w(t)$ so that $\mathbf{r}_2(w(t)) \equiv \mathbf{r}_1(t)$, then for all continuous functions $f = f(x, y, z)$,

$$\int_{\mathcal{C}_1} f(x, y, z)\, ds = \int_{\mathcal{C}_2} f(x, y, z)\, ds.$$

Note that, in particular, $w(\alpha_1) = \alpha_2$ and $w(\beta_1) = \beta_2$.

7.3 Double Integrals, Green's Theorem, and Applications

Suppose mass is painted onto a planar region. In this section we will see how the concept of "double integral" enables us to find the total mass on that region, its moments, and its

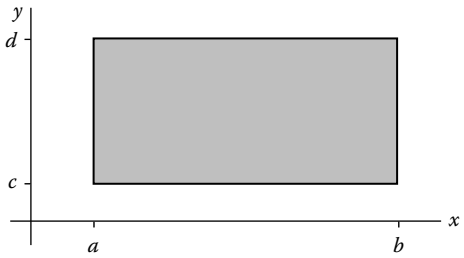


FIGURE 7.18
Rectangle.

moments of inertia. We can do this even if the mass density, ϱ, is a function of the x and y coordinates.

In this chapter, we will also see how to use polar coordinates to more easily work with a planar region that has circular geometry, for example, a disk, central sector of a disk, annulus, or sector of an annulus. We will see how many of the concepts and results of calculus for functions of a single variable generalize to functions of two or three variables.

Suppose the mass density is $\varrho = \varrho(x, y)$ and is given for (x, y) in the rectangle (Figure 7.18)

$$\mathcal{D} \triangleq \{(x, y) : \ a \le x \le b, \ c \le y \le d\}. \tag{7.19}$$

For example, mass may be painted on with a variable thickness or with a variable mixture of two materials with different densities. Break up the rectangle into many small subrectangles and assume the mass density is close to constant on each subrectangle. We can approximate the total mass on the rectangle by

$$M \approx \sum_{i=1}^{m} \sum_{j=1}^{n} \varrho(x_i^*, y_j^*) \ \triangle \ x_i \ \triangle \ y_j, \tag{7.20}$$

where (x_i^*, y_j) is a sampling point chosen in the ijth subrectangle:

$$\{(x, y) : \ x_{i-1} \le x \le x_i, \ y_{j-1} \le y \le y_j\}.$$

In (7.20), the total mass is approximated by a **double Riemann sum**.

Take the limit as both $m, n \to \infty$ and each $\triangle \ x_i, \triangle \ y_j \to 0$ to get *exact* total mass

$$M = \iint_{\mathcal{D}} \varrho(x, y) dx dy. \tag{7.21}$$

Three other notations for this are

$$M = \iint_{\mathcal{D}} \varrho(x, y) dA,$$

where dA is an **element of area**,

$$M = \int_c^d \int_a^b \varrho(x,y)dxdy,$$

and

$$M = \int_a^b \int_c^d \varrho(x,y)dydx.$$

Notice that changing the order of limits of integration from having x on the inside to having x on the outside corresponds to the element of area changing from $dxdy$ to $dydx$.

If ϱ has units of mass per unit area, then in approximation (7.20), M has units of mass because the terms in the sum have the form

$$\varrho(x_i^*, y_j^*)\left(\frac{mass}{\cancel{area}}\right) \triangle x_i(\cancel{length}) \triangle y_j(\cancel{length}).$$

As a practical matter, none of these notations tell us how to calculate the value of a double integral on a rectangle. For that, we have the following

Theorem 7.10

Suppose $\varrho(x,y)$ is continuous on the rectangle $\mathcal{D}$ given in (7.19). Then

$$M = \int_c^d \int_a^b \varrho(x,y)dxdy = \int_c^d \left(\int_a^b \varrho(x,y)dx \right) dy, \tag{7.22}$$

which is called an **iterated integral**. The latter can be evaluated using a generalization of the fundamental theorem of calculus: The inner integral, $\int_a^b \varrho(x,y)dx$, is the function of y given by

$$\int_a^b \varrho(x,y)dx = F(b,y) - F(a,y),$$

where $F(x,y)$ is any function that satisfies $\frac{\partial F}{\partial x} = \varrho(x,y)$.

The iterated integration process described in Theorem 7.10 corresponds to first integrating the mass on horizontal line segments from $(x,y) = (a,y)$ to $(x,y) = (b,y)$ and then "summing up" by integrating from $y = c$ to $y = d$.

As in Example 6.23 in Section 6.4, a function $F(x,y)$ that satisfies $\frac{\partial F}{\partial x} = \varrho(x,y)$ is called an **anti-partial derivative** with respect to x. We denote

$$F(x,y) = \int \varrho(x,y)\partial x$$

although other authors write "dx" and not "∂x."

Example 7.11

If the mass density in μg per cm^2 is $\varrho(x,y) = 1 + 0.5x + 0.2y$, find the total mass on the rectangle $0 \le x \le 1$ cm, $0 \le y \le 2$ cm depicted in Figure 7.19.

Method: The total mass is

$$M = \iint_{\mathcal{D}} \varrho dA \ \ \mu g/(\cancel{cm^2}) \times \cancel{cm^2} = \int_0^2 \int_0^1 \varrho(x,y)dxdy \ \mu g = \int_0^2 \left(\int_0^1 (1 + 0.5x + 0.2y)dx \right) dy \ \ \mu g.$$

For the inner integration, anti-partial differentiation gives

$$F(x,y) = \int_0^1 \varrho(x,y)\partial x = \int_0^1 (1 + 0.5x + 0.2y)\partial x = \Big[x + 0.25x^2 + 0.2yx\Big]_0^1$$
$$= \big(1 + 0.25 + 0.2y\big) - (0 + 0 + 0) = 1.25 + 0.2y.$$

So,

$$M = \int_0^2 \left(\int_0^1 (1 + 0.5x + 0.2y)dx \right) dy \ \mu g = \int_0^2 (1.25 + 0.2y)dy \ \mu g = \Big[1.25y + 0.1y^2\Big]_0^2 \ \mu g$$
$$= 2.5 + 0.4 - (0 + 0) \ \ \mu g = 2.9 \ \mu g. \ \bigcirc$$

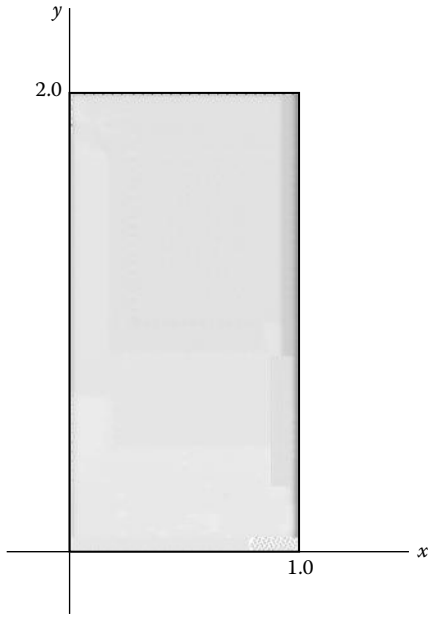


FIGURE 7.19
Variable mass density on a rectangle.

In Example 7.11, we could have integrated first with respect to y:

$$M = \int_0^1 \left(\int_0^2 (1 + 0.5x + 0.2y)dy \right) dx \ \mu g = \int_0^1 \left[y + 0.5xy + 0.1y^2 \right]_0^2 dx \ \mu g = \int_0^1 (2.4 + x)\, dx \ \mu g$$

$$= \left[2.4x + 0.5x^2 \right]_0^1 \ \mu g = 2.9 \ \mu g.$$

It is not a coincidence that we got the same final conclusion by two different orders of integration.

Theorem 7.11

(Fubini's theorem) If $f(x, y)$ is continuous on the rectangle $\mathcal{D}$ given in (7.19), then

$$\int_c^d \left(\int_a^b f(x, y)dx \right) dy = \int_a^b \left(\int_c^d f(x, y)dy \right) dx.$$

7.3.1 Double Integral as Volume

Recall from Section 7.1 that a single integral of a function of a single variable can be interpreted as the area of a region. Similarly, if $f(x, y) \geq 0$ for all (x, y) in a planar region $\mathcal{D}$, then the solid in (x, y, z)-space that is bounded on the top by the surface $z = f(x, y)$ and is bounded on the bottom by the region $\mathcal{D}$ in the plane $z = 0$ has

$$Volume = \iint_{\mathcal{D}} f(x, y)dA.$$

For example, we can recycle Example 7.11 to get the next example.

Example 7.12

Find the total volume of the solid bounded on the top by the plane $z = 1 + 0.5x + 0.2y$ and on the bottom by the rectangle $0 \leq x \leq 1$ cm, $0 \leq y \leq 2$ cm, $z = 0$. Assume z is measured in centimeters, too.

Method: The volume is

$$\text{Volume} = \int_0^2 \left(\int_0^1 (1 + 0.5x + 0.2y)dx \right) dy \ \text{cm}^3 = 2.9 \ \text{cm}^3,$$

using the integration result of Example 7.11. ○

Why are the units cm^3? Because an approximation of the volume is given by a double Riemann sum,

$$M \approx \sum_{i=1}^{m} \sum_{j=1}^{n} (\text{volume of thin parallelepiped}),$$

where the parallelepipeds have base area $(\triangle x \text{ cm})(\triangle y \text{ cm})$ and height $z = f(x^*, y^*)$ cm.

Example 7.13

Use a double Riemann sum and sampling at lower right endpoints of the subrectangles to approximate the volume in Example 7.12.

Method: The volume is

$$\text{Volume} = \iint_{\mathcal{D}} (1 + 0.5x + 0.2y)dA \text{ cm}^3.$$

For example, we may break up the region $\mathcal{D}$ into eight subrectangles, each of area (0.5 cm $\times$ 0.5 cm) and sample the integrand, $f(x,y) = 1 + 0.5x + 0.2y$, at lower right endpoints, as shown in Figure 7.20. The $\times$'s mark the points in $\mathcal{D}$ where $f(x,y)$ is sampled, that is, at $(x,y) = (0.5,0), (1,0), (0.5,0.5), (1,0.5), (0.5,1), (1,1), (0.5,1.5), (1,1.5)$. Correspondingly,

$$\begin{aligned}\text{Volume} &\approx \big(f(0.5,0) + f(1,0) + f(0.5,0.5) + f(1,0.5) + f(0.5,1) + f(1,1) \\ &\quad + f(0.5,1.5) + f(1,1.5)\big) \text{ cm} \times (0.25 \text{ cm}^2) \\ &= (1.25 + 1.5 + 1.35 + 1.6 + 1.45 + 1.7 + 1.55 + 1.8) \text{ cm} \times (0.25 \text{ cm}^2) \\ &= 3.05 \text{ cm}^3. \bigcirc\end{aligned}$$

The exact volume is 2.9 cm^3. Considering how coarsely we sampled the function's values, a relative error of $\frac{0.15}{2.9} \approx 5\%$ is pretty small!

In Section 7.4, we will see how to also calculate volumes using triple integrals.

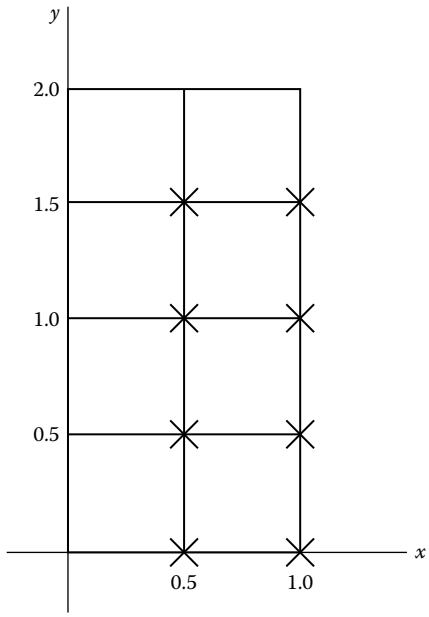


FIGURE 7.20
Sampling in a double Riemann sum.

Theorem 7.12

(Integral of a product function) If $f(x,y)=g(x)h(y)$ is a product of continuous functions on the rectangle $\mathcal{D}$ given in (7.19), then

$$\int_c^d \int_a^b g(x)h(y)dxdy = \left(\int_c^d h(y)dy\right)\left(\int_a^b g(x)dx\right).$$

Theorem 7.12 makes it faster and easier to calculate integrals in a special case.

Suppose the region is not a rectangle. We can generalize the idea of breaking up a region into horizontal or vertical line segments, as we did in writing a double integral as an iterated integral in Example 7.11. Indeed, we used this idea to find areas in calculus for a function of a single variable.

Example 7.14

Express the region shown in Figure 7.21 using inequalities.

Method: For each x in the interval $[-1,2]$, inside the region is a vertical line segment between the points $(x,y)=(x,-1+x^2)$ and $(x,1+x)$, as shown in Figure 7.21. So, the region is

$$\mathcal{D}=\left\{(x,y):\ -1\le x\le 2, -1+x^2\le y\le 1+x\right\}. \bigcirc$$

By the way, if we had not been given the picture of the region but instead we had been told it is the finite region between the curves $y=1+x$ and $y=-1+x^2$, then the first thing to do would be to find the points of intersection. How? By solving $1+x=y=-1+x^2$, that is, $1+x=-1+x^2$, that is, $0=x^2-x-2=(x+1)(x-2)$, hence $x=-1$ and $x=2$, so the points of intersection are $(x,y)=(-1,0)$ and $(2,3)$.

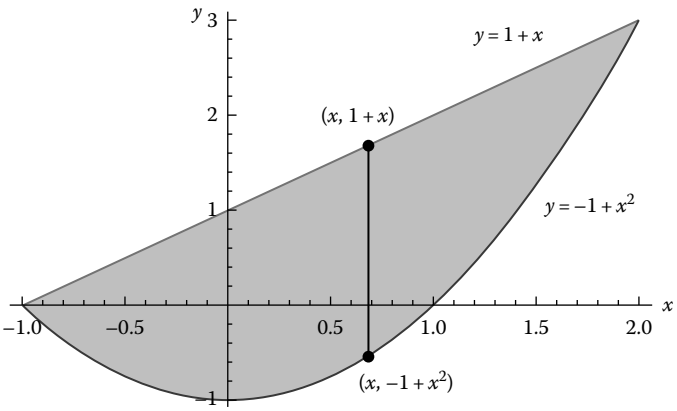


FIGURE 7.21
Example 7.14.

Example 7.15

Find the total mass on the region shown in Figure 7.21, assuming the mass density is $\varrho = 1 + 0.5x + 0.2y$.

Method: Using the result of Example 7.13,

$$M = \iint_{\mathcal{D}} (1 + 0.5x + 0.2y)dA = \int_{-1}^{2} \left(\int_{-1+x^2}^{1+x} (1 + 0.5x + 0.2y)dy \right) dx$$

$$= \int_{-1}^{2} \left[y + 0.5xy + 0.1y^2 \right]_{-1+x^2}^{1+x} dx = \int_{-1}^{2} \Big((1 + x + 0.5x(1 + x) + 0.1(1 + x)^2)$$

$$-(-1 + x^2 + 0.5x(-1 + x^2) + 0.1(-1 + x^2)^2) \Big) dx$$

$$= \cdots = \int_{-1}^{2} \left(2 + 2.2x - 0.2x^2 - 0.5x^3 - 0.1x^4 \right) dx$$

$$= \left[2x + 1.1x^2 - \frac{0.2}{3}x^3 - 0.125x^4 - 0.02x^5 \right]_{-1}^{2} = \cdots = 6.165. \bigcirc$$

Example 7.16

Find the volume of the solid tetrahedron whose vertices are $(0, 0, 0)$, $(12, 0, 0)$, $(0, 4, 0)$, and $(0, 0, 3)$.

Method: This tetrahedron, shown in Figure 7.22, is the solid bounded (a) on the bottom by a triangular region $\mathcal{D}$ in the $z = 0$ plane, that is, the xy-plane, and (b) on the top by part of a plane above $\mathcal{D}$. First, the region $\mathcal{D}$ has vertices $(0, 0)$, $(12, 0)$, and $(0, 4)$ in the xy-plane and is shown in Figure 7.23.

Using vertical line segments, we can write

$$\mathcal{D} = \left\{ (x, y) : \ 0 \le x \le 12, \ 0 \le y \le 4 - \frac{x}{3} \right\}.$$

Next, the top is part of the plane that satisfies the equation $Ax + By + Cz = E$. As in Section 6.1, we substitute in the three points on the plane to get

$$\left\{ \begin{array}{c} 12A + 0 + 0 = E \\ 0 + 4B + 0 = E \\ 0 + 0 + 3C = E \end{array} \right\}.$$

It's convenient to let $E = 12$ because that gives integer values $A = 1, B = 3, C = 4$. So, the plane is

$$x + 3y + 4z = 12.$$

So, part of the top of the solid is on the plane

$$z = 3 - \frac{x}{4} - \frac{3y}{4} \triangleq f(x, y).$$

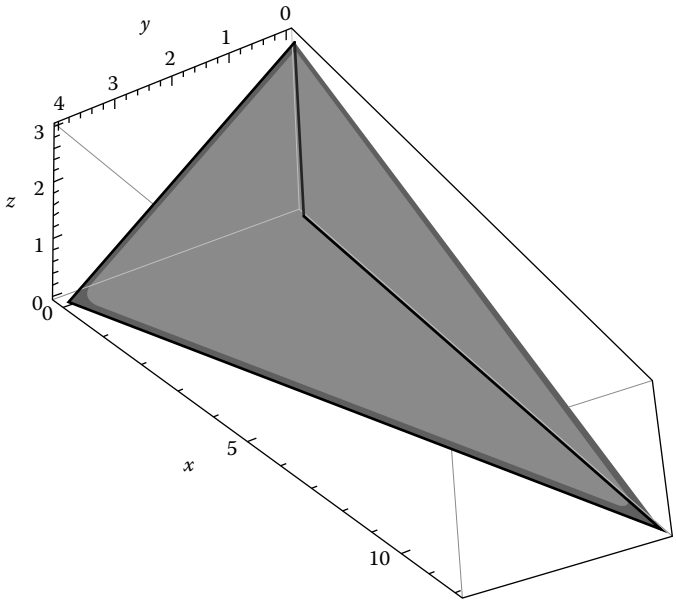


FIGURE 7.22
Solid tetrahedron in Example 7.16.

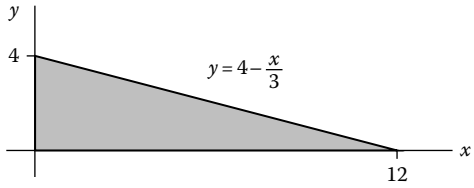


FIGURE 7.23
$\mathcal{D}$ in Example 7.16.

So, the volume of the solid is

$$\text{Volume} = \int_0^{12} \left(\int_0^{4-\frac{x}{3}} \left(3 - \frac{x}{4} - \frac{3y}{4} \right) dy \right) dx = \int_0^{12} \left[3y - \frac{xy}{4} - \frac{3y^2}{8} \right]_0^{4-\frac{x}{3}} dx$$

$$= \cdots = \int_0^{12} \left(6 - x + \frac{x^2}{24} \right) dx = \left[6x - \frac{x^2}{2} + \frac{x^3}{72} \right]_0^{12} = \cdots = 24. \bigcirc$$

The tetrahedron in Figure 7.22 was plotted by the *Mathematica* command:

```
Graphics3D[Polygon[{{12, 0, 0}, {0, 4, 0}, {0, 0, 3},{0, 0, 0}}], Axes → True],
```

after which we used the mouse to rotate the view.

A **type I** region in the xy-plane is one that can be written in the form

$$\mathcal{D} = \{\,(x,y) : \; a \le x \le b,\; \phi_1(x) \le y \le \phi_2(x)\,\},$$

where $\phi_1(x) \le \phi_2(x)$ for $a \le x \le b$. Expressing a region in the *type I* form is like writing the region as a union of vertical line segments.

A **type II** region in the xy-plane is one that can be written in the form

$$\mathcal{D} = \{\,(x,y) : \; c \le y \le d,\; \psi_1(y) \le x \le \psi_2(y)\,\},$$

where $\psi_1(y) \le \psi_2(y)$ for $c \le y \le d$. Expressing a region in the *type II* form is like writing the region as a union of horizontal line segments.

Sometimes we can write a region in both *type I* and *type II* forms. This may make our integration work possible or at least a lot easier.

Example 7.17

Rewrite

$$\int_0^5 \left(\int_{2y}^{10} e^{-x^2}\,dx \right) dy$$

as an integral on a *type I* region in order to evaluate it.

Method: It helps to draw the *type II* region $\mathcal{D} = \{(x,y) : \; 0 \le y \le 5,\; 2y \le x \le 10\}$ (Figure 7.24) and then express it as a union of vertical line segments (Figure 7.25):

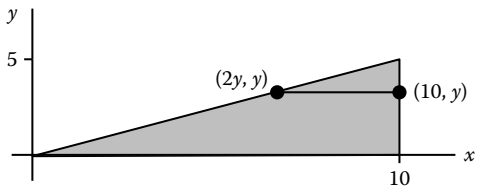


FIGURE 7.24
Horizontal line segment in $\mathcal{D}$.

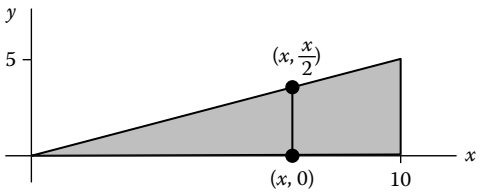


FIGURE 7.25
Vertical line segment in $\mathcal{D}$.

Because

$$\mathcal{D} = \left\{(x,y):\ 0 \le x \le 10,\ 0 \le y \le \frac{x}{2}\right\},$$

$$\int_0^5 \left(\int_{2y}^{10} e^{-x^2}dx\right) dy = \int_0^{10} \left(\int_0^{\frac{x}{2}} e^{-x^2}dy\right) dx = \int_0^{10} \left[ye^{-x^2}\right]_0^{\frac{x}{2}} dx = \int_0^{10} \frac{x}{2}e^{-x^2}dx$$

$$= \left[-\frac{1}{4}e^{-x^2}\right]_0^{10} = \frac{1}{4}(1 - e^{-100}). \bigcirc$$

Definition 7.11

The average value of a function f on a planar region $\mathcal{D}$ is

$$\bar{f} \triangleq \frac{1}{\textit{Area of } \mathcal{D}} \iint_{\mathcal{D}} f(\mathbf{r})\, dA \ = \ \frac{\iint_{\mathcal{D}} f(\mathbf{r})\, dA}{\iint_{\mathcal{D}} 1\, dA}.$$

7.3.2 Polar Coordinates

Polar coordinates are particularly useful in describing and/or integrating on regions in the plane that have a boundary curve(s) that is an arc of a circle.

Example 7.18

Describe the region shown in Figure 7.26:

(a) In polar coordinates

(b) As a *type II* region in rectangular coordinates

Method:

(a) Think of the region as a union of *radial* line segments, as shown in Figure 7.27:

$$\mathcal{D} = \{(r,\theta):\ 0 \le \theta \le \frac{3\pi}{4},\ 0 \le r \le a\}.$$

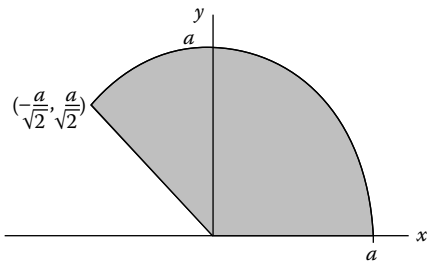


FIGURE 7.26
Wedge.

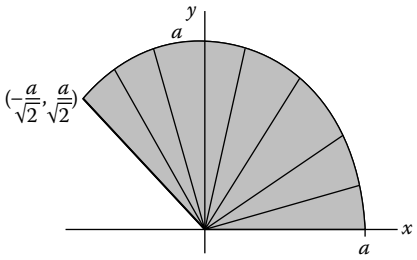


FIGURE 7.27
Wedge as a union of *radial* line segments.

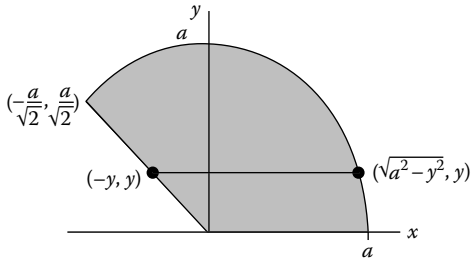


FIGURE 7.28
Wedge as union of horizontal line segments.

(b) Think of the region as a union of two domains, each of which consists of horizontal line segments:

$$\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 = \left\{(x,y):\ 0 \le y \le \frac{a}{\sqrt{2}},\ -y \le x \le \sqrt{a^2-y^2}\right\}$$
$$\cup \left\{(x,y):\ \frac{a}{\sqrt{2}} \le y \le a,\ -\sqrt{a^2-y^2} \le x \le \sqrt{a^2-y^2}\right\}. \bigcirc$$

Figure 7.28 shows a typical horizontal line segment in the set $\mathcal{D}_1$.

We want to do integrals using polar coordinates when that gives a simpler description of the region $\mathcal{D}$; but, also, sometimes polar coordinates can simplify the integrand

$$f(x,y) = f(r\cos\theta, r\sin\theta).$$

But, we also need to know the element of area, dA, expressed in polar coordinates:

$$dA = r\,dr\,d\theta,$$

as we derived in Section 6.6.

To summarize,

$$\iint\limits_{\mathcal{D} \text{ in } (x,y)} f(x,y)dA \ = \iint\limits_{\mathcal{D} \text{ in } (r,\theta)} f(r\cos\theta, r\sin\theta)\, r\,dr\,d\theta$$

Example 7.19

Find the volume of that part of the ball $0 \le x^2 + y^2 + z^2 \le a^2$ that sits above the disk $0 \le x^2 + y^2 \le 1, z = 0$, assuming $a > 1$.

Method: The region $\mathcal{D}$ is the disk $0 \le x^2 + y^2 \le 1$. The top surface of the solid is $z = +\sqrt{a^2 - x^2 - y^2} = \sqrt{a^2 - r^2}$ in polar coordinates. The volume of the solid is

$$\text{Volume} = \iint_{\mathcal{D}} z\, dA = \int_0^{2\pi} \int_0^1 \sqrt{a^2 - r^2}\, r dr\, d\theta = \left(\int_0^{2\pi} d\theta \right) \left(\int_0^1 \sqrt{a^2 - r^2}\, r dr \right).$$

After doing the first integral factor and making the substitution $w = a^2 - r^2$ in the second factor,

$$\text{Volume} = 2\pi \int_{a^2}^{a^2-1} \sqrt{w} \left(-\frac{dw}{2} \right) = 2\pi \left[-\frac{1}{3} w^{3/2} \right]_{a^2}^{a^2-1} = \frac{2\pi}{3} (a^3 - (a^2 - 1)^{3/2}). \bigcirc$$

7.3.3 Green's Theorem

Green's theorem is the first of three major theorems that relate multiple integrals to derivatives. Like the fundamental theorem of calculus (Theorem 7.4) and the fundamental theorem of line integrals (Theorem 7.8), a multiple integral on a region can be evaluated using the boundary value(s) of anti-derivatives, in some sense.

Specifically, Green's theorem says that there is an amazing and useful relationship between the integral around $\mathcal{C}$ of a planar vector field and the integral on $\mathcal{D}$ of certain of its derivatives.

Theorem 7.13

(Green's Theorem) Suppose $\mathcal{D}$ is a planar region whose boundary is a simple, closed, piecewise smooth, positively oriented curve $\mathcal{C}$. If planar vector field $\mathbf{F}(\mathbf{r}) = P(x, y)\hat{\boldsymbol{\imath}} + Q(x, y)\hat{\boldsymbol{\jmath}}$ is continuously differentiable on $\mathcal{D}$ and its boundary $\mathcal{C}$, then

$$\oint_{\mathcal{C}} \mathbf{F} \bullet d\mathbf{r} = \oint_{\mathcal{C}} (P dx + Q dy) = \iint_{\mathcal{D}} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA. \tag{7.23}$$

We will not explain this result here, but we will give references at the end of the section. Also, in the appendix, we will derive Green's theorem for any rectangle $\mathcal{D}$ and any $\mathbf{F}$ that satisfies the smoothness hypothesis. Our emphasis here will be on applying Green's theorem.

Example 7.20

(Example 7.7 in Section 7.2 again) Find the total circulation around the triangle with "nose up" shown in Figure 7.10, assuming the wind has constant speed U_0.

Method: The region $\mathcal{D}$ is the triangle shown in Figure 7.10. The total circulation is

$$\oint_{\mathcal{C}} (U_0 \cos\varphi\, \hat{\boldsymbol{\imath}} + U_0 \sin\varphi\, \hat{\boldsymbol{\jmath}}) \bullet d\mathbf{r}$$

$$= \iint_{\mathcal{D}} \left(\frac{\partial}{\partial x} [U_0 \sin\varphi] - \frac{\partial}{\partial y} [U_0 \cos\varphi] \right) dA = \iint_{\mathcal{D}} (0 - 0) dA = 0.$$

Corollary 7.2

If $\mathbf{F}$ is any constant vector, then $\oint_{\mathcal{C}} \mathbf{F} \bullet d\mathbf{r} = 0$ for any simple, closed, piecewise smooth curve $\mathcal{C}$.

Corollary 7.3

If $\mathbf{F}$ is a planar vector field that is exact and continuously differentiable on a planar region $\mathcal{D}$ and $\mathcal{C}$ is a simple, closed, piecewise smooth curve strictly inside $\mathcal{D}$, then $\oint_{\mathcal{C}} \mathbf{F} \bullet d\mathbf{r} = 0$.

Why? Because $\mathbf{F}$ is exact, it has a potential function $f = f(\mathbf{r})$, that is, $\mathbf{F} = \nabla f$. It follows that

$$P(x,y)\hat{\boldsymbol{\imath}} + Q(x,y)\hat{\boldsymbol{\jmath}} = \mathbf{F} = \frac{\partial f}{\partial x}\hat{\boldsymbol{\imath}} + \frac{\partial f}{\partial y}\hat{\boldsymbol{\jmath}}.$$

By Green's theorem,

$$\oint_{\mathcal{C}} \mathbf{F} \bullet d\mathbf{r} = \oint_{\mathcal{C}} \left(\frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy \right) = \iint_{\mathcal{D}} \left(\frac{\partial}{\partial x} \left[\frac{\partial f}{\partial y} \right] - \frac{\partial}{\partial y} \left[\frac{\partial f}{\partial x} \right] \right) dA = \iint_{\mathcal{D}} 0\, dA = 0,$$

by Clairaut's theorem (Theorem 6.2) in Section 6.4. □.

Corollary 7.4

If $\mathcal{D}$ is a planar region bounded by a simple, closed, positively oriented, piecewise smooth curve $\mathcal{C}$, then the area of $\mathcal{D}$ is given by

$$Area = \oint_{\mathcal{C}} x dy = \oint_{\mathcal{C}} (-y) dx = \frac{1}{2} \oint_{\mathcal{C}} (-y dx + x dy).$$

Why? For example, by Green's theorem,

$$\frac{1}{2} \oint_{\mathcal{C}} (-y dx + x dy) = \frac{1}{2} \iint_{\mathcal{D}} \left(\frac{\partial}{\partial x} [x] - \frac{\partial}{\partial y} [-y] \right) dA = \frac{1}{2} \iint_{\mathcal{D}} 2 dA = Area\ of\ \mathcal{D}.\ \square$$

7.3.4 Comparison with Single Variable Calculus

Suppose $\mathcal{D}$ is a *type II* planar region,

$$\mathcal{D} = \{(x, y) : \ c \leq y \leq d, \ \psi_1(y) \leq x \leq \psi_2(y)\},$$

where $\psi_1(y) \leq \psi_2(y)$ for $c \leq y \leq d$. For the sake of specificity, assume $\psi_1(d) = \psi_2(d)$.

In a single variable calculus course, we found that the area of $\mathcal{D}$ is

$$Area = \int_c^d (\psi_2(y) - \psi_1(y))dy. \tag{7.24}$$

As shown in Figure 7.29, we have $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3$, and we can parametrize the curves by

$$\begin{aligned}
&\mathcal{C}_1 : \ \mathbf{r}_1 = \psi_1(c + d - t)\hat{\boldsymbol{\imath}} + (c + d - t)\,\hat{\boldsymbol{\jmath}}, \ c \leq t \leq d, \\
&\mathcal{C}_2 : \ \mathbf{r}_2 = \psi_2(y)\hat{\boldsymbol{\imath}} + y\,\hat{\boldsymbol{\jmath}}, \ c \leq t \leq d, \\
&\mathcal{C}_3 : \ \mathbf{r}_3 = (\psi_2(c) + t(\psi_1(c) - \psi_2(c))\hat{\boldsymbol{\imath}} + c\,\hat{\boldsymbol{\jmath}}, \ 0 \leq t \leq 1.
\end{aligned}$$

Note that $\mathcal{C}_3$ is a horizontal line segment going from the point $(\psi_1(c), c)$ to the point $(\psi_2(c), c)$.

According to Corollary 7.4, we should have the area of $\mathcal{D}$ given by

$$\begin{aligned}
\text{Area} &= \int_{\mathcal{C}_1} x dy + \int_{\mathcal{C}_2} x dy + \int_{\mathcal{C}_3} x dy \\
&= \int_c^d \psi_1(c + d - t)(-dt) + \int_c^d \psi_2(y) dy + \int_0^1 (\psi_2(c) + t(\psi_1(c) - \psi_2(c))(0 \cdot dt).
\end{aligned}$$

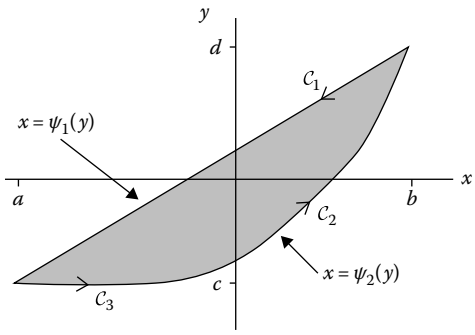


FIGURE 7.29
Area of *type II* region.

In the first integral, make the substitution $w = c + d - t$. So,

$$\text{Area} = \int_d^c \psi_1(w)(dw) + \int_c^d \psi_2(y)dy = -\int_c^d \psi_1(w)(dw) + \int_c^d \psi_2(y)dy = \int_c^d (\psi_2(y) - \psi_1(y))dy.$$

This agrees with the abstract formula (7.24) from a single variable calculus course. □

7.3.5 Green's Theorem for an Annulus

We can extend Green's theorem to a planar region that is not bounded by a simple closed curve.

Example 7.21

Extend Green's theorem to the annulus $a < r < b$ shown in Figure 7.30.

Method: We can decompose

$$\mathcal{D} \triangleq \{(r,\theta) : \ 0 \le \theta \le 2\pi, \ a < r < b\} = \mathcal{D}^+ \cup \mathcal{D}^-,$$

as shown in Figure 7.30. In each of $\mathcal{D}^{\pm}$, we can apply Green's theorem 7.13: If $\mathbf{F} = P(x,y)\hat{\boldsymbol{\imath}} + Q(x,y)\hat{\boldsymbol{\jmath}}$ is continuously differentiable on $\mathcal{D}$ and its boundary curves $\mathcal{C}_a$ and $\mathcal{C}_b$, then

$$\iint_{\mathcal{D}^+} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA = \int_{\mathcal{C}_a^+} \mathbf{F} \bullet d\mathbf{r} + \int_{\mathcal{C}_1^+} \mathbf{F} \bullet d\mathbf{r} + \int_{\mathcal{C}_b^+} \mathbf{F} \bullet d\mathbf{r} + \int_{\mathcal{C}_2^+} \mathbf{F} \bullet d\mathbf{r}$$

and

$$\iint_{\mathcal{D}^-} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA = \int_{\mathcal{C}_a^-} \mathbf{F} \bullet d\mathbf{r} + \int_{\mathcal{C}_2^-} \mathbf{F} \bullet d\mathbf{r} + \int_{\mathcal{C}_b^-} \mathbf{F} \bullet d\mathbf{r} + \int_{\mathcal{C}_1^-} \mathbf{F} \bullet d\mathbf{r}.$$

But, $\mathcal{C}_1^- = -\mathcal{C}_1^+$ and $\mathcal{C}_2^- = -\mathcal{C}_2^+$, so

$$\iint_{\mathcal{D}} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA = \iint_{\mathcal{D}^+} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA + \iint_{\mathcal{D}^-} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA$$

$$= \int_{\mathcal{C}_a^+} \mathbf{F} \bullet d\mathbf{r} + \cancel{\int_{\mathcal{C}_1^+} \mathbf{F} \bullet d\mathbf{r}} + \int_{\mathcal{C}_b^+} \mathbf{F} \bullet d\mathbf{r} + \cancel{\int_{\mathcal{C}_2^+} \mathbf{F} \bullet d\mathbf{r}} + \int_{\mathcal{C}_a^-} \mathbf{F} \bullet d\mathbf{r} + \cancel{\int_{\mathcal{C}_2^-} \mathbf{F} \bullet d\mathbf{r}} + \int_{\mathcal{C}_b^-} \mathbf{F} \bullet d\mathbf{r} + \cancel{\int_{\mathcal{C}_1^-} \mathbf{F} \bullet d\mathbf{r}},$$

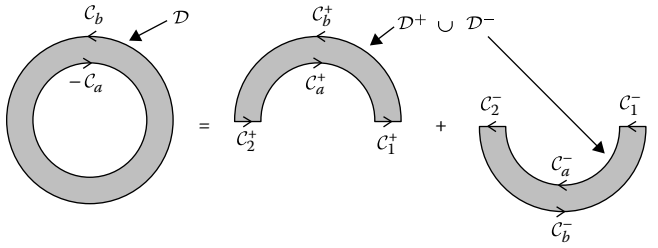


FIGURE 7.30
Decomposing an annulus.

half of the terms canceling by Theorem 7.9 in Section 7.2. So,

$$\iint_{\mathcal{D}} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA = \int_{\mathcal{C}_a^+} \mathbf{F} \bullet d\mathbf{r} + \int_{\mathcal{C}_a^-} \mathbf{F} \bullet d\mathbf{r} + \int_{\mathcal{C}_b^+} \mathbf{F} \bullet d\mathbf{r} + \int_{\mathcal{C}_b^-} \mathbf{F} \bullet d\mathbf{r}.$$

The curve $\mathcal{C}_b$ is traversed counterclockwise using $\mathcal{C}_b^+ \cup \mathcal{C}_b^-$, so

$$\int_{\mathcal{C}_b^+} \mathbf{F} \bullet d\mathbf{r} + \int_{\mathcal{C}_b^-} \mathbf{F} \bullet d\mathbf{r} = \oint_{\mathcal{C}_b} \mathbf{F} \bullet d\mathbf{r}.$$

The curve $-\mathcal{C}_a$ is traversed clockwise using $\mathcal{C}_a^+ \cup \mathcal{C}_a^-$, so

$$\int_{\mathcal{C}_a^+} \mathbf{F} \bullet d\mathbf{r} + \int_{\mathcal{C}_a^-} \mathbf{F} \bullet d\mathbf{r} = \oint_{-\mathcal{C}_a} \mathbf{F} \bullet d\mathbf{r} = -\oint_{\mathcal{C}_a} \mathbf{F} \bullet d\mathbf{r}.$$

Putting everything together,

$$\iint_{\mathcal{D}} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA = \oint_{\mathcal{C}_b} \mathbf{F} \bullet d\mathbf{r} - \oint_{\mathcal{C}_a} \mathbf{F} \bullet d\mathbf{r}. \bigcirc$$

7.3.6 Green's Theorem in Polar Coordinates

If $\mathbf{F} \triangleq F_r \hat{\mathbf{e}}_\mathbf{r} + F_\theta \hat{\mathbf{e}}_\theta$ is integrated around a curve given in polar coordinates, then we can use Green's theorem expressed in polar coordinates. To that end, use results from Section 6.7 to see that

$$\begin{aligned} P(x,y)\hat{\boldsymbol{\imath}} + Q(x,y)\hat{\boldsymbol{\jmath}} = \mathbf{F} &= F_r \hat{\mathbf{e}}_\mathbf{r} + F_\theta \hat{\mathbf{e}}_\theta = F_r(\cos\theta\,\hat{\boldsymbol{\imath}} + \sin\theta\,\hat{\boldsymbol{\jmath}}) + F_\theta(-\sin\theta\,\hat{\boldsymbol{\imath}} + \cos\theta\,\hat{\boldsymbol{\jmath}}) \\ &= (F_r\cos\theta - F_\theta\sin\theta)\hat{\boldsymbol{\imath}} + (F_r\sin\theta + F_\theta\cos\theta)\hat{\boldsymbol{\jmath}}). \end{aligned}$$

Using (6.71), we calculate

$$\begin{aligned} \frac{\partial Q}{\partial x} &= \left(\cos\theta \frac{\partial}{\partial r} - \frac{\sin\theta}{r} \frac{\partial}{\partial \theta} \right) [F_r \sin\theta + F_\theta \cos\theta] \\ &= \cos\theta \left(\frac{\partial F_r}{\partial r} \cdot \sin\theta + \frac{\partial F_\theta}{\partial r} \cdot \cos\theta \right) \\ &\quad - \frac{\sin\theta}{r} \left(\frac{\partial F_r}{\partial \theta} \cdot \sin\theta + F_r \cos\theta + \frac{\partial F_\theta}{\partial \theta} \cdot \cos\theta - F_\theta \sin\theta \right). \end{aligned}$$

Similarly,

$$\begin{aligned} -\frac{\partial P}{\partial y} &= -\left(\sin\theta \frac{\partial}{\partial r} + \frac{\cos\theta}{r} \frac{\partial}{\partial \theta} \right) [F_r \cos\theta - F_\theta \sin\theta] \\ &= -\sin\theta \left(\frac{\partial F_r}{\partial r} \cdot \cos\theta - \frac{\partial F_\theta}{\partial r} \cdot \sin\theta \right) \\ &\quad - \frac{\cos\theta}{r} \left(\frac{\partial F_r}{\partial \theta} \cdot \cos\theta - F_r \sin\theta - \frac{\partial F_\theta}{\partial \theta} \cdot \sin\theta - F_\theta \cos\theta \right) \end{aligned}$$

Summing up,

$$\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} = \cdots = \underline{\cos^2\theta \frac{\partial F_\theta}{\partial r}} - \underbrace{\frac{\sin^2\theta}{r}\frac{\partial F_r}{\partial \theta}} + \frac{\sin^2\theta}{r}F_\theta$$

$$+ \underline{\sin^2\theta \frac{\partial F_\theta}{\partial r}} - \underbrace{\frac{\cos^2\theta}{r}\frac{\partial F_r}{\partial \theta}} + \frac{\cos^2\theta}{r}F_\theta$$

$$= \cdots = \frac{1}{r}\frac{\partial}{\partial r}[rF_\theta] - \frac{1}{r}\frac{\partial F_r}{\partial \theta}.$$

So, in polar coordinates, the conclusion of Green's theorem is

$$\oint_{\mathcal{C}} \mathbf{F} \bullet d\mathbf{r} = \iint_{\mathcal{D}} \left(\frac{1}{r}\frac{\partial}{\partial r}[rF_\theta] - \frac{1}{r}\frac{\partial F_r}{\partial \theta}\right) r dr d\theta = \iint_{\mathcal{D}} \left(\frac{\partial}{\partial r}[rF_\theta] - \frac{\partial F_r}{\partial \theta}\right) dr\, d\theta. \tag{7.25}$$

Appendix: Derivation of Green's Theorem for a Rectangle

Let's see why Green's theorem is correct when $\mathcal{D}$ is a rectangle, as in (7.19), that is, $\mathcal{D} \triangleq \{(x,y) : a \le x \le b,\ c \le y \le d\}$. Let $\mathbf{F} = P(x,y)\hat{\boldsymbol{\imath}} + Q(x,y)\hat{\boldsymbol{\jmath}}$ be any continuously differentiable vector field on $\mathcal{D}$ and its boundary curve $\mathcal{C} = \mathcal{C}_1 \cup \cdots \cup \mathcal{C}_4$ shown in Figure 7.31. Parametrize the curves by

$$\mathcal{C}_1 :\ \mathbf{r}_1 = (a + t(b-a))\hat{\boldsymbol{\imath}} + c\,\hat{\boldsymbol{\jmath}},\ 0 \le t \le 1.$$

$$\mathcal{C}_2 :\ \mathbf{r}_2 = b\,\hat{\boldsymbol{\imath}} + (c + t(d-c))\hat{\boldsymbol{\jmath}},\ 0 \le t \le 1.$$

$$\mathcal{C}_3 :\ \mathbf{r}_3 = (b - t(b-a))\hat{\boldsymbol{\imath}} + d\,\hat{\boldsymbol{\jmath}},\ 0 \le t \le 1.$$

$$\mathcal{C}_4 :\ \mathbf{r}_4 = a\,\hat{\boldsymbol{\imath}} + (d - t(d-c))\hat{\boldsymbol{\jmath}},\ 0 \le t \le 1.$$

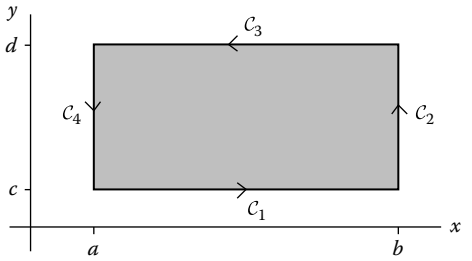


FIGURE 7.31
Green's theorem on a rectangle.

We calculate the left-hand side (LHS) of the (desired) conclusion of Green's theorem:

$$\begin{aligned}\oint_{\mathcal{C}} \mathbf{F}(\mathbf{r}) \bullet d\mathbf{r} &= \int_{\mathcal{C}_1} (P\,dx + Q\,dy) + \int_{\mathcal{C}_2} (P\,dx + Q\,dy) + \int_{\mathcal{C}_3} (P\,dx + Q\,dy) + \int_{\mathcal{C}_4} (P\,dx + Q\,dy) \\ &= \int_0^1 P(a + t(b-a), c)((b-a)dt) + \int_0^1 Q(b, c + t(d-c))((d-c)dt) \\ &\quad + \int_0^1 P(b - t(b-a), d)(-(b-a)dt) + \int_0^1 Q(a, d - t(d-c))(-(d-c)dt).\end{aligned}$$

After making the substitutions $w_1 = a + t(b-a)$, $w_2 = c + t(d-c)$, $w_3 = b - t(b-a)$, $w_4 = d - t(d-c)$, this gives

$$\oint_{\mathcal{C}} \mathbf{F}(\mathbf{r}) \bullet d\mathbf{r} = \int_a^b P(w_1, c)dw_1 + \int_c^d Q(b, w_2)dw_2 + \int_b^a P(w_3, d)dw_3 + \int_d^c Q(a, w_4)dw_4. \tag{7.26}$$

On the other side hand, we calculate the right-hand side (RHS) of the (desired) conclusion of Green's theorem:

$$\begin{aligned}\iint_{\mathcal{D}} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA &= \int_c^d \left(\int_a^b \frac{\partial Q}{\partial x}(x,y)dx\right) dy - \int_a^b \left(\int_c^d \frac{\partial P}{\partial y}(x,y)dy\right) dx \\ &= \int_c^d \Big(\big[Q(x,y)\big]_a^b\Big)\, dy - \int_a^b \Big(\big[P(x,y)\big]_c^d\Big)\, dx \\ &= \int_c^d \big(Q(b,y) - Q(a,y)\big)\, dy - \int_a^b \big(P(x,d) - P(x,c)\big)\, dx,\end{aligned}$$

that is,

$$\iint_{\mathcal{D}} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA = \int_a^b P(x,c)dx + \int_c^d Q(b,y)dy - \int_a^b P(x,d)dx - \int_c^d Q(a,y)dy. \tag{7.27}$$

This agrees with (7.26). This explains why Green's theorem is correct for any rectangle and any continuously differentiable vector field **F**. □

Learn More About It

A good reference is *Functions of Several Variables*, by Wendell Fleming, Springer-Verlag, 2nd ed., 3rd printing, © 1987. Problems 7.3.7.34 and 7.3.7.38 came from Problems 8.7.9 and 8.7.11 on page 361 of that book. The conclusion of Problem 7.3.7.4 is a small part of an Example on pp. 407–409 in *Radiative Heat Transfer*, by Michael F. Modest, McGraw-Hill, Inc., © 1993.

TABLE 7.1

$f(x, y)$ Data

x/y	10	12.5	15	17.5	20
0	1	2	3	4	5
1	2	2	3	2	1
2	3	4	3	4	3
3	2	3	2	1	2
4	1	3	4	5	4

7.3.7 Problems

1. Use four subrectangles and sampling at midpoints to approximate the volume bounded above by the surface $z = f(x, y)$ and below by the rectangle $0 \le x \le 4$, $10 \le y \le 20$ in the xy-plane, assuming data of values of $f(x, y)$ found in Table 7.1. Do state what points (x, y) you used to get data about f.
2. Evaluate $\int_0^{\frac{\pi}{6}} \int_1^3 xy \sin 2x \, dy \, dx$.
3. Evaluate $\int_1^3 \int_2^4 x^2 e^{xy} \, dy \, dx$.
4. Explain why

$$\int_{\frac{\pi}{2}-\Theta}^{\frac{\pi}{2}} \int_0^{\pi} f(\phi, \psi, \Theta) \, d\phi \, d\psi = \frac{2}{3}(\sin\Theta - \Theta\cos\Theta),$$

   where

$$f(\phi, \psi, \Theta) \triangleq \sin^3\phi \cos\psi (\sin\psi \sin\Theta - \cos\psi \cos\Theta)$$

   and Θ is a constant "scattering" angle. This double integral helps to find the "phase function" for a diffusely reflecting sphere in the study of radiative heat transfer. [Hint: Use a cosine addition formula.]
5. Evaluate $\iint_{\mathcal{D}} \frac{2}{(x+2y)^2} \, dA$, where $\mathcal{D} = \{(x, y) : \ 4 \le x \le 5, \ -1 \le y \le 0\}$.
6. Evaluate $\iint_{\mathcal{D}} \cos^2(x^2 + y^2) dA$, where $\mathcal{D} = \{(x, y) : 2 \le x^2 + y^2 \le 7\}$.
7. Evaluate $\iint_{\mathcal{D}} x \, dA$, where $\mathcal{D}$ is shown in Figure 7.32a.
8. Evaluate $\iint_{\mathcal{D}} (x - y) dA$, where $\mathcal{D}$ is shown in Figure 7.32b.
9. Evaluate $\iint_{\mathcal{D}} x \, dA$, where $\mathcal{D}$ is sketched in Figure 7.33a.
10. Evaluate $\iint_{\mathcal{D}} y \, dA$, where $\mathcal{D}$ is sketched in Figure 7.33b.
11. Evaluate $\iint_{\mathcal{D}} x \, dA$, where $\mathcal{D}$ is the finite region in the xy-plane bounded by the curves $x = y^2$ and $x = 2y + 3$.
12. Evaluate $\iint_{\mathcal{D}} y^2 e^{xy} \, dA$, where $\mathcal{D}$ is the triangle in the xy-plane whose vertices are $(0, 0)$, $(0, 2)$, $(4, 2)$. Draw an appropriate sketch of $\mathcal{D}$ showing how you decompose it into a union of line segments.

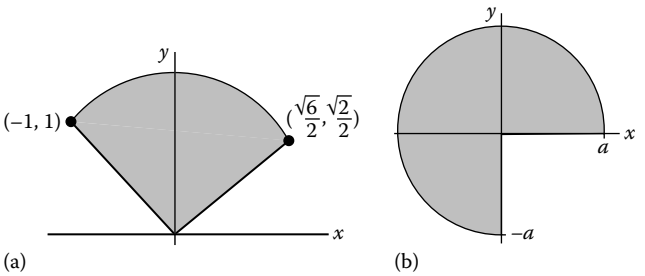


FIGURE 7.34
(a) Problem 7.3.7.15 and (b) Problem 7.3.7.16.

20. Let $\mathcal{D}$ be the half disk $0 \le x^2 + y^2 \le a^2,\ y \ge 0$. For $\mathbf{F} = y^2\,\hat{\boldsymbol{\imath}} - 2x\,\hat{\boldsymbol{\jmath}}$, calculate both sides of the conclusion of Green's theorem.
21. Evaluate

$$\oint_{\mathcal{C}} (\cos x \, \sin y \, dx + \sin x \, \cos y \, dy),$$

where $\mathcal{C}$ is the positively oriented curve that is the boundary of the triangle in the xy-plane whose vertices are $(\frac{\pi}{2}, \frac{\pi}{2})$, $(\pi, \frac{\pi}{2})$, $(\frac{\pi}{2}, \frac{3\pi}{2})$.
22. Use Green's theorem to evaluate $\oint_{\mathcal{C}} (2y\,dx + x\,dy)$, where $\mathcal{C}$ is the positively oriented curve given in polar coordinates by the cardioid $r = 2 + \sin\theta,\ 0 \le \theta \le 2\pi$.
23. Use Green's theorem to evaluate $\oint_{\mathcal{C}} (y^2\,dx + 2x^2\,dy)$, where $\mathcal{C}$ is the positively oriented boundary of the finite region $\mathcal{D}$ that lies between the curves $y = x(2 - x)$ and $y = 0$.
24. The area inside the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ is πab. Use this and Green's theorem to evaluate $\oint_{\mathcal{C}} \mathbf{F} \bullet d\mathbf{r}$, where $\mathcal{C}$ is that ellipse, traversed counterclockwise, and $\mathbf{F} \triangleq (-y + \frac{xy^2}{2})\,\hat{\boldsymbol{\imath}} + (x + \frac{x^2 y}{2})\,\hat{\boldsymbol{\jmath}}$.
25. Explain why πab is the area inside the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$.
26. Find the circulation of the vector field $\mathbf{v} = \hat{\mathbf{e}}_r - r\,\hat{\mathbf{e}}_\theta$ around the circle of radius 2 whose center is at the origin in the xy-plane. [Hint: See Problem 7.2.5.14(a).]
27. Find the area enclosed by the planar curve $\mathcal{C} : \mathbf{r} = 2\cos t\,\hat{\boldsymbol{\imath}} + \sin 2t\,\hat{\boldsymbol{\jmath}},\ 0 \le t \le 2\pi$.
28. For the special case of a disk $\mathcal{D} = \{(r,\theta) :\ 0 \le r \le a, 0 \le \theta \le 2\pi\}$, explain why (7.25) reduces to

$$\oint_{\mathcal{C}} \mathbf{F} \bullet d\mathbf{r} = \int_0^a a F_\theta(a,\theta)\,d\theta.$$

[Hints: Use $\int \frac{1}{r}\frac{\partial}{\partial r}\left[rF_\theta\right] r\,dr = rF_\theta$, $\int \left(\frac{\partial F_r}{\partial \theta}\right) d\theta = F_r$, and 2π-periodicity in θ of the function $F_r = F_r(r,\theta)$.] Note that the conclusion here equals that in Problem 7.2.5.14(a).

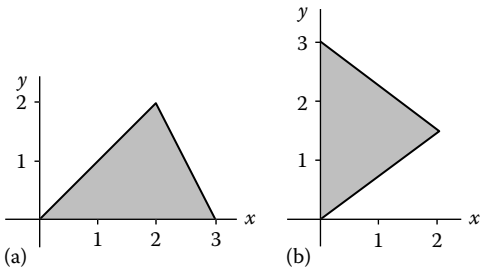


FIGURE 7.32
(a) Problem 7.3.7.7 and (b) Problem 7.3.7.8.

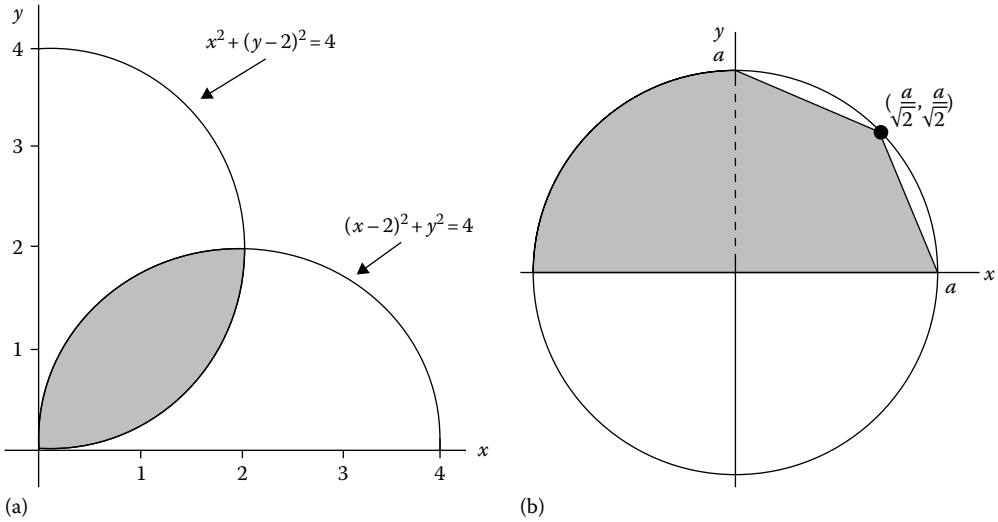


FIGURE 7.33
(a) Problem 7.3.7.9 and (b) Problem 7.3.7.10.

13. Evaluate $\int_0^2 \left(\int_x^2 e^{-y^2} dy \right) dx$.
14. Find the center of mass of the lamina that is the finite region in the xy-plane bounded by the curves $x = 1$, $y = \frac{1}{x-1}$, $x = 3$, and $y = \frac{x}{x-4}$, assuming the mass density is proportional to the distance from the y-axis.
15. Evaluate M_y for the lamina occupying planar region $\mathcal{D}$ shown in Figure 7.34a. Assume $\mathcal{D}$ is a sector of a disk and the mass density is the constant ϱ_0.
16. Find the centroid of the thin plate shown in Figure 7.34b.
17. Evaluate $\iint_{\mathcal{D}} (x^2 - y^2) dA$, where $\mathcal{D}$ is the finite region in the fourth quadrant of the xy-plane bounded by the circles $x^2 + y^2 = 1$ and $x^2 + y^2 = 4$.
18. Set up, but do not evaluate, an iterated double integral that gives the volume enclosed by the ellipsoid $4x^2 + 4y^2 + z^2 = 1$. Give an appropriate sketch.
19. Find the volume of the solid bounded below by the disk $x^2 + y^2 = a^2$ in the $z = 0$ plane and bounded above by the surface $z = 1 - y$. Assume that $a < 1$.

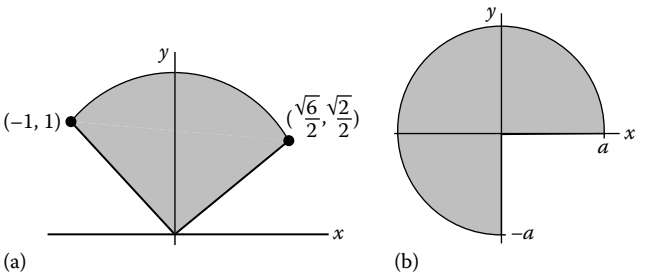


FIGURE 7.34
(a) Problem 7.3.7.15 and (b) Problem 7.3.7.16.

20. Let $\mathcal{D}$ be the half disk $0 \le x^2 + y^2 \le a^2,\ y \ge 0$. For $\mathbf{F} = y^2\,\hat{\boldsymbol{\imath}} - 2x\,\hat{\boldsymbol{\jmath}}$, calculate both sides of the conclusion of Green's theorem.
21. Evaluate

$$\oint_{\mathcal{C}} (\cos x \, \sin y \, dx + \sin x \, \cos y \, dy),$$

where $\mathcal{C}$ is the positively oriented curve that is the boundary of the triangle in the xy-plane whose vertices are $(\frac{\pi}{2}, \frac{\pi}{2})$, $(\pi, \frac{\pi}{2})$, $(\frac{\pi}{2}, \frac{3\pi}{2})$.
22. Use Green's theorem to evaluate $\oint_{\mathcal{C}} (2y\,dx + x\,dy)$, where $\mathcal{C}$ is the positively oriented curve given in polar coordinates by the cardioid $r = 2 + \sin\theta,\ 0 \le \theta \le 2\pi$.
23. Use Green's theorem to evaluate $\oint_{\mathcal{C}} (y^2\,dx + 2x^2\,dy)$, where $\mathcal{C}$ is the positively oriented boundary of the finite region $\mathcal{D}$ that lies between the curves $y = x(2 - x)$ and $y = 0$.
24. The area inside the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ is πab. Use this and Green's theorem to evaluate $\oint_{\mathcal{C}} \mathbf{F} \bullet d\mathbf{r}$, where $\mathcal{C}$ is that ellipse, traversed counterclockwise, and $\mathbf{F} \triangleq (-y + \frac{xy^2}{2})\,\hat{\boldsymbol{\imath}} + (x + \frac{x^2 y}{2})\,\hat{\boldsymbol{\jmath}}$.
25. Explain why πab is the area inside the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$.
26. Find the circulation of the vector field $\mathbf{v} = \hat{\mathbf{e}}_r - r\,\hat{\mathbf{e}}_\theta$ around the circle of radius 2 whose center is at the origin in the xy-plane. [Hint: See Problem 7.2.5.14(a).]
27. Find the area enclosed by the planar curve $\mathcal{C} : \mathbf{r} = 2\cos t\,\hat{\boldsymbol{\imath}} + \sin 2t\,\hat{\boldsymbol{\jmath}},\ 0 \le t \le 2\pi$.
28. For the special case of a disk $\mathcal{D} = \{(r,\theta) :\ 0 \le r \le a, 0 \le \theta \le 2\pi\}$, explain why (7.25) reduces to

$$\oint_{\mathcal{C}} \mathbf{F} \bullet d\mathbf{r} = \int_0^a a F_\theta(a,\theta)\,d\theta.$$

[Hints: Use $\int \frac{1}{r}\frac{\partial}{\partial r}\left[rF_\theta\right] r\,dr = rF_\theta$, $\int \left(\frac{\partial F_r}{\partial \theta}\right) d\theta = F_r$, and 2π-periodicity in θ of the function $F_r = F_r(r,\theta)$.] Note that the conclusion here equals that in Problem 7.2.5.14(a).

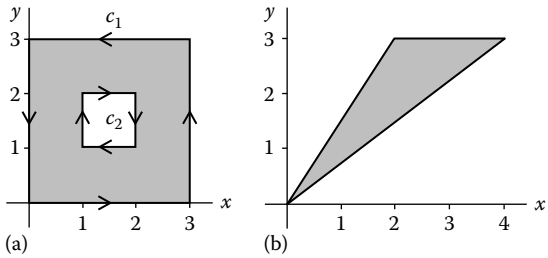


FIGURE 7.35
(a) Problem 7.3.7.30 and (b) Problem 7.3.7.35.

29. Find the circulation of the vector field $\mathbf{v} = z\,\hat{\boldsymbol{\imath}} + x\,\hat{\boldsymbol{\jmath}} - y\,\hat{\mathbf{k}}$ around the triangle whose vertices are $(0,0,0), (1,0,0), (0,5,0)$.
30. Evaluate $\oint_{\mathcal{C}_1}(y\,dx - x\,dy) + \int_{\mathcal{C}_2}(y\,dx - x\,dy)$, where $\mathcal{C}_1, \mathcal{C}_2$ are shown in Figure 7.35a.
31. $\mathcal{D}$ is the twelve square mile region $\{(x,y) : 0 \le x \le 4,\ 0 \le y \le 3\}$, where x and y are measured in miles. Above $\mathcal{D}$ is elevated land whose altitude (above sea level) measurements $f(x,y)$, in tenths of a mile, are given in Table 7.2.
    (a) Use exactly six subrectangles and sample at their midpoints to approximate the volume of land above sea level above $\mathcal{D}$.
    (b) Find the average height above sea level of the land over $\mathcal{D}$.
32. A quantity f depends on radius r and angle θ in a disk of radius 4. Data values of f are given in Table 7.3. Approximate the average value of f on the disk. [Caution: Be careful as to what function you are integrating in polar coordinates.]
33. Suppose the only things we know about a function $f = f(r,\theta)$ are the data $f(1, \frac{\pi}{8}) \approx 2, f(1, \frac{3\pi}{8}) \approx 1, f(3, \frac{\pi}{8}) \approx 3, f(3, \frac{3\pi}{8}) \approx 5$. Approximate the average value of f on the quarter disk shown in Figure 7.36. [Caution: Be careful as to what function you are integrating in polar coordinates.]
34. Let $\mathcal{C}$ be the positively oriented circle $x^2 + y^2 = 1$, $\mathcal{D}$ be the disk it encloses, and planar vector field

$$\mathbf{F} = P\,\hat{\boldsymbol{\imath}} + Q\,\hat{\boldsymbol{\jmath}} \triangleq -\frac{y}{x^2+y^2}\,\hat{\boldsymbol{\imath}} + \frac{x}{x^2+y^2}\,\hat{\boldsymbol{\jmath}}.$$

TABLE 7.2
Height $f(x,y)$ Data

x/y	**0**	**0.5**	**1**	**1.5**	**2**	**2.5**	**3**
0	0	1	2	3	3	2	1
1	1	1	1	2	3	3	2
2	2	2	3	4	4	4	3
3	3	4	4	5	6	5	4
4	2	3	3	4	5	5	4

TABLE 7.3

$f(r,\theta)$ Data

r/θ	0	$\frac{\pi}{2}$	π	$\frac{3\pi}{2}$
0	0	0	0	0
1	3	2	1	2
2	3	1	1	2
3	5	3	3	4
4	6	5	4	5

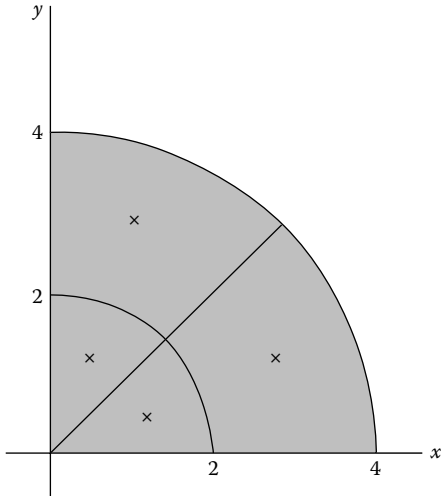


FIGURE 7.36
Problem 7.3.7.33.

Explain why $\iint_{\mathcal{D}} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA = 0$ and $\oint_{\mathcal{C}} \mathbf{F} \bullet d\mathbf{r} = 2\pi$. Why does this *not* contradict Green's theorem?

The **polar moment of inertia** for rotation of a plate around the z-axis is

$$I_0 \triangleq \iint_{\mathcal{D}} (x^2 + y^2)\varrho \, dA.$$

35. Assume the triangular lamina shown in Figure 7.35b is covered with mass whose density is $\varrho(x,y) = 10 - x + 0.1x^2 - y + 0.1y^2$ g/cm^2. Find the center of mass and the polar moment of inertia of the lamina. You may use technology to evaluate the integrals.
36. If a plate is actually rotating, then its KE of rotation is

$$\text{KE} \triangleq \frac{1}{2} \cdot I_0 \cdot (\dot{\varphi})^2.$$

Use the definition of I_0 to rewrite the KE as

$$\iint_{\mathcal{D}} \frac{1}{2} \cdot \varrho \cdot (r\dot{\varphi})^2 \, dA,$$

which makes sense as the summing up of the KE of elements of KE $\frac{1}{2} \cdot (d\varrho)(r\dot{\varphi})^2$. [Note that $r\dot{\varphi}$ is the linear velocity of a particle traveling around a circle of radius r at an angular speed of $\dot{\varphi}$.]

37. Evaluate $\int_0^2 \int_0^{\sqrt{4-x^2}} \sqrt{4-y^2}\, dy\, dx$. First, draw the region $\mathcal{D}$ on which the integration is done. After that, rewrite $\mathcal{D}$ as a type II region.
38. Here you will explain why Green's theorem is true in the special case that $\mathcal{D}$ is a planar region that can be written both as a *type I* region $\{(x,y) : a \le x \le b,\ f_1(x) \le y \le f_2(x)\}$ *and* as a *type II* region $\{(x,y) : c \le y \le d,\ g_1(y) \le x \le g_2(y)\}$. In either case, denote by $\mathcal{C}$ its positively oriented boundary. Explain why the theorems of calculus for functions of a single variable and the definition of line integrals guarantee that for any planar vector field $\mathbf{F} = P\,\hat{\boldsymbol{\imath}} + Q\,\hat{\boldsymbol{\jmath}}$,

$$\iint_{\mathcal{D}} \frac{\partial Q}{\partial x}\, dA = \oint_{\mathcal{C}} Q\, dy$$

and

$$\iint_{\mathcal{D}} \left(-\frac{\partial P}{\partial y} \right) dA = \oint_{\mathcal{C}} P\, dx.$$

39. Use Green's theorem to redo Problem 7.2.5.13.

7.4 Triple Integrals and Applications

Consider a solid box $\mathcal{V}:\ a \le x \le b, c \le y \le d,\ \alpha \le z \le \beta$. If we have mass distributed in the box with density $\varrho(x,y,z)$ in g/(cm^3), then the total mass in the box is

$$M \triangleq \int_{\alpha}^{\beta} \int_{c}^{d} \int_{a}^{b} \varrho(x,y,z) dx dy dz. \tag{7.28}$$

The reason for this is similar to why mass on a rectangle is a double integral: Break up $\mathcal{V}$ into small subboxes and assume that in each subbox, the mass density is approximately constant. This gives

$$M \approx \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{\ell} \varrho(x_i^*, y_j^*, z_k^*) \triangle x_i \triangle y_j \triangle z_k, \tag{7.29}$$

where (x_i^*, y_j^*, z_k^*) is a sampling point chosen in the *ijk*th subbox:

$$\{(x,y,z):\ x_{i-1} \le x \le x_i,\ y_{j-1} \le y \le y_j,\ z_{k-1} \le z \le z_k\}.$$

In (7.29), the mass is approximated by a **triple Riemann sum**.

Take the limit as $\max\{|\triangle x_i|, |\triangle y_j|, |\triangle z_k| : 1 \le i \le m, 1 \le j \le n, 1 \le k \le \ell\} \to 0$ and all three of $m, n, \ell \to \infty$ to get *exact* total mass given by (7.28), that is,

$$M = \iiint_{\mathcal{V}} \varrho(x,y,z)dV,$$

where dV is an element of volume.

There can be many descriptions of more general solids. For example, a solid $\mathcal{V}$ is called *type I* if it can be written in the form

$$\mathcal{V} = \{\,(x,y,z):\ (x,y) \text{ in } \mathcal{D}, \phi_1(x,y) \le z \le \phi_2(x,y)\,\},$$

where $\mathcal{D}$ is a region in the xy-plane. In this case, we get

$$M = \iiint_{\mathcal{V}} \varrho(x,y,z)dV = \iint_{\mathcal{D}} \left(\int_{\phi_1(x,y)}^{\phi_2(x,y)} \varrho(x,y,z)dz \right) dA.$$

Example 7.22

For the solid tetrahedron of Example 7.16, find the centroid (center of mass, assuming constant density) and the **polar moment of inertia** (for rotation around the z-axis), that is,

$$I_0 \triangleq \iiint_{\mathcal{V}} (x^2 + y^2)\varrho dV. \tag{7.30}$$

Method: The density function is assumed to be constant: $\varrho = \varrho_0$. From Example 7.16, we can see that this tetrahedron can be described as the *type I* solid:

$$\mathcal{V} = \left\{ (x,y,z):\ (x,y) \text{ in } \mathcal{D},\ 0 \le z \le 3 - \frac{x}{4} - \frac{3y}{4} \right\},$$

where triangular region $\mathcal{D}$ can be described as the *type I* region:

$$\mathcal{D} = \left\{ (x,y):\ 0 \le x \le 12,\ 0 \le y \le 4 - \frac{x}{3} \right\}.$$

So, the total mass is given by

$$M = \iiint_{\mathcal{V}} \varrho(x,y,z)dV = \iint_{\mathcal{D}} \left(\int_0^{3-\frac{x}{4}-\frac{3y}{4}} \varrho_0 dz \right) dA$$

$$= \varrho_0 \int_0^{12} \left(\int_0^{4-\frac{x}{3}} \left(\int_0^{3-\frac{x}{4}-\frac{3y}{4}} dz \right) dy \right) dx = \varrho_0 \int_0^{12} \left(\int_0^{4-\frac{x}{3}} \left(3 - \frac{x}{4} - \frac{3y}{4} \right) dy \right) dx$$

$$= \cdots = \varrho_0 \cdot (\textit{Volume of } \mathcal{V}) = 24\varrho_0.$$

The moments are M_{yz}, M_{zx}, M_{xy}. For example, M_{yz} is the moment with respect to the plane $x=0$, so

$$M_{yz} \triangleq M_{x=0} = \iiint_{\mathcal{V}} x\varrho(x,y,z)dV = \varrho_0 \int_0^{12} \left(\int_0^{4-\frac{x}{3}} \left(\int_0^{3-\frac{x}{4}-\frac{3y}{4}} x dz \right) dy \right) dx$$

$$= \varrho_0 \int_0^{12} \left(\int_0^{4-\frac{x}{3}} x\left(3-\frac{x}{4}-\frac{3y}{4}\right) dy \right) dx = \varrho_0 \int_0^{12} \left(\int_0^{4-\frac{x}{3}} \left(3x-\frac{1}{4}x^2-\frac{3}{4}xy\right) dy \right) dx$$

$$= \varrho_0 \int_0^{12} \left[3xy-\frac{1}{4}x^2y-\frac{3}{8}xy^2 \right]_0^{4-\frac{x}{3}} dx = \cdots = \varrho_0 \int_0^{12} \left(6x-x^2+\frac{1}{24}x^3\right) dx = \cdots = 72\varrho_0.$$

$$M_{zx} \triangleq M_{y=0} = \varrho_0 \int_0^{12} \left(\int_0^{4-\frac{x}{3}} \left(\int_0^{3-\frac{x}{4}-\frac{3y}{4}} y dz \right) dy \right) dx = \varrho_0 \int_0^{12} \left(\int_0^{4-\frac{x}{3}} \left(3y-\frac{1}{4}xy-\frac{3}{4}y^2\right) dy \right) dx$$

$$= \varrho_0 \int_0^{12} \left[\frac{3}{2}y^2-\frac{1}{8}xy^2-\frac{1}{4}y^3 \right]_0^{4-\frac{x}{3}} dx = \cdots = \varrho_0 \int_0^{12} \left(8-2x+\frac{1}{6}x^2-\frac{1}{216}x^3\right) dx = \cdots = 24\varrho_0.$$

$$M_{xy} \triangleq M_{z=0} = \varrho_0 \int_0^{12} \left(\int_0^{4-\frac{x}{3}} \left(\int_0^{3-\frac{x}{4}-\frac{3y}{4}} z dz \right) dy \right) dx$$

$$= \varrho_0 \int_0^{12} \left(\int_0^{4-\frac{x}{3}} \frac{1}{2}\left(3-\frac{1}{4}x-\frac{3}{4}y\right)^2 dy \right) dx = \cdots = \varrho_0 \int_0^{12} \left(6-\frac{3}{2}x+\frac{1}{8}x^2-\frac{1}{288}x^3\right) dx$$

$$= \cdots = 18\varrho_0.$$

So, this solid tetrahedron's centroid is at

$$(\bar{x},\bar{y},\bar{z}) \triangleq \frac{1}{M}(M_{x=0}, M_{y=0}, M_{z=0}) = \frac{1}{24\varrho_0}(72\varrho_0, 24\varrho_0, 18\varrho_0) = \left(3, 1, \frac{3}{4}\right).$$

The polar moment of inertia (for rotation around the z-axis) is

$$I_0 \triangleq \iint_{\mathcal{V}} (x^2+y^2)\varrho dV = \varrho_0 \int_0^{12} \left(\int_0^{4-\frac{x}{3}} \left(\int_0^{3-\frac{x}{4}-\frac{3y}{4}} (x^2+y^2) dz \right) dy \right) dx$$

$$= \varrho_0 \int_0^{12} \left(\int_0^{4-\frac{x}{3}} \left((x^2+y^2)\left(3-\frac{x}{4}-\frac{3y}{4}\right) \right) dy \right) dx$$

$$= \varrho_0 \int_0^{12} \left(\int_0^{4-\frac{x}{3}} \left(3x^2 - \frac{1}{4}x^3 - \frac{3}{4}x^2 y + 3y^2 - \frac{1}{4}xy^2 - \frac{3}{4}y^3 \right) dy \right) dx$$

$$= \varrho_0 \int_0^{12} \left(16 - \frac{16}{3}x + \frac{20}{3}x^2 - \frac{28}{27}x^3 + \frac{55}{1296}x^4 \right) dx = \cdots = 384\varrho_0. \bigcirc$$

Mathematica, MATLAB® with the symbolic manipulation toolbox, and many calculators make short, easy work of such iterated integrals. In these problems, the real mathematical issue is setting up the integrals rather than the integration drudgery. Indeed, the real engineering issue is knowing why it is important to calculate the various moments.

7.4.1 Cylindrical Coordinates

If a solid can be conveniently described in cylindrical coordinates, then it is usually easier to do integrals on that solid in cylindrical coordinates. An element of volume is then

$$dV = r\, dr\, d\theta\, dz.$$

In the special case of **circular symmetry**, that is, when there is no dependence on θ in the integrand or in the limits of integration, it makes sense to write $dV = r\, dr\, dz\, d\theta$ or $dV = r\, dz\, dr\, d\theta$.

Example 7.23

A typical long bone shaft shown in Figure 7.37 has close to circular symmetry and diameter that is smaller in its middle. Also, a long bone has a thin outer shell made of dense "cortical" bone material and inside there is spongy "trabecular" bone material. Such a hollow design is strong but lightweight.* Model this description of the bone's shaft and write an integral for its mass.

Method: Say the shaft's longitudinal axis is part of the z-axis, specifically $-H \le z \le H$. The circular symmetry and narrowing of the bone, $\mathcal{V}$, can be modeled by having

$$\mathcal{V} = \{(r, \theta, z) : -H \le z \le H,\ 0 \le \theta \le 2\pi, 0 \le r \le f(z)\},$$

where f decreases for $-H \le z \le 0$ and increases for $0 \le z \le H$.

The thin outer shell of cortical bone has $f(z) - \varepsilon \le r \le f(z)$, where $\varepsilon > 0$ is the thickness of the shell. The trabecular bone has $0 \le r \le f(z) - \varepsilon$. So, we can model the mass density function by

$$\varrho = \varrho(r, z) = \begin{cases} \varrho_0, & 0 \le r \le f(z) - \varepsilon \\ \varrho_1, & f(z) - \varepsilon \le r \le f(z) \end{cases}$$

in g/cm^3, assuming distances are measured in cm, and ϱ_0, ϱ_1 are constants.

* Many bicycle frames are made of similar hollow tubes of narrowing diameter.

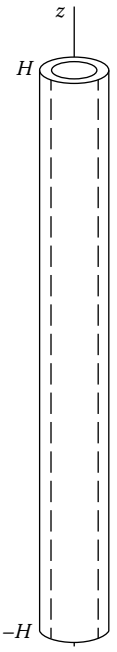


FIGURE 7.37
Shaft of a long bone.

The total mass of the bone, in g, is

$$\iiint_{\mathcal{V}} \varrho \, dV = \int_0^{2\pi} \int_{-H}^{H} \int_0^{f(z)} \varrho(r,z)\, r\, dr\, dz\, d\theta = 2\pi \int_{-H}^{H} \left(\int_0^{f(z)-\varepsilon} \varrho_0\, r\, dr + \int_{f(z)-\varepsilon}^{f(z)} \varrho_1\, r\, dr \right) dz$$

$$= 2\pi \int_{-H}^{H} \left(\varrho_0 \frac{1}{2}(f(z)-\varepsilon)^2 + \varrho_1 \frac{1}{2} \left((f(z))^2 - (f(z)-\varepsilon)^2 \right) \right) dz$$

$$= -\pi H \varepsilon^2 (\varrho_1 - \varrho_0) + 2\pi\varepsilon(\varrho_1 - \varrho_0) \int_{-H}^{H} f(z) dz + \pi \varrho_0 \int_{-H}^{H} (f(z))^2 dz. \bigcirc$$

Example 7.24

Find the z-coordinate of the center of mass of a homogeneous solid bounded below by the surface $z = 2(x^2 + y^2)$ and above by the surface $z = 6\left(1 - \frac{\sqrt{x^2+y^2}}{2}\right)$.

Method: The solid is shown in Figure 7.38. In cylindrical coordinates, the two bounding surfaces are the paraboloid $z = 2r^2$ and the cone $z = 6 - 3r$. First, let's find where the two surfaces intersect:

$$2r^2 = z = 6 - 3r \iff 0 = 2r^2 + 3r - 6.$$

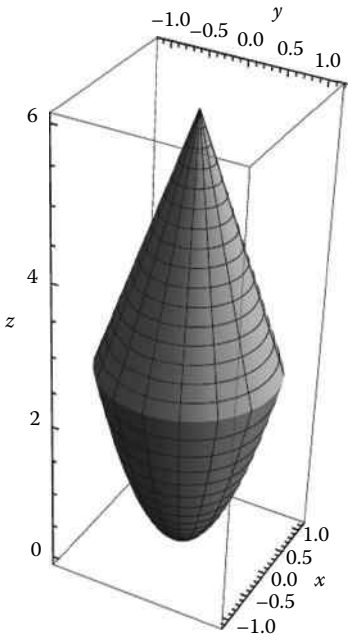


FIGURE 7.38
Example 7.24.

The latter is solvable by the quadratic formula:

$$r = \frac{-3 \pm \sqrt{57}}{4}.$$

Because $r \geq 0$ in cylindrical coordinates, the two surfaces intersect only at

$$r = \frac{-3 + \sqrt{57}}{4} \triangleq r_0. \tag{7.31}$$

If we were to shine a light down from $z = \infty$ onto the solid, its shadow in the xy-plane, that is, the $z = 0$ plane, would be the disk

$$0 \leq r \leq r_0.$$

We can think of our solid as consisting of "straws" that are vertical, that is, parallel to the z-axis, running from the bottom surface, $z = 2r^2$, to the top surface, $z = 6 - 3r$. So, in cylindrical coordinates, our solid is

$$\mathcal{V} = \{(r, \theta, z) : 0 \leq r \leq r_0,\ 2r^2 \leq z \leq 6 - 3r\}.$$

Because the solid is homogeneous, the mass density is a constant ϱ_0. The total mass of the solid is $M = \varrho_0 Volume$, that is,

$$M = \varrho_0 \int_0^{2\pi} \int_0^{r_0} \int_{2r^2}^{6-3r} r\,dz\,dr\,d\theta = 2\pi\varrho_0 \int_0^{r_0} r \cdot (6 - 3r - 2r^2)dr = 2\pi\varrho_0 \int_0^{r_0} (6r - 3r^2 - 2r^3)dr$$

$$= 2\pi\varrho_0 \left[3r^2 - r^3 - \frac{r^4}{2} \right]_0^{r_0} = \frac{\pi\varrho_0 r_0^2}{2} \left(12 - 4r_0 - 2r_0^2 \right) = \frac{\pi\varrho_0 r_0^2}{2} \left(12 - 4r_0 - (6 - 3r_0) \right).$$

So

$$M = \frac{\pi\varrho_0}{2} r_0^2(6 - r_0).$$

The moment with respect to the $z = 0$ plane is

$$M_{xy} = M_{z=0} = \varrho_0 \int_0^{2\pi} \int_0^{r_0} \int_{2r^2}^{6-3r} zrdzdrd\theta = \pi\varrho_0 \int_0^{r_0} r \cdot \left((6-3r)^2 - (2r^2)^2\right) dr$$

$$= \pi\varrho_0 \int_0^{r_0} (36r - 36r^2 + 9r^3 - 4r^5)dr = \pi\varrho_0 \left(18r_0^2 - 12r_0^3 + \frac{9r_0^4}{4} - \frac{2r_0^6}{3}\right)$$

$$= \pi\varrho_0 r_0^2 \left(18 - 12r_0 + \frac{9r_0^2}{4} - \frac{2r_0^4}{3}\right) = \cdots = \pi\varrho_0 r_0^2 \left(\frac{57}{4} - \frac{57r_0}{8}\right).$$

So, the z-coordinate of the center of mass is

$$\bar{z} = \frac{M_{xy}}{M} = \frac{\pi\varrho_0 r_0^2 \left(\frac{57}{4} - \frac{57r_0}{8}\right)}{\frac{\pi\varrho_0}{2} r_0^2(6 - r_0)} = \frac{2\left(\frac{57}{4} - \frac{57r_0}{8}\right)}{6 - r_0} = \frac{57(2 - r_0)}{4(6 - r_0)} = \frac{57\left(2 - \frac{-3+\sqrt{57}}{4}\right)}{4\left(6 - \frac{-3+\sqrt{57}}{4}\right)}.$$

$$= \frac{57(11 - \sqrt{57})}{4(27 - \sqrt{57})} \approx 2.527735. \bigcirc$$

"By eyeball," this looks reasonable, since the majority of the mass in the solid is below $z = 3$, although in the moment $M_{z=0}$, the mass higher up counts more than the mass near $z = 0$.

By the way, the surfaces bounding the solid were plotted by *Mathematica*: After defining r_0 by (7.31), we used the command

ParametricPlot3D[{{rCos[θ], rSin[θ], 2r^2},{rCos[θ], rSin[θ], 6 − 3r}}, {r, 0, r0}, {θ, 0, 2π}].

Example 7.25

$\mathcal{V}$ is the solid between the coaxial cylinders of radii 3 cm and 5 cm whose axis of symmetry is the x-axis and run from $x = -3$ cm to $x = 3$ cm. If $\mathcal{V}$ is filled with positive electric charge whose density in coulombs per cm^3 is proportional to the distance from the origin, find the total charge in $\mathcal{V}$.

Method: In an alternative version of cylindrical coordinates in which the x-axis is the axis of revolution, let $x = x$, $y = \mathrm{r}\cos\varphi$, and $z = \mathrm{r}\sin\varphi$. Then the two bounding surfaces are the cylinders $\mathrm{r} = 3$ and $\mathrm{r} = 5$, measured in cm. So,

$$\mathcal{V} = \{(x, \mathrm{r}, \varphi) : -3 \le x \le 3,\ 0 \le \varphi \le 2\pi, 3 < \mathrm{r} < 5\}.$$

Let σ be the charge density, so

$$\sigma = k\sqrt{\mathrm{r}^2 + x^2}$$

for some positive constant of proportionality k. In coulombs, the total charge in $\mathcal{V}$ is

$$Q = \iiint_{\mathcal{V}} \sigma\, dV = \int_0^{2\pi} \int_{-3}^{3} \int_3^5 k\sqrt{\mathrm{r}^2 + x^2}\, \mathrm{r}\, d\mathrm{r}\, dx\, d\varphi.$$

Using the substitution $w = \mathrm{r}^2 + x^2$, we get the anti-partial derivative

$$\int \sqrt{\mathrm{r}^2 + x^2}\, \mathrm{r}\, \partial \mathrm{r} = \int w^{1/2}\, \frac{\partial w}{2} = \frac{1}{3} w^{3/2} = \frac{1}{3}(\mathrm{r}^2 + x^2)^{3/2}.$$

There is no dependence on φ in the integrand or in the limits of integration, that is, there is circular symmetry, so

$$Q = (2\pi) \cdot \frac{k}{3} \int_{-3}^{3} \left[(\mathrm{r}^2 + x^2)^{3/2} \right]_{\mathrm{r}=3}^{\mathrm{r}=5} dx = \frac{2\pi k}{3} \int_{-3}^{3} \left((25 + x^2)^{3/2} - (9 + x^2)^{3/2} \right) dx.$$

An integral of the form $\int (a^2 + x^2)^{3/2} dx$ can be done using the inverse trigonometric substitution $x = a \tan \psi$, but it's fairly hard work. Instead, we use *Mathematica*, or a book of tables of integrals to get

$$\int (a^2 + x^2)^{3/2} dx = \sqrt{a^2 + x^2} \left(\frac{5a^2 x}{8} + \frac{x^3}{4} \right) + \frac{3a^4}{8} \ln \left| x + \sqrt{a^2 + x^2} \right|.$$

So, the total charge in $\mathcal{V}$ is

$$\begin{aligned}
Q = \frac{2\pi k}{3} \Bigg[& \sqrt{25 + x^2} \left(\frac{125x}{8} + \frac{x^3}{4} \right) + \frac{1875}{8} \ln \left| x + \sqrt{25 + x^2} \right| \\
& - \sqrt{9 + x^2} \left(\frac{45x}{8} + \frac{x^3}{4} \right) - \frac{243}{8} \ln \left| x + \sqrt{9 + x^2} \right| \Bigg]_{-3}^{3} \\
= \frac{2\pi k}{3} \Bigg(& \sqrt{34} \left(\frac{375}{8} + \frac{27}{4} \right) + \frac{1875}{8} \ln \left| 3 + \sqrt{34} \right| \\
& - \sqrt{34} \left(-\frac{375}{8} - \frac{27}{4} \right) - \frac{1875}{8} \ln \left| -3 + \sqrt{34} \right| \\
& - \sqrt{18} \left(\frac{135}{8} + \frac{27}{4} \right) - \frac{243}{8} \ln \left| 3 + \sqrt{18} \right| \Bigg) \\
& + \sqrt{18} \left(-\frac{135}{8} - \frac{27}{4} \right) + \frac{243}{8} \ln \left| -3 + \sqrt{18} \right| \Bigg) \\
= \frac{2\pi k}{3} \Bigg(& \frac{429}{4} \sqrt{34} + \frac{1875}{8} \ln \left(\frac{3 + \sqrt{34}}{-3 + \sqrt{34}} \right) - \frac{567}{4} \sqrt{2} - \frac{243}{8} \ln \left(\frac{3 + 3\sqrt{2}}{-3 + 3\sqrt{2}} \right) \Bigg) \\
= \frac{\pi k}{2} & \left(143\sqrt{34} - 189\sqrt{2} + 625 \ln \left(\frac{3 + \sqrt{34}}{5} \right) - 81 \ln(1 + \sqrt{2}) \right) \approx 1336.220\, k. \bigcirc
\end{aligned}$$

Definition 7.12

The average value of a function f on a solid $\mathcal{V}$ is

$$\bar{f} \triangleq \frac{1}{\textit{Volume of } \mathcal{V}} \iiint_{\mathcal{V}} f(\mathbf{r})\, dV = \frac{\iiint_{\mathcal{V}} f(\mathbf{r})\, dV}{\iiint_{\mathcal{V}} 1\, dV}.$$

7.4.2 Spherical Coordinates

In spherical coordinates, an element of volume is given by

$$dV = \rho^2 \sin\phi \, d\rho \, d\phi \, d\varphi,$$

as was shown in Figure 6.35 in Section 6.6.

Example 7.26

Find the z-coordinate of the centroid of the solid *ice cream cone* $\mathcal{V}$ bounded below by the surface that is the cone $\phi = \frac{\pi}{6}$ and from above by the sphere $\rho = 2\cos\phi$.

Method: Since the problem asks for the centroid, the mass density must be a constant ϱ_0, so the total mass is $M = \varrho_0 \cdot (\textit{Volume}\ \text{of}\ \mathcal{V})$.

One can think of $\mathcal{V}$ as consisting of radial straws that run from the origin to the upper surface $\rho = 2\cos\phi$, as shown in Figure 7.39. In spherical coordinates,

$$\mathcal{V} = \left\{(\rho, \phi, \theta) : \ 0 \le \theta \le 2\pi, \ 0 \le \phi \le \frac{\pi}{6}, 0 \le \rho \le 2\cos\phi\right\}.$$

The mass is given by

$$M = \iiint_{\mathcal{V}} \varrho_0 dV = \varrho_0 \int_0^{2\pi} \int_0^{\frac{\pi}{6}} \int_0^{2\cos\phi} \rho^2 \sin\phi \, d\rho d\phi d\theta = \varrho_0 \int_0^{2\pi} \int_0^{\frac{\pi}{6}} \left[\frac{\rho^3}{3}\right]_0^{2\cos\phi} \sin\phi \, d\phi d\theta$$

$$= (2\pi) \cdot \frac{8}{3} \cdot \varrho_0 \int_0^{\frac{\pi}{6}} \cos^3\phi \sin\phi \, d\phi = \frac{16\pi\varrho_0}{3} \left[-\frac{\cos^4\phi}{4}\right]_0^{\frac{\pi}{6}} = \frac{16\pi\varrho_0}{12}\left(-\cos^4\left(\frac{\pi}{6}\right) + 1\right)$$

$$= \frac{4\pi\varrho_0}{3}\left(1 - \left(\frac{\sqrt{3}}{2}\right)^4\right) = \frac{4\pi\varrho_0}{3} \cdot \frac{7}{16} = \frac{7\pi\varrho_0}{12}.$$

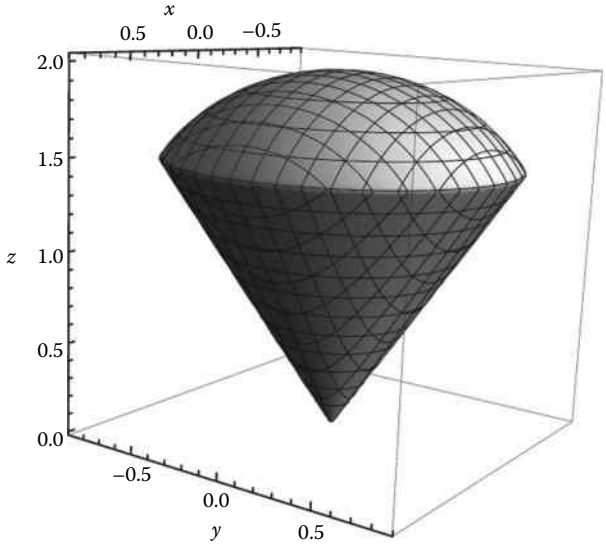


FIGURE 7.39
Solid $\mathcal{V}$ in Example 7.26.

The moment about the plane $z=0$ is

$$M_{xy} = M_{z=0} = \varrho_0 \iiint_{\mathcal{V}} z\, dV = \varrho_0 \int_0^{2\pi} \int_0^{\frac{\pi}{6}} \int_0^{2\cos\phi} (\rho\cos\phi)\rho^2 \sin\phi\, d\rho d\phi d\theta$$

$$= \varrho_0 \int_0^{2\pi} \int_0^{\frac{\pi}{6}} \left[\frac{\rho^4}{4}\right]_0^{2\cos\phi} \cos\phi\sin\phi\, d\phi d\theta = (2\pi)\cdot\frac{16}{4}\cdot\varrho_0 \int_0^{\frac{\pi}{6}} \cos^5\phi\sin\phi\, d\phi$$

$$= 8\pi\varrho_0 \cdot \left[-\frac{\cos^6\phi}{6}\right]_0^{\frac{\pi}{6}} = \frac{8\pi\varrho_0}{6}\left(-\cos^6\left(\frac{\pi}{6}\right)+1\right)$$

$$= \frac{4\pi\varrho_0}{3}\left(1-\left(\frac{\sqrt{3}}{2}\right)^6\right) = \frac{4\pi\varrho_0}{3}\cdot\frac{37}{64} = \frac{37\pi\varrho_0}{48}.$$

The z-coordinate of the centroid is at

$$\bar{z} = \frac{M_{xy}}{M} = \left(\frac{37\pi\varrho_0}{48}\right) \Big/ \left(\frac{7\pi\varrho_0}{12}\right) = \frac{37}{28}. \bigcirc$$

"By eyeball," this looks reasonable, since most of the ice cream is above $z=1$.

By the way, *Mathematica* drew the solid $\mathcal{V}$ using the command

```
RegionPlot3D[x^2 + y^2 ≤ z^2/3 && x^2 + y^2 + (z - 1)^2 ≤ 1, {x, -Sqrt[3]/2, Sqrt[3]/2}, {y, -Sqrt[3]/2, Sqrt[3]/2},
  {z, 0, 2}, AxesLabel → {"x","y","z"}]
```

followed by rotation with a mouse. This command effectively thinks of $\mathcal{V}$ as a union of vertical straws that run from the cone $z = r\sqrt{3}$ up to the sphere $z = 1+\sqrt{1-r^2}$, for (x,y) in the disk $x^2+y^2 \le \frac{3}{4}$.

Example 7.27

Evaluate

$$\mathcal{I} \triangleq \iiint_{\mathcal{V}} \sqrt{x^2+y^2+z^2}\, e^{-(x^2+y^2+z^2)/(2\alpha)} dV,$$

where

α is a positive constant

$\mathcal{V}$ is the solid that lies between the two concentric spheres $\rho=1$ and $\rho=3$ in the positive octant $x\ge 0, y\ge 0, z\ge 0$

Method: The solid is shown in Figure 7.40. Because of the spherical symmetry in the integrand, we should try using spherical coordinates. To be in the positive octant, we need (x,y) to be in the positive quadrant of the xy-plane, so $0 \le \theta \le \frac{\pi}{2}$. To be in the northern hemisphere, we need $0 \le \phi \le \frac{\pi}{2}$. So,

$$\mathcal{V} = \left\{(\rho,\phi,\theta):\ 0\le\theta\le\frac{\pi}{2},\ 0\le\phi<\frac{\pi}{2},\ 1\le\rho\le 3\right\}.$$

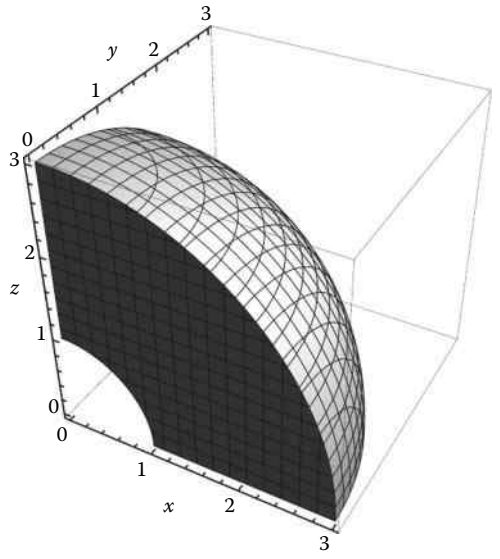


FIGURE 7.40
Solid $\mathcal{V}$ in Example 7.27.

So,

$$\mathcal{I} \triangleq \iiint_{\mathcal{V}} \sqrt{x^2+y^2+z^2}\, e^{-(x^2+y^2+z^2)/(2\alpha)} dV = \int_0^{\frac{\pi}{2}} \int_0^{\frac{\pi}{2}} \int_1^3 \rho e^{-\rho^2/(2\alpha)} \rho^2 \sin\phi \, d\rho d\phi d\theta$$

$$= (2\pi) \int_0^{\frac{\pi}{2}} \int_1^3 \rho^3 e^{-\rho^2/(2\alpha)} \sin\phi \, d\rho d\phi = 2\pi \left(\int_0^{\frac{\pi}{2}} \sin\phi \, d\phi \right) \left(\int_1^3 \rho^3 e^{-\rho^2/(2\alpha)} d\rho \right).$$

Using the substitution $\rho^2 = -2\alpha w$, so $2\rho d\rho = -2\alpha dw$, and then integration by parts, this gives

$$\mathcal{I} = 2\pi \left([-\cos\phi]_0^{\frac{\pi}{2}} \right) \left(\int_{-1/(2\alpha)}^{-9/(2\alpha)} (-2\alpha w) e^w (-\alpha dw) \right) = 2\pi \cdot 1 \cdot (2\alpha^2) \left(\left[(w-1)e^w \right]_{-1/(2\alpha)}^{-9/(2\alpha)} \right)$$

$$= 4\pi\alpha^2 \left(\left(\frac{1}{2\alpha} + 1 \right) e^{-1/(2\alpha)} - \left(\frac{9}{2\alpha} + 1 \right) e^{-9/(2\alpha)} \right). \bigcirc$$

As a reality check, is the final conclusion for $\mathcal{I}$ positive? It *should* be, because $\mathcal{I}$ is the definite integral of a nontrivial, nonnegative continuous function on a solid $\mathcal{V}$ that has positive (i.e., nonzero) volume.

If we define $F(x) \triangleq (1+x)e^{-x}$, then

$$\mathcal{I} = 4\pi\alpha^2 \left(F\left(\frac{1}{2\alpha}\right) - F\left(\frac{9}{2\alpha}\right) \right).$$

So, if we can explain why $F(x)$ is a strictly decreasing function for x in the interval $\left(\frac{1}{2\alpha}, \frac{9}{2\alpha}\right)$, then we can conclude $\mathcal{I} > 0$. In fact, $F'(x) = -xe^{-x} < 0$ at all $x > 0$, so yes, $\mathcal{I}$ passes the reality check of being positive.

By the way, *Mathematica* drew the solid $\mathcal{V}$ using the command

```
RegionPlot3D[1 ≤ x^2 + y^2 + z^2 && x^2 + y^2 + z^2 ≤ 9, {x, 0, 3}, {y, 0, 3}, {z, 0, 3},
  AxesLabel → {"x","y","z"}].
```

Example 7.28

Find the volume of the solid $\mathcal{V}$ bounded by the surfaces $z = 4$ and $\phi = \frac{\pi}{6}$.

Method: We could solve this problem as a triple integral, but it is faster and easier to use our knowledge of geometrical formulas. Our $\mathcal{V}$ is a solid, circular cone whose "base" is in the plane $z = 4$ and whose side is a finite part of the cone $\phi = \frac{\pi}{6}$. To find the radius of that base, spherical coordinates have $z = \rho\cos\phi$; hence,

$$\rho = \frac{z}{\cos\phi} = \frac{4}{\cos\frac{\pi}{6}} = \frac{4}{\frac{\sqrt{3}}{2}} = \frac{8}{\sqrt{3}}.$$

Cylindrical and spherical coordinates are connected by

$$\sqrt{x^2+y^2} = r = \rho\sin\phi = \frac{8}{\sqrt{3}}\cdot\sin\frac{\pi}{6} = \frac{8}{\sqrt{3}}\cdot\frac{1}{2} = \frac{4}{\sqrt{3}}.$$

It follows that the base of the cone has radius $R = \frac{4}{\sqrt{3}}$ and height $H = 4$, so $\mathcal{V}$ has

$$\textit{Volume} = \frac{\pi R^2 H}{3} = \frac{\pi\left(\frac{4}{\sqrt{3}}\right)^2\cdot 4}{3} = \frac{\pi\left(\frac{16}{3}\right)4}{3} = \frac{64\pi}{9}. \bigcirc$$

Learn More About It

A good reference concerning bones is *Physics of the Human Body*, by Irving P. Herman, Springer-Verlag, © 2007.

7.4.3 Problems

1. Set up an iterated integral for $\iiint_{\mathcal{V}} f(x,y,z)\,dV$, where $\mathcal{V}$ is the solid tetrahedron whose vertices are $(0,0,0)$, $(1,0,0)$, $(0,2,0)$, $(0,0,3)$. Write your final conclusion in the form $\int\_\_\int\_\_\_\_\int\_\_\_\_\_\_\_\_ d\_\_\,d\_\_\,d\_\_$ and fill in the blanks.
2. Set up an iterated integral for $\iiint_{\mathcal{V}} f(x,y,z)\,dV$, where $\mathcal{V}$ is the solid tetrahedron whose vertices are $(0,0,0)$, $(2,0,0)$, $(0,3,0)$, $(0,0,1)$. Write your final conclusion in the form $\int\_\_\int\_\_\_\_\int\_\_\_\_\_\_\_\_ d\_\_\,d\_\_\,d\_\_$ and fill in the blanks.
3. Find the y-coordinate of the centroid of the solid tetrahedron whose vertices are $(0,0,0)$, $(1,0,0)$, $(0,3,0)$, $(0,0,2)$.
4. Rewrite the triple integral in the following in the form indicated by filling in the blanks with specific expressions. Give an appropriate sketch(es) of the solid $\mathcal{V}$ and region(s) $\mathcal{D}$ for the iterated triple integral on the RHS, that is, your final conclusion.

Your final conclusion should be in the form

$$\int_0^1 \int_0^{4-4z} \int_0^{2-\frac{y}{2}-2z} f(x,y,z)\,dx\,dy\,dz = \int_{\_\_}^{\_\_} \int_{\_\_}^{\_\_} \int_{\_\_}^{\_\_} f(x,y,z)\,dz\,dx\,dy$$

with blanks filled in. The function $f(x,y,z)$ is not specified.

5. Rewrite the triple integral in the following in the form indicated by filling in the blanks with specific expressions. Give an appropriate sketch(es) of the solid $\mathcal{V}$ and region(s) $\mathcal{D}$ for the iterated triple integral on the RHS, that is, your final conclusion. Your final conclusion should be in the form

$$\int_0^4 \int_0^{2-\frac{z}{2}} \int_0^{1-\frac{x}{2}-\frac{z}{4}} f(x,y,z)\,dy\,dx\,dz = \int_{\_\_}^{\_\_} \int_{\_\_}^{\_\_} \int_{\_\_}^{\_\_} f(x,y,z)\,dz\,dy\,dx$$

with blanks filled in. The function $f(x,y,z)$ is not specified.

6. For the solid of Example 7.28, use triple integrals to find the volume and the centroid. [Hint: Symmetry will make short work of two of the coordinates of the centroid.]
7. Find the centroid of the solid that is in the first octant, that is, satisfies $x \geq 0, y \geq 0, z \geq 0$, lies above the surface $z = 1$, and lies below the surface $x^2 + y^2 + z = 4$.
8. Find the centroid of the solid that is in the first octant, that is, satisfies $x \geq 0, y \geq 0, z \geq 0$, lies above the surface $z = 0$, and lies below the surface $x^2 + y^2 + z^2 = 4$.
9. Evaluate $\iiint_{\mathcal{V}} (x^2 + y^2)\,dV$, where $\mathcal{V}$ is the solid bounded by the cylinder $x^2 + y^2 = 4$ and the planes $z = -1$ and $y + z = 2$.
10. Redo Example 7.26 using cylindrical coordinates.
11. Let $\mathcal{V}$ be the *finite type I* solid bounded above by the plane $z = 4$ and bounded below by the positive octant part of the plane $x + 3y + 4z = 12$. [The positive octant consists of all points having $x \geq 0$, $y \geq 0$, and $z \geq 0$.] Set up, with as much specificity as possible, $\iiint_{\mathcal{V}} f(x,y,z)\,dV$.
12. Let $\mathcal{V}$ be the solid bounded below by the circular paraboloid $z = x^2 + y^2$ and bounded above by the circular paraboloid $z = 4 - x^2 - y^2$. Set up, with as much specificity as possible, $\iiint_{\mathcal{V}} f(x,y,z)\,dV$.
13. Let $\mathcal{V}$ be the solid bounded below by the circular paraboloid $z = 2(x^2 + y^2)$ and bounded above by the circular paraboloid $z = 9 - x^2 - y^2$. Find the z-coordinate of the centroid.
14. Find the volume of the solid that lies above the surface $z = 3\sqrt{x^2 + y^2}$ and below the surface $z = 4$.
15. Find the polar moment of inertia (for rotation about the z-axis) of the homogeneous solid bounded by the surfaces $z = 2 - x^2 - y^2$ and $z = \sqrt{x^2 + y^2}$.
16. Find the volume of the solid $\mathcal{V} = \{(\rho, \phi, \theta) : 0 \leq \theta \leq 2\pi, 0 \leq \phi \leq \frac{\pi}{2}, 0 \leq \rho \leq 4\cos\phi\}$.
17. The solid $\mathcal{V}$ is the $x \geq 0$ half of the solid between the concentric spheres of radii 2 cm and 3 cm whose centers are at $(0,0,0)$. Suppose $\mathcal{V}$ is filled with material whose

density, in g/cm^3, is proportional to the distance from the yz-plane. Find the total mass of $\mathcal{V}$ in terms of the constant of proportionality.

18. (a) If a thin tube were drilled completely through a large solid ball, then the force of gravity on an object in that tube would have magnitude $|F| = k|z|$, for some constant k, assuming the tube runs along the z-axis, the center of the ball is at the origin, and the ball is homogeneous, that is, has constant mass density. Explain this result using a triple integral in cylindrical coordinates.
    By the way, a planet is approximately a solid ball.
    (b) This would be very difficult to implement if the ball is the Earth because of its molten metal core. Even Mars has a partially molten core. However, for the sake of argument, suppose we could do this on Mars. According to Wikipedia, Mars has a polar radius of about 3396 km and on its surface exerts a force of gravity about 0.376 that of the force of gravity on the surface of the Earth. Find the value of the constant k if the tube is drilled from pole to pole on Mars.
19. In single variable calculus, you studied the method of disks and washers for finding the volume of a solid of revolution. For example, if $f(x) \geq g(x)$ for $a \leq x \leq b$ and the finite region in the xy-plane bounded by the curves $y = f(x), y = g(x), x = a, x = b$ is rotated around the x-axis, then the volume of the solid generated is $\pi \int_a^b \left(\big(f(x)\big)^2 - \big(g(x)\big)^2 \right) dx$. Use the alternative version of cylindrical coordinates of Example 7.25 and a triple integral to explain why the "method of washers" is correct. So, you will explain why what we studied in Section 7.4 agrees with a result from single variable calculus.
20. Use the single variable calculus method of disks to check the volume calculation that is part of Example 7.24. Can we use the method of disks to do the rest of Example 7.24, that is, to find the z-coordinate of the centroid? If so, use it; if not, why not?
21. In single variable calculus, you may have studied the "method of cylindrical shells" for finding the volume of a solid of revolution. For example, if $f(x) \geq g(x)$ for $a \leq x \leq b$ and the finite region in the xy-plane bounded by the curves $y = f(x), y = g(x), x = a, x = b$ is rotated around the y-axis, then the volume of the solid generated is $2\pi \int_a^b x \big(f(x) - g(x)\big) dx$. Use an alternative version of cylindrical coordinates and a triple integral to explain why method of cylindrical shells is correct. So, you will explain why what we studied in Section 7.4 agrees with another result from single variable calculus.

7.5 Surface Integrals and Applications

Recall from Section 6.3 that the area of a parallelogram whose sides are vectors $\mathbf{A}$ and $\mathbf{B}$ is $||\mathbf{A} \times \mathbf{B}||$. Recall from Section 6.3 that a parametric surface is described by a position vector that is a function of two parameters, that is,

$$\mathcal{S}: \ \mathbf{r} = \mathbf{r}(u, v), \ (u, v) \text{ in } \mathcal{D}, \tag{7.32}$$

where $\mathcal{D}$ is a planar region. Moreover, at any point P whose position vector is $\mathbf{r}(u_0, v_0)$ on $\mathcal{S}$, the two vectors

$$\frac{\partial \mathbf{r}}{\partial u}(u_0, v_0) \text{ and } \frac{\partial \mathbf{r}}{\partial v}(u_0, v_0)$$

shown in Figure 7.41 span the tangent plane at P. An element of surface area is given by (6.64) in Section 6.6, that is,

$$dS \triangleq \left|\left| \frac{\partial \mathbf{r}}{\partial u}(u_0, v_0)\, du \times \frac{\partial \mathbf{r}}{\partial v}(u_0, v_0)\, dv \right|\right| = \left|\left| \frac{\partial \mathbf{r}}{\partial u}(u_0, v_0) \times \frac{\partial \mathbf{r}}{\partial v}(u_0, v_0) \right|\right| du\, dv. \tag{7.33}$$

Example 7.29

In Example 6.18 in Section 6.3, we considered a sphere of radius a:

$$\mathcal{S}: \ \mathbf{r} = \mathbf{r}(\phi, \theta) = a(\sin\phi\ \cos\theta\, \hat{\boldsymbol{\imath}} + \sin\phi\ \sin\theta\, \hat{\boldsymbol{\jmath}} + \cos\phi\, \hat{\boldsymbol{k}}),\ 0 \le \phi \le \pi,\ 0 \le \theta \le 2\pi,$$

that is,

$$\mathcal{S}: \ \mathbf{r} = a\hat{\mathbf{e}}_\rho,\ 0 \le \phi \le \pi,\ 0 \le \theta \le 2\pi.$$

Find the element of surface area.

Method: We calculate

$$\frac{\partial \mathbf{r}}{\partial \phi}(\phi, \theta) = a(\cos\phi\ \cos\theta\, \hat{\boldsymbol{\imath}} + \cos\phi\ \sin\theta\, \hat{\boldsymbol{\imath}} - \sin\phi\, \hat{\boldsymbol{k}}) = a\mathbf{e}_\phi$$

and

$$\frac{\partial \mathbf{r}}{\partial \theta}(\phi, \theta) = a(-\sin\phi\ \sin\theta\, \hat{\boldsymbol{\imath}} + \sin\phi\ \cos\theta\, \hat{\boldsymbol{\jmath}}) = a\sin\phi\, \hat{\mathbf{e}}_\theta;$$

hence,

$$dS = \left|\left| \frac{\partial \mathbf{r}}{\partial \phi}(\phi, \theta) \times \frac{\partial \mathbf{r}}{\partial \theta}(\phi, \theta) \right|\right| d\phi\, d\theta = \left|\left| a\mathbf{e}_\phi \times a\sin\phi\, \hat{\mathbf{e}}_\theta \right|\right| d\phi\, d\theta = a^2 \left|\left| \hat{\mathbf{e}}_\rho \right|\right| \ |\sin\phi|\ d\phi\, d\theta.$$

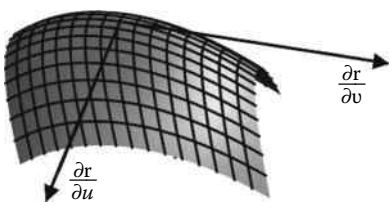


FIGURE 7.41
$\frac{\partial \mathbf{r}}{\partial u}$ and $\frac{\partial \mathbf{r}}{\partial v}$.

Because $\sin\phi \geq 0$ for $0 \leq \phi \leq \pi$, $|\sin\phi| = \sin\phi$. So, the element of surface area on a sphere of radius a is

$$dS = a^2 \sin\phi \, d\phi \, d\theta. \bigcirc$$

At $\phi = \frac{\pi}{2}$, that is, the equator, we get the largest element of surface area. Note also that because $\sin\phi \geq 0$ for $0 \leq \phi \leq \pi$, an element of surface area must be positive. (If you discover a negative element of surface area, write up a paper about it, have it published by a reputable physics journal, and then claim your Nobel Prize in physics.)

The total surface area of a surface $\mathcal{S}: \ \mathbf{r} = \mathbf{r}(u, v), \ (u, v) \ \text{in} \ \mathcal{D}$, is

$$Area(\mathcal{S}) \triangleq \iint_{\mathcal{S}} 1 \, dS \triangleq \iint_{\mathcal{D}} \left\| \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right\| du \, dv. \tag{7.34}$$

Let's check that this formula makes sense in a specific example.

Example 7.30

We will use a geometrical process to visualize the planar surface of an annulus and then see how its area is naturally calculated using formula (7.34).

As shown in Figure 7.42, let the point P be at $\mathbf{r} = a\cos\theta \, \hat{\boldsymbol{\imath}} + a\sin\theta \, \hat{\boldsymbol{\jmath}}$, that is, P is a point on a circle of radius a and center at the origin. On the radial line segment from P, put a point Q at

$$\mathbf{r} = (a\cos\theta \, \hat{\boldsymbol{\imath}} + a\sin\theta \, \hat{\boldsymbol{\jmath}}) + v(\cos\theta \, \hat{\boldsymbol{\imath}} + \sin\theta \, \hat{\boldsymbol{\jmath}}),$$

that is, Q lies on the same radial line as P, but its distance from the origin is $(a+v)$, where $0 \leq v \leq 1$. Put all such points together to form a surface,

$$\mathcal{S}: \ \mathbf{r} = (a+v)\cos\theta \, \hat{\boldsymbol{\imath}} + (a+v)\sin\theta \, \hat{\boldsymbol{\jmath}}, 0 \leq \theta \leq 2\pi, \ 0 \leq v \leq 1.$$

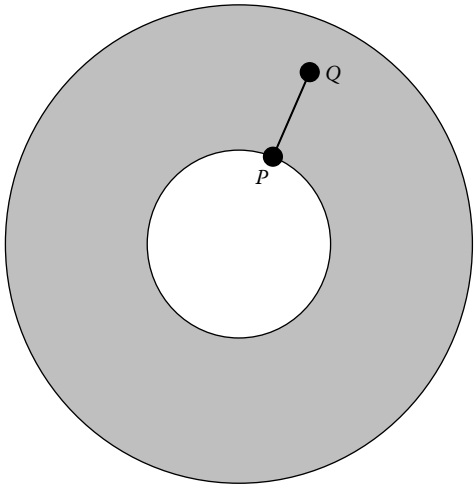


FIGURE 7.42
Annulus geometry.

In effect, we have constructed the annulus as the union of radial line segments. Find the area of the planar surface $\mathcal{S}$.

Method: We have, using the notation of cylindrical coordinates,

$$\mathcal{S}: \ \mathbf{r} = (a+v)\,\hat{\mathbf{e}}_r.$$

So,

$$dS = \left|\left|\frac{\partial \mathbf{r}}{\partial \theta} \times \frac{\partial \mathbf{r}}{\partial v}\right|\right| d\theta\, dv = \left|\left|(a+v)\hat{\mathbf{e}}_\theta \times \hat{\mathbf{e}}_r\right|\right| d\theta\, dv = \left|\left|(a+v)(-\hat{\boldsymbol{k}})\right|\right| d\theta\, dv$$

$$= |a+v|\, d\theta\, dv = (a+v)\, d\theta\, dv.$$

The total surface area of $\mathcal{S}$ is

$$Area(\mathcal{S}) = \iint_{\mathcal{S}} 1\, dS = \int_0^1 \int_0^{2\pi} (a+v)\, d\theta\, dv = 2\pi \int_0^1 (a+v) dv = 2\pi \left[av + \frac{v^2}{2}\right]_0^1 = \pi(2a+1). \bigcirc$$

Let's check this result: The area of the annulus $a \le r \le (a+1)$ is $\pi\left((a+1)^2 - a^2\right) = \pi(2a+1)$, which agrees with the conclusion of Example 7.30.

Example 7.31

In Example 7.29, we found that $dS = a^2 \sin\phi\, d\phi\, d\theta$ is the element of surface area on a sphere of radius a. Find the total surface area of the sphere.

Method: The sphere has parametrization

$$\mathcal{S}: \ \mathbf{r} = a\,\hat{\mathbf{e}}_\rho,\ 0 \le \phi \le \pi,\ 0 \le \theta \le 2\pi,$$

so $\mathcal{D} = \{(\phi, \theta): \ 0 \le \phi \le \pi,\ 0 \le \theta \le 2\pi\}$. The sphere has total

$$Surface\ area = \iint_{\mathcal{S}} 1\, dS = \int_0^{2\pi} \int_0^{\pi} a^2 \sin\phi\, d\phi d\theta = \left(\int_0^{2\pi} a^2 d\theta\right)\left(\int_0^{\pi} \sin\phi\, d\phi\right)$$

$$= 2\pi a^2\, [-\cos\phi\, d\phi]_0^{\pi} = 2\pi a^2 \cdot 2 = 4\pi a^2. \bigcirc$$

Surfaces $z = g(x, y)$

Suppose a surface is parametrized by (x, y), specifically

$$\mathcal{S}: \ \mathbf{r} = x\,\hat{\boldsymbol{\imath}} + y\,\hat{\boldsymbol{\jmath}} + g(x,y)\,\hat{\boldsymbol{k}},\ (x,y) \text{ in } \mathcal{D}.$$

Then

$$\frac{\partial \mathbf{r}}{\partial x} \times \frac{\partial \mathbf{r}}{\partial y} = \left(\hat{\boldsymbol{\imath}} + \frac{\partial g}{\partial x}\,\hat{\boldsymbol{k}}\right) \times \left(\hat{\boldsymbol{\jmath}} + \frac{\partial g}{\partial y}\,\hat{\boldsymbol{k}}\right)$$

$$= -\frac{\partial g}{\partial x}\,\hat{\boldsymbol{\imath}} - \frac{\partial g}{\partial y}\,\hat{\boldsymbol{\jmath}} + \hat{\boldsymbol{k}}. \tag{7.35}$$

So,

$$dS = \left|\left|\frac{\partial \mathbf{r}}{\partial x} \times \frac{\partial \mathbf{r}}{\partial y}\right|\right| dx\,dy = \left|\left| -\frac{\partial g}{\partial x}\,\hat{\boldsymbol{\imath}} - \frac{\partial g}{\partial y}\,\hat{\boldsymbol{\jmath}} + \hat{\boldsymbol{k}}\right|\right| dx\,dy = \sqrt{\left(\frac{\partial g}{\partial x}\right)^2 + \left(\frac{\partial g}{\partial y}\right)^2 + 1}\; dx\,dy. \tag{7.36}$$

7.5.1 Surface Integral of a Scalar-Valued Function

Over a surface

$$\mathcal{S}:\ \mathbf{r} = \mathbf{r}(u,v),\ (u,v) \text{ in } \mathcal{D},$$

the integral of a scalar-valued function $f(x,y,z) = f(\mathbf{r})$ is defined to be

$$\iint_{\mathcal{S}} f(x,y,z)\,dS \triangleq \iint_{\mathcal{D}} f\big(x(u,v), y(u,v), z(u,v)\big) \left|\left|\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v}\right|\right| du\,dv. \tag{7.37}$$

Equation (7.37) is analogous to the definition of line integrals (7.5), that is,

$$\int_{\mathcal{C}} f(\mathbf{r})\,ds \triangleq \int_{\alpha}^{\beta} f(\mathbf{r}(t))\,||\dot{\mathbf{r}}(t)||\,dt.$$

Example 7.32

Find the centroid of the portion of the plane $x+2y+3z=4$ that lies in the positive octant, that is, satisfies $x \geq 0, y \geq 0, z \geq 0$.

Method: Figure 7.43 shows our surface, $z = \frac{4}{3} - \frac{x}{3} - \frac{2y}{3}$, which can be parametrized as

$$\mathcal{S}:\ \mathbf{r} = x\,\hat{\boldsymbol{\imath}} + y\,\hat{\boldsymbol{\jmath}} + \left(\frac{4}{3} - \frac{x}{3} - \frac{2y}{3}\right)\hat{\boldsymbol{k}},\ (x,y) \text{ in } \mathcal{D},$$

where $\mathcal{D}$ is the shadow of the surface if a light is shone down on it from $z = \infty$.

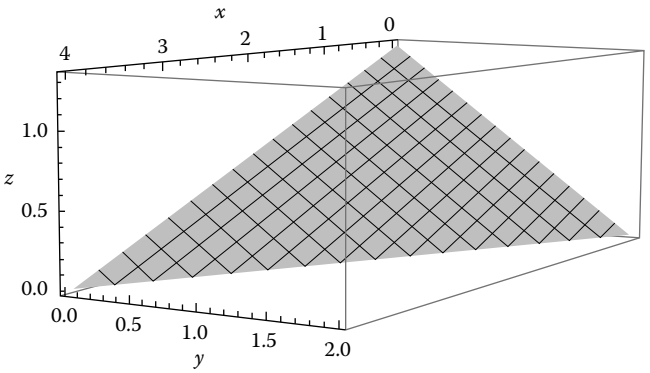


FIGURE 7.43
Surface in Example 7.32.

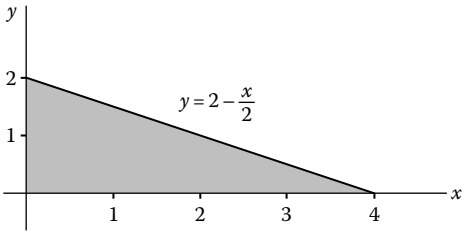


FIGURE 7.44
Triangle in Example 7.32.

Formula (7.36) calculates that $dS = \sqrt{(-\frac{1}{3})^2 + (-\frac{2}{3})^2 + 1}\, dx\, dy = \frac{\sqrt{14}}{3}\, dx\, dy$.

Since the problem asks for the centroid, the mass density must be a constant ϱ_0, so the total mass is

$$M = \iint_{\mathcal{S}} \varrho_0\, dS = \varrho_0 \iint_{\mathcal{D}} \frac{\sqrt{14}}{3}\, dx\, dy = \varrho_0\, \frac{\sqrt{14}}{3} (\textit{Area of } \mathcal{D}).$$

At this point, we must find $\mathcal{D}$: The shadow of $\mathcal{S}$ in the xy-plane is the triangle whose vertices are $(0,0)$, $(4,0)$, $(0,2)$, as shown in Figure 7.44, so

$$\mathcal{D} = \left\{(x,y) :\ 0 \le x \le 4,\ 0 \le y \le 2 - \frac{x}{2}\right\}$$

has area $\frac{1}{2}$ *base* $\times$ *height* $= 4$. So, the total mass is

$$M = \varrho_0 \frac{\sqrt{14}}{3} \cdot 4 = \frac{4\sqrt{14}}{3} \varrho_0.$$

By the way, the surface $\mathcal{S}$ is itself a triangle with vertices $(4,0,0)$, $(0,2,0)$, $(0,0,\frac{4}{3})$. In Example 6.28 in Section 6.6, we found that its *Area* $= \frac{1}{2}$ *base* $\times$ *height* $= \cdots = \frac{4\sqrt{14}}{3}$, which agrees with our calculation in this example.

Now let's calculate the moments:

$$M_{z=0} \triangleq \iint_{\mathcal{S}} z\varrho_0\, dS = \varrho_0 \int_0^4 \int_0^{2-\frac{x}{2}} z\, \frac{\sqrt{14}}{3}\, dy\, dx.$$

But, on the surface, $z = \frac{1}{3}(4 - x - 2y)$, so

$$M_{z=0} = \varrho_0 \frac{\sqrt{14}}{3} \int_0^4 \int_0^{2-\frac{x}{2}} \left(\frac{4}{3} - \frac{x}{3} - \frac{2y}{3}\right) dy\, dx = \varrho_0 \frac{\sqrt{14}}{9} \int_0^4 \left(\left[4y - xy - y^2\right]_0^{2-\frac{x}{2}}\right) dx$$

$$= \varrho_0 \frac{\sqrt{14}}{9} \int_0^4 \cdot \left(4\left(2 - \frac{x}{2}\right) - x\left(2 - \frac{x}{2}\right) - \left(2 - \frac{x}{2}\right)^2\right) dx = \varrho_0 \frac{\sqrt{14}}{9} \int_0^4 \left(4 - 2x + \frac{x^2}{4}\right) dx$$

$$= \varrho_0 \frac{\sqrt{14}}{9} \cdot \left[4x - x^2 + \frac{x^3}{12}\right]_0^4 = \frac{16\varrho_0\sqrt{14}}{27}.$$

Similarly,

$$M_{x=0} \triangleq \iint_{\mathcal{S}} x\varrho_0 \, dS = \varrho_0 \int_0^4 \int_0^{2-\frac{x}{2}} x \, \frac{\sqrt{14}}{3} \, dy \, dx = \varrho_0 \int_0^4 \frac{\sqrt{14}}{3} x \left(2 - \frac{x}{2}\right) dx$$

$$= \varrho_0 \, \frac{\sqrt{14}}{3} \cdot \left[x^2 - \frac{x^3}{6}\right]_0^4 = \frac{16\varrho_0\sqrt{14}}{9}$$

and

$$M_{y=0} \triangleq \iint_{\mathcal{S}} y\varrho_0 \, dS = \varrho_0 \int_0^4 \int_0^{2-\frac{x}{2}} y \, \frac{\sqrt{14}}{3} \, dy \, dx = \varrho_0 \int_0^4 \frac{\sqrt{14}}{3} \cdot \frac{1}{2} \left(2 - \frac{x}{2}\right)^2 dx$$

$$= \varrho_0 \, \frac{\sqrt{14}}{3} \cdot \left[-\frac{1}{3}\left(2 - \frac{x}{2}\right)^3\right]_0^4 = \frac{8\varrho_0\sqrt{14}}{9}.$$

The centroid is at

$$(\bar{x}, \bar{y}, \bar{z}) = \frac{1}{M}(M_{x=0}, M_{y=0}, M_{z=0}) = \frac{3}{4\varrho_0\sqrt{14}} \left(\frac{16\varrho_0\sqrt{14}}{9}, \frac{8\varrho_0\sqrt{14}}{9}, \frac{16\varrho_0\sqrt{14}}{27}\right) = \left(\frac{4}{3}, \frac{2}{3}, \frac{4}{9}\right). \bigcirc$$

7.5.2 Surface Integral of a Vector Field

Suppose a parametrized surface is given by

$$\mathcal{S}: \ \mathbf{r} = \mathbf{r}(u, v), \ (u, v) \text{ in } \mathcal{D}.$$

At a point P whose position vector is $\mathbf{r}(u_0, v_0)$, the two vectors

$$\frac{\partial \mathbf{r}}{\partial u}(u_0, v_0) \text{ and } \frac{\partial \mathbf{r}}{\partial v}(u_0, v_0)$$

span the tangent plane at P. It follows that the vector

$$\mathbf{n} \triangleq \frac{\partial \mathbf{r}}{\partial u}(u_0, v_0) \times \frac{\partial \mathbf{r}}{\partial v}(u_0, v_0)$$

is normal to the tangent plane at P, as long as $\mathbf{n} \neq \mathbf{0}$. For nonzero $\mathbf{n}$,

$$\widehat{\mathbf{n}} \triangleq \pm \left(\frac{1}{\left|\left| \frac{\partial \mathbf{r}}{\partial u}(u_0, v_0) \times \frac{\partial \mathbf{r}}{\partial v}(u_0, v_0) \right|\right|} \right) \frac{\partial \mathbf{r}}{\partial u}(u_0, v_0) \frac{\partial \mathbf{r}}{\partial v}(u_0, v_0)$$

is a unit vector normal to $\mathcal{S}$ at P. In some sense, $\widehat{\mathbf{n}}$ gives the direction of travel through $\mathcal{S}$ at P. Note that $\widehat{\mathbf{n}}$ is a function of the position vector $\mathbf{r}$, that is, can vary as we move along the surface, but we usually suppress the dependence on $\mathbf{r}$ from the notation for the unit normal.

Given a vector-valued function $\mathbf{F} = \mathbf{F}(\mathbf{r})$, that is, a vector field, we can define

$$\mathcal{I} \triangleq \iint_{\mathcal{S}} \big(\mathbf{F}(\mathbf{r}) \bullet \widehat{\mathbf{n}}\big) dS \triangleq \iint_{\mathcal{D}} \big(\mathbf{F}(\mathbf{r}(\mathbf{u}, \mathbf{v})) \bullet \widehat{\mathbf{n}}(u, v)\big) du \, dv \tag{7.38}$$

to be the surface integral of $\mathbf{F}$ on $\mathcal{S}$. One of the two choices for $\widehat{\mathbf{n}}$ gives

$$\mathcal{I} = \iint_{\mathcal{D}} \mathbf{F}(\mathbf{r}) \bullet \left(\frac{1}{\cancel{\left\| \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right\|}} \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right) \left(\cancel{\left\| \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right\|} du dv \right) = \iint_{\mathcal{D}} \mathbf{F}(\mathbf{r}) \bullet \left(\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right) du dv. \tag{7.39}$$

Another name for the surface integral of a vector field is the **flux of F across** $\mathcal{S}$, whose physical significance we will discuss further.

Another notation for such integrals is

$$\iint_{\mathcal{S}} \mathbf{F}(\mathbf{r}) \bullet d\mathbf{S}.$$

Here we think of $d\mathbf{S} \triangleq \widehat{\mathbf{n}}\, dS$ as a vector-valued element of surface area pointing in the direction of travel across $\mathcal{S}$.

Example 7.33

Set up an iterated double integral for the flux of a vector field out of a sphere. Be as specific as possible, for example, get a double integral with specific limits of integration.

Method: The problem is a little vague because it does not say what sphere. It makes sense to keep it simple: Assume the sphere has center at the origin and has radius a, that is, is $x^2 + y^2 + z^2 = a^2$. So,

$$\mathcal{S}: \ \mathbf{r} = a\, \hat{\mathbf{e}}_\rho, \ 0 \le \theta \le 2\pi, \ 0 \le \phi \le \pi.$$

As we saw in Example 7.29,

$$\pm \frac{\partial \mathbf{r}}{\partial \phi}(\phi, \theta) \times \frac{\partial \mathbf{r}}{\partial \theta}(\phi, \theta) = \pm a\, \mathbf{e}_\phi \times a \, \sin\phi \, \hat{\mathbf{e}}_\theta = \pm a^2 \sin\phi \, \hat{\mathbf{e}}_\rho.$$

Choose the $+$ sign to get the *outward* unit normal:

$$\widehat{\mathbf{n}} \triangleq \hat{\mathbf{e}}_\rho.$$

By (7.39), the flux of a vector field $\mathbf{F}$ out of the sphere is

$$\int_{\mathcal{S}} \mathbf{F} \bullet \left(\frac{\partial \mathbf{r}}{\partial \phi}(\phi, \theta) \times \frac{\partial \mathbf{r}}{\partial \theta}(\phi, \theta) \right) dS = \int_0^{2\pi} \int_0^{\pi} \mathbf{F} \bullet \left(a^2 \sin\phi \, \hat{\mathbf{e}}_\rho \right) d\phi \, d\theta.$$

In terms of the components of $\mathbf{F}$, that is,

$$\mathbf{F} = F_\rho \, \hat{\mathbf{e}}_\rho + F_\phi \, \mathbf{e}_\phi + F_\theta \, \hat{\mathbf{e}}_\theta,$$

the total flux out of the sphere is

$$\int_{\mathcal{S}} \mathbf{F} \bullet d\mathbf{S} = a^2 \int_0^{2\pi} \int_0^{\pi} F_\rho(a, \phi, \theta) \, \sin\phi \, d\phi \, d\theta. \ \bigcirc$$

Example 7.34

Suppose an electric charge of 1 C is at the origin of a vacuum. Find the total flux of the electric field out of spheres whose center is at the origin.

Method: Denote by ϵ_0 the electric permittivity of a vacuum. Then the electric field due to an electric charge of 1 coulomb is $\mathbf{E} = \frac{1}{4\pi\epsilon_0}||\mathbf{r}||^{-3}\mathbf{r}$, so on a sphere $\mathcal{S}_a = \{\mathbf{r} : \ ||\mathbf{r}|| = a\}$,

$$\mathbf{E} = \frac{1}{4\pi\epsilon_0} a^{-3}\,\mathbf{r} = \frac{1}{4\pi\epsilon_0} a^{-3}(a\,\hat{\mathbf{e}}_\rho) = \frac{1}{4\pi\epsilon_0} a^{-2}\,\hat{\mathbf{e}}_\rho.$$

The total flux of $\mathbf{E}$ out of $\mathcal{S}_a$ is

$$\int\limits_{\mathcal{S}_a} \mathbf{E} \bullet d\mathbf{S} = \int\limits_0^{2\pi}\int\limits_0^{\pi} \left(\frac{1}{4\pi\epsilon_0} a^{-2}\,\hat{\mathbf{e}}_\rho\right) \bullet \left(a^2 \sin\phi\,\hat{\mathbf{e}}_\rho\,d\phi\,d\theta\right) = \int\limits_0^{2\pi}\int\limits_0^{\pi} \frac{1}{4\pi\epsilon_0} \sin\phi\,d\phi\,d\theta$$

$$= \frac{1}{4\pi\epsilon_0}\left(\int\limits_0^{2\pi} d\theta\right)\left(\int\limits_0^{\pi} \sin\phi\,d\phi\right) = 2\pi\,\frac{1}{4\pi\epsilon_0}\,[-\cos\phi]_0^{\pi} = \frac{1}{\epsilon_0}. \bigcirc$$

By the way, $\frac{1}{4\pi\epsilon_0} \approx 9 \times 10^9$ m/F, where F means farads. The exact value satisfies $\frac{1}{4\pi\epsilon_0} = 10^{-7}\,c^2$, where the velocity of light in a vacuum is $c \approx 2.998 \times 10^8$ m/s.

The total flux out of such spheres does not depend on their radius! In retrospect, this makes sense because while the intensity of the electric field behaves like a^{-2}, the surface area of a sphere behaves like a^2; hence, it is not a big shock that the total flux does not depend on the radius.

Definition 7.13

A surface $\mathcal{S} : \ \mathbf{r} = \mathbf{r}(u,v), \ (u,v)$ in $\mathcal{D}$, is **piecewise smooth** if $\mathcal{D}$ can be written as the union of a finite number of subregions $\mathcal{D}_i$, $\mathbf{r}(u,v)$ is continuously differentiable on each such subregion, and $\mathbf{r}(u,v)$ is continuous on $\mathcal{D}$.

The last part of the definition that $\mathbf{r}(u,v)$ is continuous on $\mathcal{D}$ just says that as we move from (u,v) on one "patch" $\mathcal{D}_i$ to (u,v) on another patch, the corresponding points $\mathbf{r}(u,v)$ vary continuously on the surface $\mathcal{S}$.

Definition 7.14

A piecewise smooth surface $\mathcal{S} : \ \mathbf{r} = \mathbf{r}(u,v), \ (u,v)$ in $\mathcal{D}$, is

(a) **Orientable** if

$$\mathbf{n}(u,v) \triangleq \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \neq \mathbf{0} \ \text{ at all } (u,v) \text{ in } \mathcal{D}$$

(b) **Oriented** if we can choose normal vector $\mathbf{n} = \mathbf{n}(\mathbf{r}(u,v))$ that is a continuous and nonzero function of (u,v) in $\mathcal{D}$, in which case we say that **n orients** $\mathcal{S}$.

Example 7.35

If $\mathcal{S}$ is a surface of the form $z = g(x, y)$ for (x, y) in $\mathcal{D}$ and $g(x, y)$ is continuously differentiable on $\mathcal{D}$, then $\mathcal{S}$ is orientable and

$$\mathbf{n}(x, y) \triangleq -\frac{\partial g}{\partial x}\,\hat{\boldsymbol{\imath}} - \frac{\partial g}{\partial y}\,\hat{\boldsymbol{\jmath}} + \hat{\boldsymbol{k}}$$

orients $\mathcal{S}$ with the **upward orientation**.

Definition 7.15

A surface is **closed** if it is the boundary of a finite solid in $\mathbb{R}^3$.

If a surface integral is taken on a closed surface, then we may put a "circle" in the notation $\oiint_{\mathcal{S}}$. This is similar to putting a circle in the notation for a line integral on a closed curve.

Definition 7.16

A closed oriented surface enclosing a solid $\mathcal{V}$ has the **positive orientation** if the normal vector $\mathbf{n}$ that orients $\mathcal{S}$ points out of $\mathcal{V}$ everywhere on the surface. To be precise, for all P on the surface and ε sufficiently small, $\overrightarrow{OP} + \varepsilon\mathbf{n}$ is outside of $\mathcal{V}$ and $\overrightarrow{OP} - \varepsilon\mathbf{n}$ is inside $\mathcal{V}$.

Example 7.36

Find the total rate at which fluid mass is leaving the cylinder:

$$\{(x, y, z) :\ 0 \le x^2 + y^2 \le a^2 \text{ and } -H \le z \le H\}$$

if the mass density is constant and its velocity field is $\mathbf{v} = x\,\hat{\boldsymbol{\imath}} + 2\,\hat{\boldsymbol{\jmath}} + y\,\hat{\boldsymbol{k}}$.

Method: Define $\mathcal{S}$ to be the closed surface that encloses the cylinder positively oriented by outward unit normal vector $\widehat{\mathbf{n}}$. Let ϱ_0 denote the constant mass density, in units of mass per volume, and let $\mathbf{v}$ be the fluid velocity vector. The total rate at which mass is leaving the cylinder is

$$\mathcal{I} \triangleq \oiint_{\mathcal{S}} \varrho_0\,\mathbf{v} \bullet \widehat{\mathbf{n}}\,dS,$$

that is, the flux of the vector field $(\varrho_0\mathbf{v})$ out of $\mathcal{S}$. The units for each component of the vector $\varrho_0\mathbf{v}$ are (*mass/volume*) × (*length/time*), that is, (*mass per area per unit of time*).

$\mathcal{S}$ consists of three parts, each with a corresponding unit normal vector $\widehat{\mathbf{n}}$ pointing out of the cylinder: (1) the top, $\mathcal{S}_+$, that is, the surface $z = H,\ 0 \le x^2 + y^2 \le a^2$, having $\widehat{\mathbf{n}} = \hat{\boldsymbol{k}}$; (2) the bottom, $\mathcal{S}_-$, that is, the surface $z = -H,\ 0 \le x^2 + y^2 \le a^2$, having $\widehat{\mathbf{n}} = -\hat{\boldsymbol{k}}$; and (3) the side, $\widetilde{\mathcal{S}}$, that is, the surface $-H \le z \le H,\ r = a$, having $\widehat{\mathbf{n}} = \hat{\mathbf{e}}_r$ in cylindrical coordinates. (You will check the conclusions about $\widehat{\mathbf{n}}$ in Problem 7.5.3.15.)

So, we know that the total rate of mass flow out of the cylinder is

$$\mathcal{I} = \mathcal{I}_+ + \mathcal{I}_- + \widetilde{\mathcal{I}} \triangleq \varrho_0 \iint_{\mathcal{S}_+} \mathbf{v} \bullet \hat{\boldsymbol{k}}\,dS + \varrho_0 \iint_{\mathcal{S}_-} \mathbf{v} \bullet (-\hat{\boldsymbol{k}})\,dS + \varrho_0 \iint_{\widetilde{\mathcal{S}}} \mathbf{v} \bullet \hat{\mathbf{e}}_r\,dS.$$

The top and bottom surfaces are flat, so $dS = dA = r\,dr\,d\theta$ there. On the side, an element of surface area is given by the product of an element of circular arclength, $a\,d\theta$, and an element of vertical length, dz, that is, $dS = a\,d\theta\,dz$ there.

We calculate

$$\mathcal{I}_+ = \varrho_0 \int_0^{2\pi}\int_0^a \mathbf{v}(r,\theta,H) \bullet \hat{\boldsymbol{k}} r\,dr\,d\theta = \varrho_0 \int_0^{2\pi}\int_0^a (x\,\hat{\boldsymbol{\imath}} + 2\,\hat{\boldsymbol{\jmath}} + y\,\hat{\boldsymbol{k}}) \bullet \hat{\boldsymbol{k}} r\,dr\,d\theta$$

$$= \varrho_0 \int_0^{2\pi}\int_0^a yr\,dr\,d\theta = \varrho_0 \int_0^{2\pi}\int_0^a r\sin\theta\, r\,dr\,d\theta = \cdots = 0,$$

$$\mathcal{I}_- = \varrho_0 \int_0^{2\pi}\int_0^a \mathbf{v}(r,\theta,-H) \bullet (-\hat{\boldsymbol{k}}) r\,dr\,d\theta = -\varrho_0 \int_0^{2\pi}\int_0^a (x\,\hat{\boldsymbol{\imath}} + 2\,\hat{\boldsymbol{\jmath}} + y\,\hat{\boldsymbol{k}}) \bullet \hat{\boldsymbol{k}} r\,dr\,d\theta$$

$$= -\varrho_0 \int_0^{2\pi}\int_0^a yr\,dr\,d\theta = -\varrho_0 \int_0^{2\pi}\int_0^a r\sin\theta\, r\,dr\,d\theta = \cdots = 0,$$

and

$$\widetilde{\mathcal{I}} = \varrho_0 \int_{-H}^{H}\int_0^{2\pi} \mathbf{v}(a,\theta,z) \bullet \hat{\mathbf{e}}_r a\,d\theta\,dz. = \varrho_0 \int_{-H}^{H}\int_0^{2\pi} (x\,\hat{\boldsymbol{\imath}} + 2\,\hat{\boldsymbol{\jmath}} + y\,\hat{\boldsymbol{k}}) \bullet \hat{\mathbf{e}}_r\, a\,d\theta\,dz$$

$$= a\varrho_0 \int_{-H}^{H}\int_0^{2\pi} (a\cos\theta\,\hat{\boldsymbol{\imath}} + 2\,\hat{\boldsymbol{\jmath}}) \bullet (\cos\theta\,\hat{\boldsymbol{\imath}} + \sin\theta\,\hat{\boldsymbol{\jmath}}) d\theta\,dz = a\varrho_0 \int_{-H}^{H}\int_0^{2\pi} (a\cos^2\theta\, + 2\sin\theta)\,d\theta\,dz$$

$$= a\varrho_0 \left(\int_{-H}^{H} dz\right)\left(\int_0^{2\pi} (a\cos^2\theta + 2\sin\theta)\,d\theta\right) = 2aH\varrho_0 \int_0^{2\pi} \left(\frac{a(1+\cos 2\theta)}{2} + 2\sin\theta\right) d\theta$$

$$= 2aH\varrho_0 \left[\frac{a\theta}{2} + \frac{a\sin 2\theta}{4} - 2\cos\theta\right]_0^{2\pi} = 2aH\varrho_0 \cdot \pi a = 2\pi a^2 H\varrho_0.$$

So, the total rate of mass flow out of the cylinder is

$$\mathcal{I} = \mathcal{I}_+ + \mathcal{I}_- + \widetilde{\mathcal{I}} = 2\pi a^2 H\varrho_0$$

in units of *mass per unit of time*. ◯

7.5.3 Problems

1. The side wall of a twisted cylinder, shown in Figure 7.45, is parametrized by

$$\mathbf{r} = (a\cos\theta - \sin z)\hat{\boldsymbol{\imath}} + a\sin\theta\,\hat{\boldsymbol{\jmath}} + z\,\hat{\boldsymbol{k}},\ 0 \le z \le 2\pi,\ 0 \le \theta \le 2\pi,$$

where a is an unspecified positive constant. Find dS, the element of area.

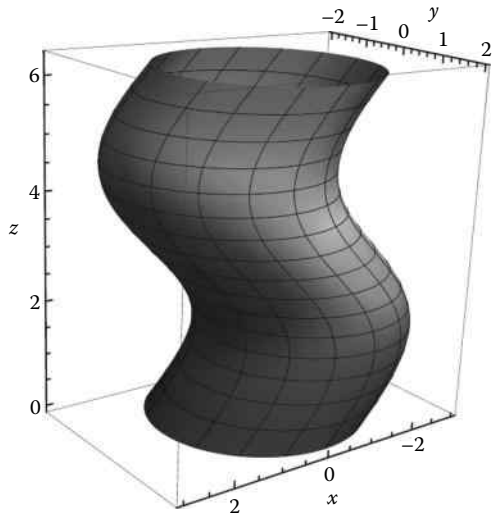


FIGURE 7.45
Twisted cylinder in Problem 7.5.3.1.

2. $\mathcal{S}$ is the top, slanted face of the solid $\mathcal{V}$ bounded by the cylinder $x^2+y^2=9$ and the planes $z=0$ and $2x+z=8$. Find the surface area of $\mathcal{S}$.
3. The solid $\mathcal{V}$ is bounded from below by the cone $\phi=\frac{\pi}{6}$ and from above by the sphere of radius 3 and center at the origin. The boundary of $\mathcal{V}$ is $\mathcal{S}$, which we can think of as being in two pieces. Find the total surface area of $\mathcal{S}$.
4. Find the surface area of the part of the paraboloid $z=\frac{x^2+y^2}{2}$ that lies below the plane $z=5$.
5. Find the centroid of the surface

$$\mathcal{S}:\ \mathbf{r}=u\,\hat{\boldsymbol{\imath}}+(2-u)\,\hat{\boldsymbol{\jmath}}+v\,\hat{\boldsymbol{k}}$$

on the domain $0\le u^2+v^2\le 1$. Sketch the surface and see whether the location of the centroid makes sense "by eyeball."
6. Let Π be the tangent plane to the surface $z^2=x^2+y-4y^2$ at the point $(x,y,z)=(3,-1,-2)$. Let $\mathcal{S}$ be the part of Π that lies in the positive octant. Find the total surface area of $\mathcal{S}$.
7. Find the z-coordinate centroid of the surface

$$\mathcal{S}:\ \mathbf{r}=v\,\hat{\boldsymbol{\imath}}+u\,\hat{\boldsymbol{\jmath}}+(2-u-2v)\hat{\boldsymbol{k}}$$

on the domain $0\le u^2+v^2\le 1$.
8. Find the z-coordinate of the centroid of the surface $\mathcal{S}$ parametrized by $\mathbf{r}=\frac{2}{3}u^{3/2}\,\hat{\boldsymbol{\imath}}+v\,\hat{\boldsymbol{\jmath}}+(u-v)\hat{\boldsymbol{k}}$ on the rectangular domain $0\le u\le 1, 0\le v\le 2$.

9. The surface $\mathcal{S}$ is the $z \geq 0$ portion of the sphere with radius a and center at the origin. Evaluate

$$\iint_{\mathcal{S}} z \, dS$$

and use that to find the z-coordinate of the centroid.

10. Use a computer algebra system to find the centroid of the surface

$$\mathcal{S}: \ \mathbf{r} = u^2 \, \hat{\boldsymbol{\imath}} + (2 - u) \, \hat{\boldsymbol{\jmath}} + v \, \hat{\boldsymbol{k}}$$

on the half disk domain $0 \leq u^2 + v^2 \leq 1, u \geq 0$. Use a computer algebra system to sketch the surface and see whether the location of the centroid makes sense "by eyeball."

11. $\mathcal{S}$ is the top, slanted face of the solid bounded by the cylinder $x^2 + y^2 = 4$ and the planes $z = 0$ and $2x + z = 5$. (a) Find the total surface area of $\mathcal{S}$, and (b) evaluate $\iint_{\mathcal{S}} z \, dS$.

12. Let $\mathbf{F} \triangleq (x + y)\hat{\boldsymbol{\imath}} + (-x + y)\hat{\boldsymbol{\jmath}}$ and $\mathcal{S}$ be the paraboloidal surface $z = 4(x^2 + y^2)$ for (x, y) in the disk $0 \leq x^2 + y^2 \leq a^2$, where a is an unspecified constant. Evaluate

$$\iint_{\mathcal{S}} \mathbf{F} \bullet \widehat{\mathbf{n}} \, dS,$$

assuming $\widehat{\mathbf{n}}$ is a normal vector giving the upward orientation to $\mathcal{S}$. [Hint: Polar coordinates may be useful in the solution.]

13. Evaluate

$$\oiint_{\mathcal{S}} (y \, \hat{\boldsymbol{\imath}} + z^2 \hat{\boldsymbol{k}}) \bullet d\mathbf{S},$$

where $\mathcal{S}$ is the closed surface enclosing the half ball $x^2 + y^2 + z^2 \leq a^2, z \geq 0$. Your final conclusion should be in terms of a.

14. Evaluate

$$\oiint_{\mathcal{S}} \left((y - x^2)\hat{\boldsymbol{\imath}} + 2xy \, \hat{\boldsymbol{\jmath}} + z \, \hat{\boldsymbol{k}} \right) \bullet \widehat{\mathbf{n}} \, dS,$$

where $\mathcal{S}$ is the closed surface enclosing the half ball $x^2 + y^2 + z^2 \leq a^2, z \geq 0$. Your final conclusion should be in terms of a.

15. Check the conclusions about $\widehat{\mathbf{n}}$ in Example 7.36 using $\mathbf{n} = \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v}$.

16. (a) Write down, but do not evaluate, an iterated double integral that gives the surface area of the portion of the surface $z = \sqrt{x^2 + 2y^2}$ that lies above the rectangle $\mathcal{D} = \{(x, y) : \ 0 \leq x \leq 3, \ 0 \leq y \leq 2\}$. Write the double integral with as much specific details as possible.

u/*v*	0	1	2	3	4	5	6	7	8	9	10	11	12
0	j − k	i − k	i + j	k	j	k	j	i + j	k	i + j	i	j	j
1	i	− k	j	i	i	j	i	− k	i + k	j + k	k	i	k
2	k	j	j	k	k	i + k	i	i + j	i + j	i + j	i	i − j	i − j
3	k	i − k	i + j	k	j	k	i + k	− k	i − j	− j	i	i + k	k
4	j	i	k	j	i	j + k	j	k	j	j	k	i − k	j
5	k	i + j	k	i	i	i + j	i + 2k	− k	i + 2j	i	j	k	k
6	− k	i − k	i	i	j	k	i	k	k	i + j	j	i + j	i + 2j

FIGURE 7.46
Values of $\mathbf{r}$ in Problem 7.5.3.17.

u/*v*	0	1	2	3	4	5	6	7	8	9	10	11	12
0	i − k	j − k	i + j	− k	i	j	i	j	i	i + k	k	j	j
1	− k	i	k	j	k	i	k	i	i + j	j	i + j	j	k
2	i − k	j − i	j	− i	j	k	j	i + j	k	i + j	i	i + j	i − j
3	i + j	k	k	i + 2j	i	i + j	i + 2k	− k	i	i	j	k	k
4	i	j	k	j	i	k	j	k	j	j	k	i − k	j
5	− k	i	j	i + k	i	j	i	− k	i	k	i	i	k
6	j	k	j	i + j	k	i + j	i	i + j	i − j	i + j	k	− k	i + 2j

FIGURE 7.47
Values of $\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v}$ in Problem 7.5.3.17.

(b) Divide $\mathcal{D}$ into six subrectangles of equal area. Sample the integrand in part (a) at six points in those six subrectangles and use those data to find an approximation of the surface area in part (a).

17. A surface $\mathcal{S}$ is parametrized by $\mathbf{r} = \mathbf{r}(u, v)$ for (u, v) in the rectangle $\mathcal{D} = \{(u, v) : 0 \leq u \leq 6, 0 \leq v \leq 12\}$. The table in Figure 7.47 gives the values of the vectors $\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v}$, and the table in Figure 7.46 gives the values of the vectors $\mathbf{r}(u, v)$, both at some values of (u, v) in $\mathcal{D}$. Let $\mathbf{F} \triangleq x\,\hat{\boldsymbol{\imath}} + y\,\hat{\boldsymbol{\jmath}} + z\,\hat{\mathbf{k}}$. Approximate $\iint_{\mathcal{S}} \mathbf{F} \bullet d\mathbf{S}$ using six subrectangles of $\mathcal{D}$. State which sampling point method you used, for example, midpoints or, upper right points.

18. Some single variable calculus courses have a topic called "surfaces of revolution": When a curve in the xy-plane is revolved around an axis in the plane, find the total surface area of the surface so generated. For example, if a curve is given in the form $y = f(x)$ for $a \leq x \leq b$ and is revolved around the x-axis, then the surface has area

$$Area = 2\pi \int_a^b |f(x)| \sqrt{1 + (f'(x))^2}\, dx.$$

Derive this result by considering the surface of revolution to be given in the form

$$\mathcal{S}: \ \mathbf{r} = x\,\hat{\boldsymbol{\imath}} + f(x)\cos\varphi\,\hat{\boldsymbol{\jmath}} + f(x)\sin\varphi\,\hat{\mathbf{k}},\ a \leq x \leq b,\ 0 \leq \varphi \leq 2\pi,$$

that is, by using the x-axis as the longitudinal axis and the alternative polar coordinates $y = \mathrm{r}\cos\varphi, z = \mathrm{r}\sin\varphi$ of Example 7.25.

19. (a) Extend the result of Problem 7.5.3.18 to the case of a surface of revolution generated by rotating a plane curve $\mathcal{C}: \ \mathbf{r} = x(t)\hat{\boldsymbol{\imath}} + y(t)\hat{\boldsymbol{\jmath}}$, $\alpha \le t \le \beta$, about the x-axis.
    (b) Apply the result of part (a) to find the surface area of the surface generated by rotating the curve $\mathcal{C}: \ \mathbf{r} = t^2\,\hat{\boldsymbol{\imath}} + \frac{t^3}{3}\,\hat{\boldsymbol{\jmath}}$, $2 \le t \le 3$, about the x-axis.
20. Let $\mathcal{V}$ be the solid whose base is in the plane $z = 0$ and whose top is the surface $z = 4 - x^2 - 3y^2$. Let $\mathcal{S}_+$ be the top surface that bounds $\mathcal{V}$ and let $\widehat{\mathbf{n}}$ be the unit vector normal to $\mathcal{S}_+$ that points out of $\mathcal{V}$. Define $\mathbf{F} = x\,\hat{\boldsymbol{\imath}} + y^2\,\hat{\boldsymbol{\jmath}} + z^3\,\hat{\boldsymbol{k}}$. (a) Find a formula for $\widehat{\mathbf{n}}$ as a function of (x, y, z). (b) Set up, with as much specificity as possible, but do not evaluate $\iint_{\mathcal{S}} \mathbf{F} \bullet \widehat{\mathbf{n}}\, dS$. (c) Set up, with as much specificity as possible, but do not evaluate $\iiint_{\mathcal{V}} div(\mathbf{F}) dV$. "As much specificity as possible" means that, for example, in part (a), writing your final conclusion in the form $\int\_\_ \int\_\_ \_\_\_\_\_ d\_\_ d\_\_$ and filling in the blanks.

7.6 Integral Theorems: Divergence, Stokes, and Applications

Some of the most powerful tools in mathematical analysis and modeling of electrical and magnetic phenomena were synthesized by George Green (Grattan-Guiness, 1995). Many other mathematicians and scientists contributed, too, both before and after him.

Definition 7.17

A solid is **nice** if it is enclosed by a closed, piecewise smooth, orientable surface.

Theorem 7.14

(**Divergence theorem** in $\mathbb{R}^3$) Suppose $\mathbf{F}$ is a continuously differentiable vector field on a solid $\mathcal{V}$ in $\mathbb{R}^3$. Suppose $\mathcal{V}$ is enclosed by a piecewise smooth, oriented parametrized surface $\mathcal{S}$ with outward-pointing unit normal vectors $\widehat{\mathbf{n}} = \widehat{\mathbf{n}}(\mathbf{r})$. Then

$$\oiint_{\mathcal{S}} \mathbf{F} \bullet \widehat{\mathbf{n}}\, dS = \iiint_{\mathcal{V}} \nabla \bullet \mathbf{F}\, dV. \tag{7.40}$$

This result is also known as **Gauss's theorem**.

Caution: Suppose the surface is $z = g(x, y)$. A common error is that when calculating the surface integral in the LHS of (7.40), we can substitute $z = g(x, y)$ into the formula for $\mathbf{F}(x, y, z)$, but we cannot substitute $z = g(x, y)$ into the formula for $\mathbf{F}(x, y, z)$ when calculating the triple integral in the RHS of (7.40).

Example 7.37

Use the divergence theorem to redo Example 7.36 in Section 7.5.

Method: The solid $\mathcal{V}$ is the solid cylinder $\{(x,y,z): \ -H \le z \le H,\ 0 \le x^2+y^2 \le a^2\}$; the closed, positively oriented surface $\mathcal{S}$ encloses $\mathcal{V}$; and $\mathbf{F} = \varrho_0(x\,\hat{\boldsymbol{\imath}} + 2\,\hat{\boldsymbol{\jmath}} + y\,\hat{\boldsymbol{k}})$. The total rate of mass flow out of the solid is

$$\oiint_{\mathcal{S}} \mathbf{F} \bullet \widehat{\mathbf{n}}\, dS = \iiint_{\mathcal{V}} \boldsymbol{\nabla} \bullet \mathbf{F}\, dV = \iiint_{\mathcal{V}} \varrho_0\, dV = \varrho_0 \bullet (\textit{Volume of } \mathcal{V}) = \varrho_0 2\pi a^2 H. \bigcirc$$

If $\boldsymbol{\nabla} \bullet \mathbf{F}$ is simple, perhaps as simple as a constant, then the divergence theorem makes short and easy work of calculating a flux integral. Even if $\boldsymbol{\nabla} \bullet \mathbf{F}$ is not simple, it might be easier to work with than $\mathbf{F}$. And, it might be a lot easier to *not* have to wade through a swamp of parametrizations of pieces of $\mathcal{S}$ and their respective outward unit normal vectors.

Example 7.38

Suppose $\mathbf{F}$ is an inverse square law force field due to a mass or electric charge at the origin. Find the total flux out of $\mathbf{F}$ out of any nice solid $\mathcal{V}$, as long as the origin is not in $\mathcal{V}$ or on its boundary $\mathcal{S}$.

Method: We calculate

$$\begin{aligned}
\boldsymbol{\nabla} \bullet \mathbf{F} &= \boldsymbol{\nabla} \bullet \left[\frac{k}{||\mathbf{r}||^3}\mathbf{r}\right] = \frac{\partial}{\partial x}\left[\frac{kx}{\rho^3}\right] + \frac{\partial}{\partial y}\left[\frac{ky}{\rho^3}\right] + \frac{\partial}{\partial z}\left[\frac{kz}{\rho^3}\right] \\
&= \frac{k}{\rho^3} + kx \cdot \left(-3\rho^{-4}\frac{\partial \rho}{\partial x}\right) + \frac{k}{\rho^3} + kx \cdot \left(-3\rho^{-4}\frac{\partial \rho}{\partial y}\right) + \frac{k}{\rho^3} + kx \cdot \left(-3rho^{-4}\frac{\partial \rho}{\partial z}\right) \\
&= k\rho^{-3} - 3kx\rho^{-4} \cdot \frac{x}{\rho} + k\rho^{-3} - 3ky\rho^{-4} \cdot \frac{y}{\rho} + k\rho^{-3} - 3kz\rho^{-4} \cdot \frac{z}{\rho} \\
&= 3k\rho^{-3} - 3k(x^2+y^2+z^2)\rho^{-5} = 3k\rho^{-3} - 3k\rho^2\rho^{-5} = 0.
\end{aligned}$$

By the divergence theorem,

$$\oiint_{\mathcal{S}} \left[\frac{k}{||\mathbf{r}||^3}\mathbf{r}\right] \bullet \widehat{\mathbf{n}}\, dS = \iiint_{\mathcal{V}} \boldsymbol{\nabla} \bullet \left[\frac{k}{||\mathbf{r}||^3}\mathbf{r}\right] dV = \iiint_{\mathcal{V}} 0\, dV = 0. \bigcirc$$

Corollary 7.5

Suppose $\mathbf{F}$ is an inverse square law force field due to a mass or electric charge at the origin. If $\mathcal{V}$ is any nice solid that does contain the origin strictly inside, then the total flux of $\mathbf{F}$ out of the boundary of $\mathcal{V}$ is $4\pi k$.

Why? Since the origin is strictly inside $\mathcal{V}$, there is a value of a sufficiently small that the ball $\mathcal{B}_a \triangleq \{\mathbf{r}: \ ||\mathbf{r}|| \le a\}$ is inside $\mathcal{V}$. Call $\mathcal{S}_a = \{\mathbf{r}: \ ||\mathbf{r}|| = a\}$, the boundary of the ball $\mathcal{B}_a$ and define the set $\widetilde{\mathcal{V}}$ to be the set of points that are in $\mathcal{V}$ that are *not* in $\mathcal{B}_a$. It follows that

$$\mathcal{V} = \mathcal{B}_a \cup \widetilde{\mathcal{V}}$$

and the boundary of $\widetilde{\mathcal{V}}$ consists of two pieces, $\mathcal{S}_a$ and $\mathcal{S}$, the boundary of the original set $\mathcal{V}$. By the divergence theorem,

$$\oiint_{\mathcal{S}_a} \mathbf{F} \bullet \widehat{\widetilde{\mathbf{n}}}\, dS + \oiint_{\mathcal{S}} \mathbf{F} \bullet \widehat{\widetilde{\mathbf{n}}}\, dS = \iiint_{\widetilde{\mathcal{V}}} \nabla \bullet \mathbf{F}\, dV = \iiint_{\widetilde{\mathcal{V}}} 0\, dV,$$

by Example 7.38. It follows that

$$\oiint_{\mathcal{S}} \mathbf{F} \bullet \widehat{\widetilde{\mathbf{n}}}\, dS = -\oiint_{\mathcal{S}_a} \mathbf{F} \bullet \widehat{\widetilde{\mathbf{n}}}\, dS.$$

As far as the solid $\widetilde{\mathcal{V}}$ is concerned, $\mathcal{S}_a$ is part of the boundary but has outward unit normal vector $\widehat{\widetilde{\mathbf{n}}}$ pointing *into* the ball $\mathcal{B}_a$. So,

$$-\oiint_{\mathcal{S}_a} \mathbf{F} \bullet \widehat{\widetilde{\mathbf{n}}}\, dS = -(\textit{flux of } \mathbf{F} \text{ into } \mathcal{B}_a) = (\textit{flux of } \mathbf{F} \text{ out of } \mathcal{B}_a).$$

Denoting by $\widehat{\mathbf{n}}$ the unit vector pointing *out* of $\mathcal{B}_a$,

$$\textit{flux of } \mathbf{F} \textit{ out of } \mathcal{B}_a = \oiint_{\mathcal{S}_a} \mathbf{F} \bullet \widehat{\mathbf{n}}\, dS = 4\pi k,$$

by the result of Example 7.34 in Section 7.5 (with $\frac{1}{4\pi\epsilon_0}$ replaced by k). In summary, this gives the desired conclusion that

$$\oiint_{\mathcal{S}} \mathbf{F} \bullet \widehat{\widetilde{\mathbf{n}}}\, dS = \textit{flux of } \mathbf{F} \textit{ out of } \mathcal{B}_a = 4\pi k. \ \square$$

Figure 7.48 shows half of a solid $\widetilde{\mathcal{V}}$, that is, half of what is left when a ball is removed from $\mathcal{V}$. In the picture's example, the outer surface $\mathcal{S}$ consists of (x, y, z) satisfying $9 = x^2 + (y^2)^{1.7} + z^2$, and the inner surface $\mathcal{S}_{\sqrt{2}}$ consists of (x, y, z) satisfying $2 = x^2 + y^2 + z^2$.

Figure 7.48 was produced by the *Mathematica* command

```
RegionPlot3D[9 ≥ x^2 + (y^2)^1.7 + z^2 && x^2 + y^2+z^2 ≥ 2, {x, −3, 0}, {y, −3, 3}, {z, −3, 3},
    AxesLabel → {"x","y","z"}]
```

followed by rotating using a mouse.

7.6.1 The Divergence Theorem in $\mathbb{R}^2$

As you will explain in Problem 7.6.4.17, Green's theorem implies that

$$\oint_{\mathcal{C}} \mathbf{F} \bullet \widehat{\mathbf{n}}\, ds = \iint_{\mathcal{D}} \nabla \bullet \mathbf{F}\, dA$$

for a nice enough planar vector field $\mathbf{F}$ and nice enough planar domain $\mathcal{D}$ enclosed by a positively oriented curve $\mathcal{C}$. Note that $\widehat{\mathbf{n}}$ is normal to the curve $\mathcal{C}$ in $\mathbb{R}^2$.

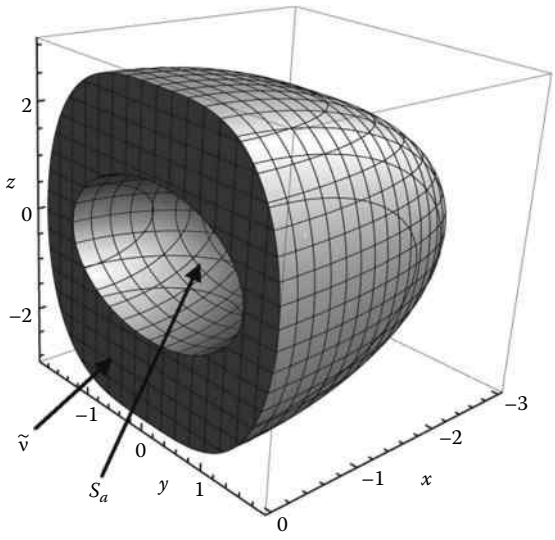


FIGURE 7.48
Half of $\widetilde{\mathcal{V}}$.

7.6.2 Euler's Conservation Equation

Example 7.39

(Euler's conservation equation) Establish the principle of conservation of mass for fluid flow:

$$\frac{\partial \varrho}{\partial t} + \nabla \bullet (\varrho \mathbf{v}) = 0. \tag{7.41}$$

Method: At first, it seems that what we are being asked to do is not connected to what we were studying: What does (7.41), an equation involving derivatives, have to do with integral theorems?

What can motivate our work is the intuition that the divergence theorem relates the derivatives of a function, specifically the divergence of a vector field, to the behavior of that function on a boundary.

Because we were told that this has something to do with flow of mass, it makes sense to recall from Example 7.36 in Section 7.5 that

$$\iint_{\mathcal{S}} \varrho\, \mathbf{v} \bullet \widehat{\mathbf{n}}\, dS$$

is the rate of mass flow across $\mathcal{S}$; where ϱ is the mass density, in units of mass per volume; $\mathbf{v}$ is the velocity vector field of the fluid, and $\mathcal{S}$ is a constant, that is, time-independent, closed, orientable surface enclosing a solid $\mathcal{V}$.

But, $M = \iiint_{\mathcal{V}} \varrho\, dV$ is the total mass of the fluid in the solid $\mathcal{V}$. So, the rate of loss of mass from $\mathcal{V}$ equals the rate of mass flow out of $\mathcal{V}$ across $\mathcal{S}$; hence,

$$\iint_{\mathcal{S}} \varrho\, \mathbf{v} \bullet \widehat{\mathbf{n}}\, dS = -\frac{dM}{dt} = -\frac{d}{dt}\left[\iiint_{\mathcal{V}} \varrho\, dV\right] = -\iiint_{\mathcal{V}} \frac{\partial \varrho}{\partial t}\, dV, \tag{7.42}$$

the latter equality being true because $\mathcal{V}$ is time independent.

On the other hand, using the divergence theorem, (7.42) can be rewritten as

$$\iiint_{\mathcal{V}} \frac{\partial \varrho}{\partial t}\, dV = -\oiint_{\mathcal{S}} \varrho \mathbf{v} \bullet \widehat{\mathbf{n}}\, dS = -\iiint_{\mathcal{V}} \boldsymbol{\nabla} \bullet (\varrho \mathbf{v})\, dV;$$

hence,

$$\iiint_{\mathcal{V}} \left(\frac{\partial \varrho}{\partial t} + \boldsymbol{\nabla} \bullet (\varrho \mathbf{v}) \right) dV = 0. \tag{7.43}$$

Because (7.43) is true for *all* solids $\mathcal{V}$ that are enclosed by a constant, closed, orientable surface $\mathcal{S}$, Euler's conservation equation (7.41) follows.* □

Recall from Section 6.7 that the vorticity of a fluid is $\boldsymbol{\omega} \triangleq \boldsymbol{\nabla} \times \mathbf{v}$, where $\mathbf{v}$ is the fluid velocity.

Example 7.40

Explain why the total flux of vorticity across any piecewise smooth, closed orientable surface is zero, assuming the vorticity is a continuously differentiable function on the surface and the solid it encloses.

Method: The total flux across such a surface $\mathcal{S}$ enclosing a nice solid $\mathcal{V}$ is

$$\oiint_{\mathcal{S}} \boldsymbol{\omega} \bullet \widehat{\mathbf{n}}\, dS = \iiint_{\mathcal{V}} \boldsymbol{\nabla} \bullet \boldsymbol{\omega}\, dV = \iiint_{\mathcal{V}} \boldsymbol{\nabla} \bullet (\boldsymbol{\nabla} \times \mathbf{v})\, dV = \iiint_{\mathcal{V}} 0\, dV = 0,$$

because from Section 6.7, we recall that the divergence of a curl of a vector field is automatically zero, as long as all the derivatives exist. ◯

Similar to Corollary 7.4 in Section 7.3, we have

Corollary 7.6

If $\mathcal{V}$ is a solid enclosed by a closed, piecewise smooth surface $\mathcal{S}$, then the volume of $\mathcal{V}$ is given by

$$Volume = \oiint_{\mathcal{S}} x\, \hat{\boldsymbol{\imath}} \bullet d\mathbf{S}.$$

Why? This follows from the divergence theorem. □

Example 7.41

Use the divergence theorem to express

$$\oiint_{\mathcal{S}} \frac{\partial u}{\partial n}\, dS$$

* Assuming $\left(\frac{\partial \varrho}{\partial t} + \boldsymbol{\nabla} \bullet (\varrho \mathbf{v}) \right)$ is continuous, it cannot be nonzero at a point P because then the integral on a sufficiently small solid ball about P would be nonzero, contradicting (7.43).

in spherical coordinates. Also, examine more specifically such integrals in the particular case of a sphere about the origin.

Method: Recall that $\frac{\partial u}{\partial n} \triangleq (\nabla u) \bullet \widehat{\mathbf{n}}$, so the divergence theorem says in general that

$$\oiint_{\mathcal{S}} \frac{\partial u}{\partial n}\, dS = \oiint_{\mathcal{S}} (\nabla u) \bullet \widehat{\mathbf{n}}\, dS = \iiint_{\mathcal{V}} \nabla \bullet (\nabla u)\, dV,$$

where $\mathcal{V}$ is the solid enclosed by the positively oriented surface $\mathcal{S}$. But,

$$\nabla \bullet (\nabla u) \triangleq \nabla^2 u$$

is the Laplacian of u. Section 6.7 gave the formula for the Laplacian in spherical coordinates, so

$$\oiint_{\mathcal{S}} \frac{\partial u}{\partial n}\, dS = \iiint_{\mathcal{V}} \nabla^2 u\, dV$$

$$= \iiint_{\mathcal{V}} \left(\frac{1}{\rho^2} \frac{\partial}{\partial \rho} \left[\rho^2 \frac{\partial u}{\partial \rho} \right] + \frac{1}{\rho^2 \sin^2 \phi} \frac{\partial}{\partial \phi} \left[\rho \sin \phi\, \frac{\partial u}{\partial \phi} \right] + \frac{1}{\rho^2 \sin^2 \phi} \frac{\partial^2 u}{\partial \theta^2} \right) dV.$$

For the specific case of a sphere about the origin, recall that the outward unit normal vector is $\widehat{\mathbf{n}} = \hat{\mathbf{e}}_\rho$, the element of volume is $\rho^2 \sin\phi\, d\rho\, d\phi\, d\theta$, and the element of surface area is $dS = a^2 \sin\phi\, d\phi\, d\theta$. So, in the case of the sphere $\mathcal{S}_a = \{\mathbf{r} : \ ||\mathbf{r}|| = a\}$ that encloses the ball $\mathcal{B}_a = \{\mathbf{r} : \ ||\mathbf{r}|| \leq a\}$,

$$\int_0^{2\pi} \int_0^{\pi} \frac{\partial u}{\partial \rho}(a, \phi, \theta)\, a^2 \sin\phi\, d\phi\, d\theta = \oiint_{\mathcal{S}_a} \frac{\partial u}{\partial \rho}\, dS = \oiint_{\mathcal{S}_a} (\nabla u) \bullet \hat{\mathbf{e}}_\rho\, dS = \oiint_{\mathcal{S}_a} \frac{\partial u}{\partial n}\, dS$$

$$= \iiint_{\mathcal{B}_a} \left(\frac{\partial}{\partial \rho} \left[\rho^2 \frac{\partial u}{\partial \rho} \right] + \frac{1}{\sin\phi} \frac{\partial}{\partial \phi} \left[\sin\phi\, \frac{\partial u}{\partial \phi} \right] + \frac{1}{\sin^2 \phi} \frac{\partial^2 u}{\partial \theta^2} \right) \sin\phi\, d\rho\, d\phi\, d\theta$$

$$= \int_0^{2\pi} \int_0^{\pi} \int_0^{a} \left(\frac{\partial}{\partial \rho} \left[\rho^2 \frac{\partial u}{\partial \rho} \right] + \frac{1}{\sin\phi} \frac{\partial}{\partial \phi} \left[\sin\phi \frac{\partial u}{\partial \phi} \right] + \frac{1}{\sin^2 \phi} \frac{\partial^2 u}{\partial \theta^2} \right) \sin\phi\, d\rho\, d\phi\, d\theta.$$

Using $\int_0^a \left(\frac{\partial}{\partial \rho} \left[\rho^2 \frac{\partial u}{\partial \rho} \right] \right) d\rho = \left[\rho^2 \frac{\partial u}{\partial \rho} \right]_0^a = a^2 \frac{\partial u}{\partial \rho}(a, \phi, \theta) - 0$, we get

$$\int_0^{2\pi} \int_0^{\pi} \frac{\partial u}{\partial \rho}(a, \phi, \theta)\, a^2 \sin\phi\, d\phi\, d\theta$$

$$= \int_0^{2\pi} \int_0^{\pi} \frac{\partial u}{\partial \rho}(a, \phi, \theta)\, a^2 \sin\phi\, d\phi\, d\theta + \int_0^{2\pi} \int_0^{\pi} \int_0^{a} \left(\frac{\partial}{\partial \phi} \left[\sin\phi \frac{\partial u}{\partial \phi} \right] + \frac{1}{\sin\phi} \frac{\partial^2 u}{\partial \theta^2} \right) d\rho\, d\phi\, d\theta.$$

It follows that

$$\int_0^{2\pi} \int_0^{\pi} \int_0^{a} \left(\frac{\partial}{\partial \phi} \left[\sin\phi \frac{\partial u}{\partial \phi} \right] + \frac{1}{\sin\phi} \frac{\partial^2 u}{\partial \theta^2} \right) d\rho\, d\phi\, d\theta = 0. \tag{7.44}$$

So, in the case of integration of $\frac{\partial u}{\partial n}$ on a sphere $\mathcal{S}_a$, the divergence theorem amounts to saying that

$$\oiint_{\mathcal{S}_a} \frac{\partial u}{\partial n}\, dS = \int_0^{2\pi} \int_0^{\pi} \frac{\partial u}{\partial \rho}(a, \phi, \theta)\, a^2 \sin\phi\, d\phi\, d\theta = \int_0^{2\pi} \int_0^{\pi} \int_0^{a} \frac{\partial}{\partial \rho} \left[\rho^2 \frac{\partial u}{\partial \rho} \right] d\rho\, d\phi\, d\theta. \bigcirc$$

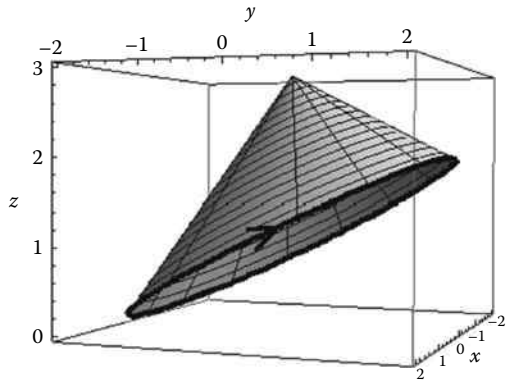


FIGURE 7.49
Positively oriented boundary curve.

We can informally "define" that $\mathcal{C}$ has "positive orientation" with respect to $\mathcal{S}$ if, as we walk facing forward in the direction of travel along $\mathcal{C}$, that is, in the direction of $\widehat{\mathbf{T}}(t)$, with our heads "pointing" in the direction of $\widehat{\mathbf{n}}$, then the surface will always be on our left, that is, our left hand will be pointing toward $\mathcal{S}$.

Here is a precise definition.

Definition 7.18

Suppose $\mathcal{S}$ is a piecewise smooth surface oriented by normal vector $\widehat{\mathbf{n}}$ and there is a simple, closed, piecewise smooth curve $\mathcal{C}: \ \mathbf{r} = \mathbf{r}(t)$ that is a boundary curve for $\mathcal{S}$. We say $\mathcal{C}$ is **positively oriented with respect to** $\mathcal{S}$ if at every point $\mathbf{r}(t)$ the unit tangent vector $\widehat{\mathbf{T}}$ has $\widehat{\mathbf{n}} \times \widehat{\mathbf{T}}$ pointing toward $\mathcal{S}$.

Example 7.42

Similar to work in Examples 6.18 in Section 6.3 and 7.29 in Section 7.5, let $\mathcal{S}$ be the "northern" hemisphere of radius a:

$$\mathcal{S}: \ \mathbf{r} = \mathbf{r}(\phi, \theta) = a(\sin\phi \ \cos\theta \, \hat{\boldsymbol{\imath}} + \sin\phi \ \sin\theta \, \hat{\boldsymbol{\jmath}} + \cos\phi \, \hat{\boldsymbol{k}}), \ 0 \le \phi \le \frac{\pi}{2}, \ 0 \le \theta \le 2\pi,$$

that is,

$$\mathcal{S}: \ \mathbf{r} = a\,\hat{\mathbf{e}}_\rho, \ 0 \le \phi \le \frac{\pi}{2}, \ 0 \le \theta \le 2\pi.$$

Study the orientation of the equator, that is, the circle $\{\mathbf{r}: \ ||\mathbf{r}|| = a\}$ with respect to the surface $\mathcal{S}$.

Method: Let P be any point on the circle, say $P = (a\cos t_0, a\sin t_0, 0)$. As the problem was stated, we were not told which of the two possible unit normal vectors

$$\pm\,\hat{\mathbf{e}}_\rho \triangleq \pm\,(\sin\phi \ \cos\theta \, \hat{\boldsymbol{\imath}} + \sin\phi \ \sin\theta \, \hat{\boldsymbol{\jmath}} + \cos\phi \, \hat{\boldsymbol{k}})$$

to use to orient $\mathcal{S}$.

If we choose the upward orientation of $\mathcal{S}$, that is,

$$\widehat{\mathbf{n}} = (\sin\phi\ \cos\theta\,\hat{\boldsymbol{\imath}} + \sin\phi\ \sin\theta\,\hat{\boldsymbol{\jmath}} + \cos\phi\,\hat{\mathbf{k}}),$$

then the counter clockwise orientation of the circle given by

$$\mathcal{C}:\ \mathbf{r} = a(\cos t\,\hat{\boldsymbol{\imath}} + \sin t\,\hat{\boldsymbol{\jmath}} + 0\,\hat{\mathbf{k}}),\ 0 \le t \le 2\pi,$$

will be positive with respect to the orientation of $\mathcal{S}$. Why? Because at P the unit tangent vector to $\mathcal{C}$ is

$$\widehat{\mathbf{T}} = -\sin t_0\,\hat{\boldsymbol{\imath}} + \cos t_0\,\hat{\boldsymbol{\jmath}} + 0\,\hat{\mathbf{k}}$$

and the unit normal vector to $\mathcal{S}$ is, noting that $\phi = \frac{\pi}{2}$ at P,

$$\widehat{\mathbf{n}} = \cos t_0\,\hat{\boldsymbol{\imath}} + \sin t_0\,\hat{\boldsymbol{\jmath}} + 0\,\hat{\mathbf{k}}.$$

Thus,

$$\widehat{\mathbf{n}} \times \widehat{\mathbf{T}} = (\cos t_0\,\hat{\boldsymbol{\imath}} + \sin t_0\,\hat{\boldsymbol{\jmath}} + 0\,\hat{\mathbf{k}}) \times (-\sin t_0\,\hat{\boldsymbol{\imath}} + \cos t_0\,\hat{\boldsymbol{\jmath}} + 0\,\hat{\mathbf{k}}) = \cdots = \hat{\mathbf{k}}$$

points toward the northern hemisphere rather than away from it.

On the other hand, if we choose the downward orientation of $\mathcal{S}$, that is,

$$\widehat{\mathbf{n}} = -(\sin\phi\ \cos\theta\,\hat{\boldsymbol{\imath}} + \sin\phi\ \sin\theta\,\hat{\boldsymbol{\jmath}} + \cos\phi\,\hat{\mathbf{k}}),$$

then the *clockwise* orientation of the circle given by

$$\mathcal{C}:\ \mathbf{r} = a(\cos\tau\,\hat{\boldsymbol{\imath}} - \sin\tau\,\hat{\boldsymbol{\jmath}} + 0\,\hat{\mathbf{k}}),\ 0 \le \tau \le 2\pi,$$

will be positive with respect to the orientation of $\mathcal{S}$. Why? Because at $P = (a\cos t_0, a\sin t_0, 0)$, that is, at $\tau = -t_0$, we calculate

$$\begin{aligned}\left.\frac{d\mathbf{r}}{d\tau}\right|_{\tau=-t_0} &= a(-\sin\tau\,\hat{\boldsymbol{\imath}} - \cos\tau\,\hat{\boldsymbol{\jmath}})\Big|_{\tau=-t_0} = a(-\sin(-t_0)\,\hat{\boldsymbol{\imath}} - \cos(-t_0)\,\hat{\boldsymbol{\jmath}}) \\ &= a(\sin(-t_0)\,\hat{\boldsymbol{\imath}} - \cos(-t_0)\,\hat{\boldsymbol{\jmath}}).\end{aligned}$$

So, the unit tangent vector to $\mathcal{C}$ is

$$\widehat{\mathbf{T}} = \sin t_0\,\hat{\boldsymbol{\imath}} - \cos t_0\,\hat{\boldsymbol{\jmath}} + 0\,\hat{\mathbf{k}}$$

and the unit normal vector to $\mathcal{S}$ is, noting that $\phi = \frac{\pi}{2}$ at P,

$$\widehat{\mathbf{n}} = -(\cos t_0\,\hat{\boldsymbol{\imath}} + \sin t_0\,\hat{\boldsymbol{\jmath}} + 0\,\hat{\mathbf{k}}),$$

and thus,

$$\widehat{\mathbf{n}} \times \widehat{\mathbf{T}} = (-\cos t_0\,\hat{\boldsymbol{\imath}} - \sin t_0\,\hat{\boldsymbol{\jmath}} + 0\,\hat{\mathbf{k}}) \times (\sin t_0\,\hat{\boldsymbol{\imath}} - \cos t_0\,\hat{\boldsymbol{\jmath}} + 0\,\hat{\mathbf{k}}) = \cdots = \hat{\mathbf{k}}$$

points toward the northern hemisphere rather than away from it. ◯

Theorem 7.15

(Stokes' theorem) Suppose $\mathcal{C}$ is a simple, closed, piecewise smooth boundary curve for $\mathcal{S}$, an oriented, piecewise smooth, parametrized surface with unit normal vectors $\widehat{\mathbf{n}}$. Suppose $\mathcal{C}$ is positively oriented with respect to $\mathcal{S}$. If $\mathbf{F}$ is a continuously differentiable vector field on an open solid containing $\mathcal{S}$, then (7.46) is true, that is,

so $\mathbf{n}$ is given by

$$\frac{\partial \mathbf{r}}{\partial r} \times \frac{\partial \mathbf{r}}{\partial \theta} = \cdots = r(\cos\theta\,\hat{\boldsymbol{\imath}} + \sin\theta\,\hat{\boldsymbol{\jmath}} + 1\,\hat{\boldsymbol{k}}). \tag{7.47}$$

or its negative. Because (7.47) has the upward orientation, we use it as the normal to $\mathcal{S}$.

So, the work is

$$W = \iint_{\mathcal{S}} \hat{\boldsymbol{\imath}} \bullet \hat{\mathbf{n}}\, dS = \iint_{\mathcal{S}} \hat{\boldsymbol{\imath}} \bullet \left(\frac{\partial \mathbf{r}}{\partial r} \times \frac{\partial \mathbf{r}}{\partial \theta}\right) dr\, d\theta = \int_0^{\frac{\pi}{2}} \int_0^2 \hat{\boldsymbol{\imath}} \bullet \Big(r(\cos\theta\,\hat{\boldsymbol{\imath}} + \sin\theta\,\hat{\boldsymbol{\jmath}} + 1\,\hat{\boldsymbol{k}})\Big)\, dr\, d\theta$$

$$= \int_0^{\frac{\pi}{2}} \int_0^2 r\cos\theta\, dr\, d\theta = \left(\int_0^{\frac{\pi}{2}} \cos\theta\, d\theta\right)\left(\int_0^2 r\, dr\right) = \left(\Big[\sin\theta\Big]_0^{\frac{\pi}{2}}\right)\left(\left[\frac{r^2}{2}\right]_0^2\right) = 1 \cdot 2 = 2. \bigcirc$$

Example 7.44

Find the total flux of ***curl***$(\mathbf{F})$ across the northern hemisphere $x^2 + y^2 + z^2 = a^2$, $z \geq 0$, oriented upward, assuming $\mathbf{F} = y^3\hat{\boldsymbol{\imath}} - x^3\hat{\boldsymbol{\jmath}} + x\,\hat{\boldsymbol{k}}$.

Method: The northern hemisphere has the circle $x^2 + y^2 = a^2$, that is, the equator, as boundary curve. As in Example 7.42, if we give the surface the upward orientation, then the boundary curve is positively oriented with respect to the surface if the curve has the counter clockwise orientation. So, let

$$\mathcal{S}:\ \mathbf{r} = a\,\hat{\mathbf{e}}_\rho,\ 0 \leq \phi \leq \frac{\pi}{2},\ 0 \leq \theta \leq 2\pi,$$

and

$$\mathcal{C}:\ \mathbf{R} = a\cos t\,\hat{\boldsymbol{\imath}} + a\sin t\,\hat{\boldsymbol{\jmath}},\ 0 \leq t \leq 2\pi.$$

Using Stokes' theorem, we get that the total flux is given by

$$\iint_{\mathcal{S}} \textbf{\textit{curl}}(\mathbf{F}) \bullet \hat{\mathbf{n}}\, dS = \oint_{\mathcal{C}} \mathbf{F} \bullet d\mathbf{R} = \int_0^{2\pi} \left(y^3\,\hat{\boldsymbol{\imath}} - x^3\,\hat{\boldsymbol{\jmath}} + x\,\hat{\boldsymbol{k}}\right) \bullet d\mathbf{R}$$

$$= \int_0^{2\pi} \left((a\sin t)^3\,\hat{\boldsymbol{\imath}} - (a\cos t)^3\,\hat{\boldsymbol{\jmath}} + a\cos t\,\hat{\boldsymbol{k}}\right) \bullet \left(-a\sin t\,\hat{\boldsymbol{\imath}} + a\cos t\,\hat{\boldsymbol{\jmath}}\right) dt$$

$$= \int_0^{2\pi} \left(-a^4\sin^4 t - a^4\cos^4 t\right) dt = -a^4 \int_0^{2\pi} \left(\sin^4 t + \cos^4 t\right) dt. \tag{7.48}$$

Some algebra and trigonometry give

$$\sin^4 t + \cos^4 t = \sin^4 t + (\cos^2 t)^2 = \sin^4 t + (1 - \sin^2 t)^2 = \sin^4 t + 1 - 2\sin^2 t + \sin^4 t$$

$$= 2\sin^4 t + (1 - 2\sin^2 t) = 2\left(\frac{1 - \cos 2t}{2}\right)^2 + \cos 2t = \frac{1}{2}(1 - \cos 2t)^2 + \cos 2t$$

$$= \left(\frac{1}{2} - \cos 2t + \frac{1}{2}\cos^2 2t\right) + \cos 2t$$

$$= \left(\frac{1}{2} - \cancel{\cos 2t} + \frac{1}{4} \cdot (1 + \cos 4t)\right) + \cancel{\cos 2t}$$

$$= \frac{3}{4} + \frac{1}{4} \cdot \cos 4t,$$

so the total flux is

$$-a^4 \int_0^{2\pi} \left(\frac{3}{4} + \frac{1}{4} \cdot \cos 4t\right) dt = -a^4 \left[\frac{3t}{4} + \frac{1}{16} \cdot \sin 4t\right]_0^{2\pi} = -\frac{3\pi a^4}{2}. \bigcirc$$

Learn More About It

Problems 7.6.4.15 and 7.6.4.16 are from *Physics, Part I*, by Robert Resnick and David Halliday, John Wiley & Sons, © 1966, specifically Problem 16.11, page 418.

7.6.4 Problems

1. Let $f(x, y, z) = 4x^2 - 3y^2 - 2z^2$. Find the total flux of ∇f out of the sphere $x^2 + y^2 + z^2 = a^2$.
2. Let $\mathcal{V}$ be the solid that lies above the surface $z = \sqrt{x^2 + y^2}$ and below the plane $z = 3$. Let $\mathcal{S}$ be the closed surface that encloses $\mathcal{V}$. Evaluate $\iint_{\mathcal{S}} \mathbf{F} \bullet d\mathbf{S}$, where $\mathbf{F}(\mathbf{r}) = \mathbf{r}$.
3. Evaluate $\iint_{\mathcal{S}} (xy\, dy\, dz - y^2\, dz\, dx + xy\, dx\, dy)$ where $\mathcal{S}$ is the surface that encloses the half cylinder $x^2 + y^2 \leq 9, y \geq 0, 0 \leq z \leq 2$. [Assume that $P\, dy\, dz + Q\, dz\, dx + R\, dx\, dy$ is another notation for $\mathbf{F} \bullet d\mathbf{S}$.]
4. Let $\mathcal{V}$ be the solid between the coaxial cylinders $x^2 + y^2 = 4$ and $x^2 + y^2 = 9$ for $0 \leq z \leq 5$. Find the total flux out of $\mathcal{V}$ for the vector field $\mathbf{F} = x^3\, \hat{\boldsymbol{\imath}} + z^2\, \hat{\boldsymbol{k}}$.
5. Suppose ϕ satisfies Poisson's equation $\nabla^2 \phi = \psi(x, y, z)$ for some unspecified function ψ. Why is

$$\iiint_{\mathcal{V}} \psi\, dV = \iint_{\mathcal{S}} \frac{\partial \phi}{\partial n}\, dS$$

   an identity for all nice solids $\mathcal{V}$ enclosed by a surface $\mathcal{S}$?
6. Suppose ϕ satisfies Poisson's equation $\nabla^2 \phi = \psi(x, y, z)$ for some unspecified function ψ. Is

$$\iiint_{\mathcal{V}} \nabla\phi \bullet \nabla\phi\, dV = \iint_{\mathcal{S}} \phi \cdot \frac{\partial \phi}{\partial n}\, dS - \iiint_{\mathcal{V}} \phi\, \psi\, dV$$

   an identity for all nice solids $\mathcal{V}$ enclosed by a surface $\mathcal{S}$? If so, why? If not, why not?
7. If $\mathbf{F} = \frac{k}{||\mathbf{r}||^3}\, \mathbf{r}$, that is, is an inverse square law force field with constant k, then we know $\nabla \bullet \mathbf{F} = 0$ for $\mathbf{r} \neq \mathbf{0}$. Why can't we use the divergence theorem to conclude that the total flux out of the sphere $\mathcal{S}_a = \{\mathbf{r} : ||\mathbf{r}|| = a\}$ is zero?
8. Define a solid ellipsoid by

$$\mathcal{V} = \{(x, y, z) : 0 \leq \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} \leq 1\}.$$

Assume $\mathcal{V}$ is made of a homogeneous material. Find the moments of inertia $I_{x=0}, I_{y=0}, I_{z=0}$ defined by, for example,

$$I_{x=0} \triangleq \iiint_{\mathcal{V}} x^2 \varrho dV.$$

[Hint: Find a simple function $\mathbf{F}$ for which $\nabla \bullet \mathbf{F} = x^2$ and use the divergence theorem for the parametric surface:

$$\mathcal{S}: \ \mathbf{r} = a \sin\phi \ \cos\theta \, \hat{\boldsymbol{\imath}} + b \sin\phi \ \sin\theta \, \hat{\boldsymbol{\jmath}} + c \cos\phi \, \hat{\mathbf{k}},$$
$$0 \le \phi \le \pi, \ 0 \le \theta \le 2\pi.]$$

9. Use Stokes' theorem to evaluate $\iint_{\mathcal{S}} \boldsymbol{curl}(xy\,\hat{\boldsymbol{\imath}} - xz\,\hat{\boldsymbol{\jmath}}) \bullet d\mathbf{S}$, where the oriented upward surface $\mathcal{S}$ is the part of the paraboloid $z = 5 - 3(x^2 + y^2)$ that lies above the plane $z = 2$.
10. Evaluate both sides of Stokes' theorem for the case of a vector field $\mathbf{F} = 3y\,\hat{\boldsymbol{\imath}} - 2x\,\hat{\boldsymbol{\jmath}} + xy\,\hat{\mathbf{k}}$, surface $\mathcal{S} = \{(x, y, z): \ x^2 + y^2 + z^2 = 4, \ z \ge 0\}$, and $\mathcal{C}$ being its boundary curve.
11. Use Stokes' theorem to find the circulation of

$$\mathbf{F} \triangleq -\frac{y^3}{3}\,\hat{\boldsymbol{\imath}} + \frac{x^3}{3}\,\hat{\boldsymbol{\jmath}} + z\,\hat{\mathbf{k}}$$

around the circle $x^2 + y^2 = 1$ in the $z = 0$ plane.
12. Redo Example 7.44 by continuing from (7.48) by noting that the curve $\mathcal{C}$ is *also* the boundary of the flat surface

$$\widetilde{\mathcal{S}}: \ \tilde{\mathbf{r}} = x\,\hat{\boldsymbol{\imath}} + y\,\hat{\boldsymbol{\jmath}} + 0\,\hat{\mathbf{k}} = r(\cos\theta\,\hat{\boldsymbol{\imath}} + \sin\theta\,\hat{\boldsymbol{\jmath}}), \ 0 \le r \le a, \ 0 \le \theta \le 2\pi$$

and evaluating

$$\textit{Total flux} = \iint_{\mathcal{S}} \boldsymbol{curl}(\mathbf{F}) \bullet \hat{\mathbf{n}}\, dS = \oint_{\mathcal{C}} \mathbf{F} \bullet d\mathbf{R} = \iint_{\widetilde{\mathcal{S}}} \boldsymbol{curl}(\mathbf{F}) \bullet \hat{\mathbf{n}}\, dS = \cdots.$$

13. If $\mathbf{v}$ is an irrotational fluid velocity vector, that is, $\nabla \times \mathbf{v} = \mathbf{0}$ everywhere, then explain why the lift around a simple, closed, piecewise smooth curve must be zero.
14. Calculate both sides of the conclusion of Stokes' theorem where $\mathcal{S}$ is the surface of the circular paraboloid $x^2 + y^2 = 7 - z$ that lies above the disk $x^2 + y^2 = 4, z = 0$ and $\mathbf{F} = xz\,\hat{\boldsymbol{\imath}} + y^2\,\hat{\boldsymbol{\jmath}} + x\,\hat{\mathbf{k}}$.
15. Suppose there is a homogeneous spherical shell of matter of total mass M centered at the origin, as depicted in Figure 7.51a. Explain why the gravitational force on an object of mass m is (a) zero at point P_0 inside the shell and (b) $-\frac{GMm}{||\mathbf{r}_1||^3}\mathbf{r}_1$ at point P_1 outside the shell. Let $\mathbf{r}_i$ be the position vector of point P_i, $i = 1, 2$.
16. Suppose there are concentric homogeneous spherical shells of matter of total masses M_1 and M_2 centered at the origin, as depicted in Figure 7.51b. Use the

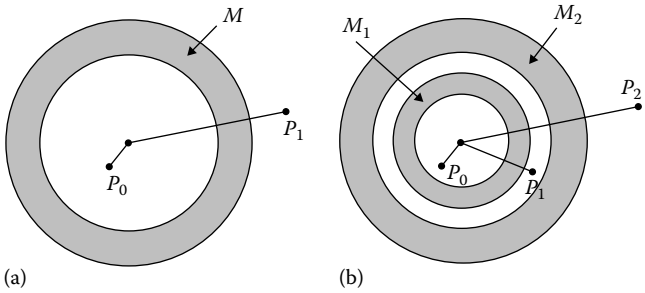


FIGURE 7.51
Problems (a) 7.6.4.15 and (b) 7.6.4.16.

results of Problem 7.6.4.15 to find the gravitational force at points (a) P_0 inside both shells, (b) point P_1 outside the inner shell but inside the outer shell, and (c) point P_2 outside both shells. Let $\mathbf{r}_i$ be the position vector of point P_i, $i = 0, 1, 2$.

17. Explain why Green's theorem implies that

$$(\star) \quad \oint_{\mathcal{C}} \mathbf{F} \bullet \widehat{\mathbf{n}}\, ds = \iint_{\mathcal{D}} \nabla \bullet \mathbf{F}\, dA$$

for a nice enough planar vector field $\mathbf{F}$ and nice enough planar domain $\mathcal{D}$ enclosed by a positively oriented curve $\mathcal{C}$, where $\widehat{\mathbf{n}}$ is normal to the curve $\mathcal{C}$ in $\mathbb{R}^2$.

Note that if $\mathbf{F} = P\,\hat{\boldsymbol{\imath}} + Q\,\hat{\boldsymbol{\jmath}}$, then $(\star)$ says that

$$\oint_{\mathcal{C}} (P\,\hat{\boldsymbol{\imath}} + Q\,\hat{\boldsymbol{\jmath}}) \bullet \widehat{\mathbf{n}}\, ds = \iint_{\mathcal{D}} \left(\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} \right) dA.$$

[Hint: Traveling along a curve $\mathcal{C} : \mathbf{r} = \mathbf{r}(t)$, $a \le t \le b$, the nonzero tangent vectors are

$$\dot{\mathbf{r}}(t) = \dot{x}(t)\,\hat{\boldsymbol{\imath}} + \dot{y}(t)\,\hat{\boldsymbol{\jmath}}.$$

Explain why $\mathbf{n} = -\dot{y}(t)\,\hat{\boldsymbol{\imath}} + \dot{x}(t)\,\hat{\boldsymbol{\jmath}}$ is normal to $\mathcal{C}$. After that, rewrite $(P\,\hat{\boldsymbol{\imath}} + Q\,\hat{\boldsymbol{\jmath}}) \bullet (-\dot{y}(t)\,\hat{\boldsymbol{\imath}} + \dot{x}(t)\,\hat{\boldsymbol{\jmath}})dt = (-Q\,\hat{\boldsymbol{\imath}} + P\,\hat{\boldsymbol{\jmath}}) \bullet (dx\,\hat{\boldsymbol{\imath}} + dy\,\hat{\boldsymbol{\jmath}}) = -Q\,dx + P\,dy$.]

18. Suppose a surface $\mathcal{S}$ and its boundary curve $\mathcal{C}$ satisfy the hypotheses of Stokes' theorem. Explain why

$$\oint_{\mathcal{C}} (f\,\nabla g) \bullet d\mathbf{r} = \iint_{\mathcal{S}} (\nabla f \times \nabla g) \bullet d\mathbf{S}$$

is an identity.

7.7 Probability Distributions

Here we will develop a foundation for the study of randomness. The most basic concept is

Definition 7.19

A **random variable** $X = X(\omega)$ is a real valued function of an independent variable ω in some set Ω.* A random variable is **discrete** if the values $X(\omega)$ can take on are a finite or **countable**† set. A random variable is **continuous** if it is not discrete.

Example 7.45

A die's six faces show the numbers $1, 2, \ldots, 6$. When we roll the die, we get a discrete random variable $X(\omega)$. The **probability** that the die shows the number i is denoted $P(\{\omega : \ X(\omega) = i\})$ or $P(X(\omega) = i)$ or $P(X = i)$ for short. If the die is fair, then $\frac{1}{6} = P(X = i)$, for $i = 1, 2, \ldots, 6$. ◯

There are deep philosophical issues in understanding what the statement $\frac{1}{6} = P(X = i)$ means. We won't go into that but will instead use the "frequency interpretation," namely, that if we roll the same fair die over and over, without end, then the ratio of the number of times the die shows i to the number of times the die is thrown will get closer and closer to $\frac{1}{6}$, as the number of throws increases without end.

The practical problem with this interpretation is that we usually have better things to do than to play endlessly with a die.

All probabilities are between 0 and 1, inclusive. It is not possible for something to happen more than 100% of the time, even though that's a phrase that is used in some commentaries about professional athletics.

Example 7.46

Suppose we fill a car's gasoline tank full and we let $X(\omega)$ denote the time $t > 0$ it takes for the tank to be 90% emptied during normal use, assuming we don't refill it until that time. Then $X(\omega)$ is a continuous random variable. While we might have a good idea about the output of the $X(\omega)$ function, there are many random influences on $X(\omega)$ in our human activities, traffic conditions, and weather conditions. We might say that $P(X(\omega) \leq 1\, month)$ is very close to 1, that is, we are confident that the tank will be 90% empty in a month or less of normal driving. ◯

---

* It is usually not enlightening to know much about the set Ω except the technicalities that there should exist a way to measure the size of some of its subsets and a well-defined concept of which sets are measurable. Those concepts are beyond the level of the course; the book by Naylor and Sell, mentioned at the end of Chapter 2, is a useful reference for the study of "measure theory."

† A set A is countable if it is like the positive integers in that we can enumerate the elements of the set as an infinite sequence $a_1, a_2, \ldots$. For example, the set of all rational numbers, $\mathbb{Q}$, is countable, but the set of all real numbers, $\mathbb{R}$, is not countable.

Example 7.47

For the fair die of Example 7.45, the CDF is

$$F(x) = \left\{ \begin{array}{ll} 0, & \text{if } x < 1 \\ \frac{1}{6}, & \text{if } 1 \le x < 2 \\ \frac{2}{6}, & \text{if } 2 \le x < 3 \\ \vdots & \\ \frac{5}{6}, & \text{if } 5 \le x < 6 \\ 1, & \text{if } x \ge 6 \end{array} \right\}. \bigcirc$$

The CDF in Example 7.47, shown in Figure 7.52, satisfies $F'(x) = 0$ everywhere except $x = 1, 2, \ldots, 6$ and that $F(x)$ has a jump of height $\frac{1}{6}$ at $x = i$ because

$$\lim_{x \to i-} F(x) = \frac{i-1}{6} \text{ and } \lim_{x \to i+} F(x) = \frac{i}{6}, \text{ for } i = 1, \ldots, 6.$$

Example 7.48

Figure 7.53 shows a hypothetical CDF of the random variable of Example 7.46, that is, the time it takes for my gas tank to be 90% emptied: This graph says that it is unusual,

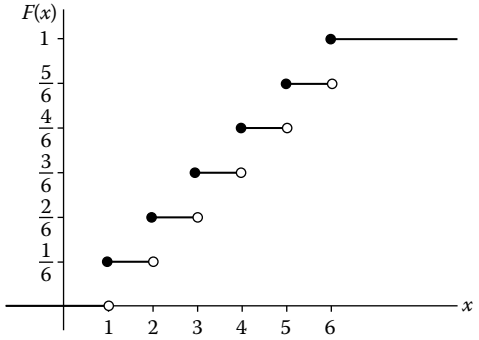


FIGURE 7.52
CDF for a fair die in Example 7.47.

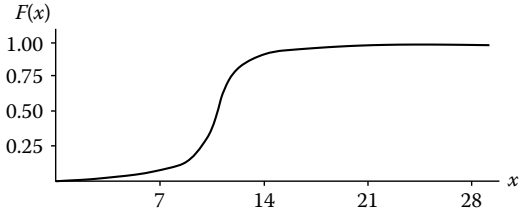


FIGURE 7.53
CDF in Example 7.48.

(b) For $k=1$, using $\alpha=\lambda$, we use integration by parts to calculate

$$\int_0^\infty x f(x)\,dx = \int_0^\infty x\,\alpha e^{-\lambda x}\,dx \triangleq \lim_{t\to\infty}\int_0^t \lambda x\, e^{-\lambda x}\,dx = \lambda \lim_{t\to\infty}\Big[-\frac{x}{\lambda}e^{-\lambda x} - \frac{1}{\lambda^2}e^{-\lambda x}\Big]_0^t$$

$$= \lambda \lim_{t\to\infty}\left(-\frac{t}{\lambda}e^{-\lambda t} - \frac{1}{\lambda^2}e^{-\lambda t} + \frac{1}{\lambda^2}\right) = \lambda\left(0+0+\frac{1}{\lambda^2}\right) = \frac{1}{\lambda}.$$

To get this result, we used L'Hôpital's rule and the assumption that λ is a positive constant:

$$\lim_{t\to\infty} te^{-\lambda t} = \lim_{t\to\infty}\frac{t}{e^{\lambda t}} = \lim_{t\to\infty}\frac{(t)'}{(e^{\lambda t})'} = \lim_{t\to\infty}\frac{1}{\lambda e^{\lambda t}} = 0.$$

While $\int_{-\infty}^{\infty} x^0 f(x)\,dx = 1$, there is no such restriction for $\int_{-\infty}^{\infty} x^k f(x)\,dx$ for $k=1,2,\ldots$.

For $k \geq 2$, we can use a reduction of order formula to find the indefinite integral

$$\int x^k e^{-\lambda x}dx = -\sum_{j=0}^{k}\frac{k!}{j!}\,x^j \lambda^{-(k-j+1)}\,e^{-\lambda x}$$

to conclude, again using L'Hôpital's rule, that

$$E[X^k] = \frac{k!}{\lambda^k}.$$

So, in this example, all of the higher moments exist. ○

In fact, for some density functions, only a few low-order moments exist.

In Calculus I, if $s(t)$ is the position of a particle on a line, then $\dot{s}(t)=v(t)$ is called the instantaneous velocity of the particle. Analogously, $f(x)=F'(x)$ can be called the instantaneous probability $P(X=x)$, that is, the probability of $X(\omega)=x$ happening.

Remarks

If X is a discrete random variable taking on only the values $\{x_1,\ldots,x_N\}$, then

(a) $\sum_{n=1}^{N} P(X=x_n)=1$ and (b) $E[X^k] \triangleq \sum_{n=1}^{N} x_n^k P(X=x_n)$.

Example 7.50

Find the first two moments of the fair die of Example 7.45.

Method: In this example, $N=6$ and $x_n=n$, for $n=1,\ldots,6$. (a) The first moment is

$$E[X] = \sum_{n=1}^{6} nP(X=x_n) = \sum_{n=1}^{6} n\cdot\frac{1}{6} = \frac{7}{2},$$

and (b) the second moment is

$$E[X^2] = \sum_{n=1}^{6} n^2 P(X=x_n) = \sum_{n=1}^{6} n^2\cdot\frac{1}{6} = \frac{91}{6}.\ \bigcirc$$

The first moment, $E[X]$, is called the **mean** of the random variable and is denoted by μ. Example 7.50(a) shows that the mean of a random variable might *not* equal *any* of its output values $X(\omega)$.

The **variance** of a random variable is defined by

$$\sigma^2 \triangleq E[(X-\mu)^2].$$

Because μ is a real number, it can be considered a constant random variable $Y(\omega) \triangleq \mu$. It's easy to see that $E[Y] = \mu$, so we can define $E[\mu] \triangleq \mu$. Using that, you will explain why

$$\sigma^2 = E[X^2] - (E[X])^2 \tag{7.49}$$

in Problem 7.7.2.3. The **standard deviation** is σ, that is, the square root of the variance.

By the way (7.49) was used in Section 2.5 when we discussed linear regression.

Example 7.51

Estimate the mean of the random variable whose hypothetical CDF we graphed in Example 7.48.

Method: We gave the graph of F, the CDF, in Figure 7.53. Based on the graph of F, we can estimate the density function $f = F'$ by using difference quotients, that is,

$$f(x_0) = F'(x_0) \approx \frac{F(x_1) - F(x_0)}{x_1 - x_0}.$$

We got the graph of f in Figure 7.54. If the region under the curve $y = f(x)$ were a lamina, we would guess that its moment $M_{x=0}$, that is, the mean μ, is somewhere between 14 and 21. This says that typically my gasoline tank is 90% empty after about 14–21 days. ◯

The **mode** of a continuous random variable is the value of x at which the density function is maximized. So, the mode is the most frequent outcome of the random variable. While the mode is, in general, different from the mean, the mode may be used as a very rough estimate for the median.

Example 7.52

Perhaps the most famous PDF is that for a **normal random variable**:

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\ \exp\left(-\frac{(x-\alpha)^2}{2\sigma^2}\right). \tag{7.50}$$

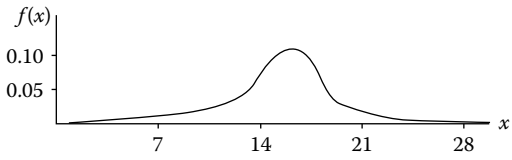


FIGURE 7.54
PDF in Example 7.51.

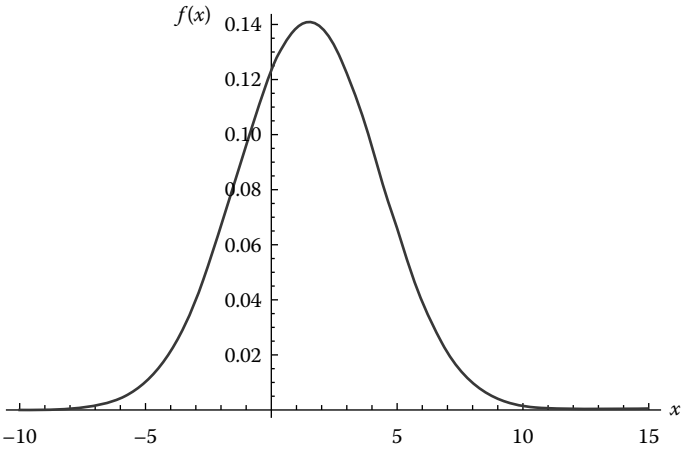


FIGURE 7.55
PDF in Example 7.52.

Find the mean of such a random variable.

Another name for a normal random variable is "Gaussian random variable," and its distribution is known as the **Gaussian distribution**.

Method: An example of the graph of such a density function, for $\alpha = 1.5$ and $\sigma = \sqrt{8}$, is shown in Figure 7.55, so we see why this is often called a "bell curve." You will show in Problem 7.7.2.8 that this density function satisfies $\int_{-\infty}^{\infty} f(x)\,dx = 1$. In this problem, we need to calculate

$$\mu = \lim_{b\to\infty} \int_{\alpha}^{\alpha+b} xf(x)\,dx + \lim_{a\to\infty} \int_{\alpha-a}^{\alpha} xf(x)\,dx \triangleq \lim_{b\to\infty} \mathcal{I}_1 + \lim_{a\to\infty} \mathcal{I}_2,$$

assuming both limits can be shown to exist. First, substituting $z = x - \alpha$ gives

$$\mathcal{I}_1 = \int_{\alpha}^{\alpha+b} xf(x)dx = \int_{\alpha}^{\alpha+b} x\frac{1}{\sqrt{2\pi}\,\sigma}\,\exp\left(-\frac{(x-\alpha)^2}{2\sigma^2}\right)dx = \int_0^b (z+\alpha)\frac{1}{\sqrt{2\pi}\,\sigma}\,\exp\left(-\frac{z^2}{2\sigma^2}\right)dz,$$

so

$$\mathcal{I}_1 = \left(\alpha\int_0^b \frac{1}{\sqrt{2\pi}\,\sigma}\,\exp\left(-\frac{z^2}{2\sigma^2}\right)dz + \int_0^b z\frac{1}{\sqrt{2\pi}\,\sigma}\,\exp\left(-\frac{z^2}{2\sigma^2}\right)dz\right).$$

Using the substitution $w = -\frac{z^2}{2\sigma^2}$ and $dw = -\frac{1}{\sigma^2}\,z\,dz$, we get

$$\int z\frac{1}{\sqrt{2\pi}\,\sigma}\,\exp\left(-\frac{z^2}{2\sigma^2}\right)dz = \int \not{z}\frac{1}{\sqrt{2\pi}\,\sigma}\,e^w\,\frac{-\sigma^2 dw}{\not{z}} = -\frac{\sigma}{\sqrt{2\pi}}\,e^w.$$

So,

$$\lim_{b\to\infty} \mathcal{I}_1 = \lim_{b\to\infty}\left(\alpha\int_0^b \frac{1}{\sqrt{2\pi}\,\sigma}\,\exp\left(-\frac{z^2}{2\sigma^2}\right)dz - \frac{\sigma}{\sqrt{2\pi}}\exp\left(-\frac{b^2}{2\sigma^2}\right) + \frac{\sigma}{\sqrt{2\pi}}\right)$$

$$= \alpha\int_0^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma}\,\exp\left(-\frac{z^2}{2\sigma^2}\right)dz + \frac{\sigma}{\sqrt{2\pi}}.$$

Similarly,

$$\mathcal{I}_2 = \int_{-a}^{0} (z+\alpha)\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{z^2}{2\sigma^2}\right)dz = \cdots = \alpha\int_{-\infty}^{0}\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{z^2}{2\sigma^2}\right)dz - \frac{\sigma}{\sqrt{2\pi}}.$$

So,

$$\begin{aligned}\mu &= \lim_{b\to\infty}\mathcal{I}_1 + \lim_{a\to\infty}\mathcal{I}_2\\ &= \alpha\int_{0}^{\infty}\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{z^2}{2\sigma^2}\right)dz + \cancel{\frac{\sigma}{\sqrt{2\pi}}} + \alpha\int_{-\infty}^{0}\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{z^2}{2\sigma^2}\right)dz - \cancel{\frac{\sigma}{\sqrt{2\pi}}}\\ &= \alpha\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{z^2}{2\sigma^2}\right)dz = \alpha\cdot 1 = \alpha.\end{aligned}$$

So, $\mu = E[X] = \alpha$ is the mean of this random variable. ○

Using similar calculations, we can show that for this random variable,

$$E[(X-\mu)^2] = \sigma^2,$$

that is, that variance of this random variable is σ^2.

Theorem 7.17

If α is a constant and $E[X]$ exists, then (a) $E[\alpha X]$ exists and equals $\alpha E[X]$, and (b) $E[\,X+\alpha\,]$ exists and equals $E[X]+\alpha$.

Definition 7.22

The **(Gaussian) error function** is defined by

$$\operatorname{erf}(x) \triangleq \frac{2}{\sqrt{\pi}}\int_{0}^{x} e^{-u^2}\,du.$$

Theorem 7.18

The error function is related to $F(x)$, the CDF of a normal random variable X with mean $=\mu=0$ and variance $=\sigma^2=\frac{1}{2}$, in the sense that

$$\operatorname{erf}(x) = P(-x \le X \le x) = F(x) - F(-x) = \int_{-x}^{x}\frac{1}{\sqrt{\pi}}e^{-u^2}\,du.$$

7.7.1 Joint Distribution

Suppose X and Y are random variables, and suppose $P(X(\omega) \leq x \text{ and } Y(\omega) \leq y)$, the probability of the simultaneous event that both $X \leq x$ and $Y \leq y$, satisfies

$$P(X(\omega) \leq x \text{ and } Y(\omega) \leq y) = \iint\limits_{\{(u,v):\, -\infty<u\leq x, -\infty<v\leq y\}} h(u,v)dudv,$$

for a function $h(x,y)$. Then we call $h(x,y)$ their **joint probability density (function)**.

Theorem 7.19

The joint density function $h(x,y)$ for random variables satisfies

$$P(a \leq X(\omega) \leq b \text{ and } c \leq Y(\omega) \leq d) = \int_c^d \int_a^b h(x,y)dxdy.$$

Definition 7.23

Two random variables X and Y are **independent** if for all intervals I and J,

$$P(\{\omega :\ X(\omega) \text{ is in } I \text{ and } Y(\omega) \text{ is in } J\}) = P(\{\omega :\ X(\omega) \text{ is in } I\}) \cdot P(\{\omega :\ Y(\omega) \text{ is in } J\}).$$

Theorem 7.20

If X and Y are independent random variables with PDF's $f(x)$ and $g(x)$, respectively, then their joint distribution function is

$$h(x,y) = f(x)g(y).$$

Theorem 7.21

If X and Y are independent random variables with PDF's $f(x)$ and $g(x)$, respectively, then the random variable $Z \triangleq X + Y$ has PDF

$$k(x) = \int_{-\infty}^{\infty} f(u)g(x-u)du. \tag{7.51}$$

[This is also called a **convolution on the real line**; it is more general than the type of convolution in Definition 4.4 in Section 4.5 we used extensively in work with Laplace transforms.]

Why? If the joint PDF of X and Y is $h(x, y)$, then we can extend its definition and the result of Theorem 7.19 to see that

$$P(Z \leq z) = P(X + Y \leq z) = \iint\limits_{\{(x,y):\, x+y \leq z\}} h(x, y)dxdy.$$

Because $x + y \leq z$ is equivalent to $y \leq z - x$, we can rewrite this as an iterated integral:

$$P(Z \leq z) = \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{z-x} h(x, y)dydx = \int\limits_{-\infty}^{\infty} \left(\int\limits_{-\infty}^{z-x} f(x)g(y)dy \right) dx.$$

Make the substitution $u = y + x$, that is, $x = u - y$, so $dy = du$. Note that $u = z$ when $y = z - x$. This gives

$$P(Z \leq z) = \int\limits_{-\infty}^{\infty} \left(\int\limits_{-\infty}^{z} f(x)g(u - x)du \right) dx.$$

But for any fixed z, the infinite strip

$$\{(u, x) : -\infty < x < \infty, -\infty < u \leq z\} = \{(x, u) : -\infty < u \leq z, -\infty < x < \infty\},$$

so $F(z)$, the CDF for Z, is given by

$$F(z) \triangleq P(Z \leq z) = \int\limits_{-\infty}^{z} \left(\int\limits_{-\infty}^{\infty} f(x)g(u - x)dx \right) du.$$

Let k be the density function for Z. It follows from the definition of PDF in Definition 7.20 that $F(z) = \int_{-\infty}^{z} k(u)du$, so

$$k(u) \triangleq \left(\int\limits_{-\infty}^{\infty} f(x)g(u - x)dx \right) \tag{7.52}$$

is a PDF for $Z = X + Y$. Reverse the roles of x and u in (7.52) to see that (7.51) is correct, that is,

$$k(x) = \int\limits_{-\infty}^{\infty} f(u)g(x - u)du. \; \square$$

Two random variables X, Y are **identically distributed** if they have the same CDF, that is, for every x, $P(X \leq x) = P(Y \leq x)$. Saying two random variables are identically distributed is not the same as saying the random variables are equal. This is related to what we said earlier, "We may have no idea *how* $X(\omega)$ functions but we may know something about what it produces."

7.7.2 Problems

1. Find the standard deviation of the fair die of Example 7.45.
2. Suppose λ is a positive constant and a random variable X has density function $f(x) = \alpha e^{-\lambda|x|}$, where α is a constant. (a) What value must α be? (b) Which moments of X exist?
3. Explain why $E[(X-\mu)^2] = E[X^2] - \mu^2$, as long as the first and second moments, $E[X]$ and $E[X^2]$, exist. [Hint: Use Theorem 7.17 and $E[(X-\mu)^2] = E[X^2 - 2\mu + \mu^2]$.]
4. Suppose X and Y are independent, identically distributed **(i. i. d.)** random variables whose density function is the one given in Example 7.49. Find the density function of $X+Y$.
5. Suppose X and Y are independent normal random variables, both of which have mean 0 and standard deviation σ. Find the density function for $X+Y$.
6. Explain why Theorem 7.18 is correct.
7. Briefly discuss how (7.49) was used in Section 2.5 when discussing linear regression.
8. Follow these steps to explain why $\int_{-\infty}^{\infty} e^{-x^2/2}dx = \sqrt{2\pi}$. Define $f(x,y) = e^{-(x^2+y^2)/2}$.

   (a) Let $\mathcal{D} = \{(r,\theta) : 0 \le \theta \le 2\pi, 0 \le r \le R\}$ and calculate that $\iint_{\mathcal{D}} f(x,y)dA = 2\pi(1 - e^{-R^2/2})$.

   (b) Let $\widetilde{\mathcal{D}} = \{(x,y) : -R \le x \le R, -R \le x \le R\}$. Explain why $|f(x,y)| \le e^{-R^2/2}$ for all (x,y) that are in the square $\widetilde{\mathcal{D}}$ but are not in the disk $\mathcal{D}$.

   (c) Explain why $(4-\pi)R^2$ is the area of the region $\mathcal{N}$ that lies inside $\widetilde{\mathcal{D}}$ but outside $\mathcal{D}$.

   (d) Use parts (b) and (c) together to explain why $\left|\iint_{\mathcal{N}} f(x,y)dA\right| \le (4-\pi)R^2 e^{-R^2/2}$.

   (e) Use L'Hôpital's rule to explain why $\lim_{R\to\infty} \iint_{\mathcal{N}} f(x,y)dA = 0$.

   (f) Explain why $\iint_{\widetilde{\mathcal{D}}} f(x,y)dA = \left(\int_{-R}^{R} e^{-u^2/2}\right)^2$.

   (g) Use parts (a) and (f) to explain why $\lim_{R\to\infty} \int_{-R}^{R} e^{-u^2/2} = \sqrt{2\pi}$.

   (h) Use part (g) to explain the convergences of $\int_0^{\infty} e^{-u^2/2} = \frac{\sqrt{2\pi}}{2}$ and $\int_{-\infty}^{\infty} e^{-u^2/2} = \sqrt{2\pi}$.

Key Terms

anti-derivative: Definition 7.2 in Section 7.1
anti-partial derivative: before Example 7.11 in Section 7.3
arclength function: Definition 7.6 in Section 7.2
average value: Defnitions 7.3 in Section 7.1, (7.9) in Section 7.2, (7.11) in Section 7.3, (7.12) in Section 7.4
chain: before (7.11) in Section 7.2
circular symmetry: before Example 7.23 in Section 7.4, Example 7.25 in Section 7.4
circulation: (7.11) in Section 7.2
closed: Definition 7.15 in Section 7.5
conservative: after Example 7.8 in Section 7.2

continuous random variable: Definition 7.19 in Section 7.7
convergent: Definition 7.4 in Section 7.1
convolution on the real line: (7.51) in Section 7.7
countable: Definition 7.19 in Section 7.7
cumulative distribution function (**CDF**): Definition 7.20(a) in Section 7.7
definite integral: before Theorem 7.1 in Section 7.1
density function: Definition 7.20(b) in Section 7.7
discrete random variable: Definition 7.19 in Section 7.7
Divergence Theorem: Theorem 7.14 in Section 7.6
divergent: Definition 7.4 in Section 7.1
double Riemann sum: (7.20) in Section 7.3
electrostatic potential difference: before Example 7.9 in Section 7.2
element of arclength: (7.4) in Section 7.2
element of area: after (7.21) in Section 7.3
Euler's Conservation Equation: Example 7.39 in Section 7.6
expectation: Definition 7.21 in Section 7.7
exponential distribution: Example 7.49 in Section 7.7
flux of F across $\mathcal{S}$: after (7.39) in Section 7.5
Fundamental Theorem of Calculus: Theorem 7.4 in Section 7.1
Fundamental Theorem of Line Integrals: Theorem 7.8 in Section 7.2
Gaussian distribution: Example 7.52 in Section 7.7
(Gaussian) error function: Definition 7.22 in Section 7.7
Gauss Theorem: Theorem 7.14 in Section 7.6
Green's Theorem: Theorem 7.13 in Section 7.3
higher moments: Definition 7.21 in Section 7.7
improper integral: Definition 7.4 in Section 7.1
identically distributed: after Theorem 7.21 in Section 7.7
indefinite integral: Definition 7.2 in Section 7.1
independent: Definition 7.23 in Section 7.7
iterated integral: (7.22) in Section 7.3
joint probability density function: Section 7.7.1
lift: (7.12) in Section 7.2
line integral: (7.5) in Section 7.2
line integral of a vector field: (7.9) in Section 7.2
magnetic permeability: before Example 7.9 in Section 7.2
mean: before (7.49)
Method of Integration by Parts: Theorem 7.7 in Section 7.1
Method of Substitution: Theorem 7.6 in Section 7.1
mode: after Example 7.51 in Section 7.7
nice solid: Definition 7.17 in Section 7.6
normal random variable: Example 7.52 in Section 7.7
opposite curve: (7.14) in Section 7.2
orientable, **oriented**, **orients**: Definition 7.14 in Section 7.5
path independent: before Example 7.8 in Section 7.2
piecewise smooth surface: Definition 7.13 in Section 7.5
polar moment of inertia: Example 7.22 in Section 7.4, Problem 7.3.7.34
positively oriented curve: after (7.10) in Section 7.2
positive orientation: Definition 7.16 in Section 7.5
positively oriented with respect to $\mathcal{S}$: Definition 7.18 in Section 7.6

potential flow: after (7.44) in Section 7.6
probability: Example 7.45 in Section 7.7
probability distribution function (PDF): Definition 7.20(b) in Section 7.7
random variable: Definition 7.19 in Section 7.7
regular partition: Theorem 7.1 in Section 7.1
Riemann sum: before Theorem 7.1 in Section 7.1
semi-infinite: before Definition 7.4 in Section 7.1
standard deviation: after (7.49)
surface integral of a scalar valued function: Section 7.5.1
surface integral of a vector field: (7.38) in Section 7.5
tangency condition: after (7.44) in Section 7.6
triple Riemann sum: (7.29) in Section 7.4
type I or **type II**: before Example 7.17 in Section 7.3
upward orientation: Example 7.35 in Section 7.5
variance: before (7.49)
work: (7.10) in Section 7.2

Mathematica Commands

ParametricPlot[{$-1 + 2$t^(3/2), $1 + $t}, {t, 0, 2}]: before Example 7.5 in Section 7.2
ParametricPlot3D[{2Cos[t], 2Sin[t], t}, {t, 0, 2π}]: Example 7.6 in Section 7.2
Graphics3D[Polygon[{{12, 0, 0}, {0, 4, 0}, {0, 0, 3}, {0, 0, 0}}], Axes $\rightarrow$ True]: after Example 7.16 in Section 7.3
ParametricPlot3D[{{rCos[θ], rSin[θ], 2r^2}, {rCos[θ], rSin[θ], $6 - 3$r}}, {r, 0, r0}, {θ], 0, 2π]}]: before Example 7.25 in Section 7.4
RegionPlot3D[x^2 + y^2 $\leq \frac{z^2}{3}$ && x^2 + y^2 + (z $-$ 1)^2 $\leq$ 1, {x, $-\frac{\sqrt{3}}{2}$, $\frac{\sqrt{3}}{2}$}, {y, $-\frac{\sqrt{3}}{2}$, $\frac{\sqrt{3}}{2}$}, {z, 0, 2}, AxesLabel $\rightarrow$ {"x", "y","z"}]: before Example 7.27 in Section 7.4
RegionPlot3D[1 $\leq$ x^2 + y^2 + z^2 && x^2 + y^2 + z^2 $\leq$ 9, {x, 0, 3}, {y, 0, 3}, {z, 0, 3}, AxesLabel $\rightarrow$ {"x","y","z"}]: before Example 7.28 in Section 7.4
RegionPlot3D[9 $\geq$ x^2 + (y^2)$^{1.7}$ + z^2 && x^2 + y^2 + z^2 $\geq$ 2, {x, -3, 0}, {y, -3, 3}, {z, -3, 3}, AxesLabel $\rightarrow$ {"x","y","z"}]: before Example 7.39 in Section 7.6

Reference

Grattan-Guiness, I. Why did George Green write his essay of 1828 on electricity and magnetism? *American Mathematical Monthly* 102, 387–396, (1995).

8

Numerical Methods I

8.1 Solving a Scalar Equation

Some engineering problems can be reduced to finding a real number x that solves a single equation

$$f(x) = 0. \tag{8.1}$$

A characteristic equation $0 = |A - \lambda I|$ is an example of such an equation. We will study **numerical methods** that produce a sequence

$$\{x_k\} = x_0, x_1, x_2, \ldots$$

of approximations for a solution $x^\star$ of (8.1).

8.1.1 Newton–Raphson Method

Given an initial guess x_0 for a solution of (8.1), we can construct a next guess, x_1, by "following the tangent line," as illustrated in Figure 8.1. We hope x_1 will be an improvement over x_0 in the sense that $0 \leq |f(x_1)| < |f(x_0)|$.

The **Newton–Raphson method**, or **Newton's method**, for short, says to take the next guess to be

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}. \tag{8.2}$$

Why? The tangent line to the curve $y = f(x)$ at the point $(x, y) = (x_0, f(x_0))$ has slope $f'(x_0)$, so the point–slope formula gives

$$f'(x_0) = \frac{y - f(x_0)}{x - x_0},$$

that is,

$$y = f(x_0) + f'(x_0)\,(x - x_0)\,.$$

Newton's method hopes that the point $(x_1, 0)$ on the tangent line happens to be closer to a point $(x^\star, f(x^\star))$ where $f(x^\star) = 0$. For $(x_1, 0)$ to be on the tangent line, we need

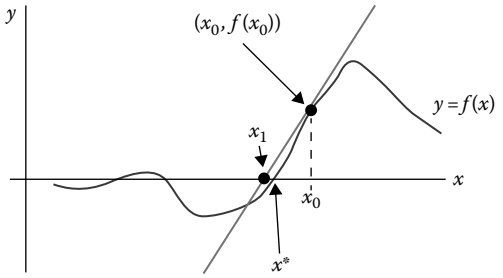


FIGURE 8.1
Newton's method.

$$0 = f(x_0) + f'(x_0)\,(x_1 - x_0)\,. \tag{8.3}$$

Solving (8.3) for x_1 gives (8.2).

Continuing from x_0 and x_1, we can construct the **iterates**

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}, \quad x_3 = x_2 - \frac{f(x_2)}{f'(x_2)}, \quad \ldots\,.$$

In general,

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)},\ k = 0, 1, 2, \ldots\,. \tag{8.4}$$

Formula (8.2) defines Newton's Method and requires one initial guess x_0.

Example 8.1

Find an approximate solution of $x^4 + 8x^2 + 21x^2 + 20x + 5$.

Method: Define $f(x) = x^4 + 8x^2 + 21x^2 + 20x + 5$. With initial guess $x_0 = -1$, Newton's method running on a calculator or a computer spreadsheet produces iterates $x_1, x_2, \ldots$ in Table 8.1. It appears that the iterates converge to $x^\star \approx -1.3819660113$ and $f(x^\star) \approx 0.0000000000$, so $x^\star$ seems to be a very good approximation of a solution of $f(x) = 0$. ◯

TABLE 8.1

Example 8.1 by Newton's Method

k	x_k	$f(x_k)$	$\triangle x_k \triangleq x_k - x_{k-1}$
0	$-1.0000000000E+00$	$-1.0000000000E+00$	
1	$-1.5000000000E+00$	$3.1250000000E-01$	$-5.0000000000E-01$
2	$-1.3750000000E+00$	$-1.9287109375E-02$	$1.2500000000E-01$
3	$-1.3819542254E+00$	$-3.2575519246E-05$	$-6.9542253521E-03$
4	$-1.3819660112E+00$	$-9.8363983625E-11$	$-1.1785862403E-05$
5	$-1.3819660113E+00$	$0.0000000000E+00$	$-3.5588421099E-11$
6	$-1.3819660113E+00$	$0.0000000000E+00$	$0.0000000000E+00$

Also, in the table we see that the changes in the iterates,

$$\triangle x_k \triangleq x_{k+1} - x_k, \tag{8.5}$$

display **quadratic convergence**, meaning that for some constant c and index K,

$$|\triangle x_{k+1}| \leq c\,|\triangle x_k|^2, \tag{8.6}$$

for all $k \geq K$. This shows that the convergence of the iterates is rapid after index K.

A typical software implementation of Newton's method stops the iteration when the absolute value of the change in the iterates, that is, $|\triangle x_k|$, is less than a specified value, ε. The requirement

$$|\triangle x_k| < \varepsilon \tag{8.7}$$

is called a **stopping rule**. An example of another stopping rule is $|f(x_k)| < \varepsilon$.

In Example 8.1, if we had used $\varepsilon = 1.E-08$ and stopping rule (8.7), then the process would have stopped at $k = 5$ and reported -1.3819660112 as an approximate solution. If, instead, we had used $\varepsilon = 1.E-04$, then the process would have stopped at $k = 4$ and reported -1.3819542254 as an approximate solution.

We see from the abstract picture in Figure 8.1 that there may be more than one solution of (8.1). Which solution Newton's method converges to can depend on the initial guess, x_0. In Example 8.1, if instead of $x_0 = -1$ we had used $x_0 = -0.8$, then Newton's method has iterates $x_k \to x^\star \approx -3.6180339887E+00$ as $k \to \infty$.

By the way, the word "iterate" is both a noun, as in "x_{k+1} is the next iterate," and a verb, as in "iterate (8.4), that is, Newton's method."

Theorem 8.1

If f is twice continuously differentiable at $x^\star$, $f(x^\star) = 0$, and $f'(x^\star) \neq 0$, then there is an open interval $\mathcal{I} = (x^\star - \delta, x^\star + \delta)$ such that if the initial guess x_0 is in $\mathcal{I}$, then the sequence of Newton's method iterates $\{x_k\}$ converges quadratically to $x^\star$ as $k \to \infty$.

At first, this theorem sounds like a good result, but it is of little practical use because it requires knowledge of a solution $x^\star$. If we already know a solution, there would be no point in finding a sequence of iterates that converges to it. Nevertheless, Theorem 8.1 is a good inspiration for a practical result.

We know from the intermediate value theorem of calculus that if a function f is continuous on a closed interval $[a, b]$ and $f(a) \cdot f(b) < 0$ then f has at least one solution in the interval (a, b), because $f(a) \cdot f(b) < 0$ implies f changes sign on $[a, b]$. Here's a result that improves on Theorem 8.1.

Theorem 8.2

Suppose f is twice continuously differentiable on an interval (a, b) and continuous on $[a, b]$ and has $f(a) \cdot f(b) < 0$. If, in addition, $f'(x) \neq 0$ on (a, b), then in (a, b) there is an $x^\star$ that

solves (8.1) and a subinterval (c, d) such that whenever the initial guess x_0 is in (c, d), then the iterates of Newton's method converge quadratically to $x^\star$ as $k \to \infty$.

This is a better theorem but it is still vague as to how to find the subinterval in which to take our initial guess. The problem of narrowing the interval where the initial guess is taken can be solved partly by using the **bisection method**: the interval $[a, b]$ can be split into two equal subintervals, $[a, 0.5(a + b)]$ and $[0.5(a + b), b]$. Assuming $f(a)f(b) < 0$, in at least one of those two intervals f will change sign, so we can use the bisection method again to split the half-sized interval into quarter-sized intervals and then into eighth-sized intervals, etc. In this way we can slowly "creep up" to a solution. When we get close enough to a solution, we can either use a stopping rule or use Newton's method to rapidly converge to a solution, $x^\star$. How close is close enough? Well, that is another theorem, but we will not state it here because it involves a complicated set of hypotheses to verify.

If, while using the bisection method, an endpoint of a subinterval happens to be a solution, then we've "hit the jackpot" and we can stop bisecting.

What if $f'(x^\star) = 0$ or $f'(x^\star)$ does not exist? Then neither Theorem 8.1 nor 8.2 applies. As a matter of logic, that does not imply that there is no solution of (8.1) nor does it imply that Newton's method does not converge to a solution.

Why, as a matter of logic? Consider the theorem that says, "If $f(x)$ is differentiable at x_0 then f is continuous at x_0." If the hypothesis fails, that is, if f is not differentiable at x_0, does that imply that f is not continuous at x_0? No. For example, $f(x) \triangleq |x|$ is not differentiable at 0, but, in spite of that, f is continuous at 0.

Regardless of the logic, it is good to have examples where $f'(x^\star)$ is not existing or is zero and Newton's method runs into trouble. The behaviors of the iterates in these examples are models for what you might actually experience in much more complicated situations.

Example 8.2

Use Newton's method to attempt to find an approximate solution of $f(x) = 0$, where

$$f(x) \triangleq \left\{ \begin{array}{rl} -\sqrt{-x}, & x < 0 \\ \sqrt{x}, & x \geq 0 \end{array} \right\}. \tag{8.8}$$

Method: For $x \neq 0$,

$$f'(x) = \left\{ \begin{array}{ll} \frac{1}{2 \cdot \sqrt{-x}}, & x < 0 \\ \frac{1}{2 \cdot \sqrt{x}}, & x > 0 \end{array} \right\},$$

so Newton's method is

$$x_{k+1} = \cdots = x_k - \left\{ \begin{array}{rl} (-\sqrt{-x_k}) \cdot 2 \cdot \sqrt{-x_k}, & \text{if } x_k < 0 \\ \sqrt{x_k} \cdot 2 \cdot \sqrt{x_k}, & \text{if } x_k > 0 \end{array} \right\} = x_k - 2x_k = -x_k,$$

as long as $x_k \neq 0$.

No matter what nonzero initial guess we take, the iterates cycle back and forth between x_0 and $-x_0$. For example, with initial guess $x_0 = 1$, the iterates cycle back and forth between 1 and -1, as shown in Figure 8.2. Newton's method is unsuccessful (Table 8.2). ◯

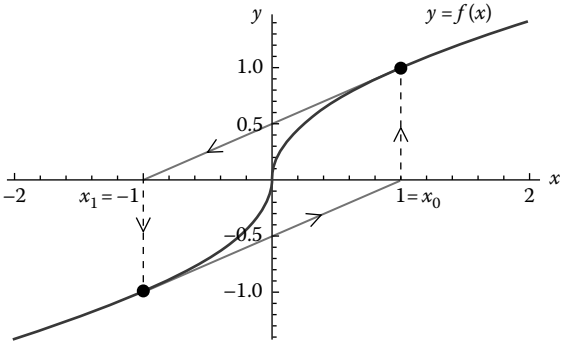


FIGURE 8.2
Example 8.2: Cycling in Newton's method.

TABLE 8.2

Example 8.2: Cycling in Newton's Method

k	x_k	$f(x_k)$	$\triangle x_k \triangleq x_k - x_{k-1}$
0	$1.0000000000E+00$	$1.000000000E+00$	
1	$-1.0000000000E+00$	$-1.000000000E+00$	$-2.0000000000E+00$
2	$1.0000000000E+00$	$1.000000000E+00$	$2.0000000000E+00$
3	$-1.0000000000E+00$	$-1.000000000E+00$	$-2.0000000000E+00$

So, what method *would* be successful in Example 8.2? The bisection method is usually slow but sure, just as the tortoise was in the fable of "the tortoise and the hare."

Example 8.3

Use the bisection method to attempt to find an approximate solution of $f(x) = 0$, where f was given in (8.8).

Method: Because $f(-1) = -1$ and $f(1) = 1$ have opposite signs, the intermediate value theorem guarantees that there is a solution of $f(x) = 0$ in the interval $[-1, 1]$. Divide this interval into two equal subintervals, $[-1, 0]$ and $[0, 1]$, and choose one of them. We hit the jackpot, because $f(0) = 0$ gives us a solution. ○

8.1.2 Modified Newton's Method

If $f(x^\star) = f'(x^\star) = 0$, then Newton's method will have difficulty for iterates $x_k \approx x^\star$. The **order** of the zero of $f(x)$ at $x^\star$ is the integer m for which $f(x^\star) = f'(x^\star) = \cdots = f^{(m)}(x^\star) = 0$, but $f^{(m+1)}(x^\star) \neq 0$, assuming there is such a finite m. If $f(x)$ is $m + 1$ times continuously differentiable at $x^\star$ and has a zero of order m at $x^\star$, then $f(x) = (x - x^\star)^m g(x)$ for some function $g(x)$ that is continuous for x near $x^\star$ and satisfies $g(x^\star) \neq 0$.

The **modified (or corrected) Newton's method** (Rall, 1966)

$$x_{k+1} = x_k - m \cdot \frac{f(x_k)}{f'(x_k)},\ k = 0, 1, 2, \ldots$$

will converge to the root $x^\star$, as long as the initial guess x_0 is close enough to $x^\star$ *and* m is the order of the zero of $f(x)$ at $x^\star$.

8.1.3 Secant Method

Suppose that a function $f(x)$ is not defined by a formula but is instead described by a physical or software process that produces a unique output y for each input x in some interval $\mathcal{I}$. For example, one software subroutine might produce the output y by some process so complicated as to make it too difficult to find a formula for f.

If we have no formula for $f(x)$, then we also have no formula for $f'(x)$, and thus, we can't use Newton's method. But, instead, we can approximate

$$f'(x_k) \approx \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}} = \frac{\triangle f_k}{\triangle x_k}, \tag{8.9}$$

where we define

$$\triangle f_k \triangleq f(x_k) - f(x_{k-1}). \tag{8.10}$$

Geometrically, we are approximating the slope of the tangent line by the slope of a secant line, so we can call the following a **secant method**:

$$x_{k+1} = x_k - \frac{f(x_k)}{\triangle f_k / \triangle x_k},\ k = 1, 2, 3, \ldots,$$

that is,

$$x_{k+1} = x_k - \frac{(x_k - x_{k-1}) f(x_k)}{f(x_k) - f(x_{k-1})} = \frac{x_{k-1} f(x_k) - x_k f(x_{k-1})}{f(x_k) - f(x_{k-1})}, k = 1, 2, 3, \ldots. \tag{8.11}$$

This method is also known as the **regula falsi**, Latin for **false position**, method. It requires *two* initial guesses x_0, x_1 in order to calculate x_2; after that, x_1, x_2 are used to calculate x_3, etc.

We see that there is danger in this method: the denominators, $f(x_k) - f(x_{k-1})$, will get close to zero if a sequence of iterates $\{x_k\}$ starts to converge to a value $x^\star$. Why? If $\{x_k\}$ starts to converge to $x^\star$, then *both* x_k and x_{k-1} will be close to $x^\star$; if $f(x)$ is continuous at $x^\star$, then the denominators, $f(x_k) - f(x_{k-1})$, will be close to $f(x^\star) - f(x^\star) = 0$.

Of course, the numerators, $(x_k - x_{k-1}) f(x_k)$, will also be close to zero if $\{x_k\}$ starts to converge. So, the secant method process involves a delicate situation of $x_{k+1} = x_k -$ (*close to zero*)/(*close to zero*).

What makes the delicate situation even more dangerous is that if two numbers, such as $f(x_k)$ and $f(x_{k-1})$, are close to each other, then their difference, $f(x_k) - f(x_{k-1})$, may behave wildly due to "round-off error." We will not attempt a deep theoretical analysis of how **round-off-error** can accumulate, but we note that numerical methods deal in numbers stored in a computer memory with only a fixed, finite number of digits of precision. For example, if the computer is forced to round-off

$$f(x_k) = 1.000000024999 \approx 1.00000002$$

and

$$f(x_{k-1}) = 1.000000005001 \approx 1.00000001,$$

then the machine acts as if $f(x_k) - f(x_{k-1}) = 1.00000002 - 1.00000001 = 1 \times 10^{-9}$. Instead, if the machine could keep all of the digits of the function values, it would calculate that

$$f(x_k) - f(x_{k-1}) = 1.000000024999 - 1.000000005001 = 0.00000001998 \approx 2 \times 10^{-9}.$$

There is a big difference between dividing by 1×10^{-9} and dividing by 2×10^{-9}! So, the numerical process may move from x_k to x_{k+1} by almost twice as much as it would if there were no round-off error. Thus, the process may produce wild swings of the iterates.

One helpful technique is to replace $f(x_k)$ by a weighted average of $f(x_k)$ and $f(x_{k-1})$:

$$x_{k+1} = x_k - \frac{(x_k - x_{k-1})\left(w \cdot f(x_k) + (1-w) \cdot f(x_{k-1})\right)}{f(x_k) - f(x_{k-1})},\ k = 1, 2, 3, \ldots, \tag{8.12}$$

where the user can adjust the constant w in the interval $[0, 1]$. Perhaps the reason that (8.12) can produce better practical results is that the weighted average $(w \cdot f(x_k) + (1-w) \cdot f(x_{k-1}))$ may help to dampen wild swings in the iterates.

In fact, the secant method in (8.11) can give **super linear convergence**, specifically

$$|\triangle x_{k+1}| \leq c\,|\triangle x_k|^{\alpha} \tag{8.13}$$

with $\alpha = \frac{1+\sqrt{5}}{2}$, under the same hypotheses under which Newton's method gave quadratic convergence, that is, with α replaced by 2. **Linear convergence** has $\alpha = 1$ and quadratic convergence has $\alpha = 2$.

In the case of linear convergence, $|\triangle x_{k+2}| \leq c\,|\triangle x_{k+1}| \leq c \cdot c\,|\triangle x_k|$, so

$$|\triangle x_k| \leq c^{k-1}\,|\triangle x_1|.$$

There is convergence if $0 \leq c < 1$.

In the case of convergence with exponent $\alpha > 1$,

$$|\triangle x_{k+2}| \leq c\,|\triangle x_{k+1}|^{\alpha} \leq c \cdot (c\,|\triangle x_k|^{\alpha})^{\alpha}, \ldots,$$

so

$$|\triangle x_k| \leq c^{(1+\alpha+\cdots+\alpha^{k-1})}\,|\triangle x_1|^{\alpha^k}.$$

8.1.4 Fixed Point Problem Iteration

Solving $x = g(x)$ for a real number x is called a **fixed-point problem** because we want to find an x that does not move under the action of $g(x)$. Pictorially, solving $x = g(x)$ is the same as finding a point of intersection of the graphs of $y = g(x)$ and $y = x$, as illustrated in Figure 8.3.

The simplest numerical method for solving $x = g(x)$ is to pick x_0, an initial guess for the solution, and iterate

$$x_{k+1} = g(x_k),\ \text{for } k = 0, 1, 2, \ldots. \tag{8.14}$$

This is called **fixed-point iteration** or the **method of successive approximations**.

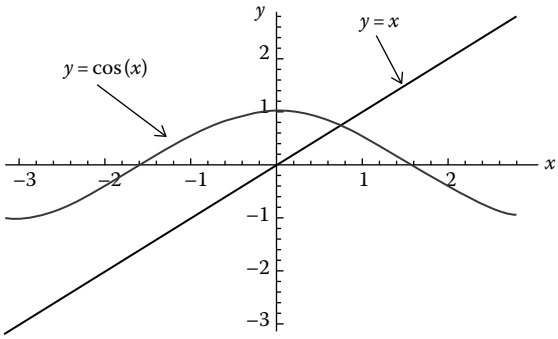


FIGURE 8.3
Fixed-point problem $x = \cos(x)$.

TABLE 8.3

Example 8.4 by Successive Approximations

k	x_k	$g(x_k)$	$\triangle x_{k+1} \triangleq x_{k+1} - x_k$
0	0.0000000000	1.0000000000	
1	1.0000000000	0.5403023059	−0.4596976941
2	0.5403023059	0.8575532158	0.3172509100
3	0.8575532158	0.6542897905	−0.2032634253
4	0.6542897905	0.7934803587	0.1391905682
5	0.7934803587	0.7013687736	−0.0921115851
6	0.7013687736	0.7639596829	0.0625909093
7	0.7639596829	0.7221024250	−0.0418572579
8	0.7221024250	0.7504177618	0.0283153367
...	...	...	...
30	0.7390822985	0.7390870427	$4.744172930 \times 10^{-6}$

Example 8.4

Find an approximate solution of $x = \cos(x)$ using the method of successive approximations.

Method: An initial guess $x_0 = 0$ produces the results in Table 8.3. We used an Excel spreadsheet for this, although we could have used a basic calculator instead. It looks as if the iterates are converging to about 0.7391 but slowly, for example, with linear convergence. ◯

8.1.5 Aitken's δ^2 Method

Aitken's δ^2 method is a **convergence acceleration** technique that speeds up linear convergence to quadratic convergence, assuming the initial guess is close enough to a simple* solution $x^\star$ for the equation $x = g(x)$.

* Meaning $g(x^\star) = x^\star$ and $|g'(x^\star)| \neq 1$.

TABLE 8.4

Example 8.5 by Aitken's δ^2 Method

k	x_k	$f(x_k)$	$\triangle x_{k+1} \triangleq x_{k+1} - x_k$
0	$0.000000000E+00$	$1.000000000E+00$	$-3.149266427E-01$
1	$6.850733573E-01$	$7.743726338E-01$	$-3.571247764E-02$
2	$7.386601562E-01$	$7.393713361E-01$	$-2.862297597E-04$
3	$7.390851064E-01$	$7.390851513E-01$	$-1.732571964E-08$
4	$7.390851340E-01$	$7.390851327E-01$	$8.816138486E-09$
5	$7.390851415E-01$	$7.390851276E-01$	$7.206302000E-09$
6	$7.390851348E-01$	$7.390851321E-01$	$1.746865097E-09$
7	$7.390851339E-01$	$7.390851328E-01$	$2.687278999E-08$
8	$7.390851596E-01$	$7.390851154E-01$	$1.763205759E-08$
9	$7.390851330E-01$	$7.390851333E-01$	$4.258583575E-08$
10	$7.390851759E-01$	$7.390851045E-01$	$2.914004560E-08$
11	$7.390851336E-01$	$7.390851330E-01$	$1.028375171E-07$
12	$7.390852358E-01$	$7.390850641E-01$	$6.925546447E-08$

Define

$$G(x) = \frac{x\,g(g(x)) - (g(x))^2}{g(g(x)) - 2g(x) + x}. \tag{8.15}$$

An iteration process using Aitken's δ^2 method is

$$x_{k+1} = G(x_k), \text{ for } k = 0, 1, 2, \ldots . \tag{8.16}$$

Example 8.5

(Example 8.4 again) Find an approximate solution of $x = \cos(x)$ using Aitken's δ^2 method.

Method: An initial guess $x_0 = 0$ produces the results in Table 8.4. We used an Excel spreadsheet for this, although we could have used a basic calculator instead. It looks as if the iterates are rapidly converging to about 0.739085, perhaps with quadratic convergence. But the error eventually stops improving and instead starts bouncing around. ○

One cycle of **Steffensen's method** uses successive approximation for two steps and then uses Aitken's δ^2 process for a step; after that, a new cycle is done, etc. This usually works better than just using Aitken's δ^2 method.

8.1.6 Newton's Method versus Other Methods

A problem posed as a fixed-point problem $x = g(x)$ is equivalent to the problem of finding a root of $f(x) = 0$, where $f(x) \triangleq x - g(x)$. So, we can use Newton's method instead of the method of successive approximations or Aitken's δ^2 method.

Example 8.6

(Example 8.4 again) Find an approximate solution of $x = \cos(x)$ using Newton's method.

Method: Define $f(x) = x - \cos x$, so solving $x = \cos x$ is equivalent to solving $f(x) = 0$. We used Newton's method in an Excel spreadsheet and got good results shown in Table 8.5. ○

8.1.7 Trouble

Unfortunately, Newton's method may not converge to a solution of $f(x) = 0$ even though a solution exists.

Perhaps even worse is that Newton's method, using the usual stopping rule, $|\triangle x_k| < \varepsilon$, may (Donavon et al., 1993) say that it has found a numerical solution when actually $f(x) = 0$ has no solution or when the numerical solution is far from an actual solution!

In spite of these cautionary notes, Newton's method is a valuable, elementary approach to solving equations. In particular, when we can verify the hypotheses of a theorem that does guarantee convergence, then we can be even more confident that Newton's method will be useful.

Learn More About It

Elementary Numerical Analysis: An Algorithmic Approach, 2*nd* ed., by S. D. Conte and Carl de Boor, McGraw-Hill, Inc. © 1972, has analysis of the convergence of the secant method and of Müller's method, which does a good job of searching for complex roots. Müller's method effectively interpolates the function $f(x)$ by a quadratic polynomial and then uses the quadratic formula to approximate a root of $f(x)$.

Analysis of Numerical Methods, by Eugene Isaacson and Herbert Bishop Keller, Dover Publications © 1994 has a proof that Aitken's δ^2 method gives quadratic convergence.

8.1.8 Problems

In problems 1–4 use (a) Newton's method and (b) the bisection method to find an approximate solution of the given equation in the given interval. Show enough steps in a

TABLE 8.5

Example 8.6 by Newton's Method

k	x_k	$f(x_k)$	$\triangle x_{k+1} \triangleq x_{k+1} - x_k$
0	$0.000000000E+00$	$-1.000000000E+00$	$-1.000000000E+00$
1	$1.000000000E+00$	$4.596976941E-01$	$2.496361322E-01$
2	$7.503638678E-01$	$1.892307382E-02$	$1.125097693E-02$
3	$7.391128909E-01$	$4.645589899E-05$	$2.775752608E-05$
4	$7.390851334E-01$	$2.847204694E-10$	$1.701233407E-10$
5	$7.390851332E-01$	$0.000000000E+00$	$0.000000000E+00$

table(s) to demonstrate how well your methods were working. For the bisection method, stop the process when $|\triangle x_k| < 10^{-7}$.

1. $e^{-1.5x} = \cos 2x$, $1 \le x \le 3$.
2. $e^{-1.5x} = \cos 2x$, $0.5 \le x \le 1$.
3. $\ln(x+3) = \cos 2x$, $-3 < x \le -2$.
4. $0 = -\frac{\pi}{2} + x \arctan(x^2)$, $1 \le x \le 2$.

In problems 5 and 6 use Newton's method to find two approximate solutions of the given equation. Show enough steps in table(s) to demonstrate how well your methods were working. Describe the convergence.

5. $\ln(x+3) = x^2 - 2$.
6. $-1 + \ln(x+3) = \cos 2x$.

In problems 7 and 8 use the secant method to find an approximate solution of the given equation.

7. $\frac{1}{1+x^2} - e^{-1.5x} = x$.
8. $1.5 \cos\left(\frac{2}{1+x^2}\right) = e^{-1.5x}$.
9. In Example 9.33 in Section 9.6 we will need to find the solutions of the equation $\tanh(\theta) = \tan(\theta)$. Find numerical approximations of the first four positive solutions for θ, that is, the four positive solutions closest to zero. Try initial guesses of θ_0 equal to 4, 7, 10, 13, respectively. (These guesses were obtained from looking at Figure 9.30.)

In problems 10 and 11 use (a) the method of successive approximations and (b) an Aitken's δ^2 method to find an approximate solution of the given equation in the given interval. Show enough steps in a table(s) to demonstrate how well your methods were working. Describe the convergence.

10. The equation of Problem 8.1.8.8.
11. The equation of Problem 8.1.8.7.
12. In Tables 8.6 and 8.7 are examples of Newton's method iteration for solving the equations $f_1(x) = 0$ and $f_2(x) = 0$, respectively. Comment on the type of convergence each table shows and speculate on the nature of the zeros of f_1, f_2 as the cause for the different types of convergence.
13. In Example 9.32 in Section 9.6 we will need to find the solutions of the equation

$$0 = \sqrt{2}\cos(\sqrt{2\lambda})\sin(\sqrt{3\lambda}) + \sqrt{3}\sin(\sqrt{2\lambda})\cos(\sqrt{3\lambda}).$$

    Find numerical approximations of the first four positive solutions for λ, that is, the four positive solutions closest to zero.
14. In Table 8.8 are examples of Newton's method and a modified Newton's method iteration with $m = 3$ for solving the equation $f(x) = 0$. Comment on the type of convergence each table shows and speculate on the nature of the zero of f as the cause for the different types of convergence.

TABLE 8.6

Problem 8.1.8.12 for $f_1(x) = 0$ by Newton's Method

k	x_k	$f(x_k)$	$\triangle x_{k+1} \triangleq x_{k+1} - x_k$
0	$1.5000000E+00$	$1.2500000E-01$	$-1.6666667E-01$
1	$1.3333333E+00$	$3.7037037E-02$	$-1.1111111E-01$
2	$1.2222222E+00$	$1.0973937E-02$	$-7.4074074E-02$
3	$1.1481481E+00$	$3.2515369E-03$	$-4.9382716E-02$
4	$1.0987654E+00$	$9.6341833E-04$	$-3.2921811E-02$
5	$1.0658436E+00$	$2.8545728E-04$	$-2.1947874E-02$
6	$1.0438957E+00$	$8.4579936E-05$	$-1.4631916E-02$
7	$1.0292638E+00$	$2.5060722E-05$	$-9.7546106E-03$
8	$1.0195092E+00$	$7.4253990E-06$	$-6.5030737E-03$
9	$1.0130061E+00$	$2.2001182E-06$	$-4.3335383E-03$

TABLE 8.7

Problem 8.1.8.12 for $f_2(x) = 0$ by Newton's Method

k	x_k	$f(x_k)$	$\triangle x_{k+1} \triangleq x_{k+1} - x_k$
0	$1.5000000E+00$	$1.1275000E-01$	$-1.5093708E-01$
1	$1.3490629E+00$	$3.0711573E-02$	$-8.4642976E-02$
2	$1.2644199E+00$	$6.8889308E-03$	$-3.3243744E-02$
3	$1.2311762E+00$	$8.3882398E-04$	$-5.3135542E-03$
4	$1.2258626E+00$	$1.9402732E-05$	$-1.2884470E-04$
5	$1.2257338E+00$	$1.1229867E-08$	$-7.4658841E-08$
6	$1.2257337E+00$	$3.7747583E-15$	$-2.5091040E-14$
7	$1.2257337E+00$	$0.0000000E+00$	$0.0000000E+00$

TABLE 8.8

Problem 8.1.8.14 for $f(x) = 0$ by Newton's Method

	Newton's			**Modified**		
k	x_k	$f(x_k)$	$\triangle x_{k+1}$	x_k	$f(x_k)$	$\triangle x_{k+1}$
0	$1.50000E+00$	$3.4864E+00$	$-5.0000E+00$	$1.50000E+00$	$3.4864E+00$	$-1.5000E+01$
1	$-3.50000E+00$	$-5.1743E+02$	$4.5455E-01$	$-1.35000E+01$	$-1.4246E+09$	$2.4194E+00$
2	$-3.04545E+00$	$-1.7988E+02$	$4.0541E-01$	$-1.10806E+01$	$-6.6486E+07$	$2.3120E+00$
3	$-2.64005E+00$	$-6.1820E+01$	$3.5346E-01$	$-8.76868E+00$	$-3.0146E+06$	$2.1642E+00$
4	$-2.28659E+00$	$-2.0959E+01$	$3.0014E-01$	$-6.60444E+00$	$-1.3000E+05$	$1.9540E+00$
5	$-1.98645E+00$	$-6.9973E+00$	$2.4745E-01$	$-4.65041E+00$	$-5.0895E+03$	$1.6467E+00$
6	$-1.73900E+00$	$-2.2971E+00$	$1.9765E-01$	$-3.00371E+00$	$-1.6218E+02$	$1.2013E+00$
7	$-1.54135E+00$	$-7.4104E-01$	$1.5287E-01$	$-1.80238E+00$	$-3.1326E+00$	$6.3306E-01$
8	$-1.38849E+00$	$-2.3504E-01$	$1.1465E-01$	$-1.16932E+00$	$-1.5629E-02$	$1.6027E-01$
9	$-1.27384E+00$	$-7.3401E-02$	$8.3644E-02$	$-1.00905E+00$	$-2.0302E-06$	$9.0184E-03$
10	$-1.19019E+00$	$-2.2620E-02$	$5.9618E-02$	$-1.00003E+00$	$-5.4657E-14$	$2.7192E-05$
11	$-1.13058E+00$	$-6.8958E-03$	$4.1710E-02$	$-1.00000E+00$	$-4.0700E-29$	$2.4647E-10$
12	$-1.08887E+00$	$-2.0849E-03$	$2.8770E-02$	$-1.00000E+00$	$0.0000E+00$	$\#DIV/0!$

8.2 Solving a System of Equations

8.2.1 Newton's Method in $\mathbb{R}^n$

To find an approximate solution of a system of n equations in n unknowns, that is,

$$\mathbf{0} = \mathbf{f}(\mathbf{x}) \triangleq \begin{Bmatrix} f_1(x_1, \ldots, x_n) \\ \vdots \\ f_n(x_1, \ldots, x_n) \end{Bmatrix}, \tag{8.17}$$

we can use a generalization of Newton's method. Recall linear approximation for a function of several variables, for example, (6.33) in Section 6.4,

$$f(x_0 + \triangle x, y_0 + \triangle y) \approx f(x_0, y_0) + \triangle x \frac{\partial f}{\partial x}(x_0, y_0) + \triangle y \frac{\partial f}{\partial y}(x_0, y_0).$$

Let $\mathbf{x}_k = [\, \mathbf{x}_{1,k} \quad \cdots \quad \mathbf{x}_{n,k} \,]^T$. For each $i = 1, \ldots, n$, linear approximation gives us

$$f_i(x_{1,k+1}, \ldots, x_{n,k+1}) \approx f_i(x_{1,k}, \ldots, x_{n,k}) + \sum_{j=1}^{n} \left(\frac{\partial f_i}{\partial x_j}(\mathbf{x}_k) \right) (x_{j,k+1} - x_{j,k}).$$

This implies that

$$\begin{bmatrix} f_1(\mathbf{x}_{k+1}) \\ \cdot \\ \cdot \\ \cdot \\ f_n(\mathbf{x}_{k+1}) \end{bmatrix} \approx \begin{bmatrix} f_1(\mathbf{x}_k)) \\ \cdot \\ \cdot \\ \cdot \\ f_n(\mathbf{x}_k) \end{bmatrix} + \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}_k) & \cdot & \cdot & \cdot & \frac{\partial f_1}{\partial x_n}(\mathbf{x}_k) \\ \cdot & \cdot & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ \frac{\partial f_n}{\partial x_1}(\mathbf{x}_k) & \cdot & \cdot & \cdot & \frac{\partial f_n}{\partial x_n}(\mathbf{x}_k) \end{bmatrix} \begin{bmatrix} x_{1,k+1} - x_{1,k}) \\ \cdot \\ \cdot \\ \cdot \\ x_{n,k+1} - x_{n,k} \end{bmatrix},$$

that is,

$$\mathbf{f}(\mathbf{x}_{k+1}) \approx \mathbf{f}(\mathbf{x}_k) + \frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x}_k) \cdot (\mathbf{x}_{k+1} - \mathbf{x}_k).$$

Note that $\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x}_k)$ is the $n \times n$ Jacobian matrix evaluated at $\mathbf{x}_k$.

As in Newton's method for a single equation, we find $\mathbf{x}_{k+1}$ by hoping that

$$\mathbf{0} = \mathbf{f}(\mathbf{x}_{k+1}),$$

that is,

$$\mathbf{0} \approx \mathbf{f}(\mathbf{x}_k) + \frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x}_k) \cdot (\mathbf{x}_{k+1} - \mathbf{x}_k),$$

that is,

$$\left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x}_k)\right)\mathbf{x}_{k+1} = \left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x}_k)\right)\mathbf{x}_k - \mathbf{f}(\mathbf{x}_k).$$

So, the generalization of Newton's method to a system of n equations in n unknowns is

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x}_k)\right)^{-1}\mathbf{f}(\mathbf{x}_k) \triangleq \mathbf{x}_k + \boldsymbol{\xi}_k. \tag{8.18}$$

We emphasize that in software implementation of Newton's method for systems, we do not usually use $\left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x}_k)\right)^{-1}$ but instead calculate $\boldsymbol{\xi}_k$ using the software's solution of the system of linear algebraic equations

$$A\boldsymbol{\xi}_k = -\mathbf{f}(\mathbf{x}_k), \quad \text{where} \quad A \triangleq \left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x}_k)\right).$$

For example, in MATLAB®,

$$\boldsymbol{\xi}_k = -A\backslash\mathbf{f}(\mathbf{x}_k).$$

Example 8.7

Find an approximate solution of the system of equations

$$\left\{\begin{array}{l} 0 = x - y + 1 \\ 0 = x^3 - y^2 + 1 \end{array}\right\}. \tag{8.19}$$

Method: Define $f(x) = x - y + 1$, $g(x, y) = x^3 - y^2 + 1$, $\mathbf{x} \triangleq \begin{bmatrix} x \\ y \end{bmatrix}$, and $\mathbf{f}(\mathbf{x}) \triangleq \begin{bmatrix} x - y + 1 \\ x^3 - y^2 + 1 \end{bmatrix}$.

With initial guess $(x_0, y_0) = (-0.6, 0.5)$, a calculator or a computer software package produces **iterates** $(x_1, y_1), (x_2, y_2), \ldots$ summarized in Table 8.9. It appears that the iterates converge to $(x^\star, y^\star) \approx (-1, 0)$, with $f(x^\star, y^\star) \approx 0$ and $g(x^\star, y^\star) \approx 0$, so $(x^\star, y^\star)$ seems to be a very good approximation of a solution of the system of equations $0 = f(x, y) = g(x, y)$. In fact, $f(-1, 0) = g(-1, 0) = 0$ exactly. ◯

Also, in the table, we see that the norms of the changes in the iterates,

$$\|(\triangle x_{k+1}, \triangle y_{k+1})\| \triangleq \sqrt{(x_{k+1} - x_k)^2 + (y_{k+1} - y_k)^2}, \tag{8.20}$$

display **quadratic convergence**, that is, for some constant c and index K,

$$\|(\triangle x_{k+1}, \triangle y_{k+1})\| \leq c\|(\triangle x_k, \triangle y_k)\|^2,$$

for all $k \geq K$.

The stopping rule is typically $\|(\triangle x_k, \triangle y_k)\| < \varepsilon$.

TABLE 8.9

Solution of Example 8.7 by Newton's Method

k	x_k	y_k	$f(x_k,y_k)$	$g(x_k,y_k)$	$\|(\triangle x_k,\triangle y_k)\|$
0	$-6.0000000E-01$	$5.0000000E-01$	$-1.0000000E-01$	$5.3400000E-01$	
1	$-8.5250000E+00$	$-7.5250000E+00$	$-1.7763568E-15$	$-6.7518533E+02$	$1.1278575E+01$
2	$-5.6281647E+00$	$-4.6281647E+00$	$0.0000000E+00$	$-1.9869899E+02$	$4.0967438E+00$
3	$-3.7228196E+00$	$-2.7228196E+00$	$0.0000000E+00$	$-5.8009740E+01$	$2.6945648E+00$
4	$-2.4891944E+00$	$-1.4891944E+00$	$-2.2204460E-16$	$-1.6640970E+01$	$1.7446095E+00$
5	$-1.7175880E+00$	$-7.1758800E-01$	$1.1102230E-16$	$-4.5820035E+00$	$1.0912163E+00$
6	$-1.2721062E+00$	$-2.7210622E-01$	$0.0000000E+00$	$-1.1326331E+00$	$6.3000637E-01$
7	$-1.0623195E+00$	$-6.2319540E-02$	$0.0000000E+00$	$-2.0273555E-01$	$2.9668317E-01$
8	$-1.0045635E+00$	$-4.5635376E-03$	$1.1102230E-16$	$-1.3774011E-02$	$8.1679321E-02$
9	$-1.0000275E+00$	$-2.7496011E-05$	$0.0000000E+00$	$-8.2491056E-05$	$6.4149316E-03$
10	$-1.0000000E+00$	$-1.0079807E-09$	$-2.2204460E-16$	$-3.0239424E-09$	$3.8883806E-05$
11	$-1.0000000E+00$	$-1.0258136E-16$	$1.1102230E-16$	$0.0000000E+00$	$1.4255000E-09$
12	$-1.0000000E+00$	$8.4409410E-18$	$0.0000000E+00$	$0.0000000E+00$	$1.1102230E-16$
13	$-1.0000000E+00$	$8.4409410E-18$	$0.0000000E+00$	$0.0000000E+00$	$0.0000000E+00$

As Newton's method works on the solution of a system $\left\{\begin{matrix} f(x,y)=0 \\ g(x,y)=0 \end{matrix}\right\}$, a change in the iterate (x,y) can get $f(x,y)$ closer to zero at the expense of $g(x,y)$ going further from zero, or vice versa. The method works on solving both equations in the long run, but in the short run, we cannot guarantee that both $f(x,y)$ and $g(x,y)$ are *always* getting closer to zero.

As in Example 8.1 in Section 8.1, there may be more than one solution of the nonlinear system of equations (8.20). Which solution Newton's method converges to can depend on the initial guess, (x_0,y_0). For example, if, instead of $(x_0,y_0)=(-0.6,0.5)$, we had used $(x_0,y_0)=(-0.6,0.6)$, then Newton's method has iterates $(x_k,y_k)\to(x^\star,y^\star)\approx(2,3)$ as $k\to\infty$, as shown in Table 8.10. In fact, $f(2,3)=g(2,3)=0$ exactly.

Again, the convergence is quadratic, specifically for $k\geq 3$.

TABLE 8.10

Second Newton's Method Solution of Example 8.7

k	x_k	y_k	$f(x_k,y_k)$	$g(x_k,y_k)$	$\|(\triangle x_k,\triangle y_k)\|$
0	$-6.0000000E-01$	$6.0000000E-01$	$-2.0000000E-01$	$4.2400000E-01$	
1	$4.9333333E+00$	$5.9333333E+00$	$8.8817842E-16$	$8.5861926E+01$	$7.6851950E+00$
2	$3.5291370E+00$	$4.5291370E+00$	$0.0000000E+00$	$2.4441641E+01$	$1.9858335E+00$
3	$2.6656624E+00$	$3.6656624E+00$	$0.0000000E+00$	$6.5044656E+00$	$1.2211375E+00$
4	$2.2005907E+00$	$3.2005907E+00$	$0.0000000E+00$	$1.4127987E+00$	$6.5771064E-01$
5	$2.0267424E+00$	$3.0267424E+00$	$0.0000000E+00$	$1.6404943E-01$	$2.4585864E-01$
6	$2.0005764E+00$	$3.0005764E+00$	$0.0000000E+00$	$3.4603058E-03$	$3.7004282E-02$
7	$2.0000003E+00$	$3.0000003E+00$	$0.0000000E+00$	$1.6602076E-06$	$8.1481894E-04$
8	$2.0000000E+00$	$3.0000000E+00$	$0.0000000E+00$	$3.8014036E-13$	$3.9131449E-07$
9	$2.0000000E+00$	$3.0000000E+00$	$0.0000000E+00$	$0.0000000E+00$	$8.9599943E-14$
10	$2.0000000E+00$	$3.0000000E+00$	$0.0000000E+00$	$0.0000000E+00$	$0.0000000E+00$

If, instead we had used $(x_0, y_0) = (-0.6, 1.02)$, then Newton's method has iterates $(x_k, y_k) \to (x^\star, y^\star) \approx (0, 1)$ as $k \to \infty$. In fact, $f(0,1) = g(0,1) = 0$ exactly. Again, the convergence is quadratic.

8.2.2 Newton–Kantorovich Theorem

We recall the concept of **norm** from Section 2.3, specifically the three properties that define the concept in (2.9) in Section 2.3:

(a) $\|\mathbf{x}\| \geq 0$, for *all* $\mathbf{x}$ in $\mathbb{R}^n$ with equality only if $\mathbf{x} = \mathbf{0}$.
(b) $\|\alpha\mathbf{x}\| = |\alpha|\,\|\mathbf{x}\|$, for *all* $\mathbf{x}$ in $\mathbb{R}^n$, and *all* scalars α.
(c) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$, for *all* $\mathbf{x}, \mathbf{y}$ in $\mathbb{R}^n$.

The **Euclidean norm** defined on $\mathbb{R}^n$ by

$$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$$

is just one of the norms that can be defined on $\mathbb{R}^n$. All of our previous works used the Euclidean norm and we will emphasize it in the following text.

Theorem 8.3

Given any norm $\|\ \|$ on $\mathbb{R}^n$ and any real, $n \times n$ matrix A, there is a constant $M \geq 0$ satisfying

$$\|A\mathbf{x}\| \leq M\|\mathbf{x}\|, \quad \text{for } \textit{all } \mathbf{x} \text{ in } \mathbb{R}^n. \tag{8.21}$$

Definition 8.1

The *smallest* such number M in the conclusion of Theorem 8.3 is called $\|A\|$ and is called the **matrix norm induced** by the vector norm $\|\ \|$ on $\mathbb{R}^n$.

Remarks:

(1) $\|A\mathbf{x}\| \leq \|A\|\,\|\mathbf{x}\|$, for all $\mathbf{x}$ in $\mathbb{R}^n$.
(2) $\|A\| = 0$ only if $A = O$, the zero operator that satisfies $O\mathbf{x} = \mathbf{0}$ for all $\mathbf{x}$ in $\mathbb{R}^n$.

As usual denote the Jacobian matrix by

$$J(\overline{\mathbf{x}}) = \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\overline{\mathbf{x}})\right]$$

and closed balls by

$$B_R(\overline{\mathbf{x}}) \triangleq \{\mathbf{x} : \|\mathbf{x} - \overline{\mathbf{x}}\| \leq R\}.$$

Theorem 8.4

(Newton–Kantorovich theorem for system (8.17)) Suppose there exist positive constants R, K_0, K_1, η, t and constant vector $\bar{\mathbf{x}}$ such that

(a) $\mathbf{f}(\mathbf{x})$ is continuously differentiable.

(b) $\|\big(J(\bar{\mathbf{x}})\big)^{-1}\| \leq K_0$.

(c) $\|J(\mathbf{x}) - J(\mathbf{y})\| \leq K_1\|\mathbf{x} - \mathbf{y}\|$, for all $\mathbf{x}$, $\mathbf{y}$ in $B_R(\bar{\mathbf{x}})$.

(d) $\|\big(J(\bar{\mathbf{x}})\big)^{-1}\mathbf{f}(\bar{\mathbf{x}})\| \leq \eta$.

(e) $h_0 \triangleq K_0K_1\eta < \frac{1}{4}$.

(f) $t\eta \leq R$, where t is the smaller root of the quadratic equation $h_0t^2 - t + 1 = 0$.

Then in $B_{t\eta}(\bar{\mathbf{x}})$, there is an $\mathbf{x}^\star$ solving (8.17) such that if the initial guess $\mathbf{x}_0$ is in $B_{t\eta}(\bar{\mathbf{x}})$, then the iterates of a modified Newton's method given by

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\bar{\mathbf{x}})\right)^{-1} \mathbf{f}(\mathbf{x}_k)$$

converge linearly to $\mathbf{x}^\star$ as $k \to \infty$.

Note that in this modified Newton's method, we can calculate the Jacobian matrix $J(\bar{\mathbf{x}})$ once and use it over and over again during the iterations. This can speed up the program's execution, at the possible cost of a slower convergence exponent α. Note that the word "modified" here has a different meaning than it did in Section 8.1.

As is usual for equation-solving methods, the closer $\bar{\mathbf{x}}$ is to a solution, the easier it is to guarantee convergence.

One difficulty with using Theorem 8.4 is that K_1 may depend on R, the size of the ball in which $\mathbf{x}$ and $\mathbf{y}$ lie, and then K_1 influences h_0 and thus η in hypothesis (f) in which R again appears. Finding R can be like a dog chasing its tail.

Theorem 8.5

(Another Newton–Kantorovich theorem for system (8.17)) Suppose there exist positive constants R, K_0, K_1, η and constant vector $\bar{\mathbf{x}}$ such that

(a) $\mathbf{f}(\mathbf{x})$ is continuously differentiable.

(b) $\|\big(J(\bar{\mathbf{x}})\big)^{-1}\| \leq K_0$.

(c) $\|J(\mathbf{x}) - J(\mathbf{y})\| \leq K_1\|\mathbf{x} - \mathbf{y}\|$, for all $\mathbf{x}$, $\mathbf{y}$ in $B_R(\bar{\mathbf{x}})$.

(d) $\|\big(J(\bar{\mathbf{x}})\big)^{-1}\mathbf{f}(\bar{\mathbf{x}})\| \leq \eta$.

(e) $K_0K_1R < \frac{2}{3}$.

(f) $\eta \leq (1 - \frac{3}{2}K_0K_1R)R$.

Then in $B_R(\bar{\mathbf{x}})$, there is an $\mathbf{x}^\star$ solving (8.17) such that if the initial guess $\mathbf{x}_0$ is in $B_R(\bar{\mathbf{x}})$, then the iterates of Newton's method, (8.18), converge quadratically to $\mathbf{x}^\star$ as $k \to \infty$.

As is usual for equation solving methods, the closer $\bar{\mathbf{x}}$ is to a solution, the easier it is to guarantee convergence.

A difficulty with using Theorem 8.5 is that we must find R that simultaneously satisfies three different hypotheses: (b), (e), and (f). Hypothesis (e) requires that R be relatively small, but then (f) requires that R be relatively large.

8.2.3 Fixed Point Problem Iteration

Similar to the case of a single scalar equation, solving $\mathbf{x} = \mathbf{g}(\mathbf{x})$ for $\mathbf{x}$ in $\mathbb{R}^n$, that is, a system of n equations in n unknowns, is a **fixed-point problem**.

Every problem of the form $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ is equivalent to the fixed-point problem $\mathbf{x} = \mathbf{g}(\mathbf{x})$ where you may define $\mathbf{g}(\mathbf{x}) = \mathbf{x} + \mathbf{f}(\mathbf{x})$ or $\mathbf{g}(\mathbf{x}) = \mathbf{x} - \mathbf{f}(\mathbf{x})$. In fact, even the seemingly simple problem of solving $A\mathbf{x} = \mathbf{b}$ can benefit tremendously from rewriting it as a fixed-point problem, as we will see in Section 8.4.

A very special case of Theorem 8.4 is when $\mathbf{f}(\mathbf{x}) = C\mathbf{x} + \mathbf{z}$ for some constant matrix C and constant vector $\mathbf{z}$. Then the Jacobian is given identically by $J(\mathbf{x}) = C$; hence, uniformity in hypothesis (c) allows R to be chosen as large as one needs to make hypothesis (f) be true. So, in this special case, the hypotheses of Theorem 8.4 simplify considerably: *define* $K_0 \triangleq \|C^{-1}\|$, $K_1 \triangleq \|C\|$, and $\eta = \|\bar{\mathbf{x}} + C^{-1}\mathbf{z}\|$. The hypotheses reduce to requiring only $h_0 = \|C\|\,\|C^{-1}\|\eta < \frac{1}{4}$; the conclusion holds in the ball $B_r(\bar{\mathbf{x}})$ where $r = t\eta = \frac{\eta}{h_0} \cdot (1 - \sqrt{1 - 4h_0})$. But this is all theoretical, and we would wonder why we would need Newton's method to solve $C\mathbf{x} + \mathbf{z}$ if we had our hands on C^{-1}!

But even this simpler hypothesis, $h_0 = \|C\|\,\|C^{-1}\|\eta < \frac{1}{4}$, is unnecessarily complicated. The **Banach contraction mapping theorem**, a result about fixed points, assures us that as long as $\|(I - C)\| < 1$, then Newton's method for $\mathbf{f}(\mathbf{x}) = C\mathbf{x} + \mathbf{z}$ will converge, no matter how far away from being correct is the initial guess $\mathbf{x}_0$, assuming there were no round-off error.

Both Theorems 8.4 and 8.5 generalize from $\mathbb{R}^n$ to infinite-dimensional, complete normed vector spaces, that is, **Banach spaces**. The word "complete" has the same Cauchy convergence meaning as in the definition of Hilbert spaces (Section 2.10). We would also need to use a generalization of the Jacobian matrix J called the **Fréchet derivative**, $D\mathbf{f}$.

8.2.4 Secant Method for Systems of Equations

If we are trying to solve the system of equations (8.17) and there is no formula for a function $f_i(\mathbf{x})$, then there is no formula for the partial derivatives $\partial f_i / \partial x_j$. As in the case of a single equation in a single unknown, we can approximate the partial derivatives. An analogue of the secant method (8.12) in Section 8.1 is

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \left[\frac{\triangle f_{i,k}}{\triangle x_{j,k}}\right]^{-1} \big(w \cdot \mathbf{f}(\mathbf{x}_k) + (1 - w) \cdot \mathbf{f}(\mathbf{x}_{k-1})\big), k = 1, 2, 3, \ldots, \tag{8.22}$$

where the approximate Jacobian matrix is given by

$$\left[\frac{\triangle f_{i,k}}{\triangle x_{j,k}}\right] \triangleq \frac{f_i(\mathbf{x}_{k-1} + (x_{j,k} - x_{j,k-1})\mathbf{e}^{(j)}) - f_i(\mathbf{x}_{k-1})}{x_{j,k} - x_{j,k-1}}. \tag{8.23}$$

Recall that $\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \ldots, \mathbf{e}^{(n)}$ are the columns of the $n \times n$ identity matrix I_n.

In order to use the secant method we have to choose *two* initial guesses, $\mathbf{x}_0$ and $\mathbf{x}_1$. We need them in order to create the first approximate Jacobian matrix in (8.23).

Also, as for the Newton–Raphson method for systems, in software implementation of the secant method in (8.22), we do not actually compute $\left[\frac{\triangle f_{i,k}}{\triangle x_{j,k}}\right]^{-1}\big(w\cdot\mathbf{f}(\mathbf{x}_k)+(1-w)\cdot\mathbf{f}(\mathbf{x}_{k-1})\big)$, but instead we use

$$\mathbf{x}_{k+1}=\mathbf{x}_k+\boldsymbol{\xi}_k,$$

where $\boldsymbol{\xi}_k$ is the software's solution of the linear algebraic system

$$\left[\frac{\triangle f_{i,k}}{\triangle x_{j,k}}\right]\boldsymbol{\xi}_k=-\big(w\cdot\mathbf{f}(\mathbf{x}_k)+(1-w)\cdot\mathbf{f}(\mathbf{x}_{k-1})\big).$$

Example 8.8

Use the secant method to find an approximate solution of the system of equations (8.19), that is,

$$\left\{\begin{array}{l}0=x-y+1\\0=x^3-y^2+1\end{array}\right\}.$$

Method: Again, define $f(x)=x-y+1$, $g(x,y)=x^3-y^2+1$,

$$\mathbf{x}\triangleq\left[\begin{array}{c}x\\y\end{array}\right],\quad\text{and}\quad\mathbf{f}(\mathbf{x})\triangleq\left[\begin{array}{c}x-y+1\\x^3-y^2+1\end{array}\right].$$

Here, the approximate Jacobian matrix $\left[\triangle f_{i,k}/\triangle x_{j,k}\right]_{\substack{1\le i\le 2\\1\le j\le 2}}$ is

$$\left[\begin{array}{ll}\dfrac{(x_k-y_{k-1}+1)-(x_{k-1}-y_{k-1}+1)}{x_k-x_{k-1}} & \dfrac{(x_{k-1}-y_k+1)-(x_{k-1}-y_{k-1}+1)}{y_k-y_{k-1}}\\[2ex]\dfrac{(x_k^3-y_{k-1}^2+1)-(x_{k-1}^3-y_{k-1}^2+1)}{x_k-x_{k-1}} & \dfrac{(x_{k-1}^3-y_k^2+1)-(x_{k-1}^3-y_{k-1}^2+1)}{y_k-y_{k-1}}\end{array}\right]$$

$$=\left[\begin{array}{ll}\dfrac{x_k-x_{k-1}}{x_k-x_{k-1}} & \dfrac{-y_k+y_{k-1}}{y_k-y_{k-1}}\\[2ex]\dfrac{x_k^3-x_{k-1}^3}{x_k-x_{k-1}} & \dfrac{-y_k^2+y_{k-1}^2}{y_k-y_{k-1}}\end{array}\right]$$

$$=\left[\begin{array}{ll}1 & -1\\x_k^2+x_{k-1}x_k+x_{k-1}^2 & -y_k+y_{k-1}\end{array}\right].$$

Instead of quadratic convergence, Table 8.11 appears to have super linear convergence. Here, for some constant c,

$$\|\triangle x_{k+1}\|\le c\,\|\triangle x_k\|^{\alpha},\tag{8.24}$$

for all $k\ge 5$, where it appears that the convergence exponent is $\alpha\approx 1.2$. ◯

TABLE 8.11

Secant Method's Solution of Example 8.8

k	x_k	y_k	$f(x_k,y_k)$	$g(x_k,y_k)$	$\|(\triangle x_k,\triangle y_k)\|$
0	$-6.0000000E-01$	$5.0000000E-01$	$-1.0000000E-01$	$5.3400000E-01$	
1	$-3.0000000E-01$	$2.5000000E-01$	$4.5000000E-01$	$9.1050000E-01$	
2	$-5.4135115E+00$	$-3.4885496E+00$	$-9.2496183E-01$	$-1.6981892E+02$	$6.3344102E+00$
3	$-3.4026652E+00$	$-1.8026881E+00$	$-5.9997710E-01$	$-4.1646186E+01$	$2.6240487E+00$
4	$-2.8546480E+00$	$-1.5871579E+00$	$-2.6749008E-01$	$-2.4781639E+01$	$5.8887703E-01$
5	$-2.4086176E+00$	$-1.3081212E+00$	$-1.0049634E-01$	$-1.4684628E+01$	$5.2612219E-01$
6	$-2.0483599E+00$	$-1.0148112E+00$	$-3.3548794E-02$	$-8.6243062E+00$	$4.6456040E-01$
7	$-1.7601519E+00$	$-7.5007227E-01$	$-1.0079643E-02$	$-5.0157962E+00$	$3.9134454E-01$
8	$-1.5325986E+00$	$-5.2990567E-01$	$-2.6929063E-03$	$-2.8806570E+00$	$3.1662889E-01$
9	$-1.3568335E+00$	$-3.5622574E-01$	$-6.0777948E-04$	$-1.6248235E+00$	$2.4709931E-01$
10	$-1.2256822E+00$	$-2.2558681E-01$	$-9.5377063E-05$	$-8.9222788E-01$	$1.8511402E-01$
11	$-1.1327300E+00$	$-1.3273356E-01$	$3.5517105E-06$	$-4.7099831E-01$	$1.3138431E-01$
12	$-1.0714720E+00$	$-7.1483673E-02$	$1.1668733E-05$	$-2.3521577E-01$	$8.6626155E-02$
...	...	...	...	...	...
30	$-1.0000000E+00$	$9.4766651E-11$	$-4.6629367E-15$	$2.8428626E-10$	$5.4491325E-10$
31	$-1.0000000E+00$	$8.8499877E-12$	$1.9984014E-15$	$2.6556091E-11$	$1.2149988E-10$
32	$-1.0000000E+00$	$-4.1666749E-12$	$1.7763568E-15$	$-1.2494672E-11$	$1.8408594E-11$

Learn More About It

The version of Newton–Kantorovich Theorem 8.4 is found on pp. 301–302 of *Elements of Functional Analysis*, by L. A. Lusternik and V. J. Sobolev, Hindustan Publishing Co., Halsted Press/John Wiley & Sons, Inc., © 1974. One can learn about the Fréchet derivative in that book, as well as Section 7.2 of *Elements of Applicable Functional Analysis*, by Charles W. Groetsch, Marcel Dekker, Inc., © 1980. The version of "another" Newton–Kantorovich Theorem 8.5 is found in Section 1.4 of *Numerical Methods for Two-Point Boundary-Value Problems*, by Herbert B. Keller, Blaisdell Publishing Co., © 1968 or Dover Publications © 1992.

8.2.5 Problems

1. For the system of equations

$$\left\{\begin{array}{l} 0 = xy^2 - y + 1 \\ 0 = x^2y + x - 0.5 \end{array}\right\},$$

(a) Write down Newton's method explicitly giving $\mathbf{x}_{k+1}$ in terms of $\mathbf{x}_k$.

(b) Using part (a) and initial guess $\mathbf{x}_0 = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}$, show work by hand to find $\mathbf{x}_1$.

In problems 2–4 use Newton's method to find two different approximate solutions of the given system of equations.

2. $\left\{\begin{array}{l} 0 = x + 2y + 4 \\ 0 = x^2 - 2x + y^2 - 24 \end{array}\right\}$
3. $\left\{\begin{array}{l} 0 = x + 2y - 9 \\ 0 = x^2 + y^2 - 4y - 21 \end{array}\right\}$
4. $\left\{\begin{array}{l} 0 = x + 2y - 6 \\ 0 = x^2 + 2x + y^2 - 2y - 23 \end{array}\right\}$

In problems 5–7 use Newton's method to find three different approximate solutions of the given system of equations.

5. $\left\{\begin{array}{l} 0 = xy^2 - 8x - y^2 + 8 \\ 0 = 2xy - x^2y \end{array}\right\}$
6. $\left\{\begin{array}{l} 0 = 2x + y - \pi \\ 0 = \sin 2x \cos y - \cos y \end{array}\right\}$
7. $\left\{\begin{array}{l} 0 = 4x - 6y - \pi \\ 0 = \cos x(1 - 2\sin y) \end{array}\right\}$
8. Use Newton's method to find an approximate solution of $\left\{\begin{array}{l} 0 = e^x + y^2 - 3y - 5 \\ 0 = x + e^{-y} + 1 \end{array}\right\}$.

In problems 9–12 use the secant method to find two approximate solutions, if possible, of the given system of equations.

9. The system of Problem 8.2.5.1.
10. The system of Problem 8.2.5.2.
11. The system of Problem 8.2.5.4.
12. The system of Problem 8.2.5.5.

8.3 Approximation of Integrals

We use numerical methods for approximating the value of a definite integral because, in general, we cannot write derive the exact value of the integral.

Suppose $f(x)$ is a function of x on a finite interval $[a, b]$. Recall that the definite integral of f on $[a, b]$ can be defined by the limit of Riemann sums, if the limit exists: If $a = x_0 < x_1 < \cdots < x_n = b$ is a partition of $[a, b]$, x_k^* are sampling points satisfying $x_{k-1} \le x_k^* \le x_k$ for $k = 1, 2, \ldots, n$, and each $\triangle x_k = x_k - x_{k-1}$ for $k = 1, 2, \ldots, n$, then

$$\sum_{k=1}^{n} f(x_k^*)\ \triangle x_k$$

is a **Riemann sum**. For an equally spaced partition,

$$h = \triangle x \triangleq \frac{b-a}{n}$$

is called the **step size**, and $x_k = a + kh$ for $k = 0, 1, \ldots, n$.

The **midpoint** approximation, M_n, is given by

$$\int_a^b f(x)\,dx \approx M_n \triangleq h \sum_{k=1}^{n} f(x_{k-0.5}), \tag{8.25}$$

where the midpoints are

$$x_{k-0.5} = x_k - 0.5h = 0.5(x_{k-1} + x_k), \text{ for } k = 1, \ldots, n.$$

Given a function f, define

$$f_k \triangleq f(x_k), \text{ for } k = 0, \ldots, n. \tag{8.26}$$

Another famous approximation you learned about in calculus is the **trapezoidal rule**,

$$\int_a^b f(x)\,dx \approx T_n \triangleq \frac{h}{2} \sum_{k=1}^{n} \left(f_{k-1} + f_k\right), \tag{8.27}$$

as illustrated in Figure 8.4. Also, if $n = 2N$ is even, then the famous **Simpson's rule** is given by

$$\int_a^b f(x)\,dx \approx S_{2N} \triangleq \frac{h}{3} \sum_{k=1}^{N} \left(f_{2k-2} + 4f_{2k-1} + f_{2k}\right). \tag{8.28}$$

In the latter, it is still true that the step size is

$$h \triangleq \frac{b-a}{n} = \frac{b-a}{2N}.$$

Any formula that gives an approximation of a definite integral is called a **quadrature rule**.

In fact there are many quadrature rules. For example, if $n = 4N$, then the **Newton–Cotes formula of closed type** of order four is given by

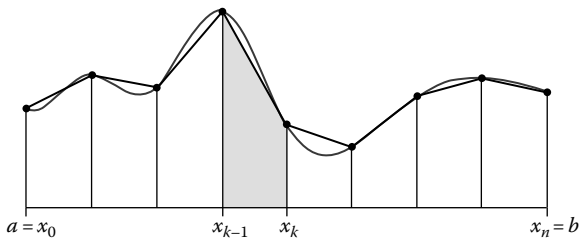


FIGURE 8.4
Approximation of an integral using areas of trapezoids.

$$\int_a^b f(x)\,dx \approx U_{4N} \triangleq \frac{2h}{45}\sum_{k=1}^{N}\left(7f_{4k-4}+32f_{4k-3}+12f_{4k-2}+32f_{4k-1}+7f_{4k}\right). \tag{8.29}$$

This is also known as **Boole's rule** (Weisstein; Boole and Moulton, 1960).

For small n, we may avoid the $\sum_{k=1}$ notation by noting that

$$\begin{aligned}
T_n &= \frac{h}{2}\left(f_0+2f_1+2f_2+\cdots+2f_{n-1}+f_n,\right)\\
S_n &= \frac{h}{3}\left(f_0+4f_1+2f_2+4f_3+2f_4+4f_5+2f_6+\cdots+4f_{n-2}+2f_{n-1}+f_n,\right)
\end{aligned}$$

and similarly for Boole's rule U_{4N}. For $n=8$, we have, in particular,

$$\begin{aligned}
T_8 &= \frac{h}{2}\left(f_0+2f_1+2f_2+\cdots+2f_7+f_8,\right)\\
S_8 &= \frac{h}{3}\left(f_0+4f_1+2f_2+4f_3+2f_4+4f_5+2f_6+4f_7+f_8,\right)\\
U_8 &= \frac{2h}{45}\left(7f_0+32f_1+12f_2+32f_3+14f_4+32f_5+12f_6+32f_7+7f_8\right).
\end{aligned} \tag{8.30}$$

Example 8.9

Use the midpoint, trapezoidal, Simpson's, and Boole's rules with (a) four, and (b) eight subintervals to approximate the integral

$$\int_{-\pi/2}^{\pi} \frac{\sin 2x}{3-e^{-|x|}}\,dx.$$

Method: (a) For $n=4$, $h=\frac{\pi-(-0.5\pi)}{4}=\frac{3\pi}{8}$, and the partition points are $x_n=a+nh=-0.5\pi+0.375\,n\pi$, and we calculated

$$\begin{aligned}
f_0 &= f(-0.5\pi)=\frac{0}{3-e^{-\pi/2}}=0,\\
f_1 &= f(-.125\pi)=\frac{-1/\sqrt{2}}{3-e^{-\pi/8}}\approx -0.3041622875,\\
f_2 &= f(.25\pi)=\frac{1}{3-e^{-\pi/4}}\approx 0.3930722012,\\
f_3 &= f(.625\pi)=\frac{-1/\sqrt{2}}{3-e^{-5\pi/8}}\approx -0.2472718569,\\
f_4 &= f(\pi)=\frac{0}{3-e^{-\pi}}=0.
\end{aligned}$$

The results we calculated for both $n=$ and $n=8$ are in Table 8.12. ◯

TABLE 8.12

Example 8.9

n	Midpoint Rule	Trapezoidal Rule	Simpson's Rule	Boole's Rule
4	$-4.1596899e-01$	$-1.86565776e-01$	$-5.5747254e-01$	$-6.7696223e-01$
8	$-3.6896831e-01$	$-3.0126738e-01$	$-3.39501251e-01$	$-3.2496983e-01$
16	$-3.4961965e-01$	$-3.3511785e-01$	$-3.4640133e-01$	$-3.4686134e-01$

By the way, we could have gotten more accurate results by noting that $f(x) = \frac{\sin 2x}{3-e^{-|x|}}$ is an odd function of x, so $\int_{-\pi/2}^{\pi/2} \frac{\sin 2x}{3-e^{-|x|}}\, dx = 0$. Thus,

$$\int_{-\pi/2}^{\pi} \frac{\sin 2x}{3-e^{-|x|}} dx = \int_{\pi/2}^{\pi} \frac{\sin 2x}{3-e^{-|x|}} dx.$$

8.3.1 Approximation Error for Quadrature Rules

One approach to using numerical methods is to use them as a black box based on their reputations for success. Another approach to using numerical methods, called **numerical analysis**, is to study their accuracy and prove mathematical theorems that guarantee their success as long as the hypotheses of the theorems are verified. The subject of numerical analysis can be as rigorous as any other mathematical subject and may involve deep techniques and results of pure mathematics.

We will use numerical methods while following a middle path, as we did in Section 8.1: We will sometimes mention a mathematical theorem that guarantees a method's success under verified hypotheses as a way to better know when a numerical method is appropriate; however, in examples we will usually not rigorously verify the hypotheses, and we will not prove the mathematical theorems.

Here we will give estimates on the errors. In the results as follows, we will use the notation

$$k_\ell \triangleq \max_{a \le x \le b} |f^{(\ell)}(x)| \tag{8.31}$$

for a function differentiable ℓ times on the interval $[a, b]$.

Theorem 8.6

Suppose f is ℓ times continuously differentiable on an interval (a, b) and continuous on $[a, b]$. Then bounds on the absolute value of the error in evaluating a definite integral are given by

(a) If $\ell \ge 2$, then $\left|M_n - \int_a^b f(x)\, dx\right| \le \frac{(b-a)}{24} k_2 h^2$.

(b) If $\ell \ge 2$, then $\left|T_n - \int_a^b f(x)\, dx\right| \le \frac{(b-a)}{12} k_2 h^2$.

(c) If $\ell \geq 4$, then $\left|S_n - \int_a^b f(x)\,dx\right| \leq \frac{(b-a)}{180}\,k_4\,h^4$.

(d) If $\ell \geq 6$, then $\left|U_n - \int_a^b f(x)\,dx\right| \leq \frac{8(b-a)}{945}\,k_6\,h^6$.

Because, for example, the second derivative of a first-degree polynomial is zero everywhere, a consequence of Theorem 8.6 is that

> The midpoint and trapezoidal rules are exact for all first-degree polynomials, Simpson's rule is exact for all third-degree polynomials, and Boole's rule is exact for all fifth-degree polynomials.

Of course, we can easily calculate an integral of a low-degree polynomial using the fundamental theorem of calculus. But the theorem and the remark after it give us some indication of just how good the quadrature rules are.

Example 8.10

Find a value of n so that Simpson's rule estimates the integral

$$\int_{-\pi}^{\pi} \frac{\cos 2x}{3 - e^{-x^2}}\,dx$$

with an error that is guaranteed to be less than 10^{-10}. For that value of n, find the Simpson's rule approximation of that integral.

Method: Using *Mathematica*TM, for the integrand $f(x) = \frac{\cos 2x}{3-e^{-x^2}}$ we graphed the function $|f''(x)|$ versus x in the interval $[-\pi, \pi]$ in order to estimate $k_4 \leq 26$. We confirmed the value using the *Mathematica* command FindMaximum[{Abs[f''''[x]], $-\pi \leq x \leq \pi$}, {x, 0}].

We want to choose n so that

$$\left|S_n - \int_{-\pi}^{\pi} \frac{\cos 2x}{3 - e^{-x^2}}\,dx\right| \leq \frac{(b-a)}{180}\,k_4\,h^4 < 10^{-10},$$

that is,

$$\frac{\pi - (-\pi)}{180} \cdot 26\,h^4 < 10^{-10}.$$

This gives $h^4 < \frac{180}{52\pi} \cdot 10^{-10}$; hence,

$$\frac{2\pi}{n} = \frac{b-a}{n} = h <\approx 0.0032398867,$$

that is,

$$n >\approx \frac{2\pi}{0.0032398867} \approx 1939.322545.$$

Since n must be even in order to use Simpson's rule, we choose $n = 1940$ or larger in order to guarantee that the error would be less than 10^{-10}. We have

$$\int_{-\pi}^{\pi} \frac{\cos 2x}{3 - e^{-x^2}}\,dx \approx S_{1940} \approx 0.113795689531577. \bigcirc$$

8.3.2 Iteration of Quadrature

In practice, we may not be able to apply an error estimate found in Theorem 8.6, either because the function values $f(x)$ are not generated by a formula but instead by a software subroutine or because the constant k_ℓ does not exist or is difficult to estimate well.

A natural approach is to pick a quadrature rule and use it for an increasing sequence of n, that is, for a finer and finer partition of the interval $[a, b]$, until the difference between successive approximations of the integral becomes less than a chosen error ε. The smallness of the change in the quadrature values is a stopping rule for the iterative process.

The following is our MATLAB "function" fSimpExtrap that iterates Simpson's rule for the function of Example 8.9; the MATLAB "function" f821 was used by fSimpExtrap because in the command window, we defined fn ='f821' and entered values for the interval endpoints a, b and the error parameter eps. In the command window, we entered

```
a = −pi/2; b = pi; eps = 10^(−10); fn ='f821'; format long e;
fSimpExtrap(a, b, fn, eps)
```

and got output

```
a = −1.570796326794897e + 00
b = 3.141592653589793e + 00
n = 8192
h = 5.752427954571154e − 04
ans = −3.449284883498875e − 01.
```

With eps $= 10^{-10}$, the stopping rule was triggered when $n = 8192$ and we estimated the definite integral to be $-3.449284883498875e - 01$. Our program used "Richardson extrapolation to the limit." This integrand is not continuously differentiable on its interval of integration, so the error guarantees of Theorem 8.6 are useless. The MATLAB program fSimpExtrap and the integrand function evaluation f821 are

```
function fSimpExtrap = fSimpExtrap(a, b, fn, eps)
a, b
format long e; sumSimp = 0; sumSimpOld = 0; delta = 999; k = 0;
while delta >= eps
  k = k + 1;
  n = 2 * k; h = (b − a)/n;
  sum = 0;
  for m = 1 : k;
    x0 = a + (m − 1) * 2 * h; x1 = x0 + h; x2 = x1 + h;
    sum = sum + feval(fn, x0) + 4 * feval(fn, x1) + feval(fn, x2);
  end
```

```
    sumSimp1 = sum * h/3;
    sumSimpOld = sumSimp;
    sumSimp = sumSimp1;
    delta = abs(sumSimpOld - sumSimp);
  end
  n, h
  % Uses Richardson extrapolation to the limit
  fSimpExtrap = (4 * sumSimp - sumSimpOld)/3;
```

and

```
function f821 = f821(x)
% Function used in Example 8.9
format long e;
f821 = sin(2 * x)/(3 - exp(-abs(x))).
```

8.3.3 Theory versus Practice

We also ran our MATLAB program fSimpExtrap for the definite integral of Example 8.10 and the stopping rule was triggered at $n = 256$, giving an approximate integral value of $I \approx 0.1137956895441768$. We've underlined the 10 digits in fSimpExtrap that agree with the conclusion of Example 8.10. We first defined a function by f820 = cos(2 * x)/(3 − exp(−x^2)).

The integral's approximate value S_{1940}, in Example 8.10, is guaranteed by mathematical theory to be correct to within 10^{-10} of the correct value. It took significantly more time to compute S_{1940} than to calculate fSimpExtrap with $n = 256.0$

The theory was overly cautious in requiring $n = 1940$, while iteration gave a good result with only $n = 256$. The difference between theory and practice is a "two-edged sword": Following the demands of theory may give software too slow to be useful, while ignoring the demands of theory may lead to software that sometimes produces misleading conclusions.

One problem with numerical quadrature is **numerical drift** in which reducing the step size h eventually becomes counter productive due to increased round-off error.

Learn More About It

The *Handbook of Mathematical Functions*, edited by Milton Abramowitz and Irene Stegun, Dover Publications © 1964, contains many quadrature rules, including Newton–Cotes formulas of both closed, and open-type and Gaussian quadrature. (By the way, in that book, Boole's rule is mistakenly referred to as Bode's rule.) *Elementary Numerical Analysis: An Algorithmic Approach*, 2*nd* ed., by S. D. Conte and Carl de Boor, McGraw-Hill, Inc. © 1972, discusses "Richardson extrapolation to the limit."

8.3.4 Problems

In problems 1–3 use Simpson's rule with $n = 8$ to approximate the length of the parametrized curve.

1. $\mathcal{C} : \mathbf{r} = \cos 2t\,\hat{\boldsymbol{i}} + \sin t\,\hat{\boldsymbol{j}} + t\,\hat{\boldsymbol{k}}, 0 \le t \le 2\pi$
2. The Zhukovskii airfoil given in (6.30), that is,

$$x(\xi,\eta) \triangleq \xi \cdot \frac{1+\xi^2+\eta^2}{2(\xi^2+\eta^2)}, y(\xi,\eta) \triangleq \eta \cdot \frac{-1+\xi^2+\eta^2}{2(\xi^2+\eta^2)},$$

   where

$$\xi(t) \triangleq -0.07 + \rho_0 \cos(t), \quad \eta(t) \triangleq 0.05 + \rho_0 \sin(t),$$

   and $\rho_0 = \sqrt{(1+0.07)^2 + (0.05)^2}$.
3. As in (8.30), write down the trapezoidal, Simpson's, and Boole's rules for $n = 12$.
4. Use the trapezoidal rule T_4 to approximate $\int_0^{\pi} \sqrt{2 - \sin^2 x}\,dx$, and give an upper bound on the error in that approximation.
5. Let $I = \int_{-\pi/4}^{\pi/3} \tan\theta\,d\theta$.
   (a) Approximate I using trapezoidal rule T_4.
   (b) Give an upper bound on the error in that approximation.
   (c) Use $\int \tan\theta\,d\theta = -\ln|\cos\theta|$ to find the exact value of I.
   (d) Show verification that your results in parts (a) and (c) agree with your result in part (b).
6. Suppose the maximum value of $|f''(x)|$ on the interval $[1,2]$ is 10 and the maximum value of $|f''''(x)|$ on the interval $[1,2]$ is 60. Find the smallest values, respectively, of n for which (a) the trapezoidal rule and (b) Simpson's rule approximates $\int_1^2 f(x)dx$ to within a guaranteed accuracy of 10^{-3}.
7. Use (a) Simpson's rule and (b) Boole' rule and as much of the vehicle speed data as possible in Table 8.13 to estimate the change in position $x(t)$ of the vehicle over the time interval $0 \le t \le 4$ s.
8. Suppose that a CAT scan of a 12-cm-long human kidney shows cross sections 1 cm apart. The cross-sectional areas, in cm^2, are 0, 4, 5, 6, 7, 7, 8, 7, 6, 6, 5, 4, 0. Use the trapezoidal rule, the Simpson's rule, and the boole's rule to approximate the volume of the kidney.

TABLE 8.13

Vehicle Speed for Problem 8.3.4.7

Time (10^{-1} s)	0	5	10	15	20	25	30	35	40
Speed (m/s)	1	4	8	13	16	20	24	27	30

8.4 Numerical Solution of $A\mathrm{x} = \mathrm{b}$

In Section 1.1 we learned a theoretical method for finding all of the solutions of $A\mathbf{x} = \mathbf{b}$, a system of m linear algebraic equations in n unknowns. This method usually works well for small systems, that is, where m and n are "small."

But how "small" is "small?" And how "well" is "well," and how often is "usually?" It turns out that even for small systems, there can be serious inaccuracies in the numerical values of the solution produced by the Gauss–Jordan or the Gaussian elimination algorithm, even if $m = n$.

In addition, the use of numerical solutions requires careful consideration of two main issues, "convergence" and "stability." We will not establish the technical definitions of these two concepts but instead discuss them informally. A numerical method is **convergent** if a limiting process results in what is a correct solution. A numerical method is **stable** if a small change in the parameters of the problem leads to a small change in the solution. In order to address the concept of stability, we must have a way of measuring both the input parameters and the solution.

From Sections 2.3 and 8.2, recall the concept of **norm**, and recall the concept of **matrix norm induced** by the vector norm $\| \ \|$ on $\mathbb{R}^n$. Recall that $\|A\|$, the norm on a matrix $\|A\|$, satisfies

$$\|A\mathbf{x}\| \leq \|A\| \ \|\mathbf{x}\|, \text{for } \textit{all } \mathbf{x} \text{ in } \mathbb{R}^n.$$

Using the concept of induced matrix norm, we can define a concept that is commonly used to study square matrices and numerical methods that use them.

Definition 8.2

Suppose $\| \ \|$ is a norm on $\mathbb{R}^n$ and A is a real, $n \times n$ matrix. We define the **standard condition number** of A by

$$\kappa(A) \triangleq \|A\| \ \|A^{-1}\|. \tag{8.32}$$

By the way, we saw $\kappa(C)$ implicitly in the context of fixed-point iteration in Section 8.2.

Let's understand how the standard condition number measures the stability of solving the linear algebraic system $A\mathbf{x} = \mathbf{b}$. Numerical computer software packages store numbers with a finite number of digits of precision, so there is always the danger of round-off error. In particular, when solving a system $A\mathbf{x} = \mathbf{b}$, the vector $\mathbf{b}$ and the matrix A stored in computer memory may differ from their measured values. In addition, their measured values may differ from their true values because all measurement processes are subject to error.

Assume that A is invertible and $\mathbf{b} \neq \mathbf{0}$. Denote by $\mathbf{x}$ the exact solution of the true system $A\mathbf{x} = \mathbf{b}$. Because A is linear, $\mathbf{x} \neq \mathbf{0}$. The standard condition number takes into account the error in the right-hand side of the system: We will assume that the system solved by computer software is

$$A\mathbf{X} = \mathbf{b} + \boldsymbol{\delta}, \tag{8.33}$$

where $\boldsymbol{\delta}$ is the error in the vector $\mathbf{b}$. Let's define $\boldsymbol{\xi} \triangleq \mathbf{X} - \mathbf{x}$, the difference between the solution of (8.33) and the exact solution of the true system.

The **absolute error** in approximating $\mathbf{b}$ by $(\mathbf{b} + \boldsymbol{\delta})$ is $\|\boldsymbol{\delta}\|$ and the corresponding **relative error** is $\|(\mathbf{b} + \boldsymbol{\delta}) - \mathbf{b}\|/\|\mathbf{b}\|$, that is,

$$\frac{\|\boldsymbol{\delta}\|}{\|\mathbf{b}\|}.$$

Similarly, the relative error in the solution is $\|\mathbf{X} - \mathbf{x}\|/\|\mathbf{x}\|$, that is,

$$\frac{\|\boldsymbol{\xi}\|}{\|\mathbf{x}\|}.$$

Using $A\mathbf{x} = \mathbf{b}$ and (8.33), we get

$$\mathbf{b} + \boldsymbol{\delta} = A\mathbf{X} = A(\mathbf{x} + \boldsymbol{\xi}) = A\mathbf{x} + A\boldsymbol{\xi} = \mathbf{b} + A\boldsymbol{\xi}; \tag{8.34}$$

hence,

$$\boldsymbol{\delta} = A\boldsymbol{\xi}.$$

So, the relative error in the solution is

$$\frac{\|\boldsymbol{\xi}\|}{\|\mathbf{x}\|} = \frac{\|\boldsymbol{\xi}\|}{\|\boldsymbol{\delta}\|} \cdot \frac{\|\boldsymbol{\delta}\|}{\|\mathbf{b}\|} \cdot \frac{\|\mathbf{b}\|}{\|\mathbf{x}\|}. \tag{8.35}$$

By remark (1) following Definition 8.1 in Section 8.2, $\|\mathbf{b}\| = \|A\mathbf{x}\| \leq \|A\|\,\|\mathbf{x}\|$. Similarly, because $A^{-1}\boldsymbol{\delta} = \boldsymbol{\xi}$, we have $\|\boldsymbol{\xi}\| = \|A^{-1}\boldsymbol{\delta}\| \leq \|A^{-1}\|\,\|\boldsymbol{\delta}\|$. So, (8.35) implies

$$\frac{\|\boldsymbol{\xi}\|}{\|\mathbf{x}\|} \leq \frac{\|A^{-1}\|\,\|\boldsymbol{\delta}\|}{\|\boldsymbol{\delta}\|} \cdot \frac{\|\boldsymbol{\delta}\|}{\|\mathbf{b}\|} \cdot \frac{\|A\|\,\|\mathbf{x}\|}{\|\mathbf{x}\|} = \|A^{-1}\|\,\|A\|\frac{\|\boldsymbol{\delta}\|}{\|\mathbf{b}\|} = \kappa(A)\frac{\|\boldsymbol{\delta}\|}{\|\mathbf{b}\|}.$$

So, the relative error in the solution is the relative error in the right-hand side of the system times a magnification factor of $\kappa(A)$.

In Problem 8.4.3.11 you will use the first remark after Definition 8.1 in Section 8.2 to explain why the matrix norm has the property that

$$\|AB\| \leq \|A\|\,\|B\|, \tag{8.36}$$

as long as the multiplication AB makes sense.

Using (8.36), it follows that

$$1 = \|I\| = \|A^{-1}A\| \leq \|A^{-1}\|\,\|A\| = \kappa(A).$$

When a matrix has the smallest possible standard condition number, that is, $\kappa(A) = 1$, we say the matrix is **perfectly conditioned**. In Problem 8.4.3.9 you will explain why orthogonal matrices are perfectly conditioned. In fact, if $Q = \left[\mathbf{q}_1 \mid \cdots \mid \mathbf{q}_n\right]$ is an orthogonal matrix, then Corollary 2.8 in Section 2.4 implies that the system $Q\mathbf{x} = \mathbf{b}$ has unique solution $\mathbf{x} = \langle \mathbf{x}, \mathbf{q}_1\rangle \mathbf{e}^1 + \cdots + \langle \mathbf{x}, \mathbf{q}_n\rangle \mathbf{e}^n$.

It can be difficult to calculate $\|A\|$ using Definition 8.1 in Section 8.2. Instead, we can recall from Section 2.10 that the Frobenius norm on matrices, defined by

$$\|A\|_F \triangleq \sqrt{\sum_{j=1}^{n}\sum_{k=1}^{n} |a_{jk}|^2},$$

satisfies $\|A\| \leq \|A\|_F$, by Theorem 2.47 in Section 2.10. So, if $\|A^{-1}\|_F\|A\|_F$ is not much bigger than one, then we know that the square matrix A is well conditioned.

8.4.1 Partial Pivoting in the Gauss–Jordan Algorithm

Recall that in the Gaussian elimination method for solving a linear system $A\mathbf{x} = \mathbf{b}$, we systematically row reduce the augmented matrix $[A|\mathbf{b}]$ to a row echelon form. In the process of doing so, we create leading entries in the pivot positions in all of the nonzero rows of a row echelon form. The Gauss–Jordan elimination algorithm goes further and row reduces the augmented matrix to RREF. Both the Gauss–Jordan and Gaussian elimination algorithms are called **direct methods** for finding the solution $\mathbf{x} = [\,x_1 \quad \ldots \quad x_n\,]^T$ because they involve reading $x_1, \ldots, x_n$ directly, or almost directly in the case of Gaussian elimination, from a row echelon form of the augmented matrix. Unfortunately, these methods do not read the solution directly from A or b.

The Gauss–Jordan and Gaussian elimination algorithms can fail to numerically converge because of round-off error accumulation. It was also discovered that, in practice, the technique of *partial pivoting* usually makes the algorithm convergent, that is, allows the algorithm to give accurate results for the solution of a system $A\mathbf{x} = \mathbf{b}$.

First, let's see where trouble lies. For clarity we will use the equals sign to denote true, that is, exact values, and the approximation sign to denote numbers rounded off to four digits of accuracy.

Example 8.11

Use the Gauss–Jordan algorithm to solve the system of equations

$$\left\{\begin{array}{r} 1.000 \times 10^{-5}x + 1.000y = 1.000 \\ 1.000x + 1.000y = 2.000 \end{array}\right\}. \tag{8.37}$$

Assume that numbers can only be stored with four digits of accuracy. Compare your result with the true solution.

Method: The Gauss–Jordan algorithm begins with

$$\left[\begin{array}{ccc} 1.000 \times 10^{-5} & 1.000 & |1.000 \\ 1.000 & 1.000 & |2.000 \end{array}\right]$$

$$\begin{array}{r} \sim \\ 10^5 R_1 \to R_1 \\ -R_1 + R_2 \to R_2 \end{array} \left[\begin{array}{ccc} 1.000 & 1.000 \times 10^5 & | \quad 1.000 \times 10^5 \\ 0.000 & -1.000 \times 10^5 & | \quad -1.000 \times 10^5 \end{array}\right].$$

The (2, 2) entry in the latter matrix is -1.000×10^5 because $1.000 - 100{,}000 \approx -100{,}000$, rounded off to four digits of precision. The (2, 3) entry in the latter matrix is -1.000×10^{-5} because $2.000 - 100{,}000 \approx -100{,}000$, rounded off to four digits of precision.

Continuing further with the Gauss–Jordan algorithm gives

$$\begin{array}{r} \sim \\ -10^{-5} R_2 \to R_2 \end{array} \left[\begin{array}{ccc} 1.000 & 1.000 \times 10^5 & | \quad 1.000 \times 10^5 \\ 0.000 & 1.000 & | \quad 1.000 \end{array}\right]$$

$$\begin{array}{r} \sim \\ -10^5 R_2 + R_1 \to R_1 \end{array} \left[\begin{array}{ccc} 1.000 & 0.000 & | \quad 0.000 \\ 0.000 & 1.000 & | \quad 1.000 \end{array}\right].$$

So, in four-digit arithmetic, the solution to the system seems to be $x \approx 0.000$, $y \approx 1.000$.

To find the true solution, we can use the inverse of a 2×2 matrix to get

$$\begin{aligned} \left[\begin{array}{c} x \\ y \end{array}\right] &= \left[\begin{array}{cc} 1.000 \times 10^{-5} & 1.000 \\ 1.000 & 1.000 \end{array}\right]^{-1} \left[\begin{array}{c} 1.000 \\ 2.000 \end{array}\right] \\ &= \frac{1}{1.000 \times 10^{-5} - 1.000} \left[\begin{array}{cc} 1.000 & -1.000 \\ -1.000 & 1.000 \times 10^{-5} \end{array}\right] \left[\begin{array}{c} 1.000 \\ 2.000 \end{array}\right] = \frac{1}{0.99999} \left[\begin{array}{c} 1 \\ 0.99998 \end{array}\right] \\ &= \left[\begin{array}{c} 1.0000100001\ldots \\ 0.9999899998\ldots \end{array}\right]. \end{aligned}$$

So, the Gauss–Jordan algorithm gave a seriously inaccurate solution. ○

The matrix $A = \left[\begin{array}{cc} 1.000 \times 10^{-5} & 1.000 \\ 1.000 & 1.000 \end{array}\right]$ in Example 8.11 is well conditioned, because we can use the Frobenius norm to conclude that

$$\kappa(A) = \|A^{-1}\| \, \|A\| \leq \|A^{-1}\|_F \|A\|_F = \frac{3 + 10^{-10}}{\sqrt{1 - 10^{-5}}}.$$

So, merely having a well-conditioned A does not guarantee that a practical implementation of the Gauss–Jordan algorithm will accurately approximate the solution of the system $A\mathbf{x} = \mathbf{b}$.

Partial pivoting says that for the augmented matrix of Example 8.11, it would have been better to first exchange the first row with the second row because the division of

the original first row by 1.000×10^{-5} led to rounding off $2.000 - 10{,}000 \approx -10{,}000$ in four-digit accuracy.

To be precise, **partial pivoting** says to do first a row exchange so that in the first column the pivot position entry is as large, in magnitude, as possible. Continuing after that, partial pivoting is done to have large magnitude numbers in pivot positions in all pivot columns.

Example 8.12

Use the Gauss–Jordan algorithm with partial pivoting to solve the system of equations of Example 8.11, that is, system (8.37). Again, assume that numbers can only be stored with four digits of accuracy. Compare your result with the true solution.

Method: The Gauss–Jordan algorithm with partial pivoting starts with

$$\begin{bmatrix} 1.000 \times 10^{-5} & 1.000 & |1.000 \\ 1.000 & 1.000 & |2.000 \end{bmatrix} \overset{\sim}{R_1 \leftrightarrow R_2} \begin{bmatrix} 1.000 & 1.000 & |2.000 \\ 1.000 \times 10^{-5} & 1.000 & |1.000 \end{bmatrix}$$

$$\overset{\sim}{-10^{-5}R_1 + R_2 \to R_2} \begin{bmatrix} 1.000 & 1.000 & |2.000 \\ 0.000 & 1.000 & |1.000 \end{bmatrix},$$

because the $(2,2)$ entry in the latter matrix is $-1.000 \times 10^{-5} \approx 1.000$, rounded off to four digits of precision.

Continuing further with the Gauss–Jordan algorithm gives

$$\overset{\sim}{-R_2 + R_1 \to R_1} \begin{bmatrix} 1.000 & 0.000 & |1.000 \\ 0.000 & 1.000 & |1.000 \end{bmatrix},$$

so, in four-digit arithmetic, the solution to the system seems to be $x \approx 1.000$, $y \approx 1.000$.

As in Example 8.11, the true solution is $x = 1.0000100001\ldots$, $y = 0.99998999989\ldots$, so the Gauss–Jordan algorithm with partial pivoting algorithm gave a very accurate solution. ○

There is a carnival game called "whack a mole" in which the player uses a mallet to knock down toy rubber moles that pop up from many holes in the playing surface. Each mole is analogous to a difficulty that "pops up" when trying to solve a problem. Knocking down one annoying issue may just lead to another annoying issue popping up.

Partial pivoting is an attempt to suppress all of the problems that can come up during the implementation of row reduction on a computer that keeps a finite number of digits of accuracy. Unfortunately, even partial pivoting is not enough to guarantee accurate results, as the next example shows.

Example 8.13

Again, assume that numbers can only be stored with four digits of accuracy. Using the Gauss–Jordan algorithm with partial pivoting to solve the system of equations

$$\left\{ \begin{array}{rcl} 1.000x & + & 1.000y = 2.000 \\ 10.00x & + & 1.000 \times 10^6 y = 1.000 \times 10^6 \end{array} \right\} \tag{8.38}$$

gives a solution, $\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0.000 \\ 1.000 \end{bmatrix}$, that is seriously inaccurate compared to the exact solution, $\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1.0000100001\ldots \\ 0.99998999989\ldots \end{bmatrix}$.

This system is a disguised version of that in Example 8.11. Because the $(2,1)$ entry is the largest in magnitude in its column, the Gauss–Jordan algorithm with partial pivoting starts with

$$\begin{bmatrix} 1.000 & 1.000 & |2.000 \\ 10.00 & 1.000 \times 10^6 & |1.000 \times 10^6 \end{bmatrix} \underset{0.1000R_1 \to R_1}{\overset{\sim}{R_1 \leftrightarrow R_2}} \begin{bmatrix} 1.000 & 1.000 \times 10^5 & |1.000 \times 10^5 \\ 1.000 & 1.000 & |2.000 \end{bmatrix}.$$

In Problem 8.4.3.12 you will check the rest of the details, including the serious inaccuracy of the solution. ◯

Example 8.13 illustrates that choosing the pivot for the first column by exchanging to the row whose entry in the first column is of largest magnitude ignores the magnitude of the other entries in that row. The technique of **implicit partial pivoting** first "normalizes" each nonzero row by also storing the row divided by its entry of largest magnitude, so each normalized row's entry of largest magnitude is 1. Then, for example, to create the pivot in the $(1,1)$ position, exchange with the first row the row whose normalized row's entry in the first column is the largest in magnitude.

For the augmented matrix of Example 8.13, the technique of *implicit* partial pivoting sees right through the phony "10.00" entry in the second row and does not do the row interchange $R_1 \leftrightarrow R_2$. In Problem 8.4.3.13 you will check for yourself that the Gauss–Jordan algorithm with implicit partial pivoting gives solution $\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1.000 \\ 1.000 \end{bmatrix}$, which is quite accurate.

8.4.2 Iterative Methods for Solving $A\mathrm{x} = \mathrm{b}$

Recall from Section 8.2 that iterative methods can be used to solve systems of equations, even nonlinear equations. Newton's method or the secant method may solve $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, while other methods may solve a fixed-point problem $\mathbf{x} = \mathbf{g}(\mathbf{x})$.

The simplest application of this to $A\mathbf{x} = \mathbf{b}$, a system of n linear algebraic equations in n unknowns, is to rewrite it as $\mathbf{0} = \mathbf{f}(\mathbf{x}) \triangleq A\mathbf{x} - \mathbf{b}$. Unfortunately, as you will see in Problem 8.4.3.10, the Jacobian matrix of this $\mathbf{f}(\mathbf{x})$ is A, so Newton's method reduces to

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x}_k)\right)^{-1} \mathbf{f}(\mathbf{x}_k) = \mathbf{x}_k - A^{-1}(A\mathbf{x}_k - \mathbf{b}) = \cdots = A^{-1}\mathbf{b}.$$

Theoretically, this gives the exact solution $\mathbf{x}_1 = A^{-1}\mathbf{b}$ no matter what initial guess $\mathbf{x}_0$ is chosen. But, if we could evaluate A^{-1} accurately why would we even bother with Newton's, method?

So, some great mathematicians found other iterative methods for approximating the solution of $A\mathbf{x} = \mathbf{b}$. These methods and their later modifications helped enable accurate numerical solution of large systems of linear algebraic equations, which in turn helped

enable accurate approximate numerical solving of partial differential equations. Not only that, but these iterative methods often make very efficient use of computer memory. And, these methods often can exploit **sparseness**, that is, presence of vastly many zeros, of matrices that are commonly used in applications to partial differential equations.

These iterative methods use the fact that finding the inverse of a diagonal or triangular matrix can be done accurately and quickly despite the issue of round-off error. First, rewrite

$$A = L + D + U,$$

where

$D = diag(a_{11}, \ldots, a_{nn})$ is the matrix of diagonal entries of A
L is the lower triangular matrix of entries taken from $A - D$
U is the upper triangular entries taken from $A - D$

Historically, the first such method was **Jacobi iteration**, which first rewrites $A\mathbf{x} = \mathbf{b}$ as $D\mathbf{x} = -(L + U)\mathbf{x} + \mathbf{b}$ and then rewrites it as the fixed-point problem

$$\mathbf{x} = D^{-1}\left(-(L + U)\mathbf{x} + \mathbf{b}\right),$$

assuming D is invertible, that is, all of the diagonal entries of A are nonzero. The method of successive approximations in Section 8.2 would say to iterate

$$\mathbf{x}_{k+1} = D^{-1}\left(-(L + U)\mathbf{x}_k + \mathbf{b}\right). \tag{8.39}$$

This method converges as long as $\|D^{-1}(L + U)\| < 1$ or as long as the spectral radius* $\rho\left(D^{-1}(L + U)\right) < 1$. It makes sense to use as the initial guess $\mathbf{x}_0 = D^{-1}\mathbf{b}$ that requires only n divisions: $\mathbf{x}_0 = \begin{bmatrix} a_{11}^{-1}b_1 & \cdots & a_{nn}^{-1}b_n \end{bmatrix}^T$.

The Jacobi method is called a **relaxation** method because we can think of the iterates as $\mathbf{x}_{k+1} = D^{-1}\mathbf{b} + (\textit{correction involving } \mathbf{x}_k)$. It's as if we think of the $k + 1$-iterate as computing the solution of $D\mathbf{x} = \mathbf{b}$, a "relaxingly" easier version of $(D+L+U)\mathbf{x} = \mathbf{b}$, but then correcting using the previously computed $\mathbf{x}_k$.

In many practical situations, Jacobi iteration converges but does so slowly.

Gauss–Seidel iteration is an old but good relaxation method that rewrites $(D+L+U)\mathbf{x} = A\mathbf{x} = \mathbf{b}$ as $(D + L)\mathbf{x} = -U\mathbf{x} + \mathbf{b}$ and iterates

$$\mathbf{x}_{k+1} = (L + D)^{-1}\left(-U\mathbf{x}_k + \mathbf{b}\right). \tag{8.40}$$

This method converges as long as $\|(D + L)^{-1}U\| < 1$ or as long as the spectral radius $\rho\left((D + L)^{-1}U\right) < 1$. It is known that in the special case when A is symmetric, that is, $A^T = A$, when it does converge the Gauss–Seidel iteration converges twice as fast as Jacobi iteration. On the other hand, when A is not symmetric, it is possible for Jacobi iteration to converge even though Gauss–Seidel does not!

The final methods we will mention are the **over-relaxation** methods of successive over-relaxation (SOR) and symmetric successive over-relaxation (SSOR).

* $\rho(A)$, the **spectral radius** of a square matrix A, is the maximum of the absolute values of its eigenvalues.

SOR is a modification of Gauss–Seidel iteration to include a **relaxation parameter** ω in the iteration:

$$\mathbf{x}_{k+1} = (\omega L + D)^{-1} (\omega(-U\mathbf{x}_k + \mathbf{b}) + (1-\omega)D\mathbf{x}_k),$$

that is,

$$\mathbf{x}_{k+1} = (\omega L + D)^{-1} (\omega\mathbf{b} + ((1-\omega)D - \omega U)\,\mathbf{x}_k). \tag{8.41}$$

Note that when $\omega = 1$, SOR is the Gauss–Seidel iteration method.

Sophisticated software packages estimate what value of the parameter ω is optimal for a given system, that is, best controls error, by converging quickly to a good approximation of the solution.

SSOR is another modification of Gauss–Seidel iteration to break up each iteration cycle into two half cycles as well as to include a relaxation parameter ω: *assuming A* is symmetric, that is, $U = L^T$,

$$\left\{ \begin{array}{c} \mathbf{x}_{k+0.5} = M_\omega^{-1} (N_\omega \mathbf{x}_k + \omega\mathbf{b}) \\ \mathbf{x}_{k+1} = \left(M_\omega^T\right)^{-1} \left(N_\omega^T \mathbf{x}_{k+0.5} + \omega\mathbf{b}\right) \end{array} \right\}, \tag{8.42}$$

where $M_\omega \triangleq D + \omega U$, $N_\omega = (1-\omega)D - \omega L$.

NOTES: It can be difficult to find the exact value of $\|A\|$, and the Frobenius norm, $\|A\|_F$, only gives an upper bound on $\|A\|$.

There is a good repair for this situation: for two other choices of norm on $\mathbf{x} = [\,x_1 \quad \cdots \quad x_n\,]^T$ in $\mathbb{R}^n$ (or $\mathbb{C}^n$), there is an exact formula for $\|A\|$.

The $\|\ \|_\infty$ norm on $\mathbb{R}^n$ is given by

$$\|\mathbf{x}\|_\infty = \max_{j=1,\ldots,n} |x_j|.$$

The induced matrix norm is defined to be the smallest number M for which it is true that

$$\|A\mathbf{x}\|_\infty \le M\|\mathbf{x}\|_\infty, \text{ for } \textit{all} \text{ unit vectors } \mathbf{x}.$$

[By unit vector $\mathbf{x}$ we mean "satisfying" $\|\mathbf{x}\|_\infty = 1$.]

It follows that for all $\mathbf{x}$ in $\mathbb{R}^n$,

$$\|A\mathbf{x}\|_\infty \le \|A\|_\infty \|\mathbf{x}\|_\infty.$$

Theorem 8.7

If $A = [a_{ij}]_{\substack{1 \le i \le n \\ 1 \le j \le n}}$, then $\|A\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|$, that is, the "maximum row sum of absolute values."

The $\| \ \|_1$ norm on $\mathbb{R}^n$ is given by

$$\|\mathbf{x}\|_1 = \sum_{j=1}^{n} |x_j|.$$

The induced matrix norm is defined to be the smallest number M for which it is true that

$$\|A\mathbf{x}\|_1 \leq M\|\mathbf{x}\|_1, \text{ for } \textit{all} \text{ unit vectors } \mathbf{x}.$$

[By unit vector $\mathbf{x}$ we mean "satisfying" $\|\mathbf{x}\|_1 = 1$.]
It follows that for all $\mathbf{x}$ in $\mathbb{R}^n$,

$$\|A\mathbf{x}\|_1 \leq \|A\|_1\|\mathbf{x}\|_1.$$

Theorem 8.8

If $A = [a_{ij}]_{\substack{1 \leq i \leq n \\ 1 \leq j \leq n}}$, then $\|A\|_1 = \max_j \sum_{i=1}^{n} |a_{ij}|$, that is, the "maximum column sum of absolute values."

Definition 8.3

An $n \times n$ matrix A is

(a) **Row diagonally dominant** if for each $i = 1, \ldots, n$,

$$|a_{ii}| > \sum_{\substack{j \\ j \neq i}} |a_{ij}|$$

(b) **Column diagonally dominant** if for each $i = 1, \ldots, n$,

$$|a_{ii}| > \sum_{\substack{i \\ i \neq j}} |a_{ij}|$$

Theorem 8.9

If A is either row or column diagonally dominant, then Jacobi iteration converges.

Learn More About It

Matrix Computations and Mathematical Software, by John R. Rice, McGraw-Hill, Inc. © 1981, is among many good resources for learning about matrix computations and software issues. Our Example 8.11 is basically the same as his Example 5.2. Rice's book, on pp. 140–141, includes a discussion of three false but commonly held beliefs about matrix iterations.

Also good are *Matrix Computations*, 3*rd* edition, by Gene H. Golub and Charles F. Van Loan, the Johns Hopkins University Press © 1996, and *Numerical Recipes: The Art of Scientific Computing*, 3*rd* edition, by William H. Press et al., Cambridge University Press © 2007.

8.4.3 Problems

In problems 1 and 2 find an upper bound on the condition number of the matrix by using the Frobenius norm.

1. $\begin{bmatrix} 1 & -2 & 3 \\ 2 & 0 & 1 \\ 3 & 1 & 0 \end{bmatrix}$

2. $\begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix}$

   [Aside: This is a 3×3 example of a **Hilbert matrix**. In general, Hilbert matrices give useful examples for testing algorithms for matrices.]

In problems 3 and 4 use the Gauss–Jordan algorithm (a) without partial pivoting, and (b) with partial pivoting, to solve the system of equations. Assume that numbers can only be stored with four digits of accuracy. Compare your result with the true solution.

3. $\left\{ \begin{array}{r} 5.000 \times 10^{-5}x + 1.000y = 1.000 \\ 1.000x + 1.000y = 2.000 \end{array} \right\}$

4. $\left\{ \begin{array}{r} 8.000 \times 10^{-5}x + 1.000y = 1.000 \\ 1.000x + 1.000y = 2.000 \end{array} \right\}$

In problems 5 and 6 use the Gauss–Jordan algorithm with (a) partial pivoting, and (b) implicit partial pivoting to solve the system of equations. Assume that numbers can only be stored with four digits of accuracy. Compare your result with the true solution.

5. $\left\{ \begin{array}{r} 1.000x + 1.000y = 2.000 \\ 5.000x + 1.000 \times 10^5 y = 1.000 \times 10^5 \end{array} \right\}$

6. $\left\{\begin{array}{ll} 1.000x + & 1.000y = 2.000 \\ 10.00x + 1.000 \times 10^5 y = & 1.000 \times 10^5 \end{array}\right\}$

In problems 7 and 8 implement an iterative method of your choice, such as Jacobi, Gauss–Seidel, SOR, or SSOR, to find an approximate solution, *if possible*, of the system of equations. Compare your result with the true solution.

7. The system of Example 8.13.
8. The system of Problem 8.4.3.4.
9. Why are all orthogonal matrices perfectly conditioned?
10. If $\mathbf{f}(\mathbf{x}) \triangleq A\mathbf{x} - \mathbf{b}$, calculate that the Jacobian matrix of $\mathbf{f}(\mathbf{x})$ is A.
11. Use the first remark after Definition 8.1 in Section 8.2 to explain why the matrix norm has the property (8.36), that is, $\|AB\| \leq \|A\| \, \|B\|$, as long as the multiplication AB makes sense: First, use that remark to note that $\|A\mathbf{x}\| \leq \|A\| \, \|\mathbf{x}\|$ for all $\mathbf{x}$. So, for all $\mathbf{x}$,

$$\|AB\mathbf{x}\| = \|A(B\mathbf{x})\| \leq \|A\| \, \|B\mathbf{x}\|.$$

    Continue by using that remark again.
12. Finish Example 8.13 by checking the rest of the details, including the serious inaccuracy of the solution.
13. Check for yourself the remarks in the second paragraph following Example 8.13 that the Gauss–Jordan algorithm with implicit partial pivoting gives solution $\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1.000 \\ 1.000 \end{bmatrix}$ and that the latter is quite accurate.

8.5 Linear Algebraic Eigenvalue Problems

We will see that there are many methods for approximating the eigenvalues and eigenvectors of a matrix. Which method is best depends on the size of the matrix, whether the matrix is real and symmetric, whether we need all of the eigenvalues, and whether we need all of the eigenvectors.

8.5.1 Elementary Method

The simplest approach to the numerical eigenvalue/eigenvector problem is to look on eigenvalues as the roots of the characteristic polynomial $\mathcal{P}(\lambda) \triangleq |A - \lambda I_n|$ and thus try to solve the equation

$$\mathcal{P}(\lambda) = 0$$

by one of the methods discussed in Section 8.1. For example, the Müller root-finding method works well for small matrices, for example, of size 20 or so. One nice thing about

this approach is that it works just as well for the generalized eigenvalue problem $A\mathbf{x} = \lambda B\mathbf{x}$, where $\mathbf{x} \neq \mathbf{0}$, by using characteristic polynomial

$$\mathcal{P}(\lambda) \triangleq |A - \lambda B|.$$

8.5.2 Power Methods

Suppose that we wanted to find the *largest* frequency of vibration of $\ddot{\mathbf{x}} = A\mathbf{x}$, a system of second-order linear constant coefficients ordinary differential equations. Here, A is a real, $n \times n$ matrix whose eigenvalues are all negative.

Suppose A is any $n \times n$ matrix whose eigenvalues in magnitude are $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$. We say A has **dominant eigenvalue** λ_1 if, in addition, $|\lambda_1| > |\lambda_2|$. If A is diagonalizable, then in theory there exists a matrix P such that $A = P\,diag(\lambda_1, \ldots, \lambda_n)P^{-1}$, and thus for any $\mathbf{x}$, we have

$$A^k\mathbf{x} = P\,diag(\lambda_1^k, \ldots, \lambda_n^k)P^{-1}\mathbf{x}.$$

Intuitively, the dominant eigenvalue of A plays a greater and greater role in $A^k\mathbf{x}$ as $k \to \infty$.

Suppose A has a real eigenvalue λ and corresponding real eigenvector $\mathbf{x}$. Then $\|A\mathbf{x}\| = \|\lambda\mathbf{x}\| = |\lambda|\,\|\mathbf{x}\|$, so $|\lambda|$ measures how much A magnifies eigenvector $\mathbf{x}$. Also, if $\mathbf{q}$ is a unit vector, that is, $\|\mathbf{q}\| = 1$, then $\mathbf{q}^T(A\mathbf{q}) = \mathbf{q}^T(\lambda\mathbf{q}) = \lambda\|\mathbf{q}\|^2 = \lambda$. Recall that if $\mathbf{q}$ is a unit vector and A is a real, symmetric matrix, then $\mathbf{q}^TA\mathbf{q}$ is the Rayleigh quotient.

The basic **power method** for finding λ_1 is to iterate the process

$$\left\{\begin{array}{l} \widehat{\mathbf{x}}_k = A\mathbf{x}_{k-1} \\ \text{Choose index } i_k \text{ such that } |\widehat{x}_{i_k,k}| = \max_{1\leq j\leq n} |\widehat{x}_{j,k}| \\ M_k \triangleq \widehat{x}_{i_k,k} \\ \mathbf{x}_k = M_k^{-1}\widehat{\mathbf{x}}_k \end{array}\right\}. \tag{8.43}$$

The first line in the process says that essentially we take powers of the matrix multiplying an initial guess $\mathbf{x}_0$. The other lines say that we normalize the iterate $\mathbf{x}_k$ so that its entry of maximum magnitude is 1. When there is a choice of more than one i_k in the second line, the process can choose i_k to change as little as possible as k increases.

If the iteration process converges, that is, *if* $\mathbf{x}_k \to \mathbf{x}_\infty$ as $k \to \infty$, then the process can probably be chosen so that there is a K and an index i such that $i_k = i$ and $x_{i,k} = 1$ for all $k \geq K$. It follows that $\|\mathbf{x}_k\| \geq 1$ for all $k \geq K$, so $\|\mathbf{x}_\infty\| \geq 1$; hence, $\mathbf{x}_\infty \neq \mathbf{0}$.

For $k \geq K$, let

$$\mu_{k+1} \triangleq \widehat{x}_{i,k+1}/x_{i,k} = \widehat{x}_{i,k+1}/1 = \widehat{x}_{i,k+1} = M_{k+1}. \tag{8.44}$$

We claim that $M_k \to \mu$ as $k \to \infty$ where μ is an eigenvalue of A with corresponding eigenvector $\mathbf{x}_\infty$. Why? (8.44) substituted into the third line of (8.43) implies that

$$M_k\mathbf{x}_{k+1} = \widehat{\mathbf{x}}_{k+1} = A\mathbf{x}_k \to A\mathbf{x}_\infty, \text{ as } k \to \infty. \tag{8.45}$$

But, in the product on the left-hand side, $M_k\mathbf{x}_{k+1}$, we know $\mathbf{x}_{k+1} \to \mathbf{x}_\infty$. It follows that $M_k \to \mu$ for some μ, and thus, (8.45) implies that

$$\mu\mathbf{x}_\infty = A\mathbf{x}_\infty.$$

Because $\mathbf{x}_\infty \neq \mathbf{0}$, it is an eigenvector of A corresponding to eigenvalue μ.

In fact, using a basis for $\mathbf{R}^n$ consisting of $\mathbf{x}_\infty$ and $n-1$ other vectors, we can justify the following result.

Theorem 8.10

If A has dominant eigenvalue λ_1 and the initial guess $\mathbf{x}$ is not in $\mathcal{E}^{\perp}_{\lambda=\lambda_1}$, then the power method converges to the dominant eigenvalue and a corresponding eigenvector. The method has linear convergence in the sense that

$$\|\mathbf{x}_k - \mathbf{x}_\infty\| \leq \alpha \cdot \left|\frac{\lambda_2}{\lambda_1}\right|^k$$

for some constant α.

In practice, the assumption that $\mathbf{x}$ is not in "$\mathcal{E}^{\perp}_{\lambda=\lambda_1}$" is not needed. Why? Because after a few iterations of the power method, the iterates will not be in $\mathcal{E}^{\perp}_{\lambda=\lambda_1}$ because of round-off error. This is one situation where round-off error actually comes to the rescue of a numerical method rather than ruins it!

Example 8.14

Use the power method to find the dominant eigenvalue and a corresponding eigenvector of $A = \begin{bmatrix} -5 & 6 \\ -3 & 4 \end{bmatrix}$.

Method: With initial guess $\mathbf{x}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, we calculate

$$\widehat{\mathbf{x}}_1 = \begin{bmatrix} -5.000000000 \\ -3.000000000 \end{bmatrix} \quad M_1 = -5.000000000 \quad \mathbf{x}_2 = \begin{bmatrix} 1 \\ 0.6000000000 \end{bmatrix},$$

$$\widehat{\mathbf{x}}_2 = \begin{bmatrix} -1.400000000 \\ -0.600000000 \end{bmatrix} \quad M_2 = -1.400000000 \quad \mathbf{x}_2 = \begin{bmatrix} 1 \\ 0.4285714286 \end{bmatrix},$$

$$\widehat{\mathbf{x}}_3 = \begin{bmatrix} -2.428571429 \\ -1.285714286 \end{bmatrix} \quad M_3 = -2.428571429 \quad \mathbf{x}_3 = \begin{bmatrix} 1 \\ 0.5294117647 \end{bmatrix},$$

$$\widehat{\mathbf{x}}_4 = \begin{bmatrix} -1.823529412 \\ -0.8823529412 \end{bmatrix} \quad M_4 = -1.823529412 \quad \mathbf{x}_4 = \begin{bmatrix} 1 \\ 0.4838709677 \end{bmatrix},$$

$$\widehat{\mathbf{x}}_5 = \begin{bmatrix} -2.096774194 \\ -1.064516129 \end{bmatrix} \quad M_5 = -2.096774194 \quad \mathbf{x}_4 = \begin{bmatrix} 1 \\ 0.5076923077 \end{bmatrix}.$$

It appears that slowly $M_k \to -2$, the dominant eigenvalue, and $\mathbf{x}_k \to \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$, a corresponding eigenvector. ◯

In the "read more about it" in this section, we refer the reader to the book by Biswa Nath Datta to find other useful methods. In addition, you can learn that using a **shift** of A, namely, replacing A by $A - \sigma I_n$, can speed up convergence of the power method by replacing $|\lambda_2/\lambda_1|$ by a significantly smaller number $|(\lambda_2 - \sigma)/(\lambda_1 - \sigma)|$. Note that A has eigenvalue λ and corresponding eigenvector $\mathbf{x}$ if, and only if, $A - \sigma I$ has eigenvalue $\lambda - \sigma$ and the same corresponding eigenvector $\mathbf{x}$. Also, it is possible to use the Rayleigh quotient to help pick a good initial guess for the power method. Further, iteration using powers of $(A - \sigma I)^{-1}$ rather than $A - \sigma I$ gives us the **inverse power method**, which also can be useful.

8.5.3 Deflation and Similarity

Recall that if P is any invertible matrix whose columns are eigenvectors of a matrix A, then $P^{-1}AP = D$ is a diagonal matrix whose diagonal entries are the eigenvalues of A. Recall also that if two matrices A and B are "similar," that is, if there is an invertible matrix P such that $P^{-1}AP = B$, then A and B's eigenvalues and their algebraic and geometric multiplicities are the same. Further, the eigenvectors of A are of the form $P\mathbf{y}$ where $\mathbf{y}$ is any eigenvector of B, as you will explain in Problem 8.5.7.10. The process $A \mapsto P^{-1}AP$ is called a **similarity transformation**.

Further, having obtained the dominant eigenvalue of A, it is possible to use **deflation of polynomials**, or **deflation**, for short. This is a fancy name for using a root λ_1 to factor the polynomial $\mathcal{P}(\lambda) = (\lambda_0 - \lambda)g(\lambda)$. Further, it involves using a similarity transformation

$$A \mapsto \left[\begin{array}{c|c} \lambda_1 & \mathbf{0}^T \\ -- & -- \\ \mathbf{0} & A_{22} \end{array}\right],$$

reducing the size of the problem from n to $n - 1$. If A is a real, symmetric matrix, then this method works well.

8.5.4 Using Similarity Transformations

Some successful iterative methods for approximating the eigenvalues and eigenvectors of A define a sequence

$$A_k \triangleq P_k^{-1} \cdots P_1^{-1} A_0 P_1 \cdots P_k, \tag{8.46}$$

so A_k is similar to A for $k = 0, 1, \ldots$. If, for some k, A_k is nearly a diagonal or triangular matrix, then the diagonal entries of A_k are approximations of the eigenvalues of A. Some of the methods use a preliminary algorithm to do one similarity transformation $A \mapsto A_0$ to enable (8.46), the second algorithm, to be computationally efficient. In other words, the preliminary algorithm prepares the matrix for the second algorithm.

These methods are like a patient detective who systematically uncovers clues that give a more and more accurate and fuller picture of the case. A detective hopes to know the truth beyond a reasonable doubt. Numerical methods used in research and development need mathematical assurances that they have controlled their errors. Historically, the software routines in LINPACK and then EISPACK were breakthroughs in providing such robust capabilities to work with matrices.

8.5.5 Background

Recall from Problem 2.4.4.17 that if $\mathbf{q}$ is a unit vector in $\mathbf{R}^n$, then the Householder matrix $Q \triangleq I - 2\mathbf{q}\mathbf{q}^T$ is both symmetric, that is, $Q^T = Q$, and is orthogonal.

Corollary 8.1

If $\mathbf{u}$ is any nonzero vector and $\beta_{\mathbf{u}} \triangleq \frac{2}{\|\mathbf{u}\|^2}$, then $Q_{\mathbf{u}} \triangleq I - \beta_{\mathbf{u}}\mathbf{u}\mathbf{u}^T$ is a Householder matrix.

Why? You will derive this in Problem 8.5.7.9. □

Recall that $\mathbf{e}^{(1)} \triangleq \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^T$. Suppose $A = \begin{bmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{bmatrix}$ is a given $n \times n$ matrix, but we want to work with a matrix B that replaces $\mathbf{a}_1$, the first column of A, by a constant times $\mathbf{e}^{(1)}$. It turns out we can use a Householder matrix to accomplish this.

Define the **extended sign function** $Sign(x)$ by

$$Sign(x) \triangleq \begin{Bmatrix} -1, & x < 0 \\ +1, & x \geq 0 \end{Bmatrix}.$$

Lemma 8.1

Suppose $\mathbf{x} = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^T$ is a real, nonzero vector. Define

$$\mathbf{u} = \mathbf{x} + (Sign(x_1)\|\mathbf{x}\|)\mathbf{e}^{(1)} \text{ and } Q = Q_{\mathbf{u}},$$

where $\| \ \|$ is the usual, Euclidean norm on $\mathbb{R}^n$. Then

$$Q\mathbf{x} = -(x_1\|\mathbf{x}\|)\mathbf{e}^{(1)}. \tag{8.47}$$

Why? You will explain this in Problem 8.5.7.8. □

Definition 8.4

A matrix $A = [a_{ij}]_{\substack{1 \le i \le n \\ 1 \le j \le n}}$ is in **upper Hessenberg form** if $a_{i,j} = 0$ for $i > j + 1$. Another name for Hessenberg matrix is **almost triangular matrix**.

For example, $A = \begin{bmatrix} \star & \star & \star & \star & \star \\ \blacksquare & \star & \star & \star & \star \\ 0 & \blacksquare & \star & \star & \star \\ 0 & 0 & \blacksquare & \star & \star \\ 0 & 0 & 0 & \blacksquare & \star \end{bmatrix}$ is in unreduced upper Hessenberg form if $\blacksquare$ stands for any nonzero real number and $\star$ stands for any real number.

Similarly, a matrix is in **lower Hessenberg form** if $a_{i,j} = 0$ for $i+1 < j$. A matrix that is in both upper and lower Hessenberg form is **tridiagonal**, meaning that $a_{ij} = 0$ for $|i - j| > 1$.

A partitioned matrix $[A_{rs}]$ with blocks $A_{rs} = O$ for $r < s$ is called upper **quasi-triangular**. For example, the partitioned matrix A in Example 1.32 in Section 1.5 is upper quasi-triangular, and the matrix A in both Problems 1.5.3.24 and 1.5.3.25 is lower quasi-triangular. All Hessenberg matrices are also quasi-triangular.

Lemma 8.2

Every real, square matrix A can be reduced to an upper Hessenberg matrix by a sequence of at most $(n - 1)$ similarity transformations. Further, if A is symmetric then that upper Hessenberg matrix is tridiagonal.

Why? For example, given an $n \times n$ matrix A, we can find an $n \times n$ Householder matrix Q so that QAQ has its first column become $\begin{bmatrix} a_{11} & \star & 0 & \cdots & 0 \end{bmatrix}^T$. How? Write A in partitioned form as

$$\left[\begin{array}{c|c} a_{11} & \mathbf{y}^T \\ \hline \mathbf{x} & A^{(n-1)} \end{array}\right].$$

By using an $(n - 1) \times (n - 1)$ Householder matrix ${}^{(n-1)}P$, we can change the $(n - 1)$-vector $\mathbf{x} \triangleq \begin{bmatrix} a_{21} & \cdots & a_{n1} \end{bmatrix}^T$ to $\begin{bmatrix} -a_{21}\|\mathbf{x}\| & 0 & \cdots & 0 \end{bmatrix}^T$. Let Q have the partitioned form

$$\left[\begin{array}{c|c} 1 & \mathbf{0}^T \\ \hline \mathbf{0} & {}^{(n-1)}P \end{array}\right],$$

where $\mathbf{0}$ is the zero vector in $\mathbb{R}^{n-1}$. In fact, $Q = I - \beta_{\mathbf{u}}\mathbf{v}\mathbf{v}^T$, where $\mathbf{v} = \begin{bmatrix} 0 & \mathbf{u}^T \end{bmatrix}^T$. The multiplication of A on the left by Q has the desired effect on the first column and the multiplication of QA on the right by Q does not change the first column. Householder matrices are both symmetric and orthogonal, so $A \mapsto QAQ$ is a similarity transformation.

If A is symmetric, then multiplication of (QA) on the right by Q also has the desirable effect of changing the first row to A to be $\begin{bmatrix} -a_{21}\|\mathbf{x}\| & 0 & \cdots & 0 \end{bmatrix}$.

Whether or not A is symmetric, the similarity transformation $A \mapsto QAQ$ reduces A to the form

$$QAQ = \left[\begin{array}{c|c} a_{11} & \mathbf{y}^T \\ \hline \star & \mathbf{z}^T \\ \hline \mathbf{0} & A^{(n-2)} \end{array}\right].$$

If $\mathbf{x} = \mathbf{0}$ (to within the precision of the machine), then we can take $Q = I$ and not bother to multiply to get $QAQ = A$.

In a similar way we can use an $(n-2)\times(n-2)$ Householder matrix ${}^{(n-2)}P$ as part of an $n\times n$ Householder matrix Q_2 so that the first two columns of $Q_2(QAQ)Q_2$ are as they should be in an upper Hessenberg form. Continuing in this way completes the explanation. □

8.5.6 QR Algorithm

Having found a similarity transformation that transforms the matrix to one in upper Hessenberg form, the next step in finding the eigenvalues and eigenvectors is to use the "QR algorithm." This method relies on the ability to find the QR factorization, as in Section 2.7, of each iterate in the process.

Suppose A_k is an $n \times n$ matrix iterate. Theoretically, there exists at least one QR factorization of A_k, namely,

$$A_k = Q_k R_k.$$

Because A is real and square, Q_k is a real, orthogonal $n \times n$ matrix and R_k is an upper triangular matrix. If A_k is invertible, then so is R_k. Because Q_k is orthogonal, the *reverse multiplication* R_kQ_k is orthogonally similar to A_k, by Problem 2.7.7.28. Remarkably, the **QR algorithm** process, specifically

$$A_{k+1} = R_kQ_k, \text{ where } A_k = Q_kR_k \text{ is a } QR - \text{factorization},$$

produces a sequence of similarity transformations, that is, (8.46), that converge to a matrix very close to a triangular matrix. As mentioned before, we can read approximate eigenvalues and eigenvectors of that triangular matrix.

The reason for the preliminary transformation to upper Hessenberg form is purely for computational efficiency, as discussed in *Matrix Computations*, by Gene H. Golub and Charles F. Van Loan, mentioned immediately as follows.

Learn More About It

Numerical Linear Algebra and Applications, 2*nd* ed., by Biswa Nath Datta, Society for Industrial and Applied Mathematics, © 2010, has much useful information about algorithms for finding eigenvalues and eigenvectors. Also, as in Section 8.4, *Matrix Computations*, 3*rd* edition, by Gene H. Golub and Charles F. Van Loan, the Johns Hopkins University Press © 1996, and *Numerical Recipes: The Art of Scientific Computing*, 3*rd* edition, by William H. Press et al., Cambridge University Press © 2007, are excellent resources.

8.5.7 Problems

In problems 1–3, reduce the given matrix to upper Hessenberg form. You may use technology, for example, *Mathematica*, to help with the arithmetic and algebra, but be sure to show all of the steps involved.

1. $\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 0 & 1 \\ 1 & 0 & -1 & 1 \\ 0 & -1 & 2 & 1 \end{bmatrix}$

2. $\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & -1 & 1 \\ 1 & 2 & 0 & 1 \\ 0 & -1 & 2 & 1 \end{bmatrix}$

3. $\begin{bmatrix} 1 & 1 & 1 & -1 \\ 1 & 1 & 0 & 1 \\ 1 & -2 & 0 & 1 \\ 0 & -1 & -2 & 1 \end{bmatrix}$

In problems 4 and 5 apply the QR algorithm for two iterations to the given matrix A_0 to find A_1 and A_2. You may use technology, for example, the *Mathematica* command QRDecomposition[A], but be sure to show all of the steps involved.

4. $\begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 0 \\ 0 & 1 & -1 \end{bmatrix}$

5. $\begin{bmatrix} 1 & 0 & -1 \\ -1 & 2 & 0 \\ 0 & -1 & 1 \end{bmatrix}$

In problems 6 and 7 apply the power method to find the dominant eigenvalue of the given matrix.

6. $\begin{bmatrix} 2 & -1 \\ 5 & -4 \end{bmatrix}$
7. $\begin{bmatrix} 3 & 1 \\ 2 & 2 \end{bmatrix}$
8. Derive the result of Lemma 8.1.
9. Derive the result of Corollary 8.1, that is, that $Q_{\mathbf{u}} \triangleq I - \beta_{\mathbf{u}} \mathbf{u}\mathbf{u}^T$ is a Householder matrix if $\mathbf{u}$ is any nonzero vector and $\beta_{\mathbf{u}} \triangleq \frac{2}{\|\mathbf{u}\|^2}$.
10. Suppose $P^{-1}AP = B$. Explain why the eigenvectors of A are of the form $P\mathbf{y}$ where $\mathbf{y}$ is any eigenvector of B.

8.6 Approximations of Derivatives

Given a function $f(x)$ and a real number a, in Section 8.3, we defined

$$x_i = a + ih, \quad x_{i-0.5} = a + (i - 0.5)h,$$
$$f_i = f(x_i), \quad \text{and } f_{i-0.5} = f(x_{i-0.5}).$$

We would prefer to use k for the index, but we use i because that is the usual notation in this topic. Also, standard notation in this topic uses k to mean something else, as we will see in (8.63) in Section 8.7, (8.68) in Section 8.7, and other places.

In addition, we can define

$$f_i' \triangleq f'(x_i).$$

Linear approximation says that

$$f'(x_i) \approx \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} = \frac{f_{i+1} - f_i}{h}.$$

This motivates defining the **forward difference operator** Δ by

$$\Delta f_i \triangleq f_{i+1} - f_i. \tag{8.48}$$

So, we can approximate

$$f_i' \approx h^{-1} \Delta f_i. \tag{8.49}$$

By definition,

$$\Delta^2 f_i \triangleq \Delta\big[\Delta f_i\big] = \Delta\big[f_{i+1} - f_i\big] = \Delta f_{i+1} - \Delta f_i = \big(f_{i+2} - f_{i+1}\big) - \big(f_{i+1} - f_i\big);$$

hence,

$$\Delta^2 f_i = f_{i+2} - 2f_{i+1} + f_i.$$

It is also useful to define the **central difference operator** δ by

$$\delta f_i \triangleq f_{i+0.5} - f_{i-0.5}. \tag{8.50}$$

Similar to the calculation of Δ^2, we can show that

$$\delta^2 f_i = f_{i+1} - 2f_i + f_{i-1}.$$

Because

$$f''(x_i) \approx \frac{f'(x_i) - f'(x_{i-1})}{x_{i+1} - x_i} \approx \frac{h^{-1}\Delta f_i - h^{-1}\Delta f_{i-1}}{h} = \frac{(f_{i+1} - f_i) - (f_i - f_{i-1})}{h^2},$$

we see that

$$f_i'' \approx h^{-2}\delta^2 f_i. \tag{8.51}$$

We know that second derivatives play an important role in science and engineering, so (8.51) explains why central difference operators are very popular.

How good are the approximations (8.49) and (8.51)? The answers come from Taylor series: because $x_{i+1} - x_i = h$,

$$f_{i+1} = f(x_{i+1}) = f(x_i) + hf'(x_i) + \frac{h^2}{2!}f''(x_i) + \frac{h^3}{3!}f'''(x_i) + \cdots.$$

Theorem 8.11

(Taylor's theorem with remainder) If $f(x)$ is $(n+1)$ times continuously differentiable on an open interval $\mathcal{I}$ and x and y are in $\mathcal{I}$, then

$$f(y) = f(x) + (y-x)f'(x) + \frac{(y-x)^2}{2!}f''(x) + \cdots + \frac{(y-x)^n}{n!}f^{(n)}(x) + \frac{(y-x)^{n+1}}{(n+1)!}f^{(n+1)}(\xi) \tag{8.52}$$

for some ξ in $\mathcal{I}$.

We will also use the **big** $\mathcal{O}$ notation.

Definition 8.5

A quantity ε is $\mathcal{O}(h^p)$ if there exists

$$\lim_{h\to 0} \frac{\varepsilon}{h^p},$$

in which case we say ε is of order p as $h \to 0$.

Define an interval by

$$\mathcal{I}_{i,\eta} \triangleq (x_i - \eta, x_i + \eta).$$

Theorem 8.12

(a) If $f(x)$ is twice continuously differentiable on an interval of the form $\mathcal{I}_{i,\eta}$ for some $\eta > h$, then

$$f'(x_i) = h^{-1}\Delta f_i + \frac{f''(\xi)}{2}h, \tag{8.53}$$

for some ξ in the interval $\mathcal{I}_{i,h}$, and

$$\left| f'(x_i) - h^{-1}\Delta f_i \right| = \mathcal{O}(h^1). \tag{8.54}$$

(b) If $f(x)$ is four times continuously differentiable on an interval of the form $\mathcal{I}_{i,\eta}$ for some $\eta > h$, then

$$f''(x_i) = h^{-2}\delta^2 f_i + \frac{f''''(\xi_+) + f''''(\xi_-)}{24}h^2, \tag{8.55}$$

for some $\xi_\pm$ in the interval $\mathcal{I}_{i,h}$, and

$$\left| f''(x_i) - h^{-2}\delta^2 f_i \right| = \mathcal{O}(h^2). \tag{8.56}$$

Why? (a) By Taylor's theorem, there exists an ξ in $\mathcal{I}_{i,h}$ such that

$$f(x_{i+1}) = f(x_i) + hf'(x_i) + \frac{f''(\xi)}{2}h^2,$$

that is,

$$f_{i+1} - f_i = hf'(x_i) + \frac{f''(\xi)}{2}h^2.$$

Rearrange terms and divide through by h to get

$$f'(x_i) = h^{-1}\left(f_{i+1} - f_i\right) + \frac{f''(\xi)}{2}h,$$

that is,

$$f'(x_i) = h^{-1}\Delta f_i + \frac{f''(\xi)}{2}h,$$

In effect, symmetry of h versus $-h$ enabled the cancellations that gave us the higher-order estimate, $\mathcal{O}(h^2)$, instead of $\mathcal{O}(h)$.

8.6.1 Problems

1. Derive the formula for $\delta\Delta\delta f_i$.

In problems 2 and 3 you are given a formula for approximating a derivative. Find the order of the error.

2. $f'(x_i) \approx (3h)^{-1}\left(2f_i - f_{i-1} - f_{i-2}\right)$.
3. $f''(x_i) \approx (3h^2)^{-1}\left(f_{i+1} - 3f_{i-1} + 2f_{i-2}\right)$.

In problems 4–7 you are given a formula for approximating a derivative. Find the value of the parameter α that makes it a valid approximation and find the order of the error of the approximation for that value of α.

4. $f'(x_i) \approx (\alpha h)^{-1}\left(f_{i+1} - 3f_i + 2f_{i-1}\right)$.
5. $f'(x_i) \approx (\alpha h)^{-1}\left(f_{i+1} - 4f_i + 3f_{i-1}\right)$.
6. $f'(x_i) \approx (\alpha h)^{-1}\left(4f_{i+1} - 3f_i - f_{i-2}\right)$.
7. $f''(x_i) \approx (\alpha h^2)^{-1}\left(2f_{i+1} - 3f_i + f_{i-2}\right)$.
8. Explain why the **central difference approximation of** f' given by

$$f'(x_i) \approx \frac{f_{i+1} - f_{i-1}}{2h},$$

   has error whose order is two.
9. We want to approximate $(\star)\quad f'(x_i) \approx af_{i-2} + bf_i + cf_{i+1}$. Choose the constants a, b, c so as to make the approximation $(\star)$ have the highest order. State what that order is.

8.7 Approximate Solutions of ODE-IVPs

Recall from Sections 3.1 and 3.2 that a scalar ODE-IVP has the form (3.19) in Section 3.2, that is,

$$\left\{\begin{array}{l} \dfrac{dy}{dt} = f(t,y) \\ \\ y(t_0) = y_0 \end{array}\right\}. \tag{8.61}$$

We will see that finding a numerical approximation of the solution of an ODE-IVP relies on approximating the derivative, $\dot{y}(t)$, so knowledge from Section 8.6 will be useful.

We use numerical methods for approximating the solution of an ODE because, in general, we cannot derive the exact solution of the ODE. Nevertheless, our examples will be for ODEs where we can write down the exact solution so we can see clearly how well the numerical methods are working.

In this section we will present some of the elementary concepts and methods in the topic of numerical solutions of ODEs, which plays a great role in applied science and engineering. For that reason, there is a vast and ever-changing "state-of-the-art" literature on this topic. The truth is that those working on computationally intensive problems should employ the services of numerical analysts just as people working on statistically intensive problems should employ statistical consultants.

Similar to the notation in Sections 8.3 and 8.6, define

$$t_i = t_0 + ih, \quad t_{i+0.5} = t_0 + (i + 0.5)h,$$
$$y_i = y(t_i), \qquad \text{and } y_{i+0.5} = y(t_{i+0.5}),$$

and, in addition,

$$f_i \triangleq f(t_i, y_i). \tag{8.62}$$

The simplest way of approximating a solution of (8.61) is **Euler's method** (EM), which uses the tangent line at the point (t_i, y_i) on the graph of $y(t)$ to approximate $y(t_{i+1})$:

$$\left\{ \begin{array}{l} y_0 = y(t_0), \quad \text{and} \\ y_{i+1} \triangleq y_i + k_1, \qquad \text{where } k_1 \triangleq hf_i = hf(t_i, y_i) \end{array} \right\}. \tag{8.63}$$

So,

$$y(t_1) \approx y_1 = y_0 + hf_0,$$
$$y(t_2) \approx y_2 = y_1 + hf_1 = y_0 + hf_0 + hf_1,$$
$$y(t_3) \approx y_3 = y_2 + hf_2 = y_0 + hf_0 + hf_1 + hf_2, \text{ etc.}$$

Note that (8.63) is a first-order nonlinear difference equation, as in Section 4.6.

After we obtain the approximate solution values at $t_1, t_2, \ldots$, we can plot those points and connect them with straight line segments. This gives a piecewise linear approximate solution graph, as illustrated in Figure 8.5.

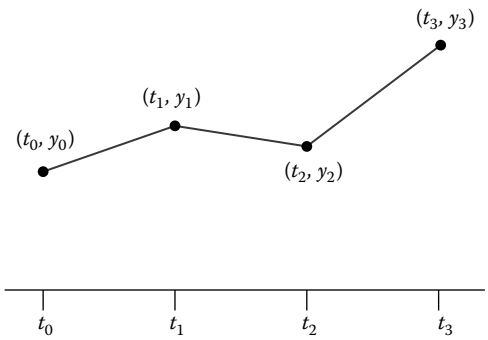


FIGURE 8.5
Numerical solution of an ODE-IVP.

A step size $h < 0$ is allowed by numerical methods for ODEs. Recall that Section 3.2's existence and uniqueness theory for (8.61) gives solutions that exist both for t less than the initial time, t_0, and for t greater than the initial time. With the exception of Problems 8.7.7.1 and 8.7.7.12, our examples and problems will be concerned with $h > 0$, although they could just as easily have $h < 0$.

To understand the performance of numerical methods for approximating the solution of an ODE-IVP, it helps to have one more notation and another background result from the first-year calculus. Define

$$\left\{ \begin{array}{c} \dot{f}_i \triangleq \dfrac{d}{dt}\Big[f(t,y(t))\Big]\Big|_{t=t_i} \\ \ddot{f}_i \triangleq \dfrac{d^2}{dt^2}\Big[f(t,y(t))\Big]\Big|_{t=t_i} \\ \vdots \qquad\qquad \vdots \\ f_i^{(n-1)} \triangleq \dfrac{d^{n-1}}{dt^{n-1}}\Big[f(t,y(t))\Big]\Big|_{t=t_i} \end{array} \right\}. \tag{8.64}$$

For example, if $\dot{y} = f(t,y) = 2ty^2$, then, using implicit differentiation, we find that

$$\dot{f}_i = \big(2y^2 + 4ty\dot{y}\big)\Big|_{t=t_i} = \big(2y^2 + 4ty \cdot 2ty^2\big)\Big|_{t=t_i} = 2y_i^2(1 + 4t_i y_i).$$

Taylor's Theorem 8.11 in Section 8.6, specifically (8.52) in Section 8.6, for the function $\dot{y}(t) \triangleq f(t,y(t))$ implies

$$y(t_{i+1}) = y(t_i) + hf_i + \frac{h^2}{2!}\dot{f}_i + \frac{h^3}{3!}\ddot{f}_i + \cdots + \frac{h^n}{n!}f_i^{(n-1)} + \mathcal{O}(h^{n+1}). \tag{8.65}$$

Recall that (3.16) in Section 3.2, or (6.36) in Section 6.4, the multivariable chain rule, implies

$$\frac{d}{dt}\big[f(t,y(t))\big] = \frac{\partial f}{\partial t}(t,y(t)) + \frac{\partial f}{\partial y}(t,y(t)) \cdot \frac{dy}{dt}(t).$$

Substituting in $\dot{y} = f(t,y)$ gives

$$\frac{d}{dt}\big[f(t,y(t))\big] = \frac{\partial f}{\partial t}(t,y(t)) + \frac{\partial f}{\partial y}(t,y(t)) \cdot f(t,y(t)). \tag{8.66}$$

So, the first definition in (8.64) can be evaluated as

$$\dot{f}_i = \frac{\partial f}{\partial t}(t_i,y_i) + \frac{\partial f}{\partial y}(t_i,y_i) \cdot f_i. \tag{8.67}$$

Suppose we have a numerical solution $y_0, y_1, y_2, \ldots$ that approximates the solution of (8.61). We define the **local error** to be

$$E = \big|y(t_{i+1}) - y_{i+1}\big|,$$

that is, the difference between the *exact* solution's value at $t = t_{i+1}$ and the numerical solution's value at $t = t_{i+1}$, under the assumption that $y_i = y(t_i)$, that is, that the numerical solution was exactly correct *before* the step of going from (t_i, y_i) to (t_{i+1}, y_{i+1}).

If we had *not* assumed that the numerical solution was exactly correct in previous steps, for example, if $y_i \neq y(t_i)$, then $\left|y(t_{i+1}) - y_{i+1}\right|$ is the **global error**.

The local error tells us how good any one step of the numerical solution process is, while the global error tells us how error propagates and compounds itself. In this section we will get results about the local error; getting results for global error is usually a lot more difficult. But, here is a "rule of thumb": if the local error is $\mathcal{O}(h^p)$, then the global error is usually $\mathcal{O}(h^{p-1})$.

Theorem 8.13

For Euler's Method (EM), the local error estimate is $\mathcal{O}(h^2)$.

Why? For $n = 1$, Taylor's theorem in the form of (8.65) gives

$$y(t_{i+1}) = y(t_i) + hf_i + \mathcal{O}(h^2).$$

Recall that when evaluating *local* error, we assume $y_i = y(t_i)$, and recall that (8.63) has $k_1 \triangleq hf_i$. The local error is

$$\begin{aligned}\left|y(t_{i+1}) - y_{i+1}\right| &= \left|\Big(y(t_i) + hf_i + \mathcal{O}(h^2)\Big) - \big(y_i + k_1\big)\right| \\ &= \left|\cancel{y(t_i)} + \cancel{hf_i} + \mathcal{O}(h^2) - \big(\cancel{y_i} + \cancel{hf_i}\big)\right| = \mathcal{O}(h^2). \ \square\end{aligned}$$

Example 8.16

The **improved tangent method** (ITM), also known as a **midpoint method** or as a **Runge–Kutta method** of order two, is

$$\left\{\begin{array}{l} y_0 = y(t_0) \\ y_{i+0.5} \triangleq y_i + k_1, \text{ where } k_1 \triangleq \frac{h}{2} f(t_i, y_i) \\ y_{i+1} \triangleq y_i + k_2, \text{ where } k_2 \triangleq hf(t_{i+0.5}, y_{i+0.5}). \end{array}\right\}. \tag{8.68}$$

Find the local error estimate.

Method: Recall that when evaluating *local* error, we assume $y_i = y(t_i)$. Equation (8.68) gives

$$y_{i+1} = y_i + hf(t_{i+0.5}, y_{i+0.5}) = y_i + hf\left(t_i + \frac{h}{2}, y_i + k_1\right). \tag{8.69}$$

Taylor series for a function of two variables, found as follows in Theorem 13.7 in Section 13.2, implies that

$$\begin{aligned} f\left(t_i + \frac{h}{2}, y_i + k_1\right) &= f(t_i, y_i) + \frac{h}{2}\frac{\partial f}{\partial t}(t_i, y_i) + k_1 \frac{\partial f}{\partial y}(t_i, y_i) + \mathcal{O}\left(\left(\frac{h}{2}\right)^2 + \frac{h}{2}k_1 + k_1^2\right) \\ &= f(t_i, y_i) + \frac{h}{2}\frac{\partial f}{\partial t}(t_i, y_i) + k_1 \frac{\partial f}{\partial y}(t_i, y_i) + \mathcal{O}\big(h^2\big). \end{aligned}$$

Combining this with (8.69) gives

$$y_{i+1} = y_i + hf\left(t_i + \frac{h}{2}, y_i + k_1\right) = y_i + h\left(f(t_i, y_i) + \frac{h}{2}\frac{\partial f}{\partial t}(t_i, y_i) + k_1\frac{\partial f}{\partial y}(t_i, y_i) + \mathcal{O}\big(h^2\big)\right).$$

But,

$$k_1 = \frac{h}{2}f(t_i, y_i) = \frac{h}{2}f_i,$$

so

$$y_{i+1} = y_i + hf_i + \frac{h^2}{2}\left(\frac{\partial f}{\partial t}(t_i, y_i)0 + \frac{\partial f}{\partial y}(t_i, y_i)\cdot f_i\right) + \mathcal{O}\big(h^3\big).$$

On the other hand, Taylor's theorem in the form of (8.65), along with (8.67), implies

$$y(t_{i+1}) = y_i + hf_i + \frac{h^2}{2}\dot{f}_i + \mathcal{O}(h^3) = y_i + hf_i + \frac{h^2}{2}\left(\frac{\partial f}{\partial t}(t_i, y_i) + \frac{\partial f}{\partial y}(t_i, y_i)\cdot f_i\right) + \mathcal{O}(h^3).$$

So, the local error is

$$\begin{aligned}\big|y(t_{i+1}) - y_{i+1}\big| = \Bigg|&\cancel{y_i} + \cancel{hf_i} + \frac{h^2}{2}\left(\cancel{\frac{\partial f}{\partial t}(t_i, y_i)} + \cancel{\frac{\partial f}{\partial y}(t_i, y_i)\cdot f_i}\right) + \mathcal{O}(h^3)\\ &- \cancel{y_i} - \cancel{hf_i} + \frac{h^2}{2}\left(\cancel{\frac{\partial f}{\partial t}(t_i, y_i)} + \cancel{\frac{\partial f}{\partial y}(t_i, y_i)f_i}\right)\Bigg|,\end{aligned}$$

that is, the local error is $\mathcal{O}(h^3)$. ◯

To calculate even higher orders, we would need further applications of the multivariable chain rule, for example,

$$\begin{aligned}\frac{d^2}{dt^2}\Big[f(t, y(t))\Big] &= \frac{d}{dt}\left[\frac{\partial f}{\partial t}(t, y(t)) + \frac{\partial f}{\partial y}(t, y(t))\cdot f(t, y(t))\right]\\ &= \frac{\partial^2 f}{\partial t^2}(t, y(t)) + 2\frac{\partial^2 f}{\partial t\partial y}(t, y(t))\cdot f(t, y(t)) + \frac{\partial f}{\partial y}(t, y(t))\cdot\frac{d}{dt}\Big[f(t, y(t))\Big]\\ &\quad + \frac{\partial^2 f}{\partial y^2}(t, y(t))\cdot\big(f(t, y(t))\big)^2.\end{aligned}$$

So, akin to (8.67),

$$\ddot{f}_i = \frac{\partial^2 f}{\partial t^2}(t_i, y_i) + 2\frac{\partial^2 f}{\partial t\partial y}(t_i, y_i)\cdot f_i + \frac{\partial f}{\partial y}(t_i, y_i)\left(\frac{\partial f}{\partial t}(t_i, y_i) + \frac{\partial f}{\partial y}(t_i, y_i)\cdot f_i\right) + \frac{\partial^2 f}{\partial y^2}(t_i, y_i)\cdot(f_i)^2,$$

that is,

$$\ddot{f}_i = \frac{\partial^2 f}{\partial t^2}(t_i, y_i) + 2\frac{\partial^2 f}{\partial t\partial y}(t_i, y_i)\cdot f_i + \frac{\partial f}{\partial y}(t_i, y_i)\cdot\dot{f}_i + \frac{\partial^2 f}{\partial y^2}(t_i, y_i)\cdot(f_i)^2. \tag{8.70}$$

8.7.1 Runge–Kutta Methods

Runge–Kutta methods were invented separately by Runge and Kutta. While their implementation by hand is a little more intricate than EM, they are much more accurate, and their implementations in software programming are not difficult.

Very popular is the **Runge–Kutta method of order four** given by

$$\left\{\begin{array}{l} y_0 = y(t_0) \\ y_{i+1} \triangleq y_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4), \text{ where} \\ k_1 \triangleq hf(t_i, y_i),\ k_2 \triangleq hf(t_{i+0.5}, y_i + \frac{1}{2}k_1), \\ k_3 \triangleq hf(t_{i+0.5}, y_i + \frac{1}{2}k_2), \text{ and} \\ k_4 \triangleq hf(t_{i+1}, y_i + k_3) \end{array}\right\}. \tag{8.71}$$

A simpler example of a Runge–Kutta method of order two is the ITM discussed in Example 8.16.

Example 8.17

Consider the ODE-IVP $\dot{y} = -y + e^{-2t}$, $y(2.0) = 1$.

(a) Use EM with step size, $h_1 = 0.25$, to estimate $y(5.0)$.

(b) Use the ITM given in (8.68) with step size, $h_2 = 0.5$, to estimate $y(5.0)$.

(c) Find the exact solution of the ODE-IVP.

(d) Compare the function value for the exact solution of the ODE-IVP with your approximate results from parts (a) and (b).

Method: Results for parts (a) and (b) are in Table 8.14. Those results were produced using a spreadsheet.

(c) The exact solution of the ODE-IVP is $y(t) = -e^{-2t} + (e^2 + e^{-2})e^{-t}$, which can be found using the integrating factor method in Section 3.2.

(d) The exact solution value is $y(5) = -e^{-10} + (e^2 + e^{-2})e^{-5} \approx 0.05065355$. The ITM with step size $h_2 = 0.5$ gave $y(5) \approx y_6 \approx 0.06033006$ and EM with step size $h_1 = 0.25$ gave $y(5) \approx y_{12} \approx 0.03260821$, so the ITM with twice as large a step size produced a significantly more accurate conclusion than EM. The honorific "improved" is truly deserved here, unlike its often inappropriate use in consumer advertising. ○

We see in Table 8.14 that, as is typical for numerical methods for approximating the solution of an ODE-IVP, the more time steps we take, the higher the relative global error becomes. This is because, for example, an error after four time steps causes two kinds of error in the fifth step to approximate y_5: (i) The global error-prone value of y_4 is substituted into the numerical method, and (ii) even if y_4 had been correct, one more local error could be made in going to the next step. Error can accumulate and propagate.

Numerical approximation of solutions of ODE-IVPs often involves a tension between (a) the need to reduce local error by reducing the step size, (b) the danger of increased round-off error because of the step size being too small, and (c) the danger of taking too

TABLE 8.14

Example 8.17: EM and ITM

i	t_i	$\approx y(t_i)$ by EM	Exact $y(t_i)$	i	t_i	$\approx y(t_i)$ by ITM
0	2.00	1.00000000	1.00000000	0	2.00	1.00000000
1	2.25	0.75457891	0.78195602			
2	2.50	0.56871143	0.61090171	1	2.50	0.62826504
3	2.75	0.42821806	0.47693148			
4	3.00	0.32218524	0.37213864	2	3.00	0.39386679
5	3.25	0.24225862	0.29024888			
6	3.50	0.18206982	0.22630505	3	3.50	0.24660862
7	3.75	0.13678034	0.17640364			
8	4.00	0.10272352	0.13747857	4	4.00	0.15429295
9	4.25	0.07712651	0.10712621			
10	4.50	0.05789575	0.08346503	5	4.50	0.09649289
11	4.75	0.04345266	0.06502389			
12	5.00	0.03260821	0.05065355	6	5.00	0.06033006

many time steps so that the program runs too slowly to produce useful results. For example, it took many decades of improved hardware and software before machines could accurately produce a forecast of the weather an hour from now in less than an hour of computations! As another example, if we need one hundred years of computing time to produce a design for a more aerodynamically efficient automobile, then it would put the company out of business.

8.7.2 Multistep Methods

Some useful methods of approximating the solution of an ODE are called **multistep methods** because they calculate y_{i+n} using $y_i, y_{i+1}, \ldots, y_{i+n-1}$. Such a multistep method is an n-th order difference equation. The **order** of a multistep method is the order of the corresponding difference equation.

One family of very popular multistep methods is called **Adams** methods. For example, the **Adams–Bashforth** method of order two is

$$\left\{ \begin{array}{l} y_0 = y(t_0) \\ y_1 \text{ calculated by some other method} \\ y_{i+1} \triangleq y_i + \frac{h}{2} \cdot \left(3f(t_i, y_i) - f(t_{i-1}, y_{i-1})\right), \; i = 1, 2, \ldots \end{array} \right\}. \tag{8.72}$$

One drawback of any multistep method is that it begins by using some other method, for example, a Runge–Kutta method, to approximate the values between y_0 and y_n, for example, y_1 for the second-order method in (8.72). When "filling in" those values, we must take care that the order of the other method's global error is no worse than the order of the local error of the multistep method or take many small steps in between, for example, $y_{0.1}, y_{0.2}, \ldots, y_{1.0}$, where $y(x_0 + \beta h) \approx y_\beta$, to achieve the same accuracy as the multistep method.

8.7.3 Predictor–Corrector Methods

These "implicit" methods use a formula of the form

$$y_{i+n} = y_{i+n-1} + g(y_{i+n}, y_{i+n-1}, \ldots, y_i)$$

that must be solved for y_{i+n} by some numerical, iterative method, for example, Newton's method. For example, an **Adams–Moulton** "corrector" implicit formula of order two is given by

$$y_{i+3} = y_{i+2} + \frac{h}{24}\left(9f_{i+3} + 19f_{i+2} - 5f_{i+1} + f_i\right),$$

where $f_j \triangleq f(t_j, y_j)$. The iteration may begin with an initial guess for y_{i+3} calculated by a predictor formula, for example, an Adams–Bashforth formula of order four given by

$$y_{i+3} = y_{i+2} + \frac{h}{24}\left(55f_{i+2} - 59f_{i+1} + 37f_i - 9f_{i-1}\right).$$

As mentioned earlier, the Adams–Bashforth formula requires filling in of intermediate values, by some other method, before it can proceed.

One advantage of corrector methods is that (Conte and de Boor, 1972) they can provide an estimate of the local error; the software package can use that estimate to decrease the step size temporarily to increase the accuracy. For this reason, Adams and other predictor–corrector methods are often used if we suspect the system is "stiff," a concept that we will briefly discuss in Section 8.7.6.

8.7.4 Systems of ODEs

All of the methods for scalar ODEs have natural generalizations to systems of aforementioned ODEs. For example, EM for a system is

$$\left\{\begin{array}{l} \mathbf{x}_0 = \mathbf{x}(t_0) \\ \mathbf{x}_{i+1} \triangleq \mathbf{x}_i + \mathbf{k}_1, \quad \text{where } \mathbf{k}_1 \triangleq h\,\mathbf{f}_i = h\,\mathbf{f}(t_i, \mathbf{x}_i) \end{array}\right\}.$$

8.7.5 Taylor's Formula Method

Perhaps one of the oldest higher-order generalization of EM is to use more terms in the Taylor series expansion (8.52) in Section 8.6. Here we will write that as

$$y(t_{i+1}) = y(t_i) + h\dot{y}(t_i) + \frac{h^2}{2!}\ddot{y}(t_i) + \cdots + \frac{h^n}{n!}y^{(n)}(t_i) + \frac{h^{n+1}}{(n+1)!}y^{(n+1)}(\xi). \tag{8.73}$$

Let y_i be the approximation for $y(t_i)$, $\dot{y}_i$ be the approximation for $\dot{y}(t_i)$, etc. The order of Taylor's method (TM) (8.73) is n, the local error is $\mathcal{O}(h^{n+1})$, and the global error is $\mathcal{O}(h^n)$.

Consider the solution of an ODE $\dot{y} = f(t, y)$. As in Section 8.6, in principle we need the chain and product rules to calculate the formulas for $\ddot{y}$, $\dddot{y}$, etc., that appear in the Taylor series.

TABLE 8.15

Example 8.18: TM and ITM

i	t_i	$\approx y(t_i)$ by TM	Exact $y(t_i)$	i	t_i	$\approx y(t_i)$ by ITM
0	1.00	−1.00000000	−1.00000000	0	1.00	−1.00000000
1	1.25	−0.68750000	−0.64000000	1	1.25	−0.68359375
2	1.50	−0.48948288	−0.44444444	2	1.50	−0.48492460
3	1.75	−0.36078143	−0.32653061	3	1.75	−0.35703326
4	2.00	−0.27470751	−0.25000000	4	2.00	−0.27194580

Example 8.18

Consider the ODE-IVP $\dot{y} = 2ty^2,\ y(1.0) = -1$.

(a) Use TM of order two with step size, $h_1 = 0.25$, to estimate $y(2.0)$.

(b) Use the ITM given in (8.68) with step size, $h_2 = 0.25$, to estimate $y(2.0)$.

(c) Find the exact solution of the ODE-IVP.

(d) Compare the function value for the exact solution of the ODE-IVP with your approximate results from parts (a) and (b).

Method: (a) $\dot{y} = f(t, y) \triangleq 2ty^2$, so

$$\ddot{y} = \frac{d}{dt}\left[2ty^2\right] = 2 \cdot 1 \cdot y^2 + 2t \cdot (2y\,\dot{y}) = 2y^2 + 2t \cdot 2y \cdot (2ty^2) = 2y^2(1 + 4t^2y).$$

So, TM of order two for this example is

$$y_{i+1} = y_i + h\dot{y}_i + \frac{h^2}{2}\,\ddot{y}_i = y_i + h \cdot 2t_i y_i^2 + \frac{h^2}{2} \cdot 2y_i^2(1 + 4t_i^2 y_i).$$

The results of TM and (b) the ITM, both with step size $h = 0.25$, are shown in Table 8.15. In particular, $y(2.0) \approx y_4 \approx -0.20439427$ for TM and $y(2.0) \approx y_4 \approx -0.27194580$ for ITM versus the exact solution value $y(2.0) = -0.25$.

(c) The method of separation of variables in Section 3.2 or integrating factor in Section 3.1 produces the exact solution $y(t) = -t^{-2}$; hence, the exact value is $y(2.0) = -0.25$.

(d) With $h = 0.25$, we see that TM is not very accurate, but the ITM is fairly accurate. One method may be more accurate than another *for a particular example*, even though they both have order two. ◯

By the way, TM of order two with $h = 0.125$ gives $y(2.0) \approx y_8 \approx -0.25381128$, with $h = 0.0625$ gives $y(2.0) \approx y_{16} \approx -0.25079251$, and with $h = 0.03125$ gives $y(2.0) \approx y_{32} \approx -0.25018258$, with very similar results for the ITM. Decreasing h seems to lead to better and better results fairly quickly.

In the realm of the "old becoming new again," the book *Numerical Recipes: The Art of Scientific Computing*, mentioned in the "read more about it" as follows, discusses a "multivalue (or Nordsieck) method" that also uses Taylor series terms.

8.7.6 Numerical Instability and Stiffness

Astute users of numerical methods notice that sometimes if the step size, h, is too large, then some of the numerical solutions of an ODE-IVP behave qualitatively in significantly

wrong ways. For example, an asymptotically stable system of ODEs could have numerical, approximate solutions that don't go to zero or even blow up as $t \to \infty$.

Throughout this section, we will assume that the LCCCHS $\dot{\mathbf{x}} = A\mathbf{x}$ is asymptotically stable. According to Theorem 5.11 in Section 5.3, that is, all eigenvalues λ of the square matrix A have negative real part.

For example, for EM for $\dot{\mathbf{x}} = A\mathbf{x}$, what is the basic cause of numerical instability? The corresponding linear homogeneous system of difference equations is $\mathbf{x}_{i+1} - \mathbf{x}_i = hA\mathbf{x}_i$, that is,

$$\mathbf{x}_{i+1} = (I + hA)\mathbf{x}_i.$$

It may be unstable, according to Theorem 5.21 in Section 5.7, because the matrix $I + hA$ may have an eigenvalue μ whose magnitude is greater than one.

Let λ_M be the eigenvalue of A whose real part is the most negative. By Theorem 2.8 in Section 2.2, the eigenvalues of $I + hA$ are $\mu = 1 + h\lambda$. Suppose $h > 0$. If

$$h \cdot \mathcal{R}e(\lambda_M) < -2,$$

then $|\mu| = |1 + h\lambda_M| > 1$, and thus, $I + hA$ would have an eigenvalue μ whose magnitude is greater than one, causing the system of difference equations to be unstable.

Also, no matter what the sign of h, if λ_0 is the eigenvalue A whose imaginary part is the greatest in magnitude, then the system of difference equations would be unstable if $|h \cdot \mathcal{I}m(\lambda_0)| > 1$.

Now let us define the concept of "stiffness." Let λ_m be the eigenvalue with the *least* negative real part. We say the system of ODEs is **stiff** if the ratio λ_M/λ_m is "large."

The trouble with stiff systems is that the demand for numerical stability implies we must take h to be smaller than $2\big(\mathcal{R}e(\lambda_M)\big)^{-1}$. But, because $\mathcal{R}e(\lambda_m)$ is small, the system will change very slowly and require many of those small time steps, which would tend to increase round-off error. The most rapidly decaying part of the solution, due to λ_M, is imposing its need for numerical stability on the rest of the system.

One way to repair the situation is to have an **adaptive step size**. Here we are getting into the territory of the methods really used by software packages. For example, the MATLAB command ode45 is a Runge–Kutta–Fehlberg method. It combines results from one of the fourth-order Runge–Kutta methods and a fifth-order Runge–Kutta method to estimate the local error. If it suspects the error might be large, it decreases the time step size. For truly informative discussions of this and also the Bulirsch–Stoer method, please see the references as follows.

Learn More About It

Further useful developments and discussion are found in *An Introduction to Numerical Methods and Analysis*, by James F. Epperson, John Wiley & Sons, Inc. © 2002, and *Numerical Recipes: The Art of Scientific Computing*, 3*rd* edition, by William H. Press et al., Cambridge University Press © 2007.

8.7.7 Problems

1. For the ODE-IVP $\dot{y} = t^2 - y^2$, $y(0) = \frac{1}{2}$.
   (a) Use EM with step size $h = 0.5$ to get an approximate solution on the interval $[0, 1]$.
   (b) Use EM with step size $h = -0.5$ to get an approximate solution on the interval $[-1, 0]$.
   (c) Use your results from parts (a) and (b) to sketch a piecewise linear approximate solution on the interval $-1 \leq t \leq 1$.
2. Consider the ODE-IVP $t\dot{y} + (t-1)y = t^2$, $y(1) = -2$.
   (a) Use EM with step size $h = 0.4$ to get an approximate solution on the interval $[1, 5]$.
   (b) Find the exact solution of the ODE-IVP.
   (c) On the same set of axes, graph the exact solution and the piecewise linear approximation using your results from part (a). Discuss how good an approximation EM gave.

In each of problems 3 and 4, you are given a desired function value for the solution of the given ODE-IVP.

   (a) Use EM, by hand with the given step size, h_1, to estimate the desired function value.
   (b) Use the Runge–Kutta method of order four given in (8.71), by hand with the given step size, h_4, to estimate the desired function value.
   (c) Find the exact solution of the ODE-IVP.
   (d) Compare the function value for the exact solution of the ODE-IVP with your approximate results from parts (a) and (b).

3. $\dot{y} = y - t$, $y(1.0) = 3$, desired function value $y(1.6)$, and (a) $h_1 = 0.2$, (b) $h_4 = 0.6$.
4. $\dot{y} = y - t$, $y(1) = 3$, desired function value $y(2.0)$, and (a) $h_1 = 0.2$, (b) $h_4 = 0.5$.

In each of problems 5–6, you are given a desired function value for the solution of the given ODE-IVP.

   (a) Use EM, by hand with the given step size, h_1, to estimate the desired function value.
   (b) Use the Runge–Kutta method of order two given in (8.68), by hand with the given step size, h_2, to estimate the desired function value.
   (c) Find the exact solution of the ODE-IVP.
   (d) Compare the function value for the exact solution of the ODE-IVP with your approximate results from parts (a) and (b).

5. $\dot{y} = -2y + e^{-t}$, $y(1.0) = 3$, desired function value $y(1.6)$, and (a) $h_1 = 0.2$, (b) $h_2 = 0.6$.
6. $\dot{y} = -2y + e^{-t}$, $y(1.0) = 3$, desired function value $y(2.0)$, and (a) $h_1 = 0.2$, (b) $h_2 = 0.5$.

In each of problems 7–9, you are given an ODE and a local error order, $\mathcal{O}(h^{n+1})$. Derive TM of order n for this ODE. Do not substitute in a specific value of h.

7. $\dot{y} = y^2 - t$, $\mathcal{O}(h^3)$.
8. $\dot{y} = t + t^2 y$, $\mathcal{O}(h^4)$.
9. $\dot{y} = ye^{\cos(t-y)}$, $\mathcal{O}(h^3)$.

In each of problems 10 and 11, you are given a desired function value for the solution of the given ODE-IVP.

(a) Derive TM of order two for this ODE. Do not substitute in a specific value of h.
(b) With step size h, now use TM to estimate the desired function value.
(c) Find the exact solution of the ODE-IVP.
(d) Compare the function value for the exact solution of the ODE-IVP with your approximate result from part (b).

10. $\dot{y} + y\cos t = t$, $y(0) = 2$, desired function value $y(\frac{1}{3})$, and (i) $h = \frac{1}{3}$, (ii) $h = \frac{1}{6}$
11. $\dot{y} = -2y + \cos 3t$, $y(0) = 1$, desired function value $y(2.0)$, and $h = 0.25$.
12. Let α be an unspecified constant. Use EM with $h = \pm 0.1$ to approximate (a) $y(1.2)$ and (b) $y(0.8)$, where $y(t)$ solves the ODE-IVP $\dot{y} = \alpha y - t$, $y(1) = 0$. Your final conclusions should be in terms of the unspecified constant α. Work by hand and show all work.
13. If we use the approximation $\dot{y}(t_i) \approx (2h)^{-1}(y_{i+1} - y_{i-1})$ and substitute that into the ODE $\dot{y} + y = 0$ then, we get the second-order difference equation $y_{i+1} + 2hy_i - y_{i-1} = 0$. Show that, no matter how small $h > 0$ is chosen, the difference equation has a solution that does not behave, as $t \to \infty$, like the solutions of the original ODE. This is an example of numerical instability.
14. If we use the usual central difference approximation of the second derivative in the ODE $\ddot{y} + 4y = 0$, we get the second-order difference equation $y_{i+1} - 2y_i + y_{i-1} + 4h^2 y_i = 0$. Show that for all h satisfying $0 < h < 1$, the difference equation has solutions that do behave like the solutions of the original ODE.
15. If we use EM for the undamped harmonic oscillator system

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \mathbf{x},$$

explain why the solutions of the system of difference equations are spirals, not the ellipses that are the solutions of the system of differential equations in $\mathbb{R}^2$.

16. If we use EM with $h = 1/n$ for the especially simple ODE-IVP $\dot{y} = f(t)$, $y(0) = y_0$, what would y_n approximate? You may assume that $f(t)$ is a continuous function. [Hints: Note that $y_1 = y_0 + hf(t_0)$, $y_2 = y_0 + hf(t_0) + hf(t_1)$. Also, think of how the definite integral was defined.]
17. The **modified EM** is

$$\left\{ \begin{array}{l} y_0 = y(t_0) \\ y_{i+1} \triangleq y_i + \frac{1}{2}(k_1 + k_2), \text{ where} \\ k_1 \triangleq hf(t_i, y_i), k_2 \triangleq hf(t_{i+1}, y_i + k_1) \end{array} \right\}.$$

With $h = 1/n$, use the modified EM for the especially simple ODE-IVP $\dot{y} = f(t)$, $y(0) = y_0$. What would y_n approximate? You may assume that $f(t)$ is a continuous function. [Hint: Note that $y_1 = y_0 + \frac{h}{2}f(t_0) + \frac{h}{2}f(t_1)$, etc.]
The **modified EM** is also known as a **Runge–Kutta method of order two**.

18. Find the local error estimate for the **improved EM**, also known as another **Runge–Kutta method of order two** given in (8.68).
19. Explain why the local error is $\mathcal{O}(h^4)$ for **Runge–Kutta method of order three**,

$$\left\{\begin{array}{l} y_0 = y(t_0) \\ y_{i+1} \triangleq y_i + \frac{1}{6}(k_1 + 4k_2 + k_3), \text{ where} \\ k_1 \triangleq hf(t_i, y_i), \\ k_2 \triangleq hf(t_{i+0.5}, y_i + \frac{1}{2}k_1), \text{ and} \\ k_3 \triangleq hf(t_{i+1}, y_i - k_1 + 2k_2) \end{array}\right\}. \tag{8.74}$$

20. Explain why the local error is $\mathcal{O}(h^4)$ for a **second Runge–Kutta method of order three**,

$$\left\{\begin{array}{l} y_0 = y(t_0) \\ y_{i+1} \triangleq y_i + \frac{1}{4}(k_1 + 3k_2), \text{ where} \\ k_1 \triangleq hf(t_i, y_i), \text{ and} \\ k_2 \triangleq hf(t_i + \frac{1}{3}h, y_i + \frac{1}{3}k_1) \end{array}\right\}. \tag{8.75}$$

21. (Small project) Explain why the local error is $\mathcal{O}(h^5)$ for the **Runge–Kutta method of order four** given in (8.71).
22. (Small project) Explain why the local error is $\mathcal{O}(h^5)$ for **Gill's method**,

$$\left\{\begin{array}{l} y_0 = y(t_0) \\ y_{i+1} \triangleq y_i + \frac{1}{6}\left(k_1 + 2\left(1 - \sqrt{\frac{1}{2}}\right)k_2 + 2\left(1 + \sqrt{\frac{1}{2}}\right)k_3 + k_4\right), \text{ where} \\ k_1 \triangleq hf(t_i, y_i), \\ k_2 \triangleq hf(t_{i+0.5}, y_i + \frac{1}{2}k_1), \\ k_3 \triangleq hf\left(t_{i+0.5}, y_i + \left(-\frac{1}{2} + \sqrt{\frac{1}{2}}\right)k_1 + \left(1 - \sqrt{\frac{1}{2}}\right)k_2\right), \text{ and} \\ k_4 \triangleq hf\left(t_{i+1}, y_i - \sqrt{\frac{1}{2}}k_2 + \left(1 + \sqrt{\frac{1}{2}}\right)k_3\right) \end{array}\right\}. \tag{8.76}$$

8.8 Approximate Solutions of Two Point BVPs

Let $' = \frac{d}{dx}$. Here is an example of a two-point boundary value problem:

$$\left\{ \begin{array}{c} y'' + \dfrac{1}{x}y' + y = 0 \\ y(1) = 0, y(3) = -1 \end{array} \right\}. \tag{8.77}$$

Because the ODE is to be satisfied over the interval $(1, 3)$, let the step size be

$$h = \frac{\text{Length of the interval}}{N} = \frac{2}{N}.$$

Denote $x_i = 1 + ih$ and $y_i = y(x_i)$, for $i = 0, 1, \ldots, N$. The *simplest* approach to approximating a solution of (8.77) is to replace the second derivative in the ODE by a central difference approximation and replace the first derivative by a forward difference approximation:

$$y''(x_i) \approx h^{-2}\delta^2 y_i = h^{-2}\left(y_{i+1} - 2y_i + y_{i-1}\right) \text{ and } y'(x_i) \approx h^{-1}\Delta y_i = h^{-1}\left(y_{i+1} - y_i\right).$$

[In homework problems you will use the central difference approximation of the first derivative, which has higher-order accuracy than the forward difference approximation.]

The approximation of the ODE becomes the **replacement equations** for this **finite difference** method: For $i = 1, \ldots, N-1$,

$$h^{-2}\left(y_{i+1} - 2y_i + y_{i-1}\right) + \frac{1}{x_i} \cdot h^{-1}\left(y_{i+1} - y_i\right) + y_i = 0,$$

that is, after multiplying through by h^2, $y_{i+1} - 2y_i + y_{i-1} + \frac{h}{x_i}\left(y_{i+1} - y_i\right) + h^2 y_i = 0$, that is,

$$y_{i-1} + \left(-2 - \frac{h}{x_i} + h^2\right) y_i + \left(1 + \frac{h}{x_i}\right) y_{i+1} = 0. \tag{8.78}$$

Our finite difference method does not have replacement equations for the ODE at the endpoints $x = x_0$ and $x = x_N$ because we do not have boundary conditions involving x_{-1} or x_{N+1}.

The boundary conditions become

$$y_0 = 0 \quad \text{and} \quad y_N = -1.$$

For $i = 1$, substitute the left-hand boundary condition and $x_1 = 1 + h$ in the replacement equation, (8.78), to get

$$0 + \left(-2 - \frac{h}{1+h} + h^2\right) y_i + \left(1 + \frac{h}{1+h}\right) y_{i+1} = 0.$$

For $i = 2, \ldots, N-2$, we keep the replacement equation, (8.78), as is.

For $i = N-1$, substitute the right-hand boundary condition and $x_{N-1} = 3 - h$ in the replacement equation, (8.78), to get

$$y_{N-2} + \left(-2 - \frac{h}{3-h} + h^2\right) y_{N-1} + \left(1 + \frac{h}{3-h}\right) \cdot (-1) = 0.$$

Define

$$\mathbf{y} \triangleq \begin{bmatrix} y_1 \\ \vdots \\ y_{N-1} \end{bmatrix}.$$

After a little further algebraic manipulations, the replacement equations along with the boundary conditions are

$$\begin{bmatrix} -2 & 1 & 0 & \cdot & \cdot & \cdot & & 0 \\ 1 & -2 & 1 & & & & & \cdot \\ 0 & 1 & -2 & \cdot & & & & \cdot \\ \cdot & & & \cdot & \cdot & & & \cdot \\ \cdot & & & & \cdot & -2 & 1 & 0 \\ \cdot & & & & & 1 & -2 & 1 \\ 0 & \cdot & \cdot & \cdot & & 0 & 1 & -2 \end{bmatrix} \mathbf{y}$$

$$+ \begin{bmatrix} -\frac{-h}{1+h} & \frac{h}{1+h} & 0 & \cdot & \cdot & \cdot & & 0 \\ 0 & \frac{-h}{1+2h} & \frac{h}{1+2h} & & & & & \cdot \\ \cdot & 0 & \frac{-h}{1+3h} & \cdot & & & & \cdot \\ \cdot & & & \cdot & \cdot & & & \cdot \\ \cdot & & & & \cdot & \frac{-h}{3-3h} & \frac{h}{3-3h} & 0 \\ & & & & & 0 & \frac{-h}{3-2h} & \frac{h}{3-2h} \\ 0 & \cdot & \cdot & \cdot & & & 0 & \frac{-h}{3-h} \end{bmatrix} \mathbf{y} + h^2 \mathbf{y} = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 + \frac{h}{3-h} \end{bmatrix}. \quad (8.79)$$

System (8.79) can be written compactly as $A\mathbf{y} = \mathbf{b}$. Notice that part of A is the famous tri-diagonal matrix we saw in Problem 4.6.4.15.

Example 8.19

Find the approximate solution of ODE-BVP (8.77) when $N = 5$.

Method: For $N = 5$, $h = 0.4$, so (8.79) is

$$\begin{bmatrix} -2 - \frac{0.4}{1.4} + 0.16 & 1 + \frac{0.4}{1.4} & 0 & 0 \\ 1 & -2 - \frac{0.4}{1.8} + 0.16 & 1 + \frac{0.4}{1.8} & 0 \\ 0 & 1 & -2 - \frac{0.4}{2.2} + 0.16 & 1 + \frac{0.4}{2.2} \\ 0 & 0 & 1 & -2 - \frac{0.4}{2.6} + 0.16 \end{bmatrix} \mathbf{y} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{3}{2.6} \end{bmatrix},$$

whose solution is

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \approx \begin{bmatrix} -0.5875491867 \\ -0.9714146554 \\ -1.158319393 \\ -1.159650930 \end{bmatrix}. \bigcirc$$

Figure 8.6 shows the dashed graph of the approximate solution, including $y_0 = 0$ and $y_5 = -1$: The dashed line segments "connect the dots" between the approximate solution graph's points $(1, 0), (1.4, -0.5875491867), \ldots, (3, -1)$. The exact solution is the solid curve.

By the way, it is possible to find the exact solution of ODE-BVP (8.77) using "Bessel functions." The ODE in (8.77) is known as **Bessel's equation of order zero** and its solutions, $J_0(x)$ and $Y_0(x)$, are known with as much accuracy as the trigonometric functions sin and cos are known. The general solution of ODE $y'' + \frac{1}{x}y' + y = 0$ is $y(x) = c_1 J_0(x) + c_2 Y_0(x)$, where c_1, c_2 are arbitrary constants. Substituting this into the boundary conditions gives

$$\left\{ \begin{array}{rl} J_0(1)c_1 + Y_0(1)c_2 = & 0 \\ J_0(3)c_1 + Y_0(3)c_2 = & -1 \end{array} \right\},$$

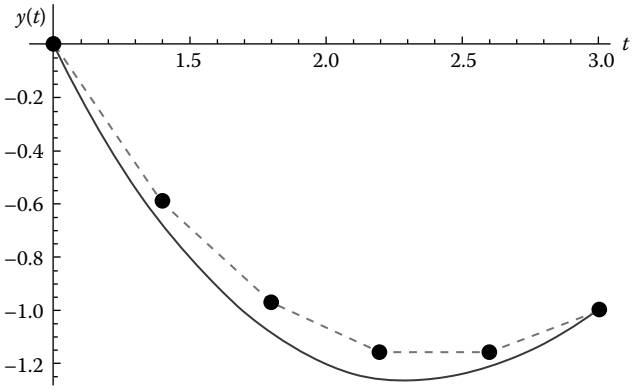


FIGURE 8.6
Numerical solution of an ODE-BVP.

a system of two algebraic equations in unknowns c_1, c_2. Using tabulated values of $J_0(1)$, $Y_0(1)$, $J_0(3)$, $Y_0(3)$, we arrive at the solution of (8.77):

$$y(x) \approx 0.2662994102\, J_0(x) - 2.469811754\, Y_0(x).$$

By the way, we chose $N = 5$ so that the matrix A would be large enough to demonstrate the details of the implementation of the finite difference replacement equations (8.79).

The approximate solution using $N = 5$ differs from the exact solution by at most 17%, which is not very bad. Ideally, we should get a numerical solution for first $N = 5$ and then $N = 10$ and then compare at the common abscissas $x = 1.4, 1.8, 2.2, 2.6$, then $N = 20$, etc., and apply a stopping rule if the numerical solutions appear to change little if h is halved.

8.8.1 An ODE-BVP Eigenvalue Problem

As we will see in Chapter 9 and especially in Chapter 11, the ODE-BVP eigenvalue problem

$$\left\{ \begin{array}{c} X'' + \lambda X = 0 \\ X(0) = 0, X(\pi) = 0 \end{array} \right\} \tag{8.80}$$

plays an important role in solving partial differential equations. By an **eigenvalue** of ODE-BVP (8.80), we mean a value of λ for which (8.80) has, on the interval $[0, \pi]$, a solution $X(x)$ that is not identically zero, that is, a **nontrivial** solution. That solution $X(x)$ is called a corresponding **eigenfunction**. As we will see in Section 9.3, the eigenvalues of (8.80) are $\lambda = 1^2, 2^2, 3^2, \ldots$, that is, $\lambda_n = n^2$ for $n = 1, 2, 3, \ldots$, and the corresponding eigenfunctions are $X_n(x) = \sin nx$ or nonzero multiples thereof.

Example 8.20

Find approximate eigenvalues and corresponding eigenfunctions of ODE-BVP (8.80) using finite differences with $N = 5$.

Method: Let $h = \frac{\pi}{5}$. For $i = 0, 1, 2, \ldots, 5$, define $x_i = ih$ and $X_i = X(x_i)$. The boundary conditions in (8.80) are replaced by $X_0 = X_5 = 0$, and the ODE is replaced by

$$h^{-2}\left(X_{i+1} - 2X_i + X_{i-1}\right) + \lambda X_i = 0,\ i = 1, \ldots, 4.$$

Define $\mathbf{X} = [\, X_1 \quad X_2 \quad X_3 \quad X_4 \,]^T$. After using the replacement boundary conditions, the system of ODE replacement equations can be written as the **discretized eigenvalue problem**

$$A\mathbf{X} + h^2\lambda\, \mathbf{X} = \mathbf{0}, \tag{8.81}$$

where the famous tridiagonal matrix is

$$A = \begin{bmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \end{bmatrix}. \tag{8.82}$$

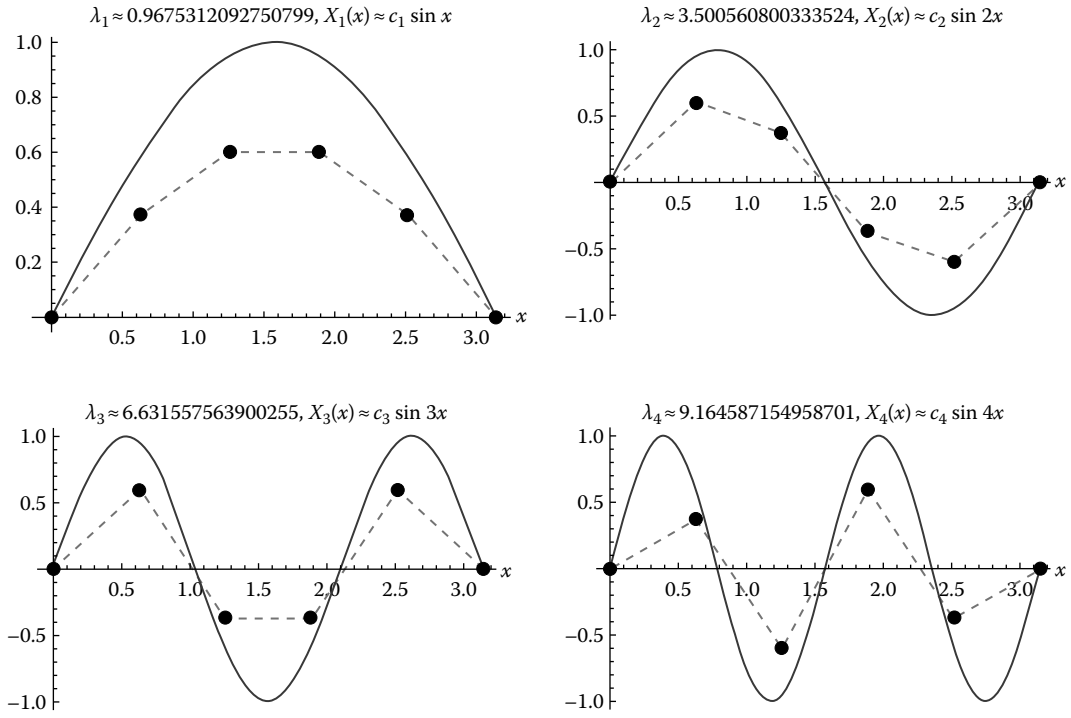


FIGURE 8.7
Numerical eigenvalues and eigenfunctions.

So, finding an approximate eigenvalue of the ODE-BVP (8.80) is equivalent to finding an eigenvalue of the linear algebra problem (8.81). We did that using the MATLAB command [V, D] = eig(A). The results for the four approximate eigenvalues are

$$\lambda_1 \approx 0.9675312092750799, \quad X_1(x) \approx c_1 \sin x,$$

$$\lambda_2 \approx 3.5005608003335249, \quad X_2(x) \approx c_2 \sin 2x,$$

$$\lambda_3 \approx 6.631557563900255, \quad X_3(x) \approx c_3 \sin 3x,$$

$$\lambda_4 \approx 9.164587154958701, \quad X_4(x) \approx c_4 \sin 4x,$$

and the corresponding numerical eigenfunctions (dashed) curves) are compared pictorially to the exact eigenfunctions (solid curves) in Figure 8.7. While the numerical eigenfunctions seem to be shrunken versions of the exact eigenfunctions, this is an artifact of the coefficients c_i. We used *Mathematica* to draw those pictures, and of course, we could have used it to find the approximate eigenvalues, too. ○

8.8.2 Using IVP Numerical Methods to Solve BVPs

The finite difference method appears to work pretty well, especially if $p(x)$, $q(x)$, the coefficients in the second-order ODE, $y'' + p(x)y' + q(x) = f(x)$, and $f(x)$, the right-hand side, are not too wildly varying functions. Even if the ODE is nonlinear, say

$$y''(x) = f\big(x, y(x), y'(x)\big),$$

we can still get finite difference replacement equations, for example,

$$h^{-2}\left(y_{i+1} - 2y_i + y_{i-1}\right) = f\left(x_i, y_i, (2h)^{-1}\left(y_{i+1} - y_{i-1}\right)\right),$$

using central difference approximations, but in general they will be a system of nonlinear equations. As we saw in Section 8.2, solving a system of nonlinear equations can be slow.

Yet another general approach is to exploit modern software packages capabilities for solving IVPs. For example, suppose we want to find an approximate numerical solution of

$$\left\{ \begin{array}{c} y'' = f(x, y, y') \\ y(a) = \alpha, y(b) = \beta \end{array} \right\}.$$

Shooting methods numerically solve the IVP

$$\left\{ \begin{array}{c} y'' = f(x, y, y') \\ y(a) = \alpha, y'(a) = z \end{array} \right\}$$

to get "solution" $y(x; z)$ and then use a numerical method to solve the scalar nonlinear equation

$$y(b; z) = \beta$$

for $z = z^\star$. The numerical solution of the original ODE-BVP is then $y(x; z^\star)$.

8.8.3 Finding Periodic Solutions of Linear Problems

Suppose a system of ODEs in $\mathbb{R}^n$ given by

$$\dot{\mathbf{x}} = A(t; \boldsymbol{\lambda})\mathbf{x} + \mathbf{f}(t) \tag{8.83}$$

is periodic with period T, that is, $A(t + \mathsf{T}; \boldsymbol{\lambda}) \equiv A(t; \boldsymbol{\lambda})$ and $\mathbf{f}(t + \mathsf{T}) \equiv \mathbf{f}(t)$. We denote the solutions of (8.83) by

$$\mathbf{x} = \mathbf{x}(t; \boldsymbol{\lambda}, \mathbf{x}_0)$$

to indicate that the solution depends on both the parameter λ or vector of parameters $\boldsymbol{\lambda}$ and the initial data $\mathbf{x}_0$. We studied such systems in Section 5.8, albeit often without parameters $\boldsymbol{\lambda}$.

If we want to find a value of $\boldsymbol{\lambda}$ for which (8.83) has a periodic solution, we can approach the problem by using an IVP numerical method to find numerical approximations of the n basic solutions

$$\mathbf{x}^{(1)}(t; \boldsymbol{\lambda}) = \mathbf{x}(t; \boldsymbol{\lambda}, \mathbf{e}^{(1)}), \quad \cdots \quad , \mathbf{x}^{(n)}(t; \boldsymbol{\lambda}) = \mathbf{x}(t; \boldsymbol{\lambda}, \mathbf{e}^{(n)})$$

of the corresponding homogeneous system $\dot{\mathbf{x}} = A(t; \boldsymbol{\lambda})\mathbf{x}$. After that, we assemble them into a numerical approximation of the principal fundamental matrix

$$X(t; \boldsymbol{\lambda}) = \left[\mathbf{x}^{(1)}(t; \boldsymbol{\lambda}) \;\vdots\; \cdots \;\vdots\; \mathbf{x}^{(n)}(t; \boldsymbol{\lambda})\right]$$

and use the method of variation of constants to find the solutions of the original ODE system (8.83). According to Theorem 5.26 in Section 5.8 (noncritical systems), if $\det\big(I - X(\mathsf{T})\big) \neq 0$, then (8.83) has exactly one periodic solution given by (5.115) in Section 5.8, that is,

$$\mathbf{x}(t) = X(t)\left(\big(I - X(\mathsf{T})\big)^{-1} X(\mathsf{T})\left(\int_0^{\mathsf{T}} (X(s))^{-1}\,\mathbf{f}(s)\,ds\right) + \int_0^{t} (X(s))^{-1}\,\mathbf{f}(s)\,ds\right).$$

In the latter's notation, we have suppressed the dependence on $\boldsymbol{\lambda}$.

In the case when $\mathbf{f}(t) \equiv \mathbf{0}$, that is, the system is homogeneous, it has a periodic solution if, and only if, the Hill's discriminant has

$$\mathsf{H}(\boldsymbol{\lambda}) \triangleq \det\big(I - X(\mathsf{T};\boldsymbol{\lambda})\big) = 0.$$

This and related problems come up in studying electronic behavior of crystals. This is a single equation to be solved for $\boldsymbol{\lambda}$, so we can use the secant method to get an approximate value(s) of $\boldsymbol{\lambda}$ for which the homogeneous system $\dot{\mathbf{x}} = A(t;\boldsymbol{\lambda})\mathbf{x}$ has a periodic solution. In effect we're using a shooting method for a linear problem. Shooting methods are discussed in the references found in the "Learn More About It" at the end of Section 8.2.

We (Turyn 1987) used this method to find an approximate **double eigenvalue** for the ODE $\ddot{y} + \pi^2\big(\lambda_0 + \lambda_1\cos(2\pi t) + 2\cos(4\pi t)\big)y = 0$, that is, a value for the parameter vector $\boldsymbol{\lambda} = [\lambda_0 \quad \lambda_1]^T$ such that the ODE has two linearly independent eigenfunctions that are periodic with period $\mathsf{T} = 1$.

8.8.4 Problems

In problems 1–5 we are given a two-point boundary value problem, where $' = \frac{d}{dx}$, and a step size, h.

(a) Find a system of replacement linear algebraic equations. For the second derivative, use the usual central difference approximation, and for a first derivative in the *ODE*, use the central difference approximation $y'(x_i) \approx \frac{1}{2h}\big(y_{i+1} - y_{i-1}\big)$.
(b) Solve that system and graph the piecewise linear approximate solution.
(c) If possible, find the exact solution, graph it, and compare it to your approximate solution in part (b).

1. $\ddot{y} + 2\sqrt{2}\cos\left(\frac{2\pi t}{3}\right)y = 0$, $y(0) = 0$, $y(1) = -1$, and $h = 0.25$.
2. $y'' + xy' - 2y = \sin(\pi x)$, $y(0) = 0$, $y(1) = 3$, and $h = 0.25$.
3. $y'' + y = x$, $y(0) = -1$, $y(2) = 5$, and $h = 0.5$.
4. $y'' + 2y = 0$, $y'(0) = 0$, $y(1) = 5$, and $h = 0.25$. [For the boundary condition at $x = 0$, use the central difference approximation $y'(0) \approx \frac{1}{2h}\big(y(h) - y(-h)\big)$.]
5. $y'' + 2y = 0$, $y'(0) = 3$, $y(1) = 5$, and $h = 0.25$. [For the boundary condition at $x = 0$, use the central difference approximation $y'(0) \approx \frac{1}{2h}\big(y(h) - y(-h)\big)$.]
6. For the ODE-BVP $(\star)$ $\ddot{y} + \dot{y} = t$, $y(0) = -1$, $y(1) = 1$, and $h = 0.25$, use the central difference approximation of $\ddot{y}$.

(a) For the first derivative term, use the central difference approximation $\dot{y}(t_i) \approx \frac{1}{2h}\left(y_{i+1} - y_{i-1}\right)$, find the system of replacement linear algebraic equations, solve them, and graph the piecewise linear approximate solution.

(b) Find and graph the exact solution of ($\star$) and compare it to the approximate solution from part (a).

(c) For the first derivative term, use the forward difference approximation $\dot{y}(t_i) \approx \frac{1}{h}\left(y_{i+1} - y_i\right)$, find the system of replacement linear algebraic equations for ($\star$), solve them, graph the piecewise linear approximate solution, and compare it to the approximate solution from part (a) and the exact solution from part (b).

7. For the two-point boundary value problem, $y'' = -x$, $y(0) - y'(0) = 1$, and $y(1) = 0$, where $' = \frac{d}{dx}$,
   (a) Find a system of replacement linear algebraic equations using step size, $h = 0.25$. For the second derivative, use the usual central difference approximation. For the first derivative in the boundary condition, use the *backward* difference approximation, $y'(0) \approx h^{-1}(y(0) - y(-h)) \approx h^{-1}(y_0 - y_{-1})$. Because this equation involves y_{-1}, also include the replacement equation for the ODE at x_0, that is, $h^{-2}(y_1 - 2y_0 + y_{-1}) = -x_0 = 0$.
   (b) Solve that system and graph the piecewise linear approximate solution.
   (c) If possible, find the exact solution, graph it, and compare it to your approximate solution in part (b).
8. Find three approximate eigenvalues and corresponding eigenfunctions of ODE-BVP $r^2R'' + rR' + \lambda R = 0$, $R(1) = R(3) = 0$, where $' = \frac{d}{dr}$, using finite differences with $N = 4$. Compare the approximate eigenvalues and corresponding eigenfunctions to the exact eigenvalues and eigenfunctions.
9. Find three approximate eigenvalues and corresponding eigenfunctions of ODE-BVP $R'' + \frac{2}{r+1}R' + \lambda\frac{1}{(r+1)^4}R = 0$, $R(0) = R(1) = 0$, where $' = \frac{d}{dr}$, using finite differences with $N = 4$.
10. If we use finite differences with $h = \frac{b-a}{N}$ to study an eigenvalue problem $y'' + p(x)y' + (q(x) + \lambda)y = 0$, $y(a) = y(b) = 0$, how many approximate eigenvalues should we expect to find?

8.9 Splines

Splines are functions that are used to (1) approximate curves and surfaces and (2) approximate solutions of differential equations. What is particularly nice about the approximation is that it has a certain degree of smoothness. For example, "cubic" splines will be twice continuously differentiable, and this makes them good for approximating a solution of a second-order ODE-BVP. So, splines are an important part of the modern toolbox of mathematical modeling of physical problems.

One overall characteristic of spline methods is that they can express an approximation of a function using a linear combination of basis functions. As such, the coefficients in the linear combination are a highly "compressed" representation of the function.

The simplest spline is the piecewise linear approximation, or "connecting the dots" on the graph of a function. This is also effectively what we used to approximate a definite integral using the trapezoidal rule illustrated in Figure 8.4.

Throughout this section we will assume that $a = x_0 < x_1 < \cdots < x_N = b$ is a partition of the interval $[a, b]$.

Definition 8.6

(a) A function $g(x)$ is **piecewise polynomial** on an interval $[a, b]$ if there is a partition of $[a, b]$ and, on each subinterval $[x_{j-1}, x_j]$, $g(x)$ is a polynomial.

(b) Suppose $k \geq 2$. A **spline** of **order** k on an interval $[a, b]$ is a function $g(x)$ that is (i) $k - 2$ times continuously differentiable on $[a, b]$, (ii) piecewise polynomial, and (iii) has each polynomial piece of degree less than or equal to $k - 1$.

Note that a spline of order 2 *is only required to be continuous on* $[a, b]$ *and piecewise linear.*

Suppose we have data points $(x_j, y_j) = (x_j, f(x_j))$ on the graph of $y = f(x)$. Define

$$L_N(x) \triangleq \left\{ \begin{array}{ll} y_0 + m_1(x - x_0), & x_0 \leq x \leq x_1 \\ y_1 + m_2(x - x_1), & x_1 \leq x \leq x_2 \\ y_2 + m_3(x - x_2), & x_2 \leq x \leq x_3 \\ \vdots & \\ y_{N-1} + m_N(x - x_{N-1}), & x_{N-1} \leq x \leq x_N \end{array} \right\}, \tag{8.84}$$

where $m_j = \frac{y_j - y_{j-1}}{h}$, for $i = 1, \ldots, N$. The piecewise linear function $L_N(x)$ is illustrated in Figure 8.8.

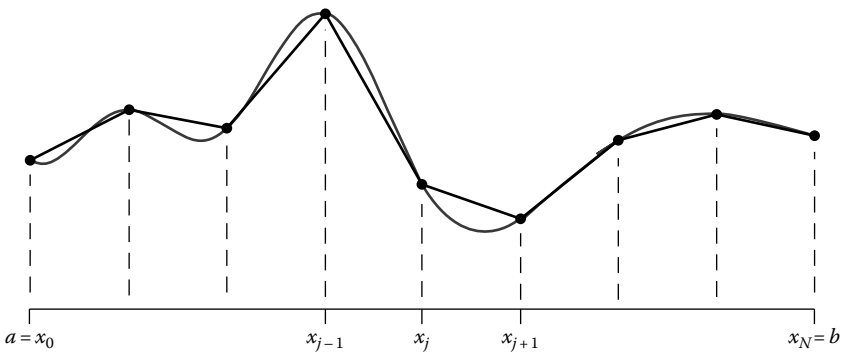


FIGURE 8.8
Approximation of a function using uniform tent–spline functions.

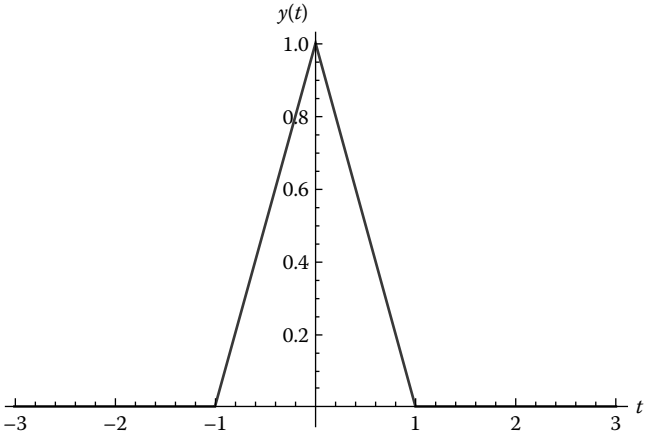


FIGURE 8.9
Tent–spline function.

In this very brief introduction to the subject, we will emphasize the **uniform** case where the abscissas are equally spaced, that is, there is a step size h with $x_j - x_{j-1} = h$ for $j = 1, \ldots, N$.

We will express the piecewise linear approximation of a function as a linear combination of basis functions. This will give a simple introduction to more complicated, powerful spline methods involving higher-order polynomials. Define a **tent function** $\phi(t)$ by

$$\phi(x) \triangleq \left\{ \begin{array}{ll} (x+1), & -1 \le x \le 0 \\ (1-x), & 0 \le x \le 1 \\ 0, & |x| \ge 1 \end{array} \right\} = \left\{ \begin{array}{ll} b_{-1}(x+1), & -1 \le x \le 0 \\ b_0(x), & 0 \le x \le 1 \\ 0, & |x| \ge 1 \end{array} \right\}, \tag{8.85}$$

where $b_{-1}(x) \triangleq x$, $b_0(x) \triangleq 1 - x$. Its graph is shown in Figure 8.9, and has the interesting property that

$$b_{-1}(x) + b_0(x) \equiv 1, \text{ for } 0 < x \le 1.$$

After the **horizontal translation** $x \mapsto (x - x_j)$, followed by the **dilation** $x \mapsto \frac{x}{h}$, we get the **uniform tent basis functions**

$$T_j(x) \triangleq \left\{ \begin{array}{ll} \left(\frac{x-x_j}{h} + 1\right), & x_{j-1} \le x \le x_j \\ \left(1 - \frac{x-x_j}{h}\right), & x_j \le x \le x_{j+1} \\ 0, & |x - x_j| \ge h \end{array} \right\} = \phi\left(\frac{x - x_j}{h}\right), \tag{8.86}$$

for $j = 0, \ldots, N+1$, where we define $x_{-1} = a - h$ and $x_{N+1} = b + h$.

Because (i) $T_j(x_j) = 1$ and (ii) $T_i(x_j) = 0$ for $i \neq j$, we can approximate $f(x)$ on the interval $[a, b]$ using

$$f(x) \approx f_N(x) \triangleq \sum_{j=0}^{N} y_j T_j(x). \tag{8.87}$$

In Problem 8.9.6.6 you will explain why (8.87) is actually the same as the piecewise linear approximation in (8.84).

Definition 8.7

A function $g(x)$ has **support interval** $\mathcal{I}$, or **support** for short, if (a) $g(x) = 0$ for all x *not* in $\mathcal{I}$ and (b) $\mathcal{I}$ is the smallest such closed interval satisfying (a).

The function $T_j(x)$ has support $[x_{j-1}, x_{j+1}]$. So, when evaluating $f_N(x)$ at an x in the interval $[a, b]$, at most two of the functions $T_j(x)$ will be nonzero in the sum. For example, if $x_{j-1} \leq x \leq x_j$, then

$$f_N(x) = y_{j-1} T_{j-1}(x) + y_j T_j(x).$$

The function $\phi(x)$ is continuous on the real line; in particular, $\phi(0^-) = \phi(0^+)$. So, each of the basis functions $T_j(x)$ is continuous on the real line, and thus, $f_N(x)$ is continuous.

8.9.1 Cubic B-Splines

Define a function by

$$\psi(x) \triangleq \tfrac{1}{6} \left\{ \begin{array}{ll} (x+2)^3, & -2 \leq x \leq -1 \\ -3(x+1)^3 + 3(x+1)^2 + 3(x+1) + 1, & -1 \leq x \leq 0 \\ -3(1-x)^3 + 3(1-x)^2 + 3(1-x) + 1, & 0 \leq x \leq 1 \\ (2-x)^3, & 1 \leq x \leq 2 \\ 0, & |x| \geq 2 \end{array} \right\} \tag{8.88}$$

$$= \left\{ \begin{array}{ll} b_{-2}(x+2), & -2 \leq x \leq -1 \\ b_{-1}(x+1), & -1 \leq x \leq 0 \\ b_0(x), & 0 \leq x \leq 1 \\ b_1(x-1), & 1 \leq x \leq 2 \\ 0, & |x| \geq 2 \end{array} \right\},$$

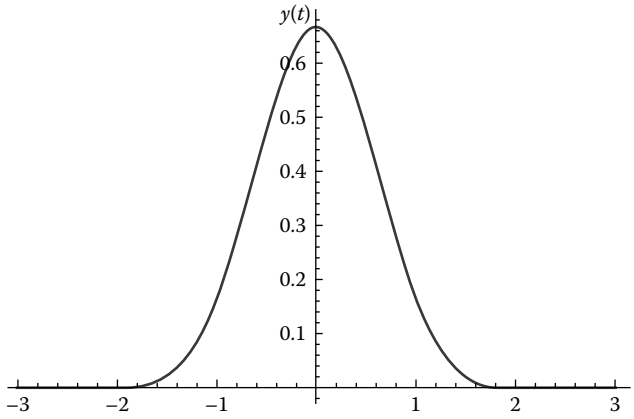


FIGURE 8.10
Cubic B-spline.

where

$$\left\{\begin{array}{l} b_{-2}(x) \triangleq \frac{1}{6}\,x^3, \\ b_{-1}(x) \triangleq \frac{1}{6}\left(-3x^3+3x^2+3x+1\right), \\ b_0(x) \triangleq \frac{1}{6}\left(3x^3-6x^2+4\right), \\ b_1(x) \triangleq \frac{1}{6}\left(-x^3+3x^2-3x+1\right) = \frac{1}{6}(1-x)^3. \end{array}\right\} \tag{8.89}$$

The graph of $\psi(x)$ is shown in Figure 8.10. It is called a **cubic B-spline** or a **blending function** and has the interesting property that

$$\psi(x-2)+\psi(x-1)+\psi(x)+\psi(x+1) = b_{-2}(x)+b_{-1}(x)+b_0(x)+b_1(x) \equiv 1, \text{ for } 0 < x \le 1.$$

After the **horizontal translation** $x \mapsto (x - x_j)$, followed by the **dilation** $x \mapsto \frac{x}{h}$, we get the **cubic spline uniform basis functions**

$$C_j(x) \triangleq \left\{\begin{array}{ll} b_{-2}\left(\frac{x-x_j}{h}+2\right), & x_{j-2} \le x \le x_{j-1} \\ b_{-1}\left(\frac{x-x_j}{h}+1\right), & x_{j-1} \le x \le x_j \\ b_0\left(\frac{x-x_j}{h}\right), & x_j \le x \le x_{j+1} \\ b_1\left(\frac{x-x_j}{h}-1\right), & x_{j+1} \le x \le x_{j+2} \\ 0, & |x-x_j| \ge 2h \end{array}\right\} = \psi\left(\frac{x-x_j}{h}\right), \tag{8.90}$$

for $j = -1, 0, \ldots, N+1$, where we define $x_{-1} = a - h$ and $x_{N+1} = b + h$.

In Problem 8.9.6.8 you will check that the function $\psi(x)$ is twice continuously differentiable on the real line. In particular, you will calculate that

$$\psi(k^+) = \psi(k^-),\ \psi'(k^+) = \psi'(k^-),\ \psi''(k^+) = \psi''(k^-),\ \text{for } k = -2, -1, -, 0, 1, 2. \tag{8.91}$$

It follows that each of the basis functions $C_j(x)$ is twice continuously differentiable on the real line, as is required of splines of order four.

We will see how to approximate a given function $f(x)$ using a linear combination of the cubic B-splines:

$$f(x) \approx f_N(x) \triangleq \sum_{j=-1}^{N+1} z_j C_j(x). \tag{8.92}$$

We include the functions $C_{-1}(x)$ because it is twice continuously differentiable at $x = x_0$ and contributes to the sum information about $f(x)$ and its first two derivatives there. Similarly we include $C_{N+1}(x)$ in order to use as much information as possible about $f(x)$ and its first two derivatives at $x = x_N$.

To solve for the z_j, it helps to calculate that

$$\begin{gathered}
\psi(-2) = 0,\ \psi(-1) = \tfrac{1}{6},\ \psi(0) = \tfrac{2}{3},\ \psi(1) = \tfrac{1}{6},\ \psi(2) = 0, \\
\psi'(-2) = 0,\ \psi'(-1) = \tfrac{1}{2},\ \psi'(0) = 0,\ \psi'(1) = -\tfrac{1}{2},\ \psi'(2) = 0, \\
\psi''(-2) = 0,\ \psi''(-1) = 1,\ \psi''(0) = -2,\ \psi''(1) = 1,\ \psi''(2) = 0.
\end{gathered}$$

By the chain rule, it follows that

$$\begin{gathered}
C_j(x_{j-2}) = 0,\ C_j(x_{j-1}) = \tfrac{1}{6},\ C_j(x_j) = \tfrac{2}{3}, C_j(x_{j+1}) = \tfrac{1}{6},\ C_j(x_{j+2}) = 0, \\
C_j'(x_{j-2}) = 0,\ C_j'(x_{j-1}) = \tfrac{1}{2h},\ C_j'(x_j) = 0, C_j'(x_{j+1}) = -\tfrac{1}{2h},\ C_j'(x_{j+2}) = 0, \\
C_j''(x_{j-2}) = 0,\ C_j''(x_{j-1}) = \tfrac{1}{h^2},\ C_j''(x_j) = -\tfrac{2}{h^2}, C_j''(x_{j+1}) = \tfrac{1}{h^2},\ C_j''(x_{j+2}) = 0.
\end{gathered} \tag{8.93}$$

In particular, at $x = x_j$, at most three of the terms in (8.92) are not zero:

$$f_N(x_j) \triangleq \sum_{j=-1}^{N+1} z_j C_j(x_j) = \frac{1}{6} z_{j-1} + \frac{2}{3} z_j + \frac{1}{6} z_{j+1}. \tag{8.94}$$

As an aside, note that this can be written in terms of the functions $b_{-2}(x), \ldots, b_1(x)$:

$$f_N(x_j) = b_{-2}(0)\, z_{j-2} + b_{-1}(0)\, z_{j-1} + b_0(0)\, z_j + b_1(0)\, z_{j+1} + b_1(1)\, z_{j+2}.$$

Similarly to (8.93) we find

$$f_N'(x_j) \triangleq \sum_{i=-1}^{N+1} z_i C_i'(x_j) = z_{j-1} C_{j-1}'(x_j) + z_{j+1} C_{j+1}'(x_j) = \frac{1}{2h} \cdot (-z_{j-1} + z_{j+1}), \tag{8.95}$$

and

$$f_N''(x_j) \triangleq \sum_{i=-1}^{N+1} z_i C_i{}'(x_j) = z_{j-1}C_{j-1}''(x_j) + z_j C_j''(x_j) + z_{j+1}C_{j+1}''(x_j) = \frac{1}{h^2} \cdot (z_{j-1} - 2z_j + z_{j+1}). \tag{8.96}$$

It follows that

$$\begin{bmatrix} f_N''(x_0) \\ f_N''(x_1) \\ f_N''(x_2) \\ \vdots \\ f_N''(x_{N-1}) \\ f_N''(x_N) \end{bmatrix} = \frac{1}{h^2} \begin{bmatrix} 1 & -2 & 1 & 0 & \cdot & \cdot & \cdot & & & 0 \\ 0 & 1 & -2 & 1 & & & & & & \cdot \\ \cdot & 0 & 1 & -2 & \cdot & & & & & \cdot \\ \cdot & & & \cdot & \cdot & \cdot & & & & \cdot \\ \cdot & & & & \cdot & \cdot & \cdot & & & \\ & & & & & \cdot & 1 & -2 & 1 & 0 \\ 0 & \cdot & \cdot & \cdot & & & 0 & 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} z_{-1} \\ z_0 \\ z_1 \\ \vdots \\ z_{N-1} \\ z_N \\ z_{N+1} \end{bmatrix}$$

and

$$\begin{bmatrix} f_N'(x_0) \\ f_N'(x_1) \\ f_N'(x_2) \\ \vdots \\ f_N'(x_{N-1}) \\ f_N'(x_N) \end{bmatrix} = \frac{1}{2h} \begin{bmatrix} -1 & 0 & 1 & 0 & \cdot & \cdot & \cdot & & & 0 \\ 0 & -1 & 0 & 1 & & & & & & \cdot \\ \cdot & 0 & -1 & 0 & \cdot & & & & & \cdot \\ \cdot & & & \cdot & \cdot & \cdot & & & & \cdot \\ \cdot & & & & \cdot & \cdot & \cdot & & & \\ & & & & & \cdot & -1 & 0 & 1 & 0 \\ 0 & \cdot & \cdot & \cdot & & & 0 & -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} z_{-1} \\ z_0 \\ z_1 \\ \vdots \\ z_{N-1} \\ z_N \\ z_{N+1} \end{bmatrix}.$$

Example 8.21

Find the approximate solution of ODE-BVP (8.77) in Section 8.8, that is,

$$\left\{ \begin{array}{c} y'' + \frac{1}{x}y' + y = 0 \\ y(1) = 0, y(3) = -1 \end{array} \right\},$$

using cubic B-splines for $N = 3$.

Method: For $N = 3$, $h = \frac{3-1}{3}$, and our approximate solution is

$$y_N(x) \triangleq \sum_{j=-1}^{4} z_j C_j(x),$$

where we will solve for the constants $z_{-1}, z_0, \ldots, z_4$. Using (8.94), the boundary conditions require

$$0 = y(1) = y_N(x_0) = \frac{1}{6}\left(z_{-1} + 4z_0 + z_1\right) \tag{8.97}$$

and

$$-1 = y(3) = y_N(x_3) = \frac{1}{6}\left(z_2 + 4z_3 + z_4\right). \tag{8.98}$$

Substituting $h = \frac{2}{3}$ and $N = 3$ into (8.94)–(8.96), the replacement equations for the ODE are, for $j = 0, 1, 2, 3$,

$$\frac{9}{4}\left(z_{j-1} - 2z_j + z_{j+1}\right) + \frac{1}{x_j}\cdot\frac{3}{4}\left(-z_{j-1} + z_{j+1}\right) + \frac{1}{6}\left(z_{j-1} + 4z_j + z_{j+1}\right) = 0. \tag{8.99}$$

We include replacement equations for the ODE at the endpoints $x = x_0$ and $x = x_N$ because, unlike the finite difference methods we used in Example 8.19 in Section 8.8, the spline functions are twice continuously differentiable at the endpoints.

Altogether, (8.97)–(8.99) give us the system

$$\begin{bmatrix}
\frac{1}{6} & \frac{4}{6} & \frac{1}{6} & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{6} & \frac{4}{6} & \frac{1}{6} \\
\frac{9}{4} - \frac{9}{4} + \frac{1}{6} & -\frac{18}{4} + \frac{4}{6} & \frac{9}{4} + \frac{9}{4} + \frac{1}{6} & 0 & 0 & 0 \\
0 & \frac{9}{4} - \frac{9}{4}\cdot\frac{3}{5} + \frac{1}{6} & -\frac{18}{4} + \frac{4}{6} & \frac{9}{4} + \frac{9}{4}\cdot\frac{3}{5} + \frac{1}{6} & 0 & 0 \\
0 & 0 & \frac{9}{4} - \frac{9}{4}\cdot\frac{3}{7} + \frac{1}{6} & -\frac{18}{4} + \frac{4}{6} & \frac{9}{4} + \frac{9}{4}\cdot\frac{3}{7} + \frac{1}{6} & 0 \\
0 & 0 & 0 & \frac{9}{4} - \frac{9}{4}\cdot\frac{3}{9} + \frac{1}{6} & -\frac{18}{4} + \frac{4}{6} & \frac{9}{4} + \frac{9}{4}\cdot\frac{3}{9} + \frac{1}{6}
\end{bmatrix}$$

$$\times \begin{bmatrix} z_{-1} \\ z_0 \\ z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{8.100}$$

The solution of (8.100) is

$$z_{-1} \approx 1.71275633, z_0 \approx -0.155705051, z_1 \approx -1.08993554, z_2 \approx -1.35065140$$

$$z_3 \approx -1.05687409, z_4 \approx -0.421852245. \bigcirc$$

While we could use a calculator to solve (8.100), it is convenient to use *Mathematica* or MATLAB to enter the correct 36 entries in the matrix of coefficients.

The solid curve in Figure 8.11 is the exact solution, as discussed in Example 8.19 in Section 8.8. The thick dashed curve is the solution using cubic B-splines. The fact that the latter solution is twice continuously differentiable is apparent as well as the excellent agreement with the exact solution.

In Example 8.21 we substituted the approximate solution into the ODE at a finite number of points x_j. This is called a **collocation method**.

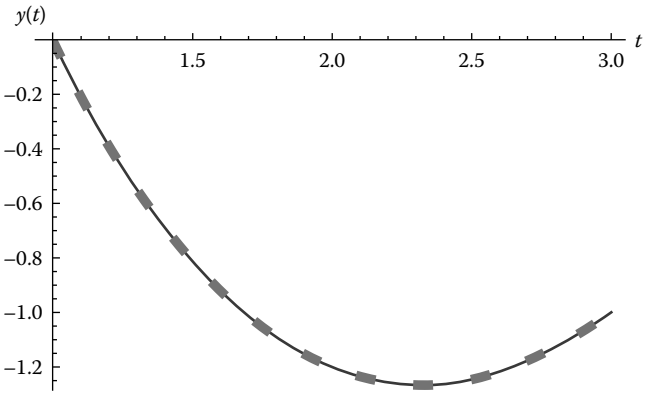


FIGURE 8.11
Solution of Example 8.21 using cubic B-splines.

8.9.2 Nonuniform Splines

The partition $a = x_0 < x_1 < \cdots < x_N = b$ need not be equally spaced. Define $h_j = x_j - x_{j-1}$ for $i = 1, \ldots, N$ and the usual data points $(x_j, y_j) = (x_j, f(x_j))$ on the graph of $y = f(x)$. We get the **tent basis functions**

$$T_j(x) \triangleq T(x - jh) = \left\{ \begin{array}{ll} \left(\dfrac{x - x_{j-1}}{x_j - x_{j-1}} \right), & x_{j-1} \le x \le x_j \\ \left(1 - \dfrac{x - x_j}{x_{j+1} - x_j} \right), & x_j \le x \le x_{j+1} \\ 0, & |x - x_j| \ge h \end{array} \right\}. \tag{8.101}$$

The uniform tent basis splines in (8.86) are a special case of (8.101), as you will verify in Problem 8.9.6.10.

We can still approximate $f(x)$ by a piecewise linear function

$$f(x) \approx f_N(x) \triangleq \sum_{j=0}^{N} y_j T_j(x), \tag{8.102}$$

as illustrated in Figure 8.12. As in the uniform case, when evaluating $f_N(x)$ at an x in the interval $[a, b]$, at most two of the functions $T_j(x)$ will be nonzero in the sum. For example, if $x_{j-1} \le x \le x_j$, then

$$f_N(x) = y_{j-1} T_{j-1}(x) + y_j T_j(x).$$

In Problem 8.9.6.9, you will explain why (8.102) is actually the same as the piecewise linear approximation in (8.84), even when the partition is not equally spaced.

Nonuniform tent–spline functions are illustrated in Figure 8.13.

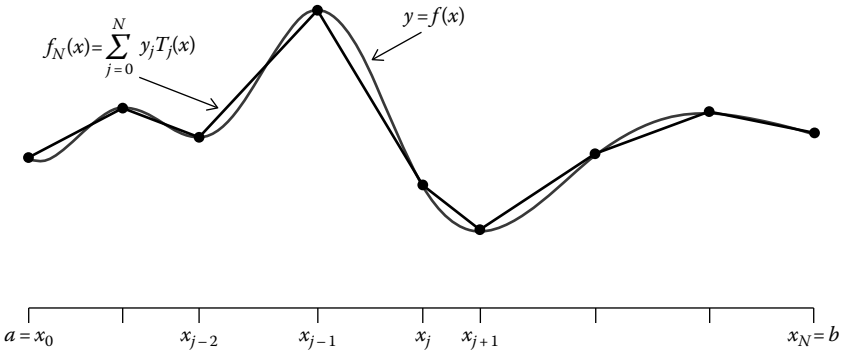


FIGURE 8.12
Approximation of a function using nonuniform tent–spline functions.

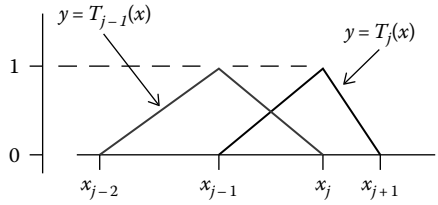


FIGURE 8.13
Nonuniform tent–spline functions $T_{j-1}(x)$ and $T_j(x)$.

Unfortunately, in the nonuniform case we cannot write $T_j(x)$ as a dilation of a translate of a simpler function, as we did in (8.86).

Indeed, the definition of higher-order splines, for example, quadratic and cubic B-splines, is much more complicated in the nonuniform case. One standard method is to use the **Cox–de Boor recursive** definition (de Boor 2001, p. 90). The spline basis functions $N_{j,k}(x)$ of order $k > 1$ satisfy

$$N_{j,k}(x) \triangleq \omega_{j,k}(x) N_{j,k-1}(x) + \big(1 - \omega_{j+1,k}(x)\big) N_{j+1,k-1}(x), \text{ for } j = 0, \ldots, N-2, \tag{8.103}$$

where

$$\omega_{j,k}(x) \triangleq \frac{x - x_j}{x_{j+k-1} - x_j}, \tag{8.104}$$

and

$$N_{j,1}(x) \triangleq \left\{ \begin{array}{ll} 1, & x_j \le x \le x_{j+1} \\ 0, & \text{all other } x \end{array} \right\}.$$

In Problem 8.9.6.7 you will explain why the recursive definition (8.103) implies that for $k = 2$,

$$N_{j,2}(x) \triangleq \left\{ \begin{array}{ll} \dfrac{x - x_j}{x_{j+1} - x_j}, & x_j \le x \le x_{j+1} \\ \dfrac{x_{j+2} - x}{x_{j+2} - x_{j+1}}, & x_{j+1} \le x \le x_{j+2} \\ 0, & \text{all other } x \end{array} \right\} = T_{j+1}(x)$$

in terms of the tent basis function we defined earlier in (8.101).

8.9.3 Spline Approximation of a Curve in $\mathbb{R}^n$

Suppose that in $\mathbb{R}^n$ we have a parametrized curve

$$\mathcal{C} : \mathbf{r} = \mathbf{r}(t), 0 \le t \le m.$$

Note that if the parameter interval is not $[0, m]$, we can change the parametrization to make it so.

The points $\mathbf{P}_0 \triangleq \mathbf{r}(0), \mathbf{P}_0 \triangleq \mathbf{r}(1), \ldots, \mathbf{P}_m \triangleq \mathbf{r}(m)$ lie on the curve $\mathcal{C}$. We will find a parametric curve that **interpolates** $\mathbf{P}_0$ and $\mathbf{P}_m$, the endpoints of the curve, but not necessarily any of the points in between.

Define "fantom" points $\mathbf{P}_{-1} = 2\mathbf{P}_0 - \mathbf{P}_1$ and $\mathbf{P}_{m+1} = 2\mathbf{P}_m - \mathbf{P}_{m-1}$.

The four points $\mathbf{P}_{-1}$, $\mathbf{P}_0$, $\mathbf{P}_1$, $\mathbf{P}_2$ determine a **B-spline curve** parametrized by

$$\mathbf{R}^{(0)}(t) \triangleq b_1(t-0)\mathbf{P}_{-1} + b_0(t-0)\mathbf{P}_0 + b_{-1}(t-0)\mathbf{P}_1 + b_{-2}(t-0)\mathbf{P}_2, 0 \le t \le 1,$$

in terms of the $b_i(t)$ functions defined in (8.89). Because of the way the fantom point $\mathbf{P}_{-1}$ was chosen, the spline curve $\mathbf{R}^{(0)}(t)$ agrees with the original curve at $t = 0$, that is, $\mathbf{R}^{(0)}(0) = \mathbf{P}_0$. In addition, the spline curve is (a) tangent to the line segment from $\mathbf{P}_0$ to $\mathbf{P}_1$, that is, $\dot{\mathbf{R}}^{(0)}(0) = \mathbf{P}_1 - \mathbf{P}_0$, at $\mathbf{P}_0$, and (b) tangent to the line segment from $\mathbf{P}_{m-1}$ to $\mathbf{P}_m$, that is, $\dot{\mathbf{R}}^{(0)}(1) = \mathbf{P}_m - \mathbf{P}_{m-1}$, at $\mathbf{P}_0$.

Next, the four points $\mathbf{P}_0, \ldots, \mathbf{P}_3$ determine a spline curve parametrized by

$$\mathbf{R}^{(1)}(t) \triangleq b_1(t-1)\mathbf{P}_0 + b_0(t-1)\mathbf{P}_1 + b_{-1}(t-1)\mathbf{P}_2 + b_{-2}(t-1)\mathbf{P}_3, 1 \le t \le 2.$$

It turns out that the "composite" curve

$$\mathbf{R}(t) = \left\{ \begin{array}{ll} \mathbf{R}^{(0)}(t), & 0 \le t \le 1 \\ \mathbf{R}^{(1)}(t), & 1 \le t \le 2 \end{array} \right\}$$

is a vector-valued spline of order four on the interval $[0, 2]$, including being twice continuously differentiable at the **join point** $\frac{1}{6}\big(\mathbf{P}_1 + 4\mathbf{P}_2 + \mathbf{P}_3\big)$ at $t = 1$.

Continuing in this way, the four points $\mathbf{P}_{i-1}, \ldots, \mathbf{P}_{i+2}$ determine a spline curve parametrized by

$$\mathbf{R}^{(i)}(t) \triangleq b_1(t-i)\mathbf{P}_{i-1} + b_0(t-i)\mathbf{P}_i + b_{-1}(t-i)\mathbf{P}_{i+1} + b_{-2}(t-i)\mathbf{P}_{i+2}, i \le t \le i+1.$$

The composite curve parametrized by

$$\mathbf{R}(t) = \begin{Bmatrix} \mathbf{R}^{(0)}(t), & 0 \le t \le 1 \\ \vdots & \vdots \\ \mathbf{R}^{(i)}(t), & i \le t \le i+1 \\ \vdots & \vdots \\ \mathbf{R}^{(m-1)}(t), & m-1 \le t \le m \end{Bmatrix}$$

is a vector-valued spline of order four on the interval $[m, 0]$. It also interpolates the endpoints using $\mathbf{R}^{(0)}(0) = \mathbf{P}_0 = \mathbf{r}(0)$ and $\mathbf{R}^{(m-1)}(m) = \mathbf{P}_m = \mathbf{r}(m)$.

Example 8.22

For "control points" $(0,1)$, $(0.9,0.9)$, $(1.2,0.1)$, $(0,0)$, $(1.2,-0.1)$, $(0.9,-0.9)$, and $(0,-1)$, find a composite curve that interpolates the first and last points in the list.

Method: Let $\mathbf{P}_0, \mathbf{P}_1, \ldots, \mathbf{P}_6$ denote the control points. Construct the fantom points $\mathbf{P}_{-1} = 2\mathbf{P}_0 - \mathbf{P}_1 = (-0.9, 1.1)$ and $\mathbf{P}_7 = 2\mathbf{P}_6 - \mathbf{P}_5 = (-0.9, -1.1)$ and then construct the composite curve's pieces $\mathbf{R}^{(0)}(t), \ldots, \mathbf{R}^{(5)}(t)$. The "control polygon" consists of the dashed line segments connecting the solid dots at $\mathbf{P}_{-1}, \mathbf{P}_0, \ldots, \mathbf{P}_7$. The results are shown in Figure 8.14. ◯

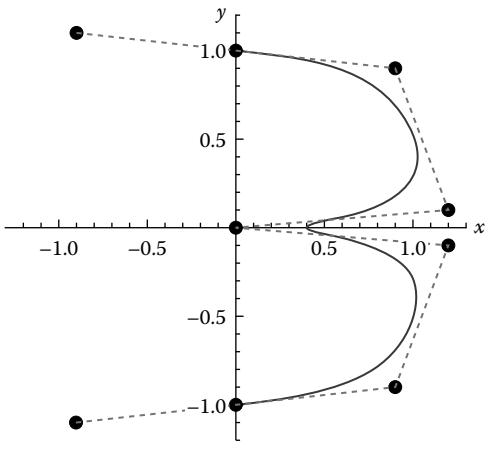


FIGURE 8.14
Composite B-spline curve.

8.9.4 Surface Splines

A parametric surface may be approximated using splines. The generalization of a control polygon is then a "control polyhedron." Given a rectangular array of control points

$$\begin{array}{ccccccc} \mathbf{P}_{00} & \mathbf{P}_{10} & \mathbf{P}_{20} & \cdot & \cdot & \cdot & \mathbf{P}_{m0} \\ \mathbf{P}_{01} & \mathbf{P}_{11} & \mathbf{P}_{21} & \cdot & \cdot & \cdot & \mathbf{P}_{m1} \\ & & \cdot & \cdot & & & \cdot \\ & & \cdot & & \cdot & & \cdot \\ & & \cdot & & & \cdot & \cdot \\ \mathbf{P}_{0n} & \mathbf{P}_{1n} & \mathbf{P}_{2n} & \cdot & \cdot & \cdot & \mathbf{P}_{mn}, \end{array}$$

we can parametrize the (i,j)th bicubic B-spline surface $\mathcal{S}^{(i,j)}$ by

$$\mathbf{R}^{(i,j)}(u,v) = \sum_{k=0}^{3}\sum_{\ell=0}^{3} b_{1-k}(u-i)b_{1-\ell}(v-j)\mathbf{P}_{i+k-1,j+\ell-1}, i \le u \le i+1, j \le v \le j+1,$$

where $1 \le i \le m-1$, $1 \le j \le n-2$. The four cubic polynomials $b_1, \ldots, b_{-2}$ were defined in (8.89).

The composite surface is the union

$$\bigcup_{i=1}^{m-2}\bigcup_{j=1}^{n-2} \mathcal{S}^{(i,j)}.$$

Similar to the work for composite cubic B-spline curves, we can introduce fantom vertices, bordering the rectangular array on all sides, so as to interpolate the boundary of the composite surface.

8.9.5 Triangular Surface Patches

The method of "finite elements" often uses an approximation of a piece of a surface by a triangular region. To understand how this relates to splines, it helps to have more background information.

Let points $\mathbf{P}_1$, $\mathbf{P}_2$, $\mathbf{P}_3$ be the three vertices of a triangle $\mathcal{T}$ in the xy–plane. Suppose the area of the triangle is A. If $\mathbf{P}$ is a point inside $\mathcal{T}$, define

$$A_i \triangleq \text{area of triangle whose vertices are} \quad \mathbf{P} \quad \text{and the} \quad \mathbf{P}_j \quad \text{for} \quad j \neq i.$$

and **areal coordinates**

$$a_i \triangleq \frac{A_i}{A}.$$

For example, a_1 is the area of the subtriangle, whose vertices are $\mathbf{P}$, $\mathbf{P}_2$, $\mathbf{P}_3$, divided by the area of the whole triangle. We have $a_1 + a_2 + a_3 = 1$ and $a_i \ge 0$ for $i = 1,2,3$. Areal

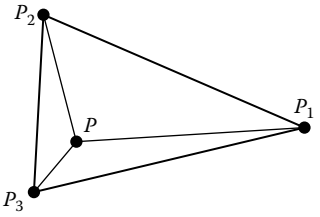


FIGURE 8.15
Triangle $\mathcal{T}$ and areal coordinates.

coordinates are the $n = 3$ version of **barycentric coordinates**. Note that for every point $\mathbf{P}$ in $\mathcal{T}$, there is a unique triple of values (a_1, a_2, a_3) for which

$$\mathbf{P} = a_1\mathbf{P}_1 + a_2\mathbf{P}_2 + a_3\mathbf{P}_3.$$

In looking at Figure 8.15, keep in mind that the triangle $\mathcal{T}$, including the point $\mathbf{P}$, lies in the xy-plane, despite the optical illusion that suggests a tetrahedron rather than the decomposition of $\mathcal{T}$ into three subtriangles.

Denote $\mathbf{a} = (a_1, a_2, a_3)$.

Denote a **multi-index** by $\mathbf{i} = (i_1, i_2, i_3)$ where i_1, i_2, i_3 are non negative integers and $i_1 + i_2 + i_3 = 1$. By convention, $0! \triangleq 1$, and we define

$$\mathbf{i}! \triangleq i_1!\, i_2!\, i_3! \quad \text{and} \quad \binom{3}{\mathbf{i}} \triangleq \frac{3!}{\mathbf{i}!} = \frac{3!}{i_1!\, i_2!\, i_3!}$$

and cubic **basis polynomials** by

$$p_{\mathbf{i}}(\mathbf{a}) \triangleq \binom{3}{\mathbf{i}} a_1^{i_1}\, a_2^{i_2}\, a_3^{i_3}.$$

There is a specific way to choose control points $\mathbf{Q}_{\mathbf{i}}$, all of which are on the boundary of the triangle except for $\mathbf{Q}_{111}$, which is at the centroid of the triangle. Additionally, $\mathbf{Q}_{300} \triangleq \mathbf{P}_1$, $\mathbf{Q}_{030} \triangleq \mathbf{P}_2$, and $\mathbf{Q}_{003} \triangleq \mathbf{P}_3$.

If $z = f(x, y)$ is a surface whose domain is $\mathcal{T}$, then the surface is interpolated by

$$z = g(\mathbf{a}) \triangleq \sum_{\mathbf{i}} f(Q_{\mathbf{i}}) p_{\mathbf{i}}(\mathbf{a}).$$

Learn More About It

Handbook on Splines for the User, by Eugene V. Shikin and Alexander I. Plis, CRC Press, Inc. © 1995, gave the results we quoted for a composite B-spline curves and surfaces. *Approximation Techniques for Engineers*, by Louis Komzsik, CRC Press, Inc. © 2007, gave the results we quoted for triangular surface patches.

8.9.6 Problems

In problems 1–5 we are given a two-point boundary value problem, where $' = \frac{d}{dx}$, and a step size, h.

(a) Use uniform cubic B-splines to find a system of replacement linear algebraic equations.
(b) Solve that system and graph the approximate solution.
(c) If possible, find the exact solution, graph it, and compare it to your approximate solution in part (b).

1. $\ddot{y} + 2\sqrt{2}\cos\left(\frac{2\pi t}{3}\right) y = 0$, $y(0) = 0$, $y(1) = -1$, and $h = 0.5$.
2. $y'' + xy' - 2y = \sin(\pi x)$, $y(0) = 0$, $y(1) = 3$, and $h = 0.5$.
3. $y'' + y = x$, $y(0) = -1$, $y(2) = 3$, and $h = 0.5$.
4. $y'' + 2y = 0$, $y'(0) = 0$, $y(1) = 5$, and $h = 0.25$.
5. $y'' = -x$, $y(0) - y'(0) = 1$, $y(1) = 0$, and $h = 0.25$.
6. Explain why (8.87) is actually the same as the piecewise linear approximation in (8.84).
7. Explain why the recursive definition (8.103) implies that for $k = 2$, $N_{j,2}(x) = T_{j+1}(x)$ in terms of the tent basis function we defined earlier in (8.101).
8. Check that the function $\psi(x)$ given in (8.88) is twice continuously differentiable on the real line by evaluating the left- and right-hand limits of $\psi(x)$, $\psi'(x)$, and $\psi''(x)$ at $x = -2, -1, 0, 1$, and 2. What is the support of $\psi(x)$?
9. Explain why (8.102) is actually the same as the piecewise linear approximation in (8.84) even when the partition is not equally spaced.
10. Explain why the uniform tent basis splines in (8.86) are a special case of (8.101).

In problems 11 and 12 construct a composite B-spline curve for the given control points.

11. $(-0.7, 0.95)$, $(-0.5, 1)$, $(0, 1)$, $(1.3, 1)$, $(0.6, 0.8)$, $(0.1, 0.33)$, $(0, -0.6)$.
12. $(0.9, 0.5)$, $(0.5, 0.9)$, $(0, 0.5)$, $(0.5, 0.3)$, $(0.9, 0.7)$, $(0.85, 0.3)$, $(0.8, 0)$.

Key Terms

absolute error: after (8.33) in Section 8.4
Adams, Adams-Bashforth method: (8.72) in Section 8.7
Adams-Moulton: after (8.72) in Section 8.7
adaptive step size: end of Section 8.7
Aitken's δ^2 method: (8.16) in Section 8.1
almost triangular matrix: Definition 8.4 in Section 8.5
areal coordinates: end of Section 8.9
B-spline curve: before Example 8.22 in Section 8.9
Banach Contraction Mapping Theorem: after Theorem 8.5 in Section 8.2
Banach spaces: after Theorem 8.5 in Section 8.2

barycentric coordinates: end of Section 8.9
basis polynomials: end of Section 8.9
Bessel's equation of order zero: after Example 8.19 in Section 8.8
big $\mathcal{O}$: Definition 8.5 in Section 8.6
bisection method: before Example 8.2 in Section 8.1
blending function: (8.89) in Section 8.9
Boole's rule: (8.50) in Section 8.6
central difference operator: (8.27) in Section 8.3
collocation method: after Example 8.21 in Section 8.9
column diagonally dominant: Definition 8.3(b) in Section 8.2
convergent: beginning of Section 8.4
convergence acceleration: before (8.16) in Section 8.1
Cox-de Boor recursive definition: (8.103) in Section 8.9
cubic B-spline: (8.89) in Section 8.9
cubic spline uniform basis functions: (8.90) in Section 8.9
deflation of polynomials: before (8.46) in Section 8.5
dilation: (8.86) in Section 8.9
direct methods: before Example 8.11 in Section 8.4
discretized eigenvalue problem: (8.81) in Section 8.8
dominant eigenvalue: before (8.43) in Section 8.5
double eigenvalue: end of Section 8.8
eigenfunction: after (8.80) in Section 8.8
eigenvalue of ODE-BVP: after (8.80) in Section 8.8
Euler's Method: (8.63) in Section 8.7
extended sign function: before Lemma 8.1 in Section 8.5
false position: (8.11) in Section 8.1
finite difference method: (8.78) in Section 8.8
fixed point problem iteration: (8.14) in Section 8.1; after Theorem 8.5 in Section 8.2
fixed point problem: before (8.14) in Section 8.1
forward difference operator: (8.48) in Section 8.6
Fréchet derivative: after Theorem 8.5 in Section 8.2
Gauss–Seidel iteration: (8.40) in Section 8.4
Gill's Method: Problem 8.7.7.23
global error: before Theorem 8.13 in Section 8.7
Hilbert matrix: Problem 8.4.3.2
horizontal translation: (8.86) in Section 8.9
Householder matrix: before Corollary 8.1 in Section 8.5
implicit partial pivoting: after Example 8.13 in Section 8.4
improved tangent method: Example 8.16 in Section 8.7
interpolates: before Example 8.22 in Section 8.9
inverse power method: after Example 8.14 in Section 8.5
iterates: before Example 8.1 in Section 8.1; after (8.19) in Section 8.2
Jacobi iteration: (8.39) in Section 8.4
join point: before Example 8.22 in Section 8.9
linear convergence: after (8.13) in Section 8.1
local error: before Theorem 8.13 in Section 8.7
lower Hessenberg form: after Definition 8.4 in Section 8.5
matrix norm induced: Definition 8.2 in Section 8.4
method of successive approximations: (8.14) in Section 8.1

midpoint approximation: (8.25) in Section 8.3
midpoint method: Example 8.16 in Section 8.7
modified Euler's method: Problem 8.7.7.17
modified (or corrected) Newton's Method: after Example 8.3 in Section 8.1
multi-index: end of Section 8.9
multistep method: before (8.72) in Section 8.7
Newton-Cotes formula of closed type: (8.29) in Section 8.3
Newton's method, Newton-Raphson method: (8.2) in Section 8.1
Newton's method in $\mathbb{R}^n$: (8.18) in Section 8.2
nontrivial solution: after (8.80) in Section 8.8
norm: Theorem 8.3 in Section 8.2
Numerical Analysis: before (8.31) in Section 8.3
numerical drift: end of Section 8.3
numerical methods: Section 8.1 beginning
order of a multistep method: before (8.72) in Section 8.7
order of a **spline**: Definition 8.6(b) in Section 8.9
order of the zero: after Example 8.3 in Section 8.1
over-relaxation methods: before (8.41) in Section 8.4
partial pivoting: before Example 8.12 in Section 8.4
perfectly conditioned: after (8.36) in Section 8.4
piecewise polynomial: Definition 8.6(a) in Section 8.9
power method: (8.43) in Section 8.5
predictor-corrector method: after (8.72) in Section 8.7
QR Algorithm: end of Section 8.5
quadratic convergence: (8.6) in Section 8.1; after (8.20) in Section 8.2
quadrature rule: after (8.28) in Section 8.3
quasi-triangular matrix: after Definition 8.4 in Section 8.5
regula falsi: (8.11) in Section 8.1
relative error: after (8.33) in Section 8.4
relaxation method: before (8.40) in Section 8.4
relaxation parameter: before (8.41) in Section 8.4
replacement equations: (8.78) in Section 8.8
round-off error: before (8.12) in Section 8.1
row diagonally dominant: Definition 8.3(a) in Section 8.4
Runge–Kutta method of order two: Example 8.16 in Section 8.7; Problem 8.7.7.17; Problem 8.7.7.18
Runge–Kutta method of order three: Problem 8.7.7.19; Problem 8.7.7.20
Runge–Kutta method of order four: (8.71) in Section 8.7, Problem 8.7.7.21
secant method: (8.11) in Section 8.1, (8.22) in Section 8.2
shift of A: after Example 8.14 in Section 8.5
shooting methods: before (8.83) in Section 8.8
similarity transformation: before (8.46) in Section 8.5
Simpson's Rule: (8.28) in Section 8.3
sparseness: before (8.39) in Section 8.4
spectral radius: after (8.39) in Section 8.4
spline: Definition 8.6(b) in Section 8.9
stable: beginning of Section 8.4
standard condition number: (8.32) in Section 8.4
step size: before (8.25) in Section 8.3

Steffensen's method: after Example 8.5 in Section 8.1
stiff: end of Section 8.7
stopping rule: (8.7) in Section 8.1
super-linear convergence: (8.13) in Section 8.1
support, **support interval**: Definition 8.7 in Section 8.9
surface splines: near end of Section 8.9
Taylor's formula method: (8.73) in Section 8.7
tent basis functions: (8.101) in Section 8.9
tent function: (8.85) in Section 8.9
trapezoidal rule: (8.27) in Section 8.3
triangular surface patches: end of Section 8.9
tri-diagonal matrix: after Definition 8.4 in Section 8.5
uniform spline: (8.84) in Section 8.9
uniform tent basis functions: (8.86) in Section 8.9
upper Hessenberg form: Definition 8.4 in Section 8.5

MATLAB® Commands

$A\backslash\mathbf{f}(\mathbf{x}_k)$: before Example 8.7 in Section 8.2
fSimpExtrap: after Example 8.10 in Section 8.3
MATLAB "function" f821: after Example 8.10 in Section 8.3
ode45: end of Section 8.7
[V, D] = eig(A): after (8.81) in Section 8.8

Mathematica Command

FindMaximum[{Abs[f''''[x]], $-\pi \leq x \leq \pi$}, {x, 0}]: Example 8.10 in Section 8.3

References

Boole, G. and Moulton, J.F. *A Treatise on the Calculus of Finite Differences*, 2nd edn. Dover, New York, 1960, p. 47.

de Boor, C. *A Practical Guide to Splines*. Springer-Verlag, New York, 2001.

Conte, S.D. and de Boor, C. *Elementary Numerical Analysis: An Algorithmie Approach*, 2nd edn. McGraw-Hill, Inc., New York, 1972.

Donovan, G.C., Miller, A.R., and Moreland, T. J. Pathological functions for Newton's method. *American Mathematical Monthly* 100 (1993), 53–58.

Rall, L.B. Convergence of the Newton process to multiple solutions. *Numerische Mathemalik* 9, 23–27, 1966.

Turyn, L. Pertubation of periodic boundary conditions. II: Bifurcation and symmetry. *Journal of Mathematical Analysis and Applications* 124, 305–326, 1987.

Weisstein, E.W. Boole's rule. From MathWorld—A Wolfram web resource. http://mathworld.wolfram.com/BoolesRule.html

9

Fourier Series

9.1 Orthogonality and Fourier Coefficients

9.1.1 Introduction

A Fourier series is a way of decomposing a function into a possibly infinite sum of "harmonic components". We call those pieces "harmonic" because they are sinusoidal functions whose frequencies are multiples of a common base frequency. Those functions are also solutions of harmonic oscillator problems. We call them components because decomposing a function in this way is like decomposing a vector into orthogonal components, as we did in Section 2.4, specifically Corollary 2.7 in Section 2.4.

Suppose $f(x)$ is a function defined only on the interval $[\,-L, L\,]$. Here, L is an unspecified positive number. The **Fourier series** "expansion," that is, decomposition into components, is given by

$$f(x) \doteq f_s(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos\left(\frac{n\pi x}{L}\right) + b_n \sin\left(\frac{n\pi x}{L}\right)\right), \tag{9.1}$$

where the real constants $a_0;\ a_1, a_2, a_3, ...;\ b_1, b_2, b_3, ...$ are called the **Fourier coefficients**.

The notation $f(x) \doteq f_s(x)$ means that f is "represented" by f_s, the Fourier series for f. We will see that $f_s(x)$ may be unequal to $f(x)$ at some value(s) of x; even worse, the Fourier series $f_s(x)$ may fail to converge at some value(s) of x. So, in principle, $f_s(x)$ and $f(x)$ may be different functions.

Note that even though the original function $f(x)$ is defined only for $-L \le x \le L$, its Fourier series $f_s(x)$ may be defined for all x and is periodic with period $2L$, that is, $f_s(x + k \cdot 2L) = f_s(x)$ for all integers k.

One way to think of the relationship of f_s to f is that it is like the relationship of your computer game's "avatar" to you. In some circumstances your avatar may behave just like you, but perhaps not always. Another way to think of the relationship is between a movie's "stunt double" and the actress she replaces: From certain viewpoints they may look and behave exactly alike. In fact, the actress may do the stunt herself so she is being her own stunt double.

The term $\frac{a_0}{2}$ can be thought of as $\frac{a_0}{2}\cos\left(\frac{0 \cdot \pi x}{L}\right)$, an "honorary" cosine term.

Another way to display (9.1) is to display the infinite series more *explicitly* using the "$+\cdots$" notation:

$$f(x) \doteq f_s(x) = \frac{a_0}{2} + a_1 \cos\left(\frac{\pi x}{L}\right) + b_1 \sin\left(\frac{\pi x}{L}\right) + a_2 \cos\left(\frac{2\pi x}{L}\right) + b_2 \sin\left(\frac{2\pi x}{L}\right) + \cdots . \tag{9.2}$$

In Calculus we worked with infinite series of constants and power series. We defined "partial sums" and studied their convergence. Similarly, here we define the **partial sums** by $s_0(x) \triangleq \frac{a_0}{2}$ and

$$s_N(x) \triangleq \frac{a_0}{2} + \sum_{n=1}^{N} \left(a_n \cos\left(\frac{n\pi x}{L}\right) + b_n \sin\left(\frac{n\pi x}{L}\right)\right), \tag{9.3}$$

for $N = 1, 2, \ldots$. In Theorem 9.1 we will give the main result which guarantees convergence of the sequence of partial sums.

If $f(x)$ is defined for x in a finite interval $[a, b]$, except possibly at finitely many such x, we will say $f(x)$ is defined* on $[a, b]$. We put an * next to the word "defined" to indicate that exceptional points are allowed. For example, an $f(x)$ that is defined* on $[0, L]$ may actually be defined only on the interval $(0, L)$.

Let us see two examples of Fourier series.

Example 9.1

Define $f(x) = (x-1)^2$ on the interval $-2 < x < 2$. The Fourier series of this function is

$$f_s(x) \triangleq \frac{7}{3} + \frac{8}{\pi} \sum_{n=1}^{\infty} (-1)^n \left(\frac{2}{n^2\pi} \cos\left(\frac{n\pi x}{2}\right) + \frac{1}{n} \sin\left(\frac{n\pi x}{2}\right)\right). \tag{9.4}$$

We will postpone to Example 9.3 the explanation for how we found this Fourier series. For now, use graphs to illustrate the convergence of the sequence of partial sums,

$$s_N(x) = \frac{7}{3} + \frac{8}{\pi} \sum_{n=1}^{N} (-1)^n \left(\frac{2}{n^2\pi} \cos\left(\frac{n\pi x}{2}\right) + \frac{1}{n} \sin\left(\frac{n\pi x}{2}\right)\right). \tag{9.5}$$

Method: Figure 9.1 depicts the convergence of the sequence of partial sums. The values of N in the graphs of the partial sums are $N = 4, 16, 128, 8192$. Notice that on the interval $-2 < x < 2$ the graphs of the partial sums get closer and closer to the graph of $f(x) = (x-1)^2$, as N gets larger and larger. But there are problems at $x = \pm 2$, where the partial sums seem to converge to vertical line segments.

Notice also that the Fourier series, and its partial sums, are periodic with period 4, so there is no relationship between the Fourier series and $(x-1)^2$ for $|x| > 2$. ◯

By the way, these graphs were plotted by *Mathematica*™.

Example 9.2

The function $f(x) = x, \ -\pi < x < \pi$ has Fourier series

$$x \doteq f_s(x) = -2 \sum_{n=1}^{\infty} \frac{(-1)^n}{n} \sin nx;$$

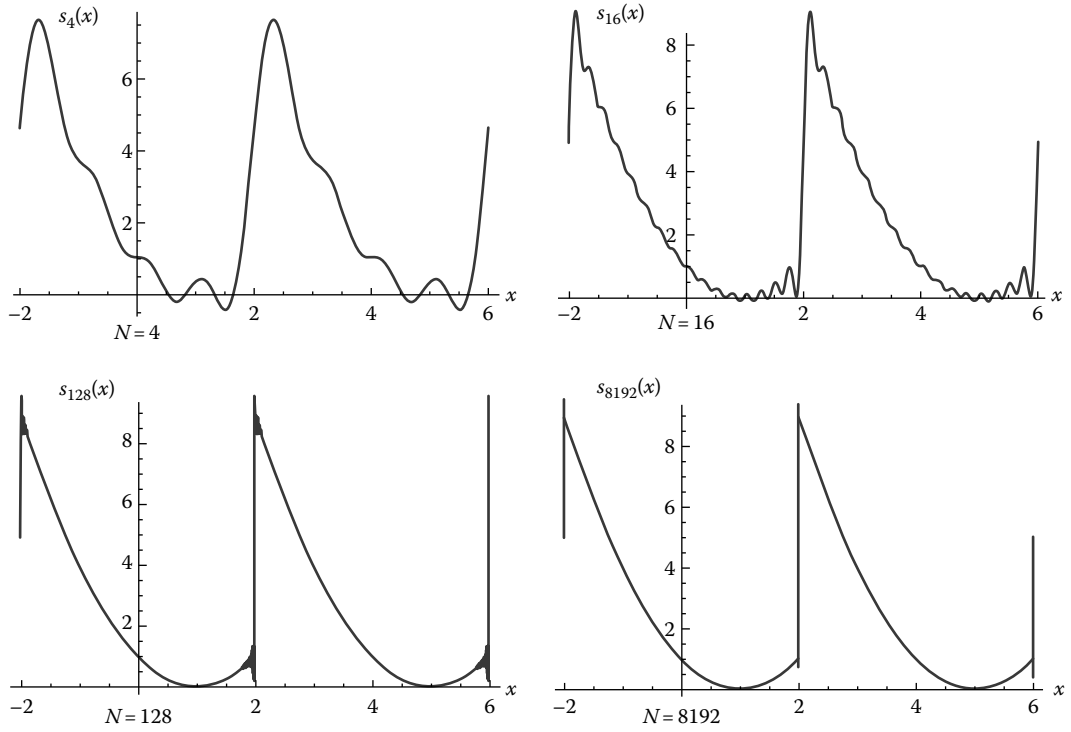


FIGURE 9.1
Convergence of sequence of partial sums for Example 9.1.

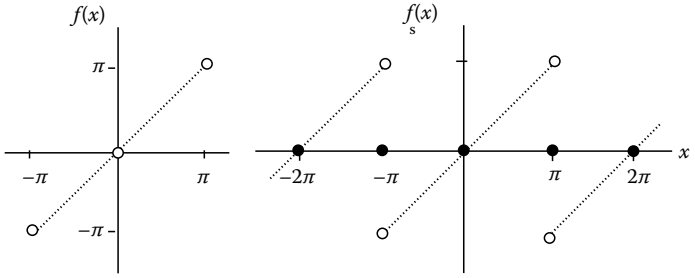


FIGURE 9.2
Fourier series for $f(x) = x$, $-\pi < x < \pi$.

we will explain why this is true in Example 9.4. Shown in Figure 9.2 are graphs of the original function, $f(x)$, which is defined only for $-\pi < x < \pi$, and its Fourier series expansion, $f_s(x)$, which is defined for all real numbers x. At $x = \pi$ all of the functions $\sin nx$ are zero, so $f_s(\pi)$ should be zero.

9.1.2 Convergence of Fourier Series

Denote

$$f(\xi^+) \triangleq \lim_{x \to \xi^+} f(x) \quad \text{and} \quad f(\xi^-) \triangleq \lim_{x \to \xi^-} f(x).$$

Definition 9.1

f has a **finite jump discontinuity** at ξ if both of $f(\xi^{\pm})$ exist as finite numbers but $f(\xi^{+}) \neq f(\xi^{-})$.

Definition 9.2

Suppose $[\,a, b\,]$ is a finite interval, that is, $-\infty < a < b < \infty$. A function $f(x)$ is **nice** on the interval $[\,a, b\,]$ if all of the following conditions are met:

(a) For all x in the interval $[\,a, b\,]$ the function values $f(x)$ and $f'(x)$ are defined* and $f(x)$ and $f'(x)$ are continuous, except at possibly a finite number of values of x, and
(b) $f(a^{+}), f(b^{-}), f'(a^{+})$, and $f'(b^{-})$ exist as finite numbers, and
(c) At all points ξ of discontinuity of f and f', the discontinuity is a finite jump.

Another name for nice is **piecewise smooth**.

Theorem 9.1

If a function $f(x)$ is nice on the interval $[\,-L, L\,]$, then

(a) At all x in $(-L, L)$, the Fourier series of f converges to

$$f_s(x) = \frac{f(x^{-}) + f(x^{+})}{2},$$

that is, the average of $f(x^{-})$ and $f(x^{+})$, and
(b) At $x = \pm L$, the Fourier series of f converges to the average of $f(-L^{+})$ and $f(L^{-})$, that is,

$$f_s(\pm L) = \frac{f(-L^{+}) + f(L^{-})}{2}.$$

If a nice function $f(x)$ is defined and continuous at x_0 then $f(x_0^{-}) = f(x_0^{+}) = f(x_0)$, hence $f_s(x_0) = f(x_0)$ at such a point of continuity. That is, the Fourier series agrees with the original function where the latter is continuous.

Also, if $f(x_0^{-}) = f(x_0^{+})$ but $f(x_0)$ is not defined, then we might as well define an "improved" version of f, say f_{Imp}, by $f_{Imp}(x_0) = f(x_0^{-}) = f(x_0^{+})$. Then $f_{Imp}(x)$ agrees with the Fourier series f_s at x_0.

9.1.3 Orthogonality and Calculating Fourier Coefficients

Recall from Corollary 2.7 in Section 2.4 that if $\mathbf{v}_1, \ldots, \mathbf{v}_k$ is an orthogonal set of nonzero vectors in $\mathbb{R}^m$ and $\mathbf{y}$ is a vector in $Span\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, then we can resolve a vector into components using the formula

$$\mathbf{y} = \sum_{i=1}^{k} \frac{1}{||\mathbf{v}_i||^2} \langle \mathbf{y},\ \mathbf{v}_i \rangle \mathbf{v}_i.$$

There is a similar recipe for Fourier series.

First, let's define an inner product for functions $f(x)$ and $g(x)$ defined for $-L < x < L$:

$$\langle f,\ g \rangle \triangleq \int_{-L}^{L} f(x)g(x)\, dx.$$

Using the trigonometric identities

$$\cos\alpha \cos\beta = \frac{1}{2}\big(\cos(\alpha - \beta) + \cos(\alpha + \beta)\big), \quad \sin\alpha \sin\beta = \frac{1}{2}\big(\cos(\alpha - \beta) - \cos(\alpha + \beta)\big),$$

and

$$\sin\alpha \cos\beta = \frac{1}{2}\big(\sin(\alpha - \beta) + \sin(\alpha + \beta)\big),$$

one can explain why for positive integers m and n we have the orthogonality relations

$$\left\langle \cos\left(\frac{m\pi x}{L}\right), \cos\left(\frac{n\pi x}{L}\right) \right\rangle = \begin{Bmatrix} L, & \text{if } m = n \\ 0, & \text{if } m \neq n \end{Bmatrix},$$

$$\left\langle \sin\left(\frac{m\pi x}{L}\right), \sin\left(\frac{n\pi x}{L}\right) \right\rangle = \begin{Bmatrix} L, & \text{if } m = n \\ 0, & \text{if } m \neq n \end{Bmatrix},$$

and

$$\left\langle \sin\left(\frac{m\pi x}{L}\right), \cos\left(\frac{n\pi x}{L}\right) \right\rangle = 0, \text{ for all } m,\ n.$$

For example, for $m \neq n$,

$$\left\langle \cos\left(\frac{m\pi x}{L}\right),\ \cos\left(\frac{n\pi x}{L}\right) \right\rangle = \int_{-L}^{L} \cos\left(\frac{m\pi x}{L}\right) \cos\left(\frac{n\pi x}{L}\right) dx$$

$$= \frac{1}{2}\int_{-L}^{L}\left(\cos\Big(\frac{(m-n)\pi x}{L}\Big)+\cos\Big(\frac{(m+n)\pi x}{L}\Big)\right)dx = \frac{1}{2}\left[\frac{\sin\left(\frac{(m-n)\pi x}{L}\right)}{-(m-n)}+\frac{\sin\left(\frac{(m+n)\pi x}{L}\right)}{-(m+n)}\right]_{-L}^{L}$$

$$= \frac{1}{2}\left(\frac{\sin\big((m-n)\pi\big)}{-(m-n)}+\frac{\sin\big((m+n)\pi\big)}{-(m+n)}-\frac{\sin\big(-(m-n)\pi\big)}{-(m-n)}-\frac{\sin\big(-(m+n)\pi\big)}{-(m+n)}\right)=0.$$

But if $m = n$,

$$\left\langle\cos\left(\frac{m\pi x}{L}\right),\cos\left(\frac{n\pi x}{L}\right)\right\rangle = \int_{-L}^{L}\cos^2\left(\frac{m\pi x}{L}\right)dx = \frac{1}{2}\int_{-L}^{L}\left(1+\cos\left(2\left(\frac{m\pi x}{L}\right)\right)\right)dx = \cdots = L.$$

Also, we can explain easily why

$$\langle 1,\,1\rangle = 2L,\quad \left\langle 1,\,\cos\left(\frac{m\pi x}{L}\right)\right\rangle = 0,\ \text{for all } m,\quad \text{and}\quad \left\langle 1,\,\sin\left(\frac{m\pi x}{L}\right)\right\rangle = 0,\ \text{for all } m.$$

Using these facts, we can find formulas for the Fourier coefficients a_0; $a_1, a_1, a_3, \ldots$; $b_1, b_1, b_3, \ldots$. In this work we will ignore issues concerning convergence of an infinite series of orthogonal functions, such as interchange of the operations of taking an infinite sum and taking an inner product. In addition, we will assume that if $f(x) \doteq f_s(x)$ then $\langle f(x), h(x)\rangle = \langle f_s(x), h(x)\rangle$ for any piecewise smooth function $h(x)$.

Suppose $f(x)$ is a nice function and has Fourier expansion given by (9.1), that is,

$$f(x) \doteq \frac{a_0}{2}+\sum_{m=1}^{\infty}a_m\cos\left(\frac{m\pi x}{L}\right)+\sum_{m=1}^{\infty}b_m\sin\left(\frac{m\pi x}{L}\right).$$

Take the inner product of both sides of this equation with $\cos\left(\frac{n\pi x}{L}\right)$ to get

$$\begin{aligned}\left\langle f(x),\cos\left(\frac{n\pi x}{L}\right)\right\rangle &= \left\langle \frac{a_0}{2}+\sum_{m=1}^{\infty}a_m\cos\left(\frac{m\pi x}{L}\right)+\sum_{m=1}^{\infty}b_m\sin\left(\frac{m\pi x}{L}\right),\ \cos\left(\frac{n\pi x}{L}\right)\right\rangle\\ &= \frac{a_0}{2}\cdot\left\langle 1,\cos\left(\frac{n\pi x}{L}\right)\right\rangle+\sum_{m=1}^{\infty}a_m\cdot\left\langle\cos\left(\frac{m\pi x}{L}\right),\cos\left(\frac{n\pi x}{L}\right)\right\rangle\\ &\quad+\sum_{m=1}^{\infty}b_m\cdot\left\langle\sin\left(\frac{m\pi x}{L}\right),\cos\left(\frac{n\pi x}{L}\right)\right\rangle\\ &= \frac{a_0}{2}\cdot 0+\sum_{m=1}^{\infty}a_m\cdot\begin{Bmatrix}L, & \text{if } m=n\\ 0, & \text{if } m\neq n\end{Bmatrix}+\sum_{m=1}^{\infty}b_m\cdot 0 = a_nL.\end{aligned}$$

So, $a_n = \frac{1}{L}\left\langle f(x),\cos\left(\frac{n\pi x}{L}\right)\right\rangle$, that is,

$$a_n = \frac{1}{L}\int_{-L}^{L}f(x)\cos\left(\frac{n\pi x}{L}\right)dx. \tag{9.6}$$

Similarly,

$$\left\langle f(x), \sin\left(\frac{n\pi x}{L}\right)\right\rangle = \frac{a_0}{2}\cdot\left\langle 1, \sin\left(\frac{n\pi x}{L}\right)\right\rangle + \sum_{m=1}^{\infty} a_m \cdot \left\langle \cos\left(\frac{m\pi x}{L}\right), \sin\left(\frac{n\pi x}{L}\right)\right\rangle$$
$$+ \sum_{m=1}^{\infty} b_m \cdot \left\langle \sin\left(\frac{m\pi x}{L}\right), \sin\left(\frac{n\pi x}{L}\right)\right\rangle$$
$$= \frac{a_0}{2}\cdot 0 + \sum_{m=1}^{\infty} a_m \cdot 0 + \sum_{m=1}^{\infty} b_m \begin{Bmatrix} L, & \text{if } m = n \\ 0, & \text{if } m \neq n \end{Bmatrix}.$$

So,

$$b_n = \frac{1}{L}\int_{-L}^{L} f(x)\sin\left(\frac{n\pi x}{L}\right) dx. \tag{9.7}$$

Also,

$$\langle f(x),\, 1\rangle = \frac{a_0}{2}\,\langle 1,\, 1\rangle + \sum_{m=1}^{\infty} a_m \cdot \left\langle \cos\left(\frac{m\pi x}{L}\right),\, 1\right\rangle + \sum_{m=1}^{\infty} b_m \cdot \left\langle \sin\left(\frac{m\pi x}{L}\right),\, 1\right\rangle$$
$$= \frac{a_0}{2}\cdot 2L + \sum_{m=1}^{\infty} a_m \cdot 0 + \sum_{m=1}^{\infty} b_m \cdot 0 = L\cdot a_0,$$

so

$$a_0 = \frac{1}{L}\int_{-L}^{L} f(x)\, dx. \tag{9.8}$$

An interesting and useful aside about the last formula is that $\frac{a_0}{2}$, the first term in the Fourier series, has a very natural physical interpretation:

$$\frac{a_0}{2} = \frac{1}{2L}\int_{-L}^{L} f(x)\, dx$$

which is the **average value** of the function f on the interval $[\,-L, L\,]$. For example, if $f(x)$ is the equilibrium temperature in a rod extending over the interval $-L \le x \le L$, then $\frac{a_0}{2}$ is the average temperature in the rod.

Formulas (9.6) through (9.8) give the Fourier coefficients for the general Fourier series (9.1) for a nice function $f(x)$ defined on the interval $-L < x < L$. The Fourier series is *unique*: Given f, (9.6) through (9.8) specify the values for the coefficients a_0 and $a_n, b_n, n = 1, 2, \ldots$

In the next problem and many other problems it is useful to note that $\sin n\pi = 0$ and $\cos(\pm n\pi) = (-1)^n$ for $n = 1, 2, \ldots$.

Example 9.3

(Example 9.1 again) Find the Fourier series for the function $f(x) = (x-1)^2$ defined only on the interval $-2 < x < 2$.

Method: Let $L = 2$. Using integration by parts and (9.6) through (9.8), we calculate

$$a_0 = \frac{1}{2}\int_{-2}^{2} f(x)dx = \frac{1}{2}\int_{-2}^{2}(x-1)^2\,dx = \frac{1}{2}\left[\frac{1}{3}(x-1)^3\right]_{-2}^{2} = \cdots = \frac{14}{3},$$

$$a_n = \frac{1}{2}\int_{-2}^{2} f(x)\cos\left(\frac{n\pi x}{2}\right)dx = \frac{1}{2}\int_{-2}^{2}(x-1)^2\cos\left(\frac{n\pi x}{2}\right)dx$$

$$= \frac{1}{2}\left[\frac{(x-1)^2\sin\left(\frac{n\pi x}{2}\right)}{\frac{n\pi}{2}} + \frac{2(x-1)\cos\left(\frac{n\pi x}{2}\right)}{\left(\frac{n\pi}{2}\right)^2} - \frac{2\sin\left(\frac{n\pi x}{2}\right)}{\left(\frac{n\pi}{2}\right)^3}\right]_{-2}^{2}$$

$$= \frac{1}{2}\left(\left(0 + \frac{2(-1)^n}{\left(\frac{n\pi}{2}\right)^2} - 0\right) - \left(0 + \frac{2(-3)(-1)^n}{\left(\frac{n\pi}{2}\right)^2} - 0\right)\right) = \frac{16}{n^2\pi^2}(-1)^n,$$

and

$$b_n = \frac{1}{2}\int_{-2}^{2} f(x)\sin\left(\frac{n\pi x}{2}\right)dx = \frac{1}{2}\int_{-2}^{2}(x-1)^2\sin\left(\frac{n\pi x}{2}\right)dx$$

$$= \frac{1}{2}\left[\frac{(x-1)^2\cos\left(\frac{n\pi x}{2}\right)}{-\frac{n\pi}{2}} + \frac{2(x-1)\sin\left(\frac{n\pi x}{2}\right)}{\left(\frac{n\pi}{2}\right)^2} + \frac{2\cos\left(\frac{n\pi x}{2}\right)}{\left(\frac{n\pi}{2}\right)^3}\right]_{-2}^{2}$$

$$= \frac{1}{2}\left(\left(\frac{(-1)^n}{-\frac{n\pi}{2}} + 0 + \cancel{\frac{2(-1)^n}{\left(\frac{n\pi}{2}\right)^3}}\right) - \left(\frac{9(-1)^n}{-\frac{n\pi}{2}} + 0 + \cancel{\frac{2(-1)^n}{\left(\frac{n\pi}{2}\right)^3}}\right)\right) = \frac{8}{n\pi}(-1)^n,$$

The Fourier series of this function is

$$(x-1)^2 \doteq f_s(x) = \frac{7}{3} + \frac{8}{\pi}\sum_{n=1}^{\infty}(-1)^n\left(\frac{2}{n^2\pi}\cos\left(\frac{n\pi x}{2}\right) + \frac{1}{n}\sin\left(\frac{n\pi x}{2}\right)\right). \bigcirc$$

9.1.4 Even and Odd Functions and Their Fourier Series Coefficients

A function F is **even** if $F(-x) \equiv F(x)$ whenever both x and $-x$ are in the domain of F. For example, an even power of x is an even function, $\cos\alpha x$ is an even function for any constant α, and $|x|$ is an even function of x. In fact, if g is any function, then $g(|x|)$ is an even function of x.

Thinking graphically, a function is even if its graph $y = f(x)$ is symmetric about the y-axis, that is, (x, y) is on the graph if and only if $(-x, y)$ is on the graph, as illustrated in Figure 9.3.

A function F is **odd** if $F(-x) \equiv F(x)$ whenever both x and $-x$ are in the domain of F. For example, an odd power of x is an odd function, $\sin\alpha x$ is an odd function for any constant α, and $x|x|$ is an odd function of x. In fact, if g is any function, then $xg(|x|)$ is an odd function of x.

Thinking graphically, a function is odd if its graph $y = f(x)$ is symmetric about the origin, that is, (x, y) is on the graph if and only if $(-x, -y)$ is on the graph, as illustrated in Figure 9.4.

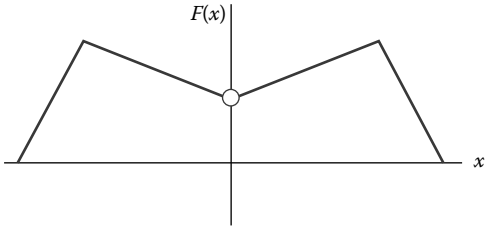


FIGURE 9.3
Even function picture.

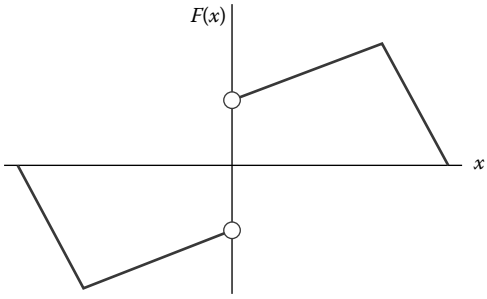


FIGURE 9.4
Odd function picture.

Theorem 9.2

(a) If F is an even function, then for any value of L, $\int_{-L}^{L} F(x)\,dx = 2\int_{0}^{L} F(x)\,dx$.

(b) If F is an odd function, then for any value of L, $\int_{-L}^{L} F(x)\,dx = 0$.

Why?

(a) To see this, write $\int_{-L}^{L} F(x)\,dx = \int_{-L}^{0} F(x)\,dx + \int_{0}^{L} F(x)\,dx$; for the former, make the substitution $z = -x$ to see that $\int_{-L}^{0} F(x)\,dx = \int_{0}^{L} F(x)\,dx$.

(b) To see this, write $\int_{-L}^{L} F(x)\,dx = \int_{-L}^{0} F(x)\,dx + \int_{0}^{L} F(x)\,dx$; for the former, make the substitution $z = -x$ to see that $\int_{-L}^{0} F(x)\,dx = -\int_{0}^{L} F(x)\,dx$. □

Using these observations, we note that

Theorem 9.3

(a) If f is an odd function, then its Fourier coefficients $a_0, a_1, \ldots, a_n, \ldots$ are zero, that is, the Fourier series of f on $[-L, L]$ consists only of sine function terms, that is,

$$f(x) \doteq f_s(x) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right), \text{ where } b_n = \frac{2}{L}\int_0^L f(x)\sin\left(\frac{n\pi x}{L}\right)dx.$$

(b) If f is an even function, then its Fourier coefficients $b_1, \ldots, b_n, \ldots$ are zero, that is, the Fourier series of f on $[-L, L]$ consists only of the constant term and the cosine function terms, that is,

$$f(x) \doteq f_s(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right),$$

where

$$a_0 = \frac{2}{L}\int_0^L f(x)\,dx \quad \text{and} \quad a_n = \frac{2}{L}\int_0^L f(x)\cos\left(\frac{n\pi x}{L}\right)dx.$$

Why? (a) If f is odd, then $F(x) \triangleq f(x)\cos\left(\frac{n\pi x}{L}\right)$ is odd for all n, so Theorem 9.2(b) implies that $a_n = 0$ for all n. The formula for b_n follows from Theorem 9.2(a).

For (b) you will use similar reasoning to explain this in Problem 9.1.8.23. □

By Theorem 9.3, to find the Fourier series of an odd or even function we have fewer integrations to do and we need only integrate over the interval $[0, L]$.

This suggests that we should replace a given function by, say, an even function, if we could. For example, sometimes we expect the solution of a physical problem to have the symmetry property of being an even function. This might require some caution; if we were wrong to assume the solution is even, we might find no solution or an incorrect "solution" or miss finding some of the correct solutions.

Example 9.4

Show that the function $f(x) = x, -\pi < x < \pi$ has Fourier series

$$x \doteq f_s(x) = \sum_{n=1}^{\infty} (-1)^{n+1}\,\frac{2}{n}\sin nx.$$

Method: With $L = \pi$, Theorem 9.3 says that f being an odd function implies its Fourier series consists only of sine terms, specifically

$$x \doteq \sum_{n=1}^{\infty} b_n \sin nx$$

and the coefficients are, after using integration by parts,

$$b_n = \frac{2}{\pi}\int_0^{\pi} x\,\sin nx\,dx = \frac{2}{\pi}\left[\frac{x\cos nx}{-n} + \frac{\sin nx}{n^2}\right]_0^{\pi} = \frac{2}{\pi}\left(\frac{-\pi\cos n\pi + 0}{n} + \frac{0-0}{n^2}\right)$$

$$= \frac{-2(-1)^n}{n}.$$

The result follows. ○

Example 9.5

For the square wave function with period $2L$, that is, given by

$$f(x) = \begin{Bmatrix} -1, & -L < x < 0 \\ 1, & 0 < x < L \end{Bmatrix},$$

(a) Find the Fourier series expansion, and

(b) Sketch the graphs of the original function and its Fourier series.

Method: (a) With $L = \pi$, Theorem 9.3 says that f being an odd function implies its Fourier series consists only of sine terms, specifically

$$f(x) \doteq f_s(x) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right).$$

We calculate

$$b_n = \frac{2}{L}\int_0^L f(x)\sin\left(\frac{n\pi x}{L}\right)dx = \frac{2}{L}\int_0^L 1\cdot\sin\left(\frac{n\pi x}{L}\right)dx = \frac{2}{L}\left[\frac{\cos\left(\frac{n\pi x}{L}\right)}{-\frac{n\pi}{L}}\right]_0^L = \frac{2}{n\pi}\left(1-(-1)^n\right).$$

The Fourier series of the square wave is given by

$$f(x) \doteq f_s(x) = \frac{2}{\pi}\sum_{n=1}^{\infty}\frac{1-(-1)^n}{n}\sin\left(\frac{n\pi x}{L}\right). \tag{9.9}$$

Note that $(1-(1)^n)$ is zero for even n and $(1-(1)^n)$ equals two for odd n. Odd integers can be expressed in the form $n = 2k-1$, so we can rewrite the Fourier series (9.9) as

$$f(x) \doteq f_s(x) = \frac{4}{\pi}\sum_{k=1}^{\infty}\frac{1}{2k-1}\sin\left(\frac{(2k-1)\pi x}{L}\right).$$

(b) Shown in Figure 9.5 are graphs of the original function, $f(x)$, which is defined only for $-L < x < 0$ and $0 < x < L$, and its Fourier series expansion, $f_s(x)$, which is defined for all real numbers x. The Fourier series graph, that is, for $f_s(x)$, has filled circles at $x = nL$, for all integers n. These are values of x where the original function, $f(x)$, had jump discontinuities. As we see from Theorem 9.1, this "filling in" of the function is a result of using a Fourier series expansion. ◯

Square waves are commonly used in engineering, so their Fourier series is a useful tool in our toolbox of mathematical results.

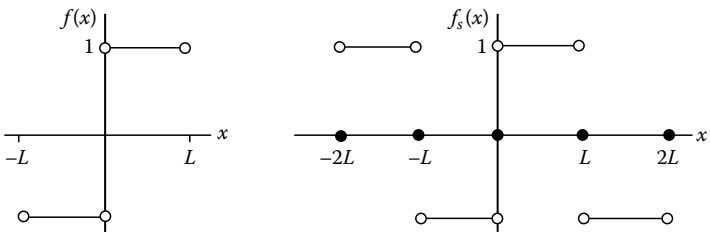


FIGURE 9.5
Example 9.1: Fourier series for square wave.

9.1.5 Finding the Fourier Series in a Special Case

Example 9.6

Find the Fourier series for the function $f(x) = 2\sin 2x - \cos 5x$ defined for $-\pi < x < \pi$.

Method: This can be done very easily: Because $L = \pi$, the Fourier series given by (9.1) should be in the form

$$2\sin 2x - \cos 5x = f(x) \doteq f_s(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos nx + \sum_{n=1}^{\infty} b_n \sin nx.$$

So, this function $f(x)$ was already *given* as a Fourier series! There is no need to calculate the integrals used in formulas (9.6) through (9.8). But if we did, by uniqueness of the Fourier series of $f(x)$, we would find that $a_0 = 0$, $a_5 = -1$, $b_2 = 2$, and all other a_n's and b_n's are zero. The Fourier series of $f(x) = 2\sin 2x - \cos 5x, \ -\pi < x < \pi$, is

$$f_s(x) = 2\sin 2x - \cos 5x. \bigcirc$$

If we wanted to do Example 9.6 by calculating integrals, the work would be tedious because we would need to do separate calculations for $n = 2$ and $n = 5$.

Note that although the original function is defined only for $-\pi < x < \pi$, its Fourier series is defined for all x and periodic with period 2π.

9.1.6 Periodic Extension of a Function Given for $0 < x < 2L$

Suppose $f(x)$ is given for $0 < x < 2L$ *and* we are told that f is periodic with period $2L$, that is, $f(x - 2kL) = f(x)$ at all x in the domain of f and all integers k. This implicitly gives the values of $f(x)$ for all x with $-\infty < x < \infty$, with the possible exception of a discrete set of points. For example, $f(-\frac{L}{2}) = f\left(-\frac{L}{2} + 2L\right) = f\left(\frac{3L}{2}\right)$, by periodicity.

Now suppose that we are given a function $f(x)$ defined* *only* for x in the interval $[\,0, 2L\,]$. We can **extend** f **periodically** to be defined for all x, or at least all x that are an integer multiple (possibly negative) of $2L$ away from the places where f is defined* on $[\,0, 2L\,]$. The picture "says a thousand words." In symbols, the **periodic extension** of f is a function $\widetilde{f}$ defined by

$$\widetilde{f}(x) \triangleq f(x - k2L),$$

where k is chosen so that $0 \le x - k2L \le 2L$; note that k may depend on x. For example, if $L = \pi$, then $\widetilde{f}\left(\frac{31\pi}{6}\right) \triangleq f\left(\frac{7\pi}{6}\right)$, by choosing $k = 2$, and $\widetilde{f}\left(-\frac{31\pi}{6}\right) \triangleq f\left(\frac{5\pi}{6}\right)$, by choosing $k = -3$.

Because the original function $f(x)$ is given on the interval $[\,0, 2L\,]$ and its periodic extension is given on the whole real line, we can use values of $\widetilde{f}(x)$ for $-L < x < L$ and thus, we can calculate the Fourier series of $\widetilde{f}(x)$ on the interval $-L < x < L$. Because of the periodicity of $\widetilde{f}(x)$ we see, as in Figure 9.6, that the values of the Fourier series of $\widetilde{f}(x)$ for $-L < x < 0$ give us the values of the Fourier series of $f(x)$ for $-L < x < 0$.

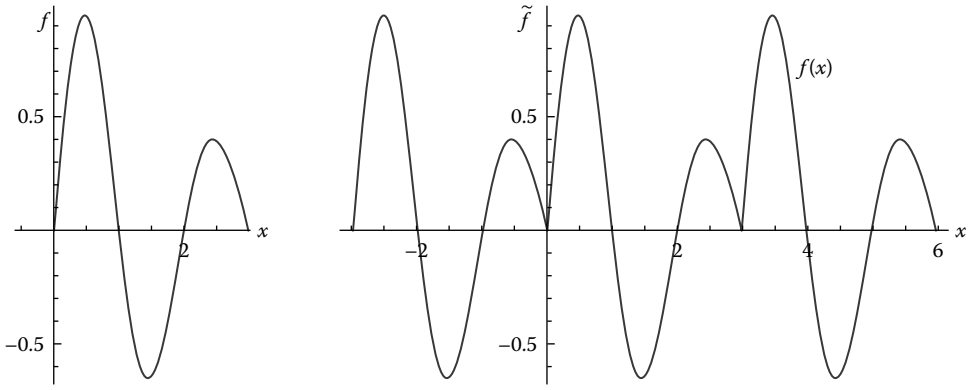


FIGURE 9.6
Periodic extension of a function.

Example 9.7

For the function defined by $f(x) = \begin{Bmatrix} \sin 2x, & 0 < x < \pi \\ 0, & \pi < x < 2\pi \end{Bmatrix}$, (a) sketch the graph of the periodic extension of f on the interval $[-\pi, \pi]$, (b) sketch the graph of the Fourier series of f on an interval of length two periods, and (c) find the Fourier series, that is, evaluate the Fourier coefficients.

Method:

(a) Let $\widetilde{f}$ be the periodic extension of f whose period is 2π. Then

$$\widetilde{f}(x) = \begin{Bmatrix} 0, & -\pi < x < 0 \\ \sin 2x, & 0 < x < \pi \end{Bmatrix}$$

and $L = \pi$. The sketch of the graph of the original function and the periodic extension, graphed only on the interval $[-\pi, \pi]$, is in Figure 9.7.

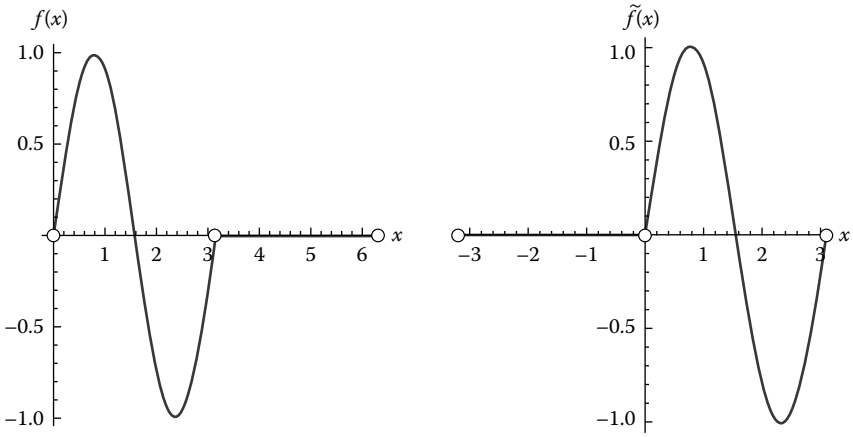


FIGURE 9.7
Graphs for Example 9.7(a).

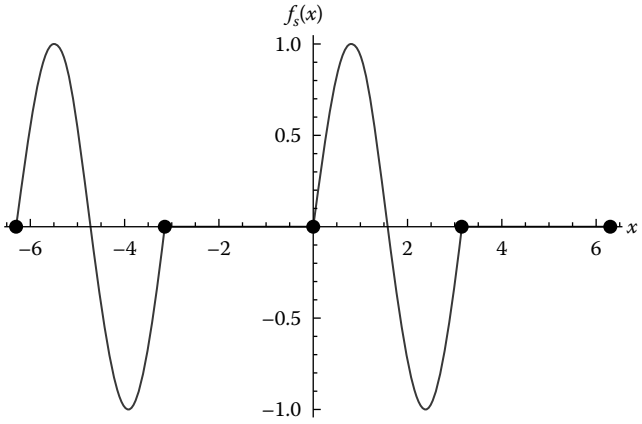


FIGURE 9.8
Graph for Example 9.7(b).

(b) The sketch of the graph of the Fourier series of f on an interval of length two periods is in Figure 9.8.

(c) We calculate

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\,dx = \frac{1}{\pi}\left(\int_{-\pi}^{0} 0\cdot dx + \int_{0}^{\pi} \sin 2x\,dx\right) = \frac{1}{\pi}\left[\frac{\cos 2x}{-2}\right]_0^{\pi} = 0.$$

For $n \neq 2$,

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos nx\,dx = \frac{1}{\pi}\int_{0}^{\pi} \sin 2x\cos nx\,dx$$

$$= \frac{1}{\pi}\int_{0}^{\pi}\frac{1}{2}\Big(\sin\big((2-n)x\big) + \sin\big((2+n)x\big)\Big)\,dx = \frac{1}{2\pi}\left[\frac{\cos\big((2-n)x\big)}{-(2-n)} + \frac{\cos\big((2+n)x\big)}{-(2+n)}\right]_0^{\pi}$$

$$= \frac{1}{2\pi}\left(\left(\frac{(-1)^n}{-(2-n)} + \frac{(-1)^n}{-(2+n)}\right) - \left(\frac{1}{-(2-n)} + \frac{1}{-(2+n)}\right)\right)$$

$$= \frac{4}{2\pi(4-n^2)}\left(1-(-1)^n\right),$$

and similarly

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin nx\,dx = \frac{1}{\pi}\int_{0}^{\pi} \sin nx\sin 2x\,dx = \frac{1}{\pi}\int_{0}^{\pi}\frac{1}{2}\Big(\cos\big((n-2)x\big) - \cos\big((n+2)x\big)\Big)\,dx$$

$$= \frac{1}{2\pi}\left[\frac{\sin\big((n-2)x\big)}{n-2} - \frac{\sin\big((n+2)x\big)}{n+2}\right]_0^{\pi} = 0.$$

For $n = 2$, we evaluate

$$a_2 = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos 2x\,dx = \frac{1}{\pi}\int_{0}^{\pi} \cos 2x\sin 2x\,dx = \frac{1}{\pi}\int_{0}^{\pi}\frac{1}{2}\,\sin 4x\,dx = \frac{1}{2\pi}\left[\frac{\cos 4x}{-4}\right]_0^{\pi} = 0,$$

and

$$b_2 = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin 2x\,dx = \frac{1}{\pi}\int_{0}^{\pi} \sin 2x \sin 2x\,dx = \frac{1}{\pi}\int_{0}^{\pi} \frac{1-\cos 4x}{2}dx$$
$$= \frac{1}{2\pi}\left[x - \frac{\sin 4x}{4}\right]_0^{\pi} = \frac{1}{2}.$$

The Fourier series of the original function f is the same as the Fourier series of the periodic extension $\widetilde{f}$, that is,

$$f(x) \doteq f_s(x) = \frac{1}{2}\sin 2x - \frac{2}{\pi}\sum_{n\neq 2}^{\infty} \frac{1-(-1)^n}{n^2-4}\cos nx. \bigcirc \tag{9.10}$$

Again, $(1-(1)^n)$ is zero for even n. Odd integers can be expressed in the form $n = 2k-1$, so we can rewrite the Fourier series (9.10) as

$$f(x) \doteq f_s(x) = \frac{1}{2}\sin 2x - \frac{4}{\pi}\sum_{k=1}^{\infty} \frac{1}{(2k-1)^2-4}\cos\big((2k-1)x\big).$$

Remark

In the above example, we used the values of $\widetilde{f}(x) = f(x-2L)$ for $-L < x < 0$ in order to find the Fourier coefficients for $\widetilde{f}$ and thus for f. Alternatively, we could use the result of Problem 9.1.8.22 to calculate the Fourier coefficients using the values of $f(x)$ for $0 < x < 2L$:

$$a_0 = \frac{1}{L}\int_0^{2L} f(x)\,dx, \quad a_n = \frac{1}{L}\int_0^{2L} f(x)\cos\left(\frac{n\pi x}{L}\right)dx, \quad b_n = \frac{1}{L}\int_0^{2L} f(x)\sin\left(\frac{n\pi x}{L}\right)dx. \tag{9.11}$$

9.1.7 Other Kinds of Fourier Series

In Section 9.4 we will learn about the "complex Fourier series" defined by

$$f(x) \doteq f_{\mathbb{C}}(x) = \sum_{n=-\infty}^{\infty} c_n\, e^{in\pi x/L}$$

with complex coefficients c_n, for a nice function $f(x)$ defined* on the interval $[-L, L]$. This will turn out to be completely equivalent to the "real" form of Fourier series defined in (9.1).

In Section 9.5 we will learn about the "discrete Fourier transform" (*DFT*) defined by

$$F_\ell \triangleq \frac{1}{\sqrt{N}}\sum_{k=0}^{N-1} f_k\, e^{-i2\pi k\ell/N}, \text{ for } \ell = 0, 1, \ldots, N-1,$$

with complex coefficients f_k, for a function $f(t)$ defined at $t = 0, t_1, \ldots, t_{N-1}$. This is a very popular tool for analysis of discrete time systems, including for signal analysis, and also

a useful tool for data analysis and manipulation. The *DFT* can be motivated by complex Fourier series.

In Section 9.4 we will learn about the "Fourier transform" defined by

$$\mathcal{F}[f(t)] = F(\omega) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t)\, e^{-i\omega t}\, dt$$

defined for a nice function $f(x)$ defined on the interval $(-\infty, \infty)$. We will see that it can be *motivated* by taking complex Fourier series on $(-L, L)$ and letting $L \to \infty$. It will turn out that Fourier transforms can help us to solve partial differential equations. Fourier transforms are also important in signal processing.

Learn More About It

Linear Operator Theory in Engineering and Science, (vol. 40 of Applied Mathematical Sciences), by Arch W. Naylor and George R. Sell, Springer-Verlag, © 1982, has a useful exposition of "Integration and Measure Theory" in its Appendix D.

9.1.8 Problems

1. Find the Fourier series of the function $f(x)$ whose graph is sketched in Figure 9.9.
2. What is the Fourier representation of $f(x) = 2 + \sin x$, defined for $-\pi < x < \pi$?
3. Evaluate the appropriate integrals for the Fourier coefficients to explain why the representation $x \doteq \sum_{n=1}^{\infty} \frac{2(-1)^{n+1}L}{n\pi} \sin\left(\frac{n\pi x}{L}\right)$ is true on the interval $[-L, L]$.

For problems 4–6, a function is given over one period. Sketch the graphs of (a) the given function, (b) the periodic extension of the function over an interval of length two periods, and (c) the Fourier series over an interval of length two periods. For (c), use open and closed circles at appropriate places on the graph.

4. $f(x) = \begin{cases} 0, & -\pi < x < 0 \\ \frac{x}{\pi}, & 0 < x < \pi \end{cases}$
5. $f(x) = \begin{cases} 0, & -\pi < x < -\frac{\pi}{2} \\ \cos x, & -\frac{\pi}{2} < x < \pi \end{cases}$

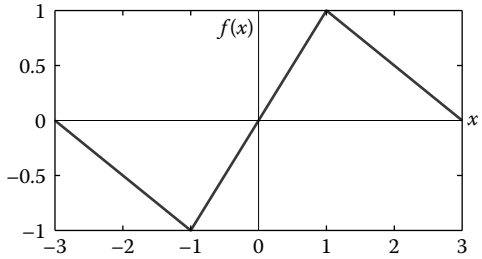


FIGURE 9.9
Graph of function in Problem 9.1.8.1.

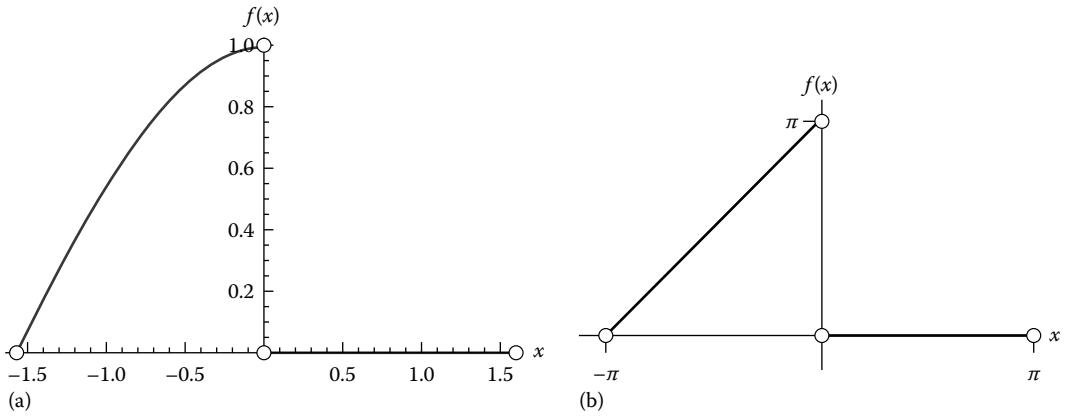


FIGURE 9.10
Graph of functions in Problems (a) 9.1.8.7 and (b) 9.1.8.8.

6. $f(t) = \begin{cases} -1, & -\pi < t < -\frac{\pi}{2} \\ \sin t, & -\frac{\pi}{2} < t < \pi \end{cases}$

For problems 7 and 8, a function is sketched over one period. Sketch the graphs of (a) the periodic extension of the function over an interval of length two periods and (b) the Fourier series over an interval of length two periods. For (a) and (b), use open and closed circles at appropriate places on the graph.

7. The partly sinusoidal function whose graph is sketched over one period in Figure 9.10a
8. The function whose graph is sketched over one period in Figure 9.10b

For problems 9–14, find the Fourier series representation of the given functions that are defined* over one period. Some of these problems can be done without evaluating integrals, for example, by using trigonometric function identities.

9. $f(x) = |x|, \ -L < x < L$
10. $f(x) = \cos^2 x, \ -\pi < x < \pi$
11. $f(x) = 1 + \sin(x - \frac{\pi}{3}) \cos^2 x, \ -\frac{\pi}{2} < x < \frac{\pi}{2}$
12. $f(x) = |\sin(x)|, \ -\pi < x < \pi$
13. $f(x) = x \cos x, \ -\pi < x < \pi$
14. The function whose graph is sketched over one period in Figure 9.11
15. The "sigmoidal" function f is sketched over one period in Figure 9.12a.
    (a) Find the Fourier series of f.
    (b) Sketch the graph of the Fourier series of f over an interval of length exactly two periods.
    (c) Find the function g that consists of the first five *nonzero* terms of the Fourier series of f.
    (d) Sketch the graphs of f and g over an interval of length exactly one period.

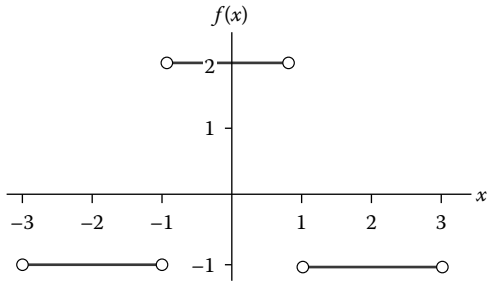


FIGURE 9.11
Graph of function in Problem 9.1.8.14.

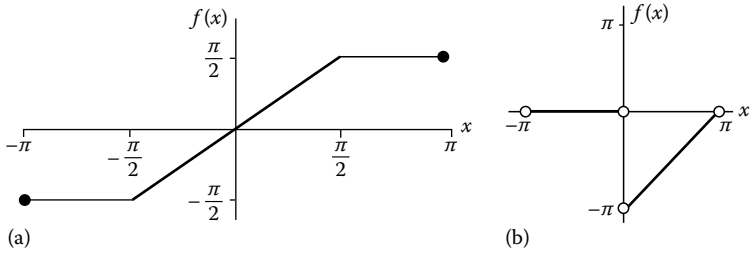


FIGURE 9.12
Graphs of functions in (a) Problems 9.1.8.15 and (b) 9.1.8.16.

For each of problems 16–21, the function given is defined* over one period. (a) Find the Fourier series representation, and (b) sketch the graph of the Fourier series over an interval of length exactly two periods.

16. The function $f(x)$ sketched over one period, $-\pi \le x \le \pi$, in Figure 9.12b
17. $f(x) = e^{-x}, -L < x < L$. You may find useful the facts in (3.9) in Section 3.1, (3.10) in Section 3.1, that is, that

$$\int e^{\alpha x} \cos \beta)\, dx = \frac{e^{\alpha x}}{\alpha^2 + \beta^2} (\beta \sin \beta x + \alpha \cos \beta x),$$

$$\int e^{\alpha x} \sin \beta x\, dx = \frac{e^{\alpha x}}{\alpha^2 + \beta^2} (-\beta \cos \beta x + \alpha \sin \beta x).$$

18. (i) $f(x) = |\pi - x|, -\pi < x < \pi$,
(ii) $f(x) = \pi - |x|, -\pi < x < \pi$
19. $f(x) = \begin{cases} x, & -\pi < x < 0 \\ \dfrac{x}{2}, & 0 < x < \pi \end{cases}$
20. The function $f(x)$ sketched over one period in Figure 9.13
21. The function $f(x)$ sketched over one period in Figure 9.14

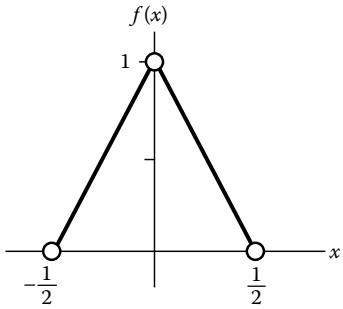


FIGURE 9.13
Graph of function in Problem 9.1.8.20.

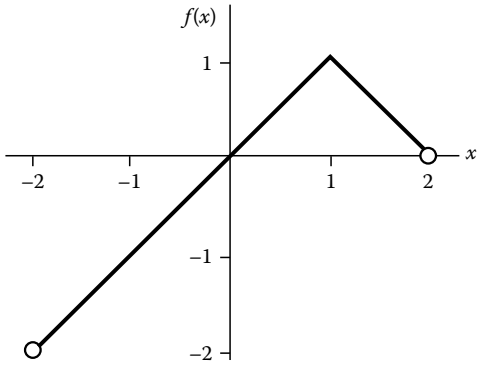


FIGURE 9.14
Graph of function in Problem 9.1.8.21.

22. Show that if $F(x)$ is periodic with period $2L$ and $[\,a, b\,]$ is *any* interval of length $2L$, that is, $b - a = 2L$, then $\int_a^b F(x)\,dx = \int_{-L}^{L} F(x)\,dx$. [This result is useful because one might find it more convenient to integrate on $[\,a, b\,]$ rather than on $[\,-L, L\,]$].
23. Suppose $f(x)$ is even. Explain why Theorem 9.3(b) is true, that is, the Fourier series of f on $[\,-L, L\,]$ consists only of the constant term and the cosine function terms.
24. Suppose $f(x)$ is periodic with period 2π and satisfies $f(x - \pi) = -f(x)$ for all x. Explain why all the even-indexed Fourier coefficients are zero.

9.2 Fourier Cosine and Sine Series

A "Fourier cosine series" or a "Fourier sine series" is a way to analyze a function $f(x)$ that is given only on the interval $[\,0, L\,]$, that is, only the right half of the interval $[\,-L, L\,]$. For this reason a "Fourier sine series" or a "Fourier cosine series" is called a **half-range expansion**. Such half-range expansions are very useful in the process of solving some partial differential equations by a method called *separation of variables*.

We will see that both the ideas and calculations of Fourier cosine and Fourier sine series are based on the same for Fourier series.

9.2.1 Fourier Cosine Series

If a function $f(x)$ is defined* *only* on the interval $[\,0, L\,]$, we may decompose $f(x)$ into its **Fourier cosine series**, a possibly infinite sum:

$$f(x) \doteq \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right), \tag{9.12}$$

where the constants $a_0; a_1, a_2, a_3, \ldots$ are called the **Fourier cosine coefficients**.

This corresponds to a graphical process of

- First *extending* the function f *evenly* to give a function $\breve{f}$ defined* on the interval $[\,-L, L\,]$,
- Next, finding the Fourier series of the even function $\breve{f}$ defined* on the interval $[\,-L, L\,]$,
- Next, noting that there are no sine terms in the Fourier series of $\breve{f}$, and
- Finally, seeing that the Fourier series of $\breve{f}$ gives the Fourier cosine series for the original function f.

As for a Fourier series, an averaging formula gives the values of the Fourier cosine series, $f_{cos}(x)$, at jump points and endpoints of the graph of the function $\breve{f}(x)$.

In general, the Fourier cosine series of a given function f defined* on the interval $[\,0, L\,]$ is the Fourier series of the *even* function $\breve{f}$ defined* on the interval $[\,-L, L\,]$. So,

$$f(x) \doteq f_{cos}(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right),$$

where

$$a_0 = \frac{1}{L}\int_{-L}^{L} \breve{f}(x)\,dx = \frac{2}{L}\int_{0}^{L} f(x)\,dx,$$

and

$$a_n = \frac{1}{L}\int_{-L}^{L} \breve{f}(x)\cos\left(\frac{n\pi x}{L}\right)dx = \frac{2}{L}\int_{0}^{L} f(x)\cos\left(\frac{n\pi x}{L}\right)dx,\ n = 1, 2, \ldots.$$

Even though the existence of the Fourier cosine series of f is accomplished using its even extension, $\breve{f}$, when we calculate the Fourier cosine coefficients we only need to use integrals on the interval $[\,0, L\,]$. So, we need only use values of the original function f, not the artificially created function $\breve{f}$.

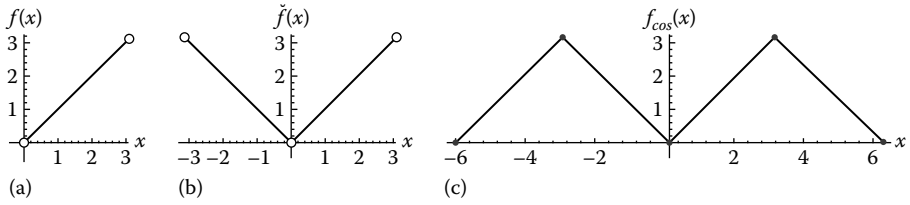


FIGURE 9.15
Graphs of (a) $f(x)$, (b) $\breve{f}(x)$, (c) $f_{\cos}$ in Example 9.8.

Example 9.8

(Graphically) Suppose $f(x) = x$ for $0 < x < L$. Show successively the graphs of (a) $f(x)$ for $0 < x < L$, (b) $\breve{f}(x)$ for $-L < x < L$, and (c) $f_{cos}(x)$, the Fourier cosine series of f on the interval $-L \le x \le 3L$, that is, graphed for two periods of the Fourier cosine series.

Method: Figure 9.15 gives the complete graphical solution. ○

Example 9.9

Find the Fourier cosine series of $f(x) = x$ defined for $0 < x < \pi$.

Method: First, note that $L = \pi$ in this example. The Fourier cosine series of f defined on the interval $(0, \pi)$ is given in the form

$$f(x) \doteq f_{cos}(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos nx,$$

where

$$a_0 = \frac{2}{\pi}\int_0^{\pi} f(x)\,dx, \text{ and } a_n = \frac{2}{\pi}\int_0^{\pi} f(x)\cos nx\,dx.$$

So, we calculate

$$a_0 = \frac{2}{\pi}\int_0^{\pi} x\,dx = \frac{2}{\pi}\left[\frac{1}{2}x^2\right]_0^{\pi} = \pi,$$

$$a_n = \frac{2}{\pi}\int_0^{\pi} x\cos nx\,dx = \frac{2}{\pi}\Big[\frac{x\sin nx}{n} + \frac{\cos nx}{n^2}\Big]_0^{\pi} = \cdots = \frac{2\big((-1)^n - 1\big)}{n^2\pi}.$$

The Fourier cosine series expansion of f is

$$f(x) \doteq f_{cos}(x) = \frac{\pi}{2} + \sum_{n=1}^{\infty} \frac{2\big((-1)^n - 1\big)}{n^2\pi}\cos nx = \frac{\pi}{2} - \frac{4}{\pi}\sum_{k=1}^{\infty} \frac{1}{(2k-1)^2}\cos\big((2k-1)x\big).$$ ○

Recall the concept of a nice function from Definition 9.2 in Section 9.1.

The convergence of the Fourier cosine series of f follows from results about convergence of the Fourier series of $\breve{f}$, the even extension of f defined* on $[\,-L, L\,]$.

Theorem 9.4

If a function f is nice on the interval $[\,0, L\,]$, then at all x in $[\,0, L\,]$ the Fourier cosine series of f converges to the average of $\breve{f}(x^-)$ and $\breve{f}(x^+)$, hence

$$f_{cos}(x) = \frac{f(x^-) + f(x^+)}{2},\ 0 < x < L, \quad f_{cos}(0) = f(0^+), \quad \text{and} \quad f_{cos}(L) = f(L^-).$$

Note that the function $f_{cos}(x)$ is even and periodic with period $2L$.

If a nice function f has $f(x)$ defined and continuous at some x value in the interval $0 < x < L$, then $\breve{f}(x^-) = \breve{f}(x^+) = f(x)$, hence $f_{cos}(x) = f(x)$ at such a point of continuity. The Fourier cosine series agrees with the original function wherever the latter is continuous.

Note also that if f is not defined or not continuous at some x value in the interval $0 < x < L$, but, nevertheless, $f(x^-) = f(x^+)$, then we might as well define an "improved" version of f, say f_{Imp}, by $f_{Imp}(x) = f(x^-) = f(x^+)$, and the Fourier cosine series of f agrees with f_{Imp} at that x.

9.2.2 Fourier Sine Series

If a function $f(x)$ is defined* *only* on the interval $[\,0, L\,]$, we may decompose $f(x)$ into its **Fourier sine series**, a possibly infinite sum of sine functions:

$$f(x) \doteq \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right), \tag{9.13}$$

where the constants $b_1, b_2, b_3, \ldots$ are called the **Fourier sine coefficients**.

This corresponds to a graphical process of

- First *extending* the function f *oddly* to give a function $\breve{f}$ defined* on the interval $[\,-L, L\,]$,
- Next, finding the Fourier series of the odd function $\breve{f}$ defined* on the interval $[\,-L, L\,]$,
- Next, noting that there are no constant and no cosine terms in the Fourier series of $\breve{f}$, and
- Finally, seeing that the Fourier series of $\breve{f}$ gives the Fourier sine series for the original function f.

As usual, an averaging formula gives the values of the Fourier sine series, $f_{sin}(x)$, at jump points and endpoints of the graph of the original $\breve{f}(x)$.

In general, the Fourier sine series of a given function f defined* on the interval $[\,0, L\,]$ is the Fourier series of the *odd* function $\breve{f}$ defined* on the interval $[\,-L, L\,]$. So,

$$f(x) \doteq f_{sin}(x) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right),$$

where

$$b_n = \frac{1}{L}\int_{-L}^{L} \check{f}(x)\sin\left(\frac{n\pi x}{L}\right)dx = \frac{2}{L}\int_{0}^{L} f(x)\sin\left(\frac{n\pi x}{L}\right)dx,\ n = 1, 2, \ldots\ .$$

Even though the existence of the Fourier sine series of f is accomplished using its odd extension, $\check{f}$, when we calculate the Fourier sine coefficients we only need to use integrals on the interval $[\,0, L\,]$. So, we need only use values of the original function f, not the artificially created function $\check{f}$.

Example 9.10

(Graphically) Suppose $f(x) = x$ for $0 < x < L$. Show successively the graphs of (a) $f(x)$ for $0 < x < L$, (b) $\check{f}(x)$ for $-L < x < L$, and (c) $f_{sin}(x)$, the Fourier sine series of f on the interval $-L \le x \le 3L$, that is, graphed for two periods of the Fourier sine series.

Method: Figure 9.16 gives the complete graphical solution. ◯

Example 9.11

Find the Fourier sine series of $f(x) = x$ defined for $0 < x < \pi$.

Method: First, note that $L = \pi$ in this example. The Fourier sine series of f defined* on the interval $(0, \pi)$ is the same as the Fourier series of $\check{f}$, the odd extension of f, defined* on the interval $(-\pi, \pi)$. Because $\check{f}(x)$ is odd on $(-\pi, \pi)$, its Fourier series is given in the form

$$\check{f}(x) \doteq \sum_{n=1}^{\infty} b_n \sin nx,$$

where oddness of $\check{f}$ guarantees that

$$b_n = \frac{2}{\pi}\int_{0}^{\pi} f(x)\sin nx\,dx.$$

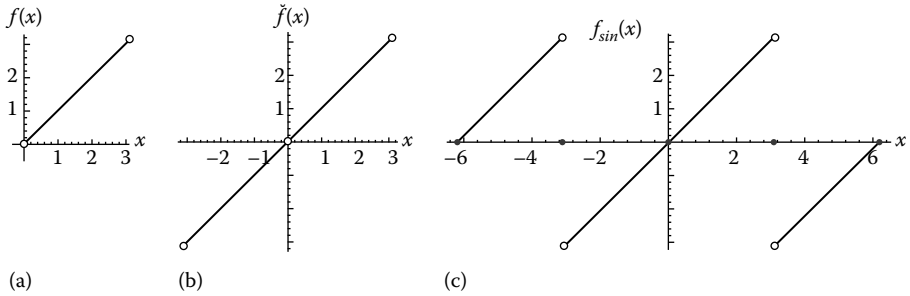


FIGURE 9.16
Graphs of (a) $f(x)$, (b) $f(x)$, (c) f_{sin} in Example 9.10.

So, we calculate

$$b_n = \frac{2}{\pi}\int_0^{\pi} x\sin nx\,dx = \frac{2}{\pi}\left[\frac{x\cos nx}{-n} + \frac{\sin nx}{n^2}\right]_0^{\pi} = \cdots = \frac{2(-1)^{n+1}}{n}.$$

The Fourier sine series expansion of f is

$$f(x) \doteq f_{sin}(x) = -2\cdot\sum_{n=1}^{\infty}\frac{(-1)^n}{n}\sin nx. \bigcirc$$

As to the convergence of the Fourier sine series of f, results about this follow from results about convergence of the Fourier series of $\breve{f}$, the odd extension of f defined* on $[-L, L]$.

Theorem 9.5

If a function f is nice on the interval $[0, L]$, then at all x in $[0, L]$ the Fourier sine series of f converges to the average of $\breve{f}(x^-)$ and $\breve{f}(x^+)$, hence

$$f_{sin}(x) = \frac{f(x^-)+f(x^+)}{2},\ 0 < x < L, \quad \text{and} \quad f_{sin}(0) = f_{sin}(L) = 0.$$

Note that the function $f_{sin}(x)$ is odd and periodic with period $2L$.

If a nice function $f(x)$ is defined* and continuous at x_0 in the interval $(0, L)$, then $\breve{f}(x_0^-) = \breve{f}(x_0^+) = f(x_0)$, hence $f_{sin}(x_0) = f(x_0)$ at such a point of continuity. The Fourier sine series agrees with the original function where the latter is continuous, except possibly at the endpoints $x = 0, x = L$.

Note also that if f is not defined or not continuous at some x value in the interval $0 < x < L$, but, nevertheless, $f(x^-) = f(x^+)$, then we might as well define an "improved" version of f, say f_{Imp}, by $f_{Imp}(x) = f(x^-) = f(x^+)$, and then the Fourier sine series of f agrees with f_{Imp} at that x.

By the way, it's interesting to see the graph of the Fourier series of the function defined by $f(x) = -\sin\left(\frac{x}{2}\right)$ on the interval $0 < x < \pi$. Note that $f_{sin}(x) = -\sin\left(\frac{x}{2}\right)$ only for the intervals $\ldots, -5\pi < x < -3\pi,\ -\pi < x < \pi,\ 3\pi < x < 5\pi, \ldots$, which make up about half of the real line. See Figure 9.17.

Remarks

1. If a function $f(x)$ is given *only* on the interval $[0, L]$, we may decompose $f(x)$ into either a Fourier cosine series or into a Fourier sine series. In an applied context, usually we are interested in one or the other but not both at the same time, unless we have some reason to contrast the two series.

   In particular, given a function f defined* on $[0, L]$, it makes no sense to add the Fourier cosine series and the Fourier sine series. In fact, on the interval $[0, L]$, $f_{cos}(x) + f_{sin}(x) \doteq 2f(x)$, not $f(x)$.
2. Suppose a function is a linear combination of sine functions whose *least* period is $2L$, for example, $f(x) = 2\sin x - \sin 3x$, whose least period is 2π. [Note that f is also

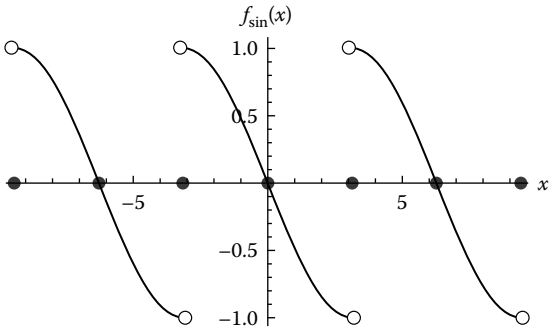


FIGURE 9.17
Graphs of $f_{\sin}$.

periodic with period 4π, 6π, etc.] As we saw in Example 9.6 in Section 9.1, there are several ways to calculate the Fourier series coefficients of $\check{f}$, the odd extension of f: We could calculate the usual integrals, we could use orthogonality, or we could note that

$$f(x) = 2\sin x - \sin 3x \doteq f_{sin}(x) = \sum_{n=1}^{\infty} b_n \sin nx$$

explains why f was already given in the form of its Fourier sine series on $[\,0,\,\pi\,]$, with $b_1 = 2$, $b_3 = -1$, and all other $b_n = 0$. In short,

$$f(x) \doteq f_{sin}(x) = 2\sin x - \sin 3x.$$

Example 9.12

For the function $f(x) = \cos\left(\frac{x}{2}\right)$, defined for $0 < x < 2\pi$, find the Fourier cosine series and Fourier sine series and graph them on an interval of length exactly two periods.

Method:

(a) For a function on the interval $0 < x < 2\pi$, the Fourier cosine series has the form

$$f(x) \doteq f_{cos}(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{nx}{2}\right).$$

So, $f(x) = \cos\left(\frac{x}{2}\right)$ is its own Fourier cosine series, with $a_1 = 1$; $a_n = 0$, for $n = 0, 2, 3, \ldots$. In short,

$$f(x) \doteq f_{cos}(x) = \cos\left(\frac{x}{2}\right).$$

(b) For a function on the interval $0 < x < 2\pi$, the Fourier sine series has the form

$$f(x) \doteq f_{sin}(x) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{nx}{2}\right).$$

We calculate

$$b_n = \frac{2}{2\pi}\int_0^{2\pi} \cos\left(\frac{x}{2}\right)\sin\left(\frac{nx}{2}\right)dx = \frac{2}{2\pi}\cdot\frac{1}{2}\int_0^{2\pi}\left(\sin\left(\frac{(n+1)x}{2}\right) + \sin\left(\frac{(n-1)x}{2}\right)\right)dx.$$

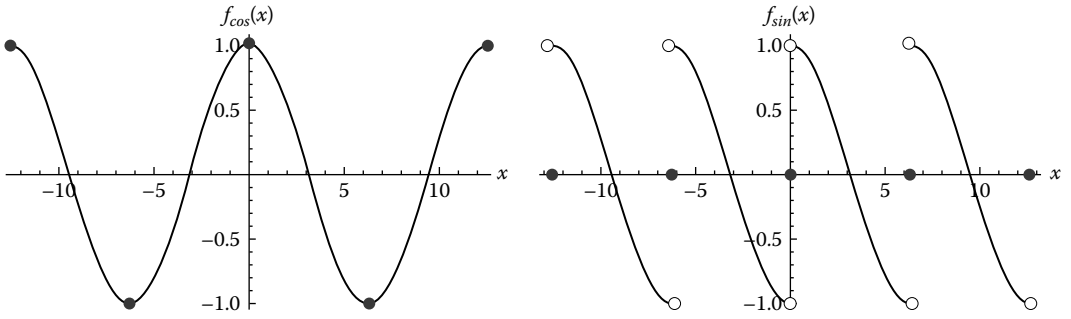


FIGURE 9.18
Example 9.12.

For $n \neq 1$,

$$b_n = \frac{1}{2\pi}\left[\frac{\cos\left(\frac{(n+1)x}{2}\right)}{-\frac{n+1}{2}} + \frac{\cos\left(\frac{(n-1)x}{2}\right)}{-\frac{n-1}{2}}\right]_0^{2\pi} = \cdots = \frac{4n\big((-1)^n - 1\big)}{2\pi(n^2-1)}.$$

For $n = 1$, we have

$$b_1 = \frac{2}{2\pi}\int_0^{2\pi} \cos\left(\frac{x}{2}\right)\sin\left(\frac{x}{2}\right)dx = \frac{2}{2\pi}\cdot\frac{1}{2}\int_0^{2\pi}\sin x\,dx = \frac{1}{2\pi}\big[-\cos x\big]_0^{2\pi} = 0.$$

So, noting that the only nonzero terms involve even n, that is, n that can be written in the form $n = 2k$, we have

$$f(x) \doteq f_{sin}(x) = \frac{8}{\pi}\sum_{k=1}^{\infty}\frac{k}{4k^2-1}\sin kx.$$

(c) (d) The graphs of the Fourier cosine series and the Fourier sine series, each on an interval of length exactly two periods, are given in Figure 9.18. ◯

9.2.3 Fourier Analysis and Oscillations

To solve an ordinary differential equation (ODE) of the form $L[y] = f(t)$, we can decompose $f(t)$ in a Fourier series, Fourier cosine series, or Fourier sine series. This will naturally decompose the solution $y(t)$ in a series. This method is useful for handling an arbitrary function $f(t)$ and is natural in the sense that in the study of oscillations it is a common technique to Fourier analyze either a forcing function or the solution, or both.

This method can be thought of as an alternative to finding a solution in the form of a convolution. A convolution may be elegant in its simple form but may still require a lot of further study, as in Example 4.34 in Section 4.5, in order to say much about the behavior of the solution.

Example 9.13

Use Fourier analysis to solve the IVP of Example 4.34 in Section 4.5 in the particular case of $T = 2\pi$, that is,

$$\left\{\begin{array}{l} \dot{y} + 2y = f(t) \\ \\ y(0) = 0 \end{array}\right\},$$

where the square wave $f(t)$ is periodic with period 2π, that is, is given over one period by

$$f(t) = \left\{\begin{array}{ll} 1, & 0 < t < \pi \\ -1, & \pi < t < 2\pi \end{array}\right\},$$

and compare the solution to that found in Chapter 4.

Method: Recall from Example 9.5 in Section 9.1 that, with $L = \pi$,

$$f(t) \doteq f_s(t) = \frac{4}{\pi} \sum_{k=1}^{\infty} \frac{1}{2k-1} \sin\big((2k-1)t\big).$$

A convergence theorem, specifically Theorem 9.1 in Section 9.1, says that $f(t) = f_s(t)$ except at t being an integer multiple of π.

Our ODE can be rewritten as

$$\dot{y} + 2y = \frac{4}{\pi} \sum_{k=1}^{\infty} \frac{1}{2k-1} \sin\big((2k-1)t\big).$$

Look for a particular solution in the form

$$y_p(t) = \frac{4}{\pi} \sum_{k=1}^{\infty} \frac{1}{2k-1}\, y_{k,p}(t),$$

where $y_{k,p}(t)$ satisfies

$$\dot{y}_{k,p} + 2y_{k,p} = \sin\big((2k-1)t\big),\ k = 1, 2, \ldots . \tag{9.14}$$

The method of undetermined coefficients for (9.14) gives solution form

$$y_{k,p}(t) = A\cos\big((2k-1)t\big) + B\sin\big((2k-1)t\big),$$

where A, B are constants to be determined. Substitute this into ODE (9.14) and sort terms to get

$$\begin{aligned} &\underline{0}\cos\big((2k-1)t\big) + \underline{1}\sin\big((2k-1)t\big) = \dot{y}_{k,p} + 2y_{k,p} \\ &\quad = \big(2A + (2k-1)B\big)\cos\big((2k-1)t\big) + \big(-(2k-1)A + 2B\big)\sin\big((2k-1)t\big). \end{aligned}$$

So, we need

$$\begin{aligned} \begin{bmatrix} A \\ B \end{bmatrix} &= \begin{bmatrix} 2 & 2k-1 \\ -(2k-1) & 2 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \frac{1}{4 + (2k-1)^2} \begin{bmatrix} 2 & -(2k-1) \\ 2k-1 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ &= \frac{1}{4 + (2k-1)^2} \begin{bmatrix} -(2k-1) \\ 2 \end{bmatrix}. \end{aligned}$$

The solutions, for $k = 1, 2, \ldots$, are

$$y_{k,p}(t) = \frac{1}{4 + (2k-1)^2} \left(-(2k-1)\cos\big((2k-1)t\big) + 2\sin\big((2k-1)t\big)\right).$$

Thus

$$y_p(t) = \frac{4}{\pi}\sum_{k=1}^{\infty}\frac{1}{2k-1}\,y_{k,p}(t)$$
$$= \frac{4}{\pi}\sum_{k=1}^{\infty}\frac{1}{(2k-1)(4+(2k-1)^2)}\left(-(2k-1)\cos\left((2k-1)t\right)+2\sin\left((2k-1)t\right)\right).$$

The general solution of the ODE is

$$y(t) = y_h(t) + y_p(t) = c_1e^{-2t} + y_p(t).$$

The initial condition yields

$$0 = y(0) = c_1 + y_p(0) = c_1 + \frac{4}{\pi}\sum_{k=1}^{\infty}\frac{1}{(2k-1)(4+(2k-1)^2)}\left(-(2k-1)\right),$$

hence

$$c_1 = \frac{4}{\pi}\sum_{k=1}^{\infty}\frac{1}{(2k-1)(4+(2k-1)^2)}\cdot\left(2k-1\right).$$

By adding two infinite series together, the solution of the ODE is

$$y(t) = \frac{4}{\pi}\sum_{k=1}^{\infty}\frac{1}{(2k-1)(4+(2k-1)^2)}$$
$$\times\left((2k-1)e^{-2t}-(2k-1)\cos\left((2k-1)t\right)+2\sin\left((2k-1)t\right)\right).$$

Figure 9.19(a) and (b) show graphs of the approximation of the solution $y(t)$ by a finite sum

$$y_N(t) \triangleq \frac{4}{\pi}\sum_{k=1}^{N}\frac{1}{(2k-1)(4+(2k-1)^2)}$$
$$\times\left((2k-1)e^{-2t}-(2k-1)\cos\left((2k-1)t\right)+2\sin\left((2k-1)t\right)\right).$$

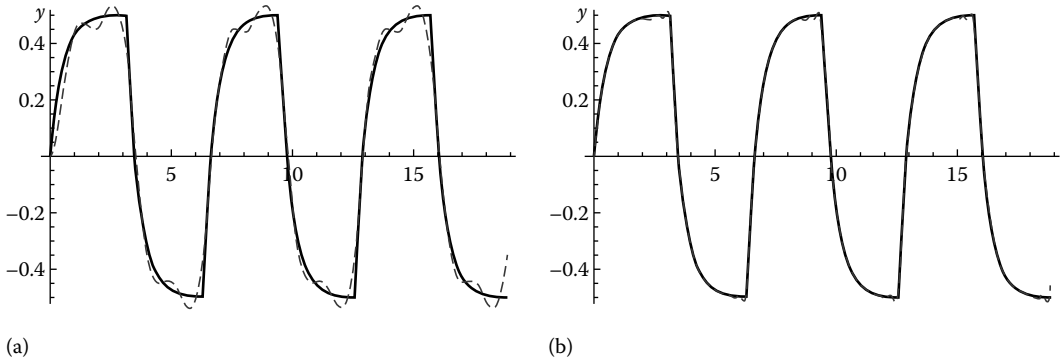


FIGURE 9.19
Fourier analysis and oscillations in Example 9.13. (a) Approximate solution (dashed) $y_2(t)$ and exact solution and (b) approximate solution (dashed) $y_8(t)$ and exact solution.

In each picture the approximation is drawn as a dashed curve and the exact solution (see Example 4.34) in Section 4.5 is drawn as a solid curve. In Figure 9.19a the approximation uses $N = 2$ and in Figure 9.19b the approximation uses $N = 8$. The latter appears to be correct to within 1%. Even with $N = 2$ the approximation is pretty good.

As we discussed in Example 5.41 in Section 5.8, the exact solution is not periodic with period 2π but there is a periodic steady-state solution. The approximate solution $y_N(t)$ is also not periodic with period 2π but the corresponding approximate steady-state solution

$$y_{N,S}(t) \triangleq \frac{4}{\pi} \sum_{k=1}^{N} \frac{1}{(2k-1)(4+(2k-1)^2)} \left(-(2k-1)\cos\left((2k-1)t\right) + 2\sin\left((2k-1)t\right)\right).$$

is close to the exact steady-state solution. ○

9.2.4 Problems

For problems 1–3, for the given functions that are defined* only on the interval given or shown, (a) find the Fourier sine series representation, (b) find the Fourier cosine series representation, and (c) sketch the graphs of those two series, each over an interval of length exactly two periods.

1. The function $f(x) = \frac{1}{2} - \cos 2x$ defined for $0 < x < \pi$.
2. The function whose graph is sketched in Figure 9.20.
3. The function whose graph is sketched in Figure 9.21.

For problems 4–6, for the given functions that are defined* only on the interval given or shown, find the Fourier sine series representation and sketch the graph of the series over an interval of length exactly two periods.

4. The function $f(x) = \cos 3x$ defined for $0 < x < \frac{\pi}{2}$
5. The function defined* on $[\,0,\, L\,]$ by

$$f(x) = \begin{Bmatrix} 2x & 0 < x < \frac{L}{2} \\ L - x & \frac{L}{2} < x < L \end{Bmatrix}.$$

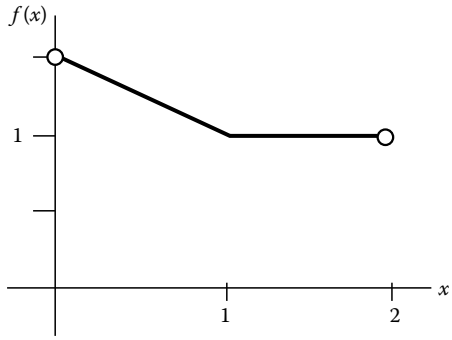


FIGURE 9.20
Function given in Problem 9.2.4.2.

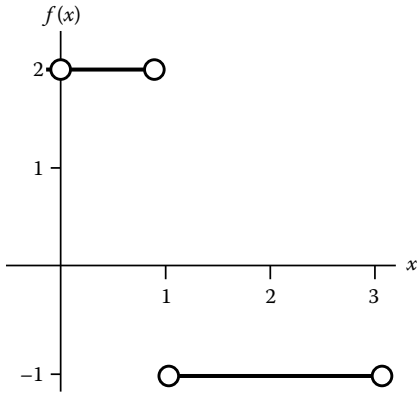


FIGURE 9.21
Graph of function in Problem 9.2.4.3.

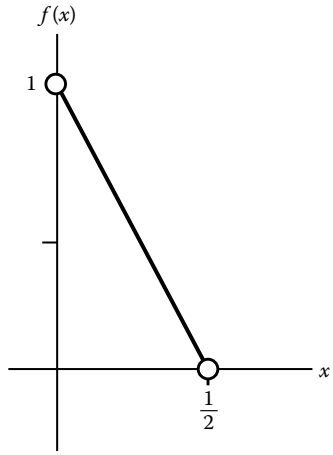


FIGURE 9.22
Graph of function in Problems 9.2.4.6 and 9.2.4.9.

6. The function whose graph is sketched in Figure 9.22.

For problems 7–9, for the given functions that are defined* only on the interval given or shown, find the Fourier cosine series representation and sketch the graph of the series over an interval of length exactly two periods.

7. The function $f(x) = \sin x$ defined for $0 < x < \pi$.
8. The function defined by

$$f(x) = \begin{Bmatrix} 2x & 0 < x < \frac{L}{2} \\ L - x & \frac{L}{2} < x < L \end{Bmatrix}.$$

9. The function whose graph is sketched in Figure 9.22.
10. Suppose $f(x)$ is "even about $x = \frac{L}{2}$," that is, $f(x + \frac{L}{2}) = f(x - \frac{L}{2})$ for $0 < x < L$, and let $\sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right)$ be its Fourier sine series. Explain why $b_n = 0$ for all even n.

In problems 11–13, use Fourier analysis to solve the IVP where $f(t)$ is as in Example 9.13. Give graphs of the exact solution and an approximate solution. [Note: *Mathematica* can plot the graph of a function defined in terms of a definite integral from $u = 0$ to $u = t$, so for the purpose only of plotting the graph of the solution it is not necessary to analyze the convolution as in Example 4.34 in Section 4.5.]

11. $\dot{y} + 3y = f(t),\ y(0) = 0.$
12. $\dot{y} + 3y = f(t),\ y(0) = 1.$
13. $\ddot{y} + 2y = f(t),\ y(0) = 0,\ \dot{y}(0) = 1.$
14. Suppose $f(x)$ is an even function. (a) If $g(x)$ is an even function, explain why $f(g(x))$ is even. (b) if $g(x)$ is an odd function, explain why $f(g(x))$ is even.

9.3 Generalized Fourier Series

Given a function f defined* on the interval $[\,0, L\,]$ we saw that we can expand f in a Fourier cosine series or a Fourier sine series. In this section we will both learn about other series expansions and see how the Fourier sine series and Fourier cosine series expansions are special cases of a general result. As a particularly useful application of the more general result we will learn about some types of expansions that are used in solving Chapter 11 problems involving heat conduction in a one dimensional rod and problems in polar, cylindrical, and spherical coordinates.

Expanding a function as a Fourier sine series

$$f(x) \doteq \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right),\ 0 < x < L, \tag{9.15}$$

or as a Fourier cosine series

$$f(x) \doteq \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right),\ 0 < x < L, \tag{9.16}$$

are unified by both being half-range expansions based on Fourier series.

As we will see, both (9.15) and (9.16) are special cases of a method called "generalized Fourier series expansions." We unify all of these special cases using eigenvalue problems for an ODE on an interval and boundary conditions (BCs) at the endpoints of that interval. We began our study of eigenvalue problems in Section 2.1 when we discussed eigenvalues of a matrix, and we will see how eigenvalue problems for an ODE-BVP (boundary value problem) have some concepts that are analogous to those of the eigenvalue problem for a matrix.

We will start with the special case that reiterates the Fourier sine series expansion. Consider ODE-BVP

$$\left\{\begin{matrix} X''(x) + \lambda X(x) = 0,\ 0 < x < L \\ X(0) = X(L) = 0 \end{matrix}\right\}. \tag{9.17}$$

By an **eigenfunction** of (9.17) we mean a function $X = X(x)$ that (i) satisfies ODE $X''(x) + \lambda X(x) = 0$ on the interval $0 < x < L$, (ii) satisfies the boundary conditions $X(0) = 0$ and $X(L) = 0$, and (iii) is not the identically zero function on $0 < x < L$. Condition (iii) is analogous to the requirement that a matrix's eigenvector not be the zero vector. By an **eigenvalue** of (9.17) we mean a value of λ for which (9.17) has an eigenfunction. This is analogous to saying an eigenvalue of a matrix A is a value of λ for which the homogeneous system of linear algebraic equations $A\mathbf{x} - \lambda\mathbf{x} = \mathbf{0}$ has an eigenvector, that is, a nonzero solution vector $\mathbf{x}$.

It will turn out that the terms $\sin\left(\frac{n\pi x}{L}\right)$ are eigenfunctions for (9.17), so Fourier sine series are related to eigenvalue problem (9.17).

Example 9.14

Explain why ODE-BVP (9.17) has eigenvalues $\lambda_n = \left(\frac{n\pi}{L}\right)^2$ and corresponding eigenfunctions $X_n(x) = \sin\left(\frac{n\pi x}{L}\right)$, for $n = 1, 2, \ldots$

Method: To find the eigenvalues, we must consider three separate cases: $\lambda = 0$, $\lambda > 0$, and $\lambda < 0$.

Case 1: If $\lambda = 0$, then the differential equation $X''(x) + \lambda X(x) = 0$ is $X''(x) = 0$, whose solutions are $X = c_1 + c_2 x$, for arbitrary constants c_1, c_2. Applying the first BC gives $0 = X(0) = c_1 + c_2 \cdot 0 = c_1$, hence $X = c_2 x$. Applying the second BC gives $0 = X(L) = c_2 L$, so $c_2 = 0$. So, $X = c_1 + c_2 x = 0 + 0 \cdot x \equiv 0$. When $\lambda = 0$, the ODE-BVP has only the trivial solution for the function $X(x)$. So, $\lambda = 0$ is not an eigenvalue for this problem.

Case 2: If $\lambda > 0$, it will turn out to be convenient to rewrite $\lambda = \omega^2$, where $\omega \triangleq \sqrt{\lambda}$. Then the differential equation $X''(x) + \lambda X(x) = 0$ is $X''(x) + \omega^2 X(x) = 0$, the undamped harmonic oscillator differential equation of Section 3.3 whose solutions are $X = c_1 \cos\omega x + c_2 \sin\omega x$, for constants c_1, c_2. Applying the first BC gives $0 = X(0) = c_1 1 + c_2 0 = c_1$. So, $X = c_2 \sin\omega x$. Applying the second BC gives $0 = X(L) = c_2 \sin(\omega L)$, so either $c_2 = 0$ or $\sin(\omega L) = 0$. We cannot allow $c_2 = 0$ because then $X = c_2 \sin\omega x = 0 \cdot \sin\omega x \equiv 0$. So, we have the trivial solution for the function $X(x)$ unless ω satisfies the "characteristic equation"

$$\sin(\omega L) = 0. \tag{9.18}$$

But, trigonometry tells us that there are infinitely many values of ω that make this true: $\omega = \frac{n\pi}{L}$, any integer n.

Now, $n = 0$ is an integer but gives $\omega = 0$, hence $\lambda = 0^2$, which is not allowed in Case 2. Additionally, while any integer $n < 0$, say $n = -m$, does give $\omega = \frac{-m\pi}{L}$ that satisfies $\sin(\omega L) = 0$, it turns out that $n < 0$ gives no eigenfunction for $X(x)$ beyond the ones we get for $n > 0$. Why? Because, if $n = -m$ then $X(x) = \sin\omega x = \sin\left(\frac{-m\pi x}{L}\right) = -\sin\left(\frac{m\pi x}{L}\right)$, which essentially duplicates the eigenfunction $\sin\left(\frac{m\pi x}{L}\right)$, and $\left(-\frac{m\pi}{L}\right)^2 = \left(\frac{m\pi}{L}\right)^2$.

The case $\lambda > 0$ gives *eigenvalues* $\lambda_n = \left(\frac{n\pi}{L}\right)^2$, $n = 1, 2, \ldots$, and corresponding *eigenfunctions* $X_n(x) = \sin\left(\frac{n\pi x}{L}\right)$, $n = 1, 2, \ldots$

Case 3: If $\lambda < 0$, it will turn out to be convenient to rewrite $\lambda = -\omega^2$, where $\omega \triangleq \sqrt{-\lambda}$. Then the differential equation $X''(x) + \lambda X(x) = 0$ is $X''(x) - \omega^2 X(x) = 0$, whose solutions are $X = c_1 \cosh(\omega x) + c_2 \sinh(\omega x)$, for arbitrary constants c_1, c_2. Applying the first BC gives $0 = X(0) = c_1 1 + c_2 0 = c_1$. So, $X = c_2 \sinh(\omega x)$. Applying the second BC gives $0 = X(L) = c_2 \sinh(\omega L)$, so either $c_2 = 0$ or $\sinh(\omega L) = 0$. We cannot allow $c_2 = 0$ because then $X = c_2 \sinh(\omega x) = 0 \sinh(\omega x) \equiv 0$. So, we have the trivial solution for the function $X(x)$ unless we insist that $\sinh(\omega L) = 0$. But, there is <u>no</u> value of ω that make

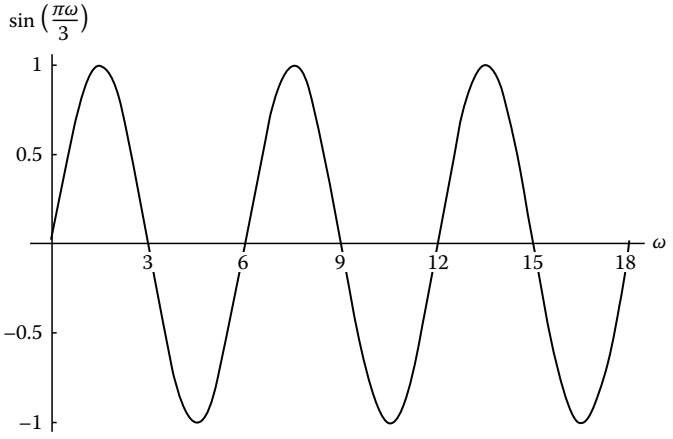


FIGURE 9.23
Finding eigenvalues.

this true, because $\sinh(\theta) > 0$ for all $\theta > 0$. So, Case 3 gives no further eigenvalue or eigenfunction. ◯

Shown in Figure 9.23 is a graph of $y = \sin(\omega L)$ versus ω in a specific example, that of $L = \frac{\pi}{3}$. For this value of L that there are infinitely many roots ω of the **characteristic equation**, $\sin(\omega L) = 0$, specifically $\omega = 3, 6, 9, \ldots$

Recall the concept of a nice function from Definition 9.2 in Section 9.1. A nice function automatically satisfies the **square integrability condition**

$$\int_0^L |f(x)|^2 dx < \infty. \tag{9.19}$$

Remark

Each nice function $f(x)$ can be represented as Fourier sine series

$$f(x) \doteq \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right),\ 0 < x < L,$$

corresponding to BCs $X(0) = X(L) = 0$, where

$$b_n = \frac{\int_0^L f(x) \sin\left(\frac{n\pi x}{L}\right) dx}{\int_0^L |\sin\left(\frac{n\pi x}{L}\right)|^2 dx} = \frac{2}{L} \int_0^L f(x) \sin\left(\frac{n\pi x}{L}\right) dx.$$

The formula for the coefficients b_n followed from the orthogonality relation

$$\int_0^L \sin\left(\frac{n\pi x}{L}\right)\sin\left(\frac{n'\pi x}{L}\right)dx = 0, \text{ for } n \neq n'.$$

9.3.1 Other Boundary Conditions

Now, instead of the boundary conditions $X(0) = 0, X(L) = 0$, there are three other simple pairs of boundary conditions:

1. $X'(0) = 0,\ X'(L) = 0$
2. $X(0) = 0,\ X'(L) = 0$
3. $X'(0) = 0,\ X(L) = 0.$

It turns out that for each pair of boundary conditions, ODE $X'' + \lambda X = 0,\ 0 < x < L$ gives an infinite number of eigenvalues and corresponding eigenfunctions.

The eigenvalue problem $X'' + \lambda X = 0,\ X'(0) = X'(L) = 0$ gives eigenfunctions 1 and $\cos\left(\frac{n\pi x}{L}\right), n = 1, 2, \ldots$, and when we expand a nice function in those eigenfunctions we get the familiar Fourier cosine series expansion

$$f(x) \doteq \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right),\ 0 < x < L.$$

The results are summarized in Table 9.1, at the end of this section. In Problems 9.3.3.2 and 9.3.3.3 you will use some of the same kind of analysis as we did above.

The eigenvalue problem $X'' + \lambda X = 0,\ X(0) = X'(L) = 0$ gives eigenfunctions

$$\sin\left(\frac{(n-\frac{1}{2})\pi x}{L}\right),$$

and when we expand a nice function in those eigenfunctions we get a **generalized Fourier series expansion**

$$f(x) \doteq \sum_{n=1}^{\infty} c_n \sin\left(\frac{(n-\frac{1}{2})\pi x}{L}\right),\ 0 < x < L,$$

where

$$c_n = \frac{\int_0^L f(x)\sin\left(\frac{(n-\frac{1}{2})\pi x}{L}\right)dx}{\int_0^L \left|\sin\left(\frac{(n-\frac{1}{2})\pi x}{L}\right)\right|^2 dx} = \frac{2}{L}\int_0^L f(x)\sin\left(\frac{(n-\frac{1}{2})\pi x}{L}\right)dx.$$

TABLE 9.1

Eigenvalues and Eigenfunctions for $X'' + \lambda X = 0$

Boundary Conditions\Values of λ	**λ > 0**	**λ = 0**	**λ < 0**
$X(0) = X(L) = 0$	$\lambda_n = \left(\frac{n\pi}{L}\right)^2$,	None	None
	$X_n(x) = \sin\left(\frac{n\pi x}{L}\right)$, $n = 1, 2, \ldots$		
$X'(0) = X'(L) = 0$	$\lambda_n = \left(\frac{n\pi}{L}\right)^2$,	$\lambda_0 = 0$,	None
	$X_n(x) = \cos\left(\frac{n\pi x}{L}\right)$, $n = 1, 2, \ldots$	$X_0(x) \equiv 1$	
$X(0) = X'(L) = 0$	$\lambda_n = \left(\frac{(n-\frac{1}{2})\pi}{L}\right)^2$,	None	None
	$X_n(x) = \sin\left(\frac{(n-\frac{1}{2})\pi x}{L}\right)$, $n = 1, 2, \ldots$		
$X'(0) = X(L) = 0$	$\lambda_n = \left(\frac{(n-\frac{1}{2})\pi}{L}\right)^2$,	None	None
	$X_n(x) = \cos\left(\frac{(n-\frac{1}{2})\pi x}{L}\right)$, $n = 1, 2, \ldots$		

The eigenvalue problem $X'' + \lambda X = 0,\ X'(0) = X(L) = 0$ gives eigenfunctions

$$\cos\left(\frac{(n-\frac{1}{2})\pi x}{L}\right),$$

and when we expand a nice function in those eigenfunctions we get a *generalized* Fourier series expansion

$$f(x) \doteq \sum_{n=1}^{\infty} d_n \cos\left(\frac{(n-\frac{1}{2})\pi x}{L}\right),\ 0 < x < L,$$

where

$$d_n = \frac{\int_0^L f(x)\cos\left(\frac{(n-\frac{1}{2})\pi x}{L}\right)dx}{\int_0^L \left|\cos\left(\frac{(n-\frac{1}{2})\pi x}{L}\right)\right|^2 dx} = \frac{2}{L}\int_0^L f(x)\cos\left(\frac{(n-\frac{1}{2})\pi x}{L}\right)dx.$$

The results are summarized in Table 9.1.

So far, we've mentioned ODE $X'' + \lambda X = 0$ along with four choices of BCs, which give Fourier sine series, Fourier cosine series, and two generalized Fourier series, respectively. In Chapter 11 we will use each of these four types of expansions repeatedly to help solve partial differential equations models.

In fact, this can be generalized even more, in anticipation of even more partial differential equation models. We'll start with generalizing the boundary conditions: consider ODE-BVP

$$X''(x) + \lambda X(x) = 0,\ 0 < x < L, \tag{9.20}$$

$$\left.\begin{array}{l} \epsilon_0 X(0) - \epsilon_1 X'(0) = 0, \\ \\ \gamma_0 X(L) + \gamma_1 X'(L) = 0 \end{array}\right\}, \tag{9.21}$$

where $\epsilon_0, \epsilon_1, \gamma_0, \gamma_1$ are given real constants (scalars).

As we will discuss in Section 10.2, boundary conditions (9.21) include many interesting physical models for problems of heat conduction in a thin rod. Let $X(x)$ be the equilibrium temperature distribution. Here are three examples:

1. If $\epsilon_0 = 1$ and $\epsilon_1 = 0$, the BC at $x = 0$ is $X(0) = 0$, which specifies the temperature to be zero there.
2. If $\epsilon_0 = 0$ and $\epsilon_1 = 1$, the BC at $x = 0$ is $X'(0) = 0$, which specifies the heat flux to be zero there, that is, specifies that the left end is insulated.
3. If $\epsilon_0 = h$ and $\epsilon_1 = 1$, the BC at $x = 0$ is $hX(0) = X'(0)$, which specifies that heat is flowing out of the rod at a rate proportional to the amount that the temperature exceeds 0°C.

Theorem 9.6

For eigenvalue problem (9.20) and (9.21), there are infinitely many eigenvalues

$$\lambda_1 < \lambda_2 < \lambda_3 < \cdots \to \infty \tag{9.22}$$

and for each eigenvalue λ_n the only eigenfunctions are multiples of

$$X_n(x) = \cos\left(\sqrt{\lambda_n}\, x - \delta_n\right). \tag{9.23}$$

The positive eigenvalues are found by solving a characteristic equation

$$0 = (\epsilon_0\gamma_0 - \lambda\,\epsilon_1\gamma_1)\sin\left(\sqrt{\lambda}\, L\right) + (\epsilon_0\gamma_1 + \epsilon_1\gamma_0)\sqrt{\lambda}\,\cos\left(\sqrt{\lambda}\, L\right) \tag{9.24}$$

and the eigenfunctions satisfy the orthogonality relation

$$0 = \int_0^L \cos\left(\sqrt{\lambda_n}\, x - \delta_n\right)\cos\left(\sqrt{\lambda_{n'}}\, x - \delta_{n'}\right) dx, \text{ for } n \neq n'. \tag{9.25}$$

Further, we have a generalized Fourier expansion for nice functions

$$f(x) \doteq \sum_{n=1}^{\infty} A_n \cos\left(\sqrt{\lambda_n}\, x - \delta_n\right),\ 0 < x < L,$$

where

$$A_n = \frac{\int_0^L f(x) \cos\left(\sqrt{\lambda_n}\, x - \delta_n\right) dx}{\int_0^L |\cos\left(\sqrt{\lambda_n}\, x - \delta_n\right)|^2 dx}.$$

Finally, at most one eigenvalue can be negative, and, if $\epsilon_0, \epsilon_1, \gamma_0, \gamma_1 \geq 0$ and not both of ϵ_0 and γ_0 are zero, then all eigenvalues are positive.

As we'll see in Section 10.2, the case where $\epsilon_0, \epsilon_1, \gamma_0, \gamma_1 \geq 0$ is physically correct for heat flow because heat tends to flow from a warm place to a cool place.

Example 9.15

Illustrate the results of Theorem 9.6, when $\epsilon_0 = 0, \epsilon_1 = 1, \gamma_0 = 0, \gamma_1 = 1$, that is, the BCs which specify that both ends are insulated.

Method: In Problem 9.3.3.1, you will explain why for these BCs there is an eigenvalue $\lambda_1 = 0$ and there is no negative eigenvalue. The characteristic equation satisfied by all positive eigenvalues is

$$0 = -\lambda \sin\left(\sqrt{\lambda}\, L\right).$$

According to Theorem 9.6, the positive eigenvalues are

$$\left(\frac{\pi}{L}\right)^2 < \left(\frac{2\pi}{L}\right)^2 < \cdots \to \infty.$$

To find the corresponding eigenfunctions, substitute $\lambda = \left(\frac{n\pi}{L}\right)^2$ into the ODE-BVP: The ODE becomes $X''(x) + \left(\frac{n\pi}{L}\right)^2 X(x)$, whose solutions are $X(x) = c_1 \cos\left(\frac{n\pi x}{L}\right) + c_2 \sin\left(\frac{n\pi x}{L}\right)$, which must also satisfy the BCs. Since $X'(x) = \frac{n\pi}{L}\left(-c_1 \sin\left(\frac{n\pi x}{L}\right) + c_2 \cos\left(\frac{n\pi x}{L}\right)\right)$, the BCs imply $0 = X'(0) = \frac{n\pi}{L} c_2$ and $0 = X'(L) = \frac{n\pi}{L}\left(-c_1 \sin n\pi + c_2 \cos n\pi\right) = \frac{n\pi}{L}(-1)^n c_2$, either of which yields $c_2 = 0$. That leaves $X = X(x) = c_1 \left(\cos\left(\frac{n\pi x}{L}\right)\right)$, that is, $X(x)$ is a multiple of a single eigenfunction $X_n(x) \triangleq \cos\left(\frac{(n-1)\pi x}{L}\right)$, for $n = 2, 3, \ldots$.

Putting everything together, we summarize the results for this ODE-BVP: the eigenvalues are

$$0 < \left(\frac{\pi}{L}\right)^2 < \left(\frac{2\pi}{L}\right)^2 < \cdots \to \infty$$

and the corresponding eigenfunctions are $1, \cos\left(\frac{\pi x}{L}\right), \cos\left(\frac{2\pi x}{L}\right), \ldots$. ○

Example 9.16

Illustrate the results of Theorem 9.6 for the ODE-BVP eigenvalue problem

$$\left\{\begin{array}{c} X''(x)+\lambda X(x)=0,\ 0<x<L \\ hX(0)-X'(0)=0, \\ X'(L)=0 \end{array}\right\}, \tag{9.26}$$

where $h>0$ is a given heat transfer coefficient.

Method: First, $\lambda=0$ is not an eigenvalue. Why? If $\lambda=0$, the solutions of ODE $X''+0\cdot X=0$ are $X=c_1+c_2x$. When this is substituted into the two BCs we get $hc_1-c_2=0$ and $c_1+Lc_2=0$, which yields $c_1=c_2=0$. Why? Because the determinant $\begin{vmatrix} h & -1 \\ 1 & L \end{vmatrix}=hL+1>0$ follows from $h>0$ and $L>0$. Also, in Problem 9.3.3.4 you will explain why there is no eigenfunction when $\lambda<0$.

When $\lambda>0$, the solutions of the ODE are $X=c_1\cos(\sqrt{\lambda}\,x)+c_2\sin(\sqrt{\lambda}\,x)$. The BCs require

$$hc_1-c_2\sqrt{\lambda}=0 \quad \text{and} \quad \sqrt{\lambda}\left(-c_1\sin\left(\sqrt{\lambda}\,L\right)+c_2\cos\left(\sqrt{\lambda}\,L\right)\right)=0.$$

Because $h>0$, the first equation gives $c_1=\frac{\sqrt{\lambda}}{h}c_2$. Substituting that into the second equation gives

$$0=\sqrt{\lambda}\left(-\frac{\sqrt{\lambda}}{h}\sin\left(\sqrt{\lambda}\,L\right)+\cos\left(\sqrt{\lambda}\,L\right)\right)c_2.$$

Because an eigenfunction must be not identically zero, we need $c_2\neq 0$ in order to have an eigenvalue. So, λ must satisfy the characteristic equation

$$0=-\frac{\sqrt{\lambda}}{h}\sin\left(\sqrt{\lambda}\,L\right)+\cos\left(\sqrt{\lambda}\,L\right),$$

that is,

$$\frac{\sqrt{\lambda}}{h}\sin\left(\sqrt{\lambda}\,L\right)=\cos\left(\sqrt{\lambda}\,L\right).$$

Dividing through by $\cos(\sqrt{\lambda}\,L)$ and multiplying through by $\frac{h}{\sqrt{\lambda}}$ gives

$$\tan\left(\sqrt{\lambda}\,L\right)=\frac{h}{\sqrt{\lambda}}. \tag{9.27}$$

It is convenient to substitute $\theta=\sqrt{\lambda}\,L$ into (9.27) to get the equivalent equation

$$\tan(\theta)=\frac{hL}{\theta}.$$

The **graphical method** of analysis says to graph the functions $y=\tan(\theta)$ versus θ and $y=\frac{hL}{\theta}$ versus θ on the same axes, so there are infinitely many points of intersection θ_n, and in fact $\theta_n\approx(n-1)\pi$, as $n\to\infty$. The eigenvalues of the original problem are $\lambda_n\triangleq\left(\frac{\theta_n}{L}\right)^2$.

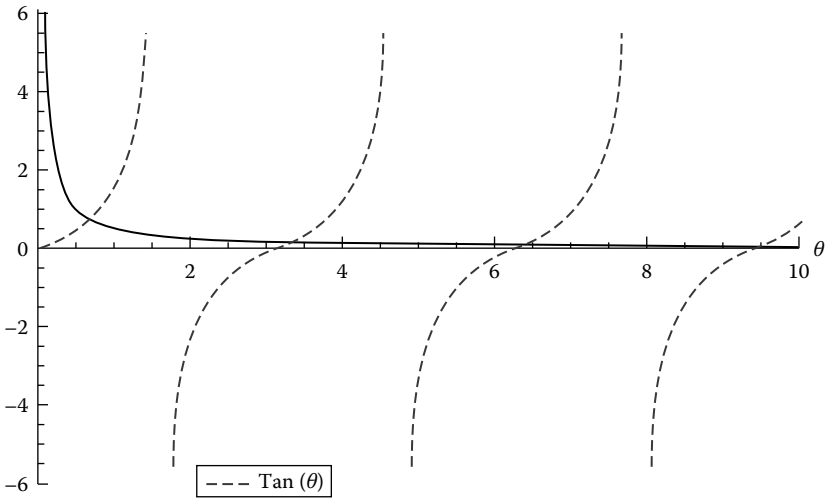


FIGURE 9.24
Graphical method for finding eigenvalues.

As a more specific example, consider the case where $h = \frac{1}{4}$ and $L = 2$, as graphically analyzed in Figure 9.24. The picture suggests that the minimum eigenvalue, λ_1, is approximately 0.7. We proceed to get a better approximation by using Newton's iterative method for finding a root, discussed in Section 8.1. If we define $f(\theta) \triangleq \theta \tan\theta - 0.5$ and take as our initial guess $\theta_0 = 0.7$, then Newton's Method says to take

$$\theta_{n+1} \triangleq \theta_n - \frac{f(\theta_n)}{f'(\theta_n)} = \theta_n - \frac{\theta_n \tan\theta_n - 0.5}{\tan\theta_n + \theta_n \sec^2\theta_n}.$$

A calculator or spreadsheet gives successive iterates $\theta_1 \approx 0.656053888613$, $\theta_2 \approx 0.653281389255$, $\theta_3 \approx 0.653271187232$, $\theta_4 \approx 0.653271187094$, $\theta_5 \approx 0.653271187094$, so $L = 2$ gives the minimum eigenvalue $\lambda_1 = \left(\frac{\theta_5}{2}\right)^2 \approx 0.106690810972$. We will revisit this in Example 9.34 in Section 9.7.

In Problem 9.3.3.5, you will establish why all positive eigenvalues for (9.20) and (9.31) satisfy the characteristic equation (9.24). ○

Table 9.1 gives results about the eigenvalues and eigenfunctions for ODE $X''(x) + \lambda X(x) = 0$ for the four most common sets of separated homogeneous boundary conditions. Each of the four groups of entries corresponds to a different set of physical boundary conditions.

9.3.2 Periodic Boundary Conditions and the Full Fourier Series

Consider ODE-BVP eigenvalue problem

$$\left\{ \begin{array}{c} X''(x) + \lambda X(x) = 0, \ -\pi < x < \pi \\ \\ X(-\pi) - X(\pi) = 0, \\ \\ X'(-\pi) - X'(\pi) = 0 \end{array} \right\}. \tag{9.28}$$

The boundary conditions here are unlike the ones we have studied because each of these two **periodic boundary conditions** mix the behavior of the function $X(x)$ at two points, $\pm\pi$.

If $\lambda = 0$, the ODE in (9.28) becomes $X''(x) = 0$, whose solutions are $X(x) = c_1 + c_2 x$. The first boundary condition requires $c_1 - c_2\pi - (c_1 + c_2\pi) = X(-\pi) - X(\pi) = 0$, hence $-2c_2\pi = 0$, hence $c_2 = 0$. We see that $X(x) \equiv c_1$ satisfies both boundary conditions, so $X_0(x) \equiv 1$ is an eigenfunction corresponding to eigenvalue $\lambda = 0$.

If $\lambda > 0$, the ODE in (9.28) has solutions $X(x) = c_1 \cos(\sqrt{\lambda}\, x) + c_2 \sin(\sqrt{\lambda}\, x)$. The first boundary condition requires

$$0 = X(-\pi) - X(\pi) = c_1 \cos\left(-\sqrt{\lambda}\,\pi\right) + c_2 \sin\left(-\sqrt{\lambda}\,\pi\right) - \left(c_1 \cos\left(\sqrt{\lambda}\,\pi\right) + c_2 \sin\left(\sqrt{\lambda}\,\pi\right)\right),$$

hence $-2c_2 \sin(\sqrt{\lambda}\,\pi) = 0$, after using the facts that cosine is an even function and sine is an odd function. Since $X'(x) = -\sqrt{\lambda}\, c_1 \sin(\sqrt{\lambda}\, x) + \sqrt{\lambda}\, c_2 \cos(\sqrt{\lambda}\, x)$, the second boundary condition requires

$$\begin{aligned} 0 &= X'(-\pi) - X'(\pi) \\ &= \sqrt{\lambda}\left(-c_1 \sin\left(-\sqrt{\lambda}\,\pi\right) + c_2 \cos\left(-\sqrt{\lambda}\,\pi\right) - \left(-c_1 \sin\left(\sqrt{\lambda}\,\pi\right) + c_2 \cos\left(\sqrt{\lambda}\,\pi\right)\right)\right), \end{aligned}$$

hence $2c_1 \sin(\sqrt{\lambda}\,\pi) = 0$. So, λ is an eigenvalue if, and only if, $\sin(\sqrt{\lambda}\,\pi) = 0$. For $n = 1, 2, \ldots$, $\lambda = \lambda_n = n^2$ are eigenvalues, with corresponding eigenfunctions $X(x) = c_1 \cos nx + c_2 \sin nx$, for arbitrary constants c_1, c_2.

Please check for yourself that $\lambda < 0$ does not provide any eigenvalues.

We may write "$\{\cos nx, \sin nx\}$ are the eigenfunctions," meaning that both $\cos nx$ and $\sin nx$ are eigenfunctions corresponding to the same eigenvalue $\lambda_n = n^2$.

So, from ODE-BVP (9.28) we get the full Fourier series expansion of a function on the interval $[-\pi, \pi]$.

9.3.3 Problems

Near the beginning of this section, we studied ODE-BVP (9.17), that is,

$$\left\{\begin{array}{c} X''(x) + \lambda X(x) = 0,\ 0 < x < L \\ X(0) = X(L) = 0 \end{array}\right\}.$$

Using the three cases $\lambda = 0$, $\lambda > 0$, and $\lambda < 0$, we explained why (9.18) has infinitely many solutions for the *eigenvalues* $\lambda_n = \left(\frac{n\pi}{L}\right)^2$, $n = 1, 2, \ldots$ and corresponding *eigenfunctions* $X_n(x) = \sin\left(\frac{n\pi x}{L}\right)$, $n = 1, 2, \ldots$, and for no other values of λ is there a nontrivial solution for $X(x)$, that is, *not* identically zero function.

In problems 1–3 you will use a similar analysis of three cases to establish the other lines of Table 9.1.

1. Establish the second group of entries of Table 9.1.
2. Establish the third group of entries of Table 9.1.
3. Establish the fourth group of entries of Table 9.1.

4. Explain why ODE-BVP eigenvalue problem (9.26) has no eigenvalue $\lambda < 0$ by substituting into the two BCs the solutions of the ODE, in the form $X = c_1 \cosh(\omega x) + c_2 \sinh(\omega x)$, where $\omega \triangleq \sqrt{-\lambda}$.
5. Use analysis similar to that of Example 9.16 to explain why all of the positive eigenvalues for (9.20) and (9.21) satisfy the characteristic equation (9.24). Hint: Substitute the general solution of the ODE into the two BCs (9.21) and write the equations as a homogeneous system $A \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, where A is 2×2 matrix.
6. Find all eigenvalues λ_n and corresponding eigenfunctions $X_n(x)$ for ODE-BVP

$$\left\{ \begin{array}{c} 0 = X''(x) + 2X'(x) + (\lambda + 1)X(x), 0 < x < L, \\ \\ X(0) = 0,\ X(L) = 0 \end{array} \right\}.$$

7. Find all nonnegative eigenvalues λ_n and corresponding eigenfunctions $X_n(x)$ for ODE-BVP

$$\left\{ \begin{array}{c} 0 = X''(x) + 2X'(x) + (\lambda + 1)X(x), 0 < x < L, \\ \\ X'(0) = X'(L) = 0. \end{array} \right\}.$$

8. The function $f(x) = \sin x$ is defined only for $0 < x < \frac{\pi}{2}$. Find the generalized Fourier series for $f(x)$ expanded in the functions $\sin\big((2n-1)x\big), n = 1, 2, \ldots$
9. The function $f(x) = \sin x$ is defined only for $0 < x < \pi$. Find the generalized Fourier series for $f(x)$ expanded in the functions $\cos\big((n - \frac{1}{2})x\big), n = 1, 2, \ldots$
10. Assume that $h > 0$ is a constant. Use a graphical analysis to explain the existence of infinitely many eigenvalues for the problem $X''(x) + \lambda X(x) = 0,\ X'(0) = 0,\ X(\pi) + hX'(\pi) = 0$.
11. Assume that $\mu > 0$. For the ODE-BVP

$$(\star) \quad \left\{ \begin{array}{c} X'' + 2\mu X' + 2\mu^2 X(x) = 0,\ 0 < x < L \\ \\ X(0) = 0, X'(L) = 0 \end{array} \right\},$$

find an equation in μ whose solutions give eigenvalues for $(\star)$. Solve the equation, that is, find the eigenvalues.

9.4 Complex Fourier Series and Fourier Transform

Recall that a function defined* and square integrable on the interval $[\,-L, L\,]$ has Fourier series representation (9.1) in Section 9.1, that is,

$$f(x) \doteq f_s(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right),$$

where

$$a_0 = \frac{1}{L}\int_{-L}^{L} f(x)dx, \quad a_n = \frac{1}{L}\int_{-L}^{L} f(x)\cos\left(\frac{n\pi x}{L}\right)dx, \quad b_n = \frac{1}{L}\int_{-L}^{L} f(x)\sin\left(\frac{n\pi x}{L}\right)dx.$$

Let us use what we know about complex functions:

$$\cos\theta = \frac{1}{2}\left(e^{i\theta} + e^{-i\theta}\right) \quad \text{and} \quad \sin\theta = \frac{1}{2i}\left(e^{i\theta} - e^{-i\theta}\right) = \frac{i}{2}\left(-e^{i\theta} + e^{-i\theta}\right),$$

the latter by using the fact that $\frac{1}{i} = -i$. We can rewrite the Fourier series representation as

$$\begin{aligned} f(x) \doteq f_s(x) &= \frac{a_0}{2} + \frac{1}{2}\sum_{n=1}^{\infty} a_n\left(e^{in\pi x/L} + e^{-in\pi x/L}\right) + \frac{1}{2}\sum_{n=1}^{\infty} ib_n\left(-e^{in\pi x/L} + e^{-in\pi x/L}\right) \\ &= \frac{a_0}{2} + \frac{1}{2}\sum_{n=1}^{\infty}\left(a_n - ib_n\right)e^{in\pi x/L} + \frac{1}{2}\sum_{n=1}^{\infty}\left(a_n + ib_n\right)e^{-in\pi x/L}. \end{aligned}$$

Define complex numbers

$$c_0 = \frac{1}{2}a_0, \quad c_n = \frac{1}{2}\left(a_n - ib_n\right), \quad c_{-n} = \frac{1}{2}\left(a_n + ib_n\right).$$

We have the **complex Fourier series** representation

$$f(x) \doteq f_{\mathbb{C}}(x) = \sum_{n=-\infty}^{\infty} c_n\, e^{in\pi x/L}. \tag{9.29}$$

Define an inner product on complex valued functions by

$$\left\langle f(x), g(x)\right\rangle \triangleq \int_{-L}^{L} f(x)\overline{g(x)}\,dx, \tag{9.30}$$

where $\bar{z}$ denotes the complex conjugate of z. We have

$$\left\langle e^{in\pi x/L}, e^{im\pi x/L}\right\rangle = \int_{-L}^{L} e^{in\pi x/L}\ \overline{e^{im\pi x/L}}\,dx = \int_{-L}^{L} e^{in\pi x/L}\ e^{-im\pi x/L}\,dx = \int_{-L}^{L} e^{i(n-m)\pi x/L}\,dx.$$

For $n \neq m$,

$$\begin{aligned} \left\langle e^{in\pi x/L}, e^{im\pi x/L}\right\rangle &= \left[\frac{e^{i(n-m)\pi x/L}}{(n-m)\pi/L}\right]_{-L}^{L} = \frac{L}{(n-m)\pi}\left(e^{i(n-m)\pi} - e^{-i(n-m)\pi}\right) \\ &= \frac{L}{(n-m)\pi}\left((-1)^{n-m} - (-1)^{-(n-m)}\right) = 0, \end{aligned}$$

because $(-1)^{-(n-m)} = (-1)^{n-m}$ follows from $\left((-1)^{n-m}\right)^2 = 1$ for all integers n, m.

On the other hand, if $n = m$, then

$$\left\langle e^{in\pi x/L}, e^{im\pi x/L} \right\rangle = \int_{-L}^{L} e^{i(n-n)\pi x/L}\, dx = \int_{-L}^{L} 1\, dx = 2L.$$

To summarize, $\{e^{in\pi x/L}\}_{n=1}^{\infty}$ is an orthogonal set of functions. Formally, that is, without rigorous proof,

$$\left\langle f(x), e^{im\pi x/L} \right\rangle = \left\langle \sum_{n=-\infty}^{\infty} c_n\, e^{in\pi x/L},\ e^{im\pi x/L} \right\rangle = \sum_{n=-\infty}^{\infty} c_n \begin{Bmatrix} 2L, & n = m \\ 0, & n \neq m \end{Bmatrix} = c_m \cdot 2L.$$

It follows that the **complex Fourier series coefficients** in (9.29) are given by

$$c_n = \frac{1}{2L} \int_{-L}^{L} f(x)\, e^{-in\pi x/L}\, dx = \frac{1}{2L} \int_{0}^{2L} f(\xi)\, e^{-in\pi \xi/L}\, d\xi. \tag{9.31}$$

The latter follows from the result of Problem 9.1.8.22, which says that the integral is the same over all intervals of length $2L$. So,

$$f(x) \doteq f_{\mathbb{C}}(x) = \frac{1}{2L} \sum_{n=-\infty}^{\infty} \left(\int_{0}^{2L} f(\xi)\, e^{-in\pi \xi/L}\, d\xi \right) e^{in\pi x/L} \tag{9.32}$$

Example 9.17

Find the complex Fourier series of the function defined only by

$$f(x) = \begin{Bmatrix} \sin 2x, & -\pi < x < 0 \\ 0, & 0 < x < \pi \end{Bmatrix}.$$

Method: Let $L = \pi$.

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\, e^{-inx}\, dx = \frac{1}{2\pi} \int_{-\pi}^{0} \sin 2x\, e^{-inx}\, dx = \frac{1}{2\pi} \int_{-\pi}^{0} \frac{1}{2i} \left(e^{i2x} - e^{-i2x} \right) e^{-inx}\, dx,$$

that is,

$$c_n = \frac{1}{4\pi i} \int_{-\pi}^{0} \left(e^{-i(n-2)x} - e^{-i(n+2)x} \right) dx. \tag{9.33}$$

We have to consider separately the cases $n = 2$, $n = -2$, and $|n| \neq 2$. For $n = 2$,

$$c_2 = \frac{1}{4\pi i} \int_{-\pi}^{0} \left(1 - e^{-i4x} \right) dx = \frac{1}{4\pi i} \left[x - \frac{1}{-i4} e^{-i4x} \right]_{-\pi}^{0} = \cdots = \frac{1}{4\pi i} \cdot \pi = -\frac{i}{4}.$$

For $n = -2$,

$$c_{-2} = \frac{1}{4\pi i}\int_{-\pi}^{0}\left(e^{i4x} - 1\right)dx = -c_2 = \frac{i}{4}.$$

For $|n| \neq 2$, (9.33) implies

$$c_n = \frac{1}{4\pi i}\left[\frac{e^{-i(n-2)x}}{-i(n-2)} - \frac{e^{-i(n+2)x}}{-i(n+2)}\right]_{-\pi}^{0}.$$

Using $i^2 = -1$,

$$c_n = \frac{1}{4\pi}\left(\frac{1}{(n-2)}\left(1-(-1)^{-(n-2)}\right) - \frac{1}{(n+2)}\left(1-(-1)^{-(n+2)}\right)\right)$$

$$= \frac{1}{4\pi}\left(1-(-1)^n\right)\cdot\left(\frac{1}{n-2} - \frac{1}{n+2}\right) = \frac{4}{4\pi(n^2-4)}\left(1-(-1)^n\right).$$

For $n = even$, $\left(1-(-1)^n\right) = 0$; for $n = odd = 2k+1$, $\left(1-(-1)^n\right) = 2$. So, the complex Fourier series of $f(x)$ is

$$f(x) \doteq f_{\mathbb{C}}(x) = \frac{i}{4}\cdot\left(e^{-i2x} - e^{i2x}\right) + \frac{2}{\pi}\sum_{k=-\infty}^{\infty}\frac{1}{(2k+1)^2-4}\,e^{i(2k+1)x}. \bigcirc$$

Example 9.18

Find the complex Fourier series of the function defined only by

$$f(x) = 3 - 2\cos 2x + \sin 5x, \quad -\pi < x < \pi.$$

Method: Let $L = \pi$. Instead of calculating complex Fourier coefficients we can directly write

$$f(x) = 3 - 2\cdot\frac{1}{2}\left(e^{i2x} + e^{-i2x}\right) + \frac{i}{2}\left(-e^{i5x} + e^{-i5x}\right),$$

that is,

$$f(x) = \frac{i}{2}\,e^{-i5x} - e^{-i2x} + 3 - e^{i2x} - \frac{i}{2}\,e^{i5x}. \bigcirc$$

Note that Example 9.18's method couldn't have been used in Example 9.17 because its function $f(x)$ is not simply $\sin 2x$ on the whole interval $-\pi < x < \pi$.

9.4.1 The Fourier Transform

Definition 9.3

A function $f(x)$ is **absolutely integrable** on $\mathbb{R}$ if the improper integral $\int_{-\infty}^{\infty}|f(x)|\,dx$ converges.

Definition 9.4

If $f(x)$ is absolutely integrable on $\mathbb{R}$ then its **Fourier transform** is defined by the improper integral

$$F(\omega) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x)\, e^{-i\omega x}\, dx. \tag{9.34}$$

Note that $F(\omega)$ is a function of the parameter ω that appears in the integrand.

Theorem 9.7

(Fourier Inversion Theorem) If $f(x)$ is absolutely integrable on $\mathbb{R}$ and is piecewise smooth on every finite interval, then

$$f(x) \doteq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega)\, e^{i\omega x}\, d\omega, \tag{9.35}$$

in the sense that at every x_0

$$\frac{f(x_0^+) + f(x_0^-)}{2} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega)\, e^{i\omega x_0}\, d\omega. \tag{9.36}$$

If, in addition, $f(x)$ is continuous at x_0 then

$$f(x_0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega)\, e^{i\omega x_0}\, d\omega.$$

Note that the conclusion of Theorem 9.7 at a point where $f(x)$ has a finite jump discontinuity is reminiscent of the result for convergence of Fourier series.

The Fourier transform and its inversion theorem are very useful for solving some physical problems, as we will see in Section 17.2. A reference for rigorous proofs for the existence of the Fourier transform and its inversion theorem will be given in the "Learn More About It" found at the end of this section.

In the real world we deal with finite, albeit indefinitely large intervals, so letting $L \to \infty$ for intervals $[\,-L, L\,]$ motivates the Fourier transform on $(-\infty, \infty)$.

For fixed positive L, define

$$\omega_n = \frac{n\pi}{L} \quad \text{and} \quad \triangle\omega \triangleq \omega_{n+1} - \omega_n = \frac{\pi}{L}.$$

The complex Fourier series on $[-L, L]$ given in (9.29) can be rewritten as

$$f(x) \doteq f_{\mathbb{C}}(x) = \sum_{n=-\infty}^{\infty} c_n \, e^{i\omega_n x} \tag{9.37}$$

where

$$c_n = \frac{1}{2L} \int_{-L}^{L} f(x) \, e^{-i\omega_n x} \, dx.$$

We will explain why (9.37) is a Riemann sum for the integral in (9.35): Rewrite (9.37) as

$$f(x) \doteq f_{\mathbb{C}}(x) = \sum_{n=-\infty}^{\infty} \frac{L}{\pi} \cdot c_n \, e^{i\omega_n x} \triangle\omega = \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} \left(\sqrt{2\pi} \cdot \frac{L}{\pi} \cdot \frac{1}{2L} \int_{-L}^{L} f(x) \, e^{-i\omega_n x} dx \right) e^{i\omega_n x} \triangle\omega,$$

that is,

$$f(x) \doteq f_{\mathbb{C}}(x) = \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} \int_{-L}^{L} f(x) \, e^{-i\omega_n x} \, dx \right) e^{i\omega_n x} \triangle\omega.$$

So,

$$f(x) \doteq f_{\mathbb{C}}(x) = \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} F_L(\omega_n) e^{i\omega_n x} \triangle\omega, \tag{9.38}$$

where

$$F_L(\omega) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-L}^{L} f(x) \, e^{-i\omega x} \, dx.$$

By definition of principal value,

$$\lim_{L\to\infty} F_L(\omega) = P.V. \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) \, e^{-i\omega x} \, d\omega = P.V. \, F(\omega),$$

where $F(\omega)$, the Fourier transform, was defined in (9.34).

In a non-rigorous way, take the limit of (9.38), as $L \to \infty$, to get

$$f(x) \doteq f_{\mathbb{C}}(x) = P.V. \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} F(\omega_n) e^{i\omega_n x} \triangle\omega \, . \tag{9.39}$$

Also in a non-rigorous way, take the limit of (9.39), as $\triangle\omega \to 0^+$, to get

$$f(x) \doteq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega) e^{i\omega x}\, d\omega.$$

The above calculations motivate why we can think of the Fourier transform as a continuous version of complex Fourier series.

Just as the symbol $\mathcal{L}$ stands for the Laplace transform operator and $\mathcal{L}^{-1}$ for the inverse Laplace transform operator, we denote the Fourier transform operator and its inverse operator by

$$\mathcal{F}[f(t)] = F(\omega) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t)\, e^{-i\omega t} dt \tag{9.40}$$

and

$$f(t) = \mathcal{F}^{-1}[\,F(\omega)] \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega)\, e^{i\omega t} dt. \tag{9.41}$$

We caution that there are many slightly different definitions of the Fourier transform and corresponding inverse transform in current use. For example, some authors dispense with the factor $\frac{1}{\sqrt{2\pi}}$ in the Fourier transform but then have a factor of $\frac{1}{2\pi}$ in the inverse Fourier transform. Other authors replace $e^{-i\omega t}$ by $e^{-i2\pi\omega t}$ and $e^{i\omega t}$ by $e^{i2\pi\omega t}$, in which case they do not need the factor of $\frac{1}{\sqrt{2\pi}}$ in either the Fourier or inverse Fourier transform.

Example 9.19

Let a be a positive constant. Find the Fourier transform of the **finite pulse**

$$f(t) = \begin{Bmatrix} 1, & -a < t < a \\ 0, & |t| > a \end{Bmatrix}.$$

Method: $\mathcal{F}[f(t)\,](\omega) = \frac{1}{\sqrt{2\pi}} \left(\int_{-\infty}^{-a} 0\, e^{-i\omega t}\, dt + \int_{-a}^{a} 1\, e^{-i\omega t} dt + \int_{a}^{\infty} 0\, e^{-i\omega t} dt \right)$, so for $\omega \neq 0$,

$$\mathcal{F}[f(t)\,](\omega) = \frac{1}{\sqrt{2\pi}} \left[\frac{e^{-i\omega t}}{-i\omega} \right]_{-a}^{a} = \frac{1}{\sqrt{2\pi}} \cdot \frac{2}{\omega} \cdot \frac{-1}{2i} \left(e^{-ia\omega} - e^{ia\omega} \right) = \sqrt{\frac{2}{\pi}} \cdot \frac{\sin(a\omega)}{\omega}.$$

For $\omega = 0$,

$$\mathcal{F}[f(t)\,](0) = \frac{1}{\sqrt{2\pi}} \int_{-a}^{a} 1\, dt = \frac{1}{\sqrt{2\pi}} \Big[t \Big]_{-a}^{a} = \sqrt{\frac{2}{\pi}} \cdot a.$$

So,

$$\mathcal{F}[f(t)\,](\omega) = a \cdot \sqrt{\frac{2}{\pi}} \cdot \begin{Bmatrix} \dfrac{\sin(a\omega)}{a\omega}, & \omega \neq 0 \\ 1, & \omega = 0 \end{Bmatrix}. \bigcirc$$

Note that in Example 9.19, the function $f(t)$ has a jump discontinuity but its Fourier transform, the function $F(\omega)$, is continuous everywhere. This is why people say that the operation of taking Fourier transform is "smoothing."

Example 9.20

(Frequency shift) (a) For all absolutely integrable functions $f(t)$ and constants a explain why

$$(a)\ \ \mathcal{F}[\, e^{iat} f(t)\,](\omega) = F(\omega - a) \quad \text{and} \quad (b)\ \ \mathcal{F}^{-1}[\, F(\omega - a)\,] = e^{iat} f(t), \tag{9.42}$$

where $F(\omega) = \mathcal{F}[f(t)\,](\omega)$ is assumed to exist.

Method: (a) Because there exists $\mathcal{F}[f(t)\,](\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t)\, e^{-i\omega t} dt$, we calculate that

$$\mathcal{F}[\, e^{iat} f(t)\,](\omega) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} f(t)\, e^{iat}\, e^{-i\omega t}\, dt = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} f(t)\, e^{-i(\omega - a)t}\, dt = F(\omega - a).$$

(b) Follows from taking the inverse transform of the result in part (a). ○

Example 9.21

Find the Fourier transform of an example of a **finite wave train** given by

$$f(x) = \begin{Bmatrix} \sin 2x, & 0 \le x \le 3\pi \\ \\ 0, & \text{all other } x \end{Bmatrix}.$$

Method: Using a well-known indefinite integration formula for an exponential function times a sine function, we calculate that, for $\omega \neq \pm 2$,

$$\begin{aligned} \mathcal{F}[f(x)](\omega) &\triangleq \frac{1}{\sqrt{2\pi}} \int\limits_{0}^{3\pi} \sin 2x\, e^{-i\omega x} dx = \frac{1}{\sqrt{2\pi}} \left[\frac{e^{-i\omega x}}{(-i\omega)^2 + 2^2} \big(-i\omega \sin 2x - 2\cos 2x \big) \right]_0^{3\pi} \\ &= \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{4 - \omega^2} \cdot \left(-2e^{-i3\omega\pi} + 2 \right) = \sqrt{\frac{2}{\pi}} \cdot \frac{1 - e^{-i3\omega\pi}}{4 - \omega^2}. \end{aligned}$$

For $\omega = \pm 2$,

$$\mathcal{F}[f(x)](\pm 2) \triangleq \frac{1}{\sqrt{2\pi}} \int\limits_{0}^{3\pi} \frac{i}{2} \left(-e^{i2x} + e^{-i2x} \right) e^{\mp i2x}\, dx = \cdots = \sqrt{\frac{2}{\pi}} \cdot \left(\mp \frac{3\pi i}{4} \right).$$

So

$$\mathcal{F}[f(x)\,](\omega) = \sqrt{\frac{2}{\pi}} \cdot \begin{Bmatrix} \dfrac{1 - e^{-i3\omega\pi}}{4 - \omega^2}, & |\omega| \neq 2 \\ \\ \mp \dfrac{3\pi i}{4}, & \omega = \pm 2 \end{Bmatrix}. \bigcirc$$

By the way, using L'Hôpital's Rule we can check that the Fourier transform in Example 9.21 is continuous at $\omega = \pm 2$.

Theorem 9.8

Suppose a is a positive constant and $F(\omega) = \mathcal{F}[f(x)](\omega)$ exists. Then

$$\mathcal{F}[f(ax)](\omega) = \frac{1}{a} F\left(\frac{\omega}{a}\right) \quad \text{and} \quad \mathcal{F}^{-1}\left[F\left(\frac{\omega}{a}\right)\right](x) = a\,\mathcal{F}^{-1}[F(\omega)](x). \tag{9.43}$$

Why? A non-rigorous explanation, that is, one that does not consider issues of convergence, uses the change of variables $y = ax$ to get

$$\mathcal{F}[f(ax)](\omega) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(ax) e^{-i\omega x}\, dx = \cdots = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(y) e^{-i\omega(y/a)} (a^{-1} dy) = a^{-1} F\left(\frac{\omega}{a}\right). \square$$

The next result doubles the usefulness of tables of Fourier transforms because we can reverse the roles of Fourier transform and inverse Fourier transform. This was not possible for the Laplace transform.

Theorem 9.9

$$(a)\ \ \mathcal{F}^{-1}[F(\omega)](x) = \mathcal{F}[F(\omega)](-x) \quad \text{and} \quad (b)\ \ \mathcal{F}[f(x)](\omega) = \mathcal{F}^{-1}[f(x)](-\omega), \tag{9.44}$$

where $\mathcal{F}[f(x)](\omega)$ and $\mathcal{F}[F(\omega)](x)$ are assumed to exist.

Why? The only difference in *form* between the definitions of $\mathcal{F}$ and $\mathcal{F}^{-1}$ is $e^{-i\omega x}$ versus $e^{i\omega x}$. $\square$

9.4.2 Convolution

Define **convolution** on the real line by

$$(f * g)(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(x - \xi) f(\xi)\, d\xi.$$

Note that, in general, rigorous calculation of the improper integral in a convolution involves limiting processes.

For any constant a, define a function for $-\infty < t < \infty$ by

$$Step(t - a) \triangleq \begin{Bmatrix} 1, & t > a \\ \\ 0, & t < a \end{Bmatrix}. \tag{9.45}$$

Note that the function $step(t-a)$, as opposed to $Step(t-a)$, is defined only for $0 \le t < \infty$. We used step$(t-a)$ in the context of Laplace transforms in Chapter 4.

Example 9.22

Find $Step(t-a) * Step(t-b)$, assuming a and b are constants with $a < b$.

Method: $Step(t-a) * Step(t-b) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} Step\big((t-\xi)-a\big)\, Step(\xi - b)\, d\xi$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \begin{Bmatrix} 1, & t-a-\xi > 0 \\ 0, & t-a-\xi < 0 \end{Bmatrix} \begin{Bmatrix} 1, & \xi - b > 0 \\ 0, & \xi - b < 0 \end{Bmatrix} d\xi$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \begin{Bmatrix} 1, & \xi < t-a \\ 0, & \xi > t-a \end{Bmatrix} \begin{Bmatrix} 1, & \xi > b \\ 0, & \xi < b \end{Bmatrix} d\xi = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \begin{Bmatrix} 1, & b < \xi < t-a \\ 0, & \xi > t-a \text{ or } \xi < b \end{Bmatrix}.$$

We analyze this by considering two cases.
Case (1): If $b < t-a$, that is, $t > a+b$, then

$$Step(t-a) * Step(t-b) = \frac{1}{\sqrt{2\pi}} \int_{b}^{t-a} 1\, d\xi = \frac{1}{\sqrt{2\pi}} \cdot \big((t-a)-b\big).$$

Case (2): If $b > t-a$, that is, $t < a+b$, then $Step(t-a) * Step(t-b) = \frac{1}{\sqrt{2\pi}} \int_{b}^{t-a} 0\, d\xi = 0$.
So,

$$Step(t-a) * Step(t-b) = \frac{1}{\sqrt{2\pi}} \big(t-(a+b)\big) Step\big(t-(a+b)\big). \bigcirc \tag{9.46}$$

By the way, the function on the right-hand side of (9.46) is called a **ramp function**. A graph of such a function is shown in Figure 9.25.

Theorem 9.10

(Convolution)

$$\mathcal{F}[\,(f*g)(x)\,] = \mathcal{F}[f(x)\,] \cdot \mathcal{F}[\,g(x)\,] \quad \text{and} \quad (f*g)(x) = \mathcal{F}^{-1}[\,F(\omega)\,G(\omega)\,], \tag{9.47}$$

assuming $F(\omega) \triangleq \mathcal{F}[f(x)\,](\omega)$ and $G(\omega) \triangleq \mathcal{F}[\,g(x)\,](\omega)$ exist.

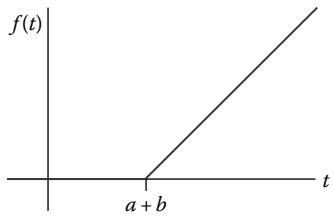


FIGURE 9.25
Example 9.22: Ramp function.

TABLE 9.2

Fourier Transforms: $a > 0, \beta > 0, k > 0, \tau > 0, c, t_0$ Are Nonzero Constants

	$f(x)$	$F(\omega) = \mathcal{F}[f(x)]$
F.1	$f'(x)$	$i\omega F(\omega)$
F.2	$e^{iat} f(t)$	$F(\omega - a)$
F.3	$f(t) = \begin{cases} 1, & -a < t < a \\ 0, & \lvert t\rvert > a \end{cases}$	$a \cdot \sqrt{\frac{2}{\pi}} \cdot \begin{cases} \frac{\sin(a\omega)}{a\omega}, & \omega \neq 0 \\ 1, & \omega = 0 \end{cases}$
F.4	$\operatorname{sinc}(t) \triangleq \frac{\sin t}{t},\ t \neq 0$	$\sqrt{\frac{\pi}{2}} \cdot \begin{cases} 1, & -1 < \omega < 1 \\ \frac{1}{2}, & \omega = \pm 1 \\ 0, & \lvert\omega\rvert > 1 \end{cases}$
F.5	$\frac{1}{a^2+t^2}$	$\sqrt{\frac{\pi}{2}} \cdot \frac{e^{-\lvert a\omega\rvert}}{\lvert a\rvert}$
F.6	$\frac{e^{-\lvert ax\rvert}}{\lvert a\rvert}$	$\sqrt{\frac{2}{\pi}} \cdot \frac{1}{a^2+\omega^2}$
F.7	$e^{-x^2/2}$	$e^{-\omega^2/2}$
F.8	$f(ax)$	$\frac{1}{a} F\left(\frac{\omega}{a}\right)$
F.9	$e^{-\beta x^2}$	$\frac{1}{\sqrt{2\beta}} \cdot e^{-\omega^2/(4\beta)}$
F.10	$f(t - t_0)$	$e^{-i\omega t_0} F(\omega)$
F.11	$e^{-at}\, Step(t)$, where $a > 0$	$\frac{1}{\sqrt{2\pi}} \frac{1}{a+i\omega}$
F.12	$(f * g)(t)$	$F(\omega)G(\omega)$
F.13	$f(x) \cos kx$	$\frac{1}{2}\left(F(\omega - k) + F(\omega + k)\right)$
F.14	$f(x) \sin kx$	$\frac{1}{i2}\left(F(\omega - k) - F(\omega + k)\right)$
F.15	$f(x) = \begin{cases} \sin kx, & 0 \leq x \leq \tau \\ 0, & \text{all other } x \end{cases}$	$\sqrt{\frac{2}{\pi}} \cdot \frac{1}{k^2-\omega^2}\left(e^{-i\omega\tau}\left(-i\omega \sin k\tau - k \cos k\tau\right) + k\right)$

You will give a non-rigorous explanation for Theorem 9.10, that is, Table 9.2 entry F.12, in Problem 17.1.2.8.

Note that the function $Step(t-a)$ is not absolutely integrable on the real line so its Fourier transform is not defined. So we could not use Theorem 9.10 to do Example 9.22. However, the mathematical subject of the "theory of distributions" (Friedman 1991) can make rigorous use of Fourier transforms of functions that are not absolutely integrable as well as the "delta" function.

Theorem 9.11

(Time delay)

$$\mathcal{F}[f(t-t_0)\,](\omega) = e^{-i\omega t_0}\mathcal{F}[f(t)\,](\omega), \tag{9.48}$$

where $\mathcal{F}[f(t)\,](\omega)$ is assumed to exist.

Learn More About It

A mathematically rigorous proof of the Fourier inversion theorem, Theorem 9.7, is found in *Complex Variables*, 2*nd* edn., by Stephen D. Fisher, Dover Publications © 1999.

9.4.3 Problems

In problems 1–4 find the complex Fourier series of the given function defined* only on the given interval.

1. $f(x) = \begin{cases} \cos 3x, & -\pi < x < 0 \\ 0, & 0 < x < \pi \end{cases}$
2. $f(t) = \begin{cases} -1, & 0 \le t \le 4 \\ 2, & 4 \le t \le 6 \end{cases}$
3. $f(x) = 1 - i3\sin 2x + \frac{1}{2}\cos 5x, \ -\pi < x < \pi.$
4. $f(t) = e^{i\pi t/2}, -1 \le t \le 1.$

In problems 5–8 find the Fourier transform of the given function.

5. $f(t) = Step(t-1) - Step(t-3)$, where $Step(t-a)$ was defined in (9.45).
6. $f(t) \triangleq \begin{cases} t, & 0 < t < 1 \\ 2-t, & 1 < t < 2 \end{cases}$
7. $f(t) \triangleq \begin{cases} e^{-\alpha t}\sin(\omega_0 t), & t \ge 0 \\ 0, & -\infty < t < 0 \end{cases}$, where α is a positive constant.
8. $f(t) \triangleq \begin{cases} \frac{1}{\nu}\left(1+\cos(2\pi\nu t)\right), & |t| < \frac{1}{2\nu} \\ 0, & \text{all other } t \end{cases}$, where ν is a positive constant.
9. Suppose a is a positive constant. Use the definition of Fourier transform to explain table entry F.11 in Table 17.1 in Section 17.1, that is, why for $a > 0$, $\mathcal{F}[\,e^{-at}\,Step(t)\,] = \dfrac{1}{\sqrt{2\pi}}\,\dfrac{1}{a+i\omega}$.

10. Find the Fourier transform of $tf(t)$ in terms of the Fourier transform of the unspecified function $f(t)$.
11. Find the inverse Fourier transform of $\frac{1-e^{-i5\pi\omega}}{4-\omega^2}$.

In problems 12–14 use the method of Theorem 9.8 and/or the results of previous examples to find the Fourier or inverse Fourier transform of the given function.

12. $\mathcal{F}\big[f(t)\,](\omega)$, where

$$f(t) \triangleq \begin{Bmatrix} \sin x, & 0 \le x \le 6\pi \\ 0, & \text{all other } x \end{Bmatrix}$$

13. $\mathcal{F}^{-1}\big[\,\frac{\sin b\omega}{\omega}\,\big]$, where b is a nonzero constant
14. $\mathcal{F}\big[\,e^{-x^2}\,\big]$.
15. Recall that the finite pulse

$$p(t) \triangleq \begin{Bmatrix} 1, & -1 < t < 1 \\ 0, & \text{all other } t \end{Bmatrix}$$

has $\mathcal{F}[\,p(t)\,] = \sqrt{\frac{2}{\pi}} \cdot \frac{\sin\omega}{\omega}$. Use Theorem 9.11 and a geometric series to help find the Fourier transform of

$$g(t) \triangleq \sum_{n=1}^{\infty} e^{-n} p(t-2n).$$

16. Use the results of Example 9.20 and Theorem 9.9 to explain why Theorem 9.10 is correct.
17. Use Table 9.2's entry F.2 to explain why Table 9.2's entries (a) F.13 and (b) F.14 are true.
18. Explain why the Fourier transform of the function $f(t) \triangleq \begin{Bmatrix} 1, & -b < t < 0 \\ -1, & 0 < t < b \\ 0, & |t| > b \end{Bmatrix}$ is

$$\mathcal{F}[f(t)\,](\omega) = \sqrt{\frac{2}{\pi}} \cdot \begin{Bmatrix} \dfrac{\cos(\omega b) - 1}{i\omega b}, & \omega \neq 0 \\ 0, & \omega = 0 \end{Bmatrix}.$$

19. Use the result of Problem 9.4.3.18 to find the Fourier transform of $g(t) \triangleq \dfrac{\cos(bt) - 1}{bt}$.

9.5 Discrete Fourier and Fast Fourier Transforms

Suppose $f(t)$ is periodic with period N, where N is a positive, even integer. Define the **finite Fourier transform** of $f(t)$ to be

$$\mathcal{F}^{(N)}[f(t)](\omega) \triangleq \frac{1}{\sqrt{N}} \int_0^N f(t) e^{-i2\pi t\omega/N}\, dt. \tag{9.49}$$

Let $L = \frac{N}{2}$. We recall that a complex Fourier series on interval $\left[-\frac{N}{2}, \frac{N}{2}\right]$, or on $[0, N]$, is given by

$$f(t) \doteq \sum_{n=-\infty}^{\infty} c_n e^{i2\pi t n/N},$$

where the coefficients are

$$c_n = \frac{1}{N} \int_0^N f(t) e^{-i2\pi t n/N}\, dt = \frac{1}{\sqrt{N}} \mathcal{F}^{(N)}[f(t)](n),\ n = 0, \pm 1, \pm 2, \ldots .$$

So, the finite Fourier transform is related to a complex Fourier series, with inversion theorem resulting in

$$f(t) \doteq \frac{1}{\sqrt{N}} \sum_{n=-\infty}^{\infty} \mathcal{F}^{(N)}[f(t)](n) e^{i2\pi t n/N}.$$

If we replace the integral in (9.49) by a Riemann sum later we will get a "discrete Fourier transform" in (9.50) below. For any positive integer N, partition the interval $0 \le t \le N$ into N equal subintervals of length $\triangle t = 1$:

$$0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = N.$$

Given a function $f(t)$ defined on the interval $[0, N]$, denote

$$f_k = f(t_k),\ k = 0, 1, \ldots, N-1,$$

and form the vector

$$\mathbf{f} \triangleq \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_{N-1} \end{bmatrix},$$

which lies in $\mathbb{C}^N$, or in $\mathbb{R}^N$ if $f(t)$ is real-valued.

We may also consider vectors to be sequences, that is, we may write

$$\mathbf{f} = \{f_k\}_{k=0}^{N-1}.$$

Another name for a finite or infinite sequence of function values is a **time series**.

We define $\mathbf{F} = DFT[\,\mathbf{f}\,]$ to be the vector $\mathbf{F} \triangleq \begin{bmatrix} F_0 \\ F_1 \\ \vdots \\ F_{N-1} \end{bmatrix}$, whose components are given by

$$F_\ell \triangleq \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} f_k\, e^{-i2\pi k\ell/N}, \text{ for } \ell = 0, 1, \ldots, N-1, \tag{9.50}$$

which is called the **analysis equation**. This defines the **Discrete Fourier Transform** (*DFT*).

Note that F_ℓ is a Riemann sum approximation for $\mathcal{F}^{(N)}[f(t)\,](\ell)$, with sampling at left end points.

Note that

$$e^{-i2\pi k\ell/N} = \left(e^{-i2\pi/N}\right)^{k\ell} = \omega^{-k\ell},$$

where $\omega \triangleq e^{i2\pi/N}$ is an "N-th root of unity," that is, satisfies $\omega^N = 1$. So,

$$\mathbf{F} = DFT[\,\mathbf{f}\,] = \frac{1}{\sqrt{N}} \begin{bmatrix} f_0 & +f_1+ & \cdots & +f_{N-1} \\ f_0 & +\omega^{-1}f_1+ & \cdots & +\omega^{-(N-1)}f_{N-1} \\ f_0 & +\omega^{-2}f_1+ & \cdots & +\omega^{-2(N-1)}f_{N-1} \\ & & \cdot & \\ & & \cdot & \\ & & \cdot & \\ f_0 & +\omega^{-(N-1)}f_1+ & \cdots & +\omega^{-(N-1)(N-1)}f_{N-1} \end{bmatrix} \mathbf{f}$$

$$= \frac{1}{\sqrt{N}} \begin{bmatrix} 1 & 1 & . & . & . & 1 \\ 1 & \omega^{-1} & . & . & . & \omega^{-(N-1)} \\ 1 & \omega^{-2} & & & & \omega^{-2(N-1)} \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ 1 & \omega^{-(N-1)} & . & . & . & \omega^{-(N-1)(N-1)} \end{bmatrix} \mathbf{f}.$$

Implicitly this defines a symmetric matrix

$$F = [F_{\ell k}]_{\substack{0 \le \ell \le N-1 \\ 0 \le k \le N-1}} \triangleq \frac{1}{\sqrt{N}} [\omega^{k\ell}]_{\substack{0 \le \ell \le N-1 \\ 0 \le k \le N-1}} =$$

$$= \frac{1}{\sqrt{N}} \begin{bmatrix} 1 & 1 & 1 & . & . & . & 1 \\ 1 & \omega^1 & \omega^2 & & & & \omega^{N-1} \\ 1 & \omega^2 & \omega^4 & & & & \omega^{2(N-1)} \\ . & & & . & & & . \\ . & & & & . & & . \\ . & & & & & . & . \\ 1 & \omega^{N-1} & \omega^{2(N-1)} & . & . & . & \omega^{(N-1)(N-1)} \end{bmatrix}.$$

Because $\omega^{-1} = e^{-i2\pi/N}$ is the complex conjugate of ω, the discrete Fourier transform is given by

$$\mathbf{F} = DFT[\,\mathbf{f}\,] = F^*\mathbf{f},$$

where $F^* \triangleq \overline{(F^T)}$, that is, the complex conjugate of the transpose of F is called the **Hermitian conjugate** of the matrix F. (In Section 2.10 we called this the adjoint operator.)

The discrete Fourier transform defines a map $\mathbf{f} \mapsto \mathbf{F}$ on $\mathbb{C}^n$. In the subject of discrete systems theory, the map (9.50), that is, $\mathbf{f} \mapsto F^*\mathbf{f}$, is an example of a **digital filter**.

An alternative way of writing F takes advantage of the fact that $\omega^N = 1$, hence $\omega^{2(N-1)} = \omega^{N-2}$, $\omega^{3(N-1)} = \omega^{N-3}$, etc., so

$$F = \frac{1}{\sqrt{N}} \begin{bmatrix} 1 & 1 & 1 & . & . & . & 1 \\ 1 & \omega^1 & \omega^2 & & & & \omega^{N-1} \\ 1 & \omega^2 & \omega^4 & & & & \omega^{N-2} \\ . & & & . & & & . \\ . & & & & . & & . \\ . & & & & & . & . \\ 1 & \omega^{N-1} & \omega^{N-2} & . & . & . & \omega \end{bmatrix}. \tag{9.51}$$

F has many interesting properties:

Theorem 9.12

$$F^T = F, \quad \left(F^*\right)^T = F^*, \quad \text{and} \quad F = \overline{F^*}.$$

Example 9.23

Explain why F is a **unitary** matrix, that is,

$$FF^* = F^*F = I. \tag{9.52}$$

The latter is the generalization to complex matrices of the definition of "orthogonal matrix."

Method: We want to explain why the dot product of the ℓ-th row of F with the k-column of F^* is 1 if $k = \ell$ and 0 if $k \neq \ell$. That dot product is

$$\frac{1}{\sqrt{N}}\begin{bmatrix}1\\ \omega^{\ell}\\ \omega^{2\ell}\\ \vdots\\ \omega^{\ell(N-1)}\end{bmatrix} \bullet \frac{1}{\sqrt{N}}\begin{bmatrix}1\\ \omega^{-k}\\ \omega^{-2k}\\ \vdots\\ \omega^{-k(N-1)}\end{bmatrix} = \frac{1}{N}\sum_{n=0}^{N-1}\omega^{\ell n}\,\omega^{-kn} = \frac{1}{N}\sum_{n=0}^{N-1}\omega^{(\ell-k)n} = \frac{1}{N}\sum_{n=0}^{N-1}\left(\omega^{\ell-k}\right)^{n}. \tag{9.53}$$

If $\ell = k$, then (9.53) reduces to $\frac{1}{N}\cdot N = 1$. If $\ell \neq k$, then $\omega^{(\ell-k)} \neq 1$ so (9.53) is a geometric series whose sum is

$$\frac{1}{N}\cdot\frac{1-\left(\omega^{\ell-k}\right)^{N}}{1-\omega^{\ell-k}} = \frac{1}{N}\cdot\frac{1-1}{1-\omega^{\ell-k}} = 0,$$

because $\omega^{\ell-k}$ is also an N-th root of unity. ◯

From (9.52), that is, the conclusion of Example 9.23, we conclude that

$$F^{-1} = F^{*}.$$

This in turn implies that

$$F\,\mathbf{F} = \mathbf{f}, \tag{9.54}$$

which expresses the inverse discrete transform $\mathbf{F} \mapsto \mathbf{f}$ in terms of the matrix F.

While (9.50) is called the **analysis equation**, (9.54), that is,

$$f_k = \frac{1}{\sqrt{N}}\sum_{\ell=0}^{N-1} F_\ell\, e^{i2\pi k\ell/N}, \text{ for } k = 0, 1, \ldots, N-1, \tag{9.55}$$

is called the **synthesis equation**.

Example 9.24

When $N = 2$,

$$F = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & e^{i\pi}\end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}$$

is real and an orthogonal matrix. A related matrix is

$$H_2 \triangleq \sqrt{2}\,F = \begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix},$$

which satisfies $H_2 H_2^T = H_2^T H_2 = 2I = N \cdot I$ and has all of its entries being ± 1. H_2 is called the **Hadamard matrix** of order two. The question of which higher order Hadamard matrices exist is an open subject and has great implications for coding theory.

Example 9.25

For general N find the discrete Fourier transform of $\cos\left(\frac{4\pi t}{N}\right)$.

Method: The problem as stated has some ambiguity because the function $f(t) \triangleq \cos\left(\frac{4\pi t}{N}\right)$ is periodic with period being any positive integer multiple of $N/2$, including $2N/2$, $3N/2$, etc. The simplest choice is to make the N in the definition of $f(t)$'s denominator be the same N as in the definition of the discrete Fourier transform.

It helps to use Euler's formula to rewrite

$$f(t) = \frac{1}{2}\left(e^{i4\pi t/N} + e^{-i4\pi t/N}\right)$$

and thus

$$f_k = f(k) = \frac{1}{2}\left(e^{i4\pi k/N} + e^{-i4\pi k/N}\right) = \frac{1}{2}\left(\omega^{2k} + \omega^{-2k}\right),$$

where $\omega \triangleq e^{i2\pi/N}$.

Because $1 = \omega^N = \omega^m\,\omega^{N-m}$, for any positive integer m, we have a "circularity" or "aliasing" property:

$$\omega^{-m} = \omega^{N-m}. \tag{9.56}$$

It follows that

$$f_k = \frac{1}{2}\left(\omega^{2k} + \omega^{N-2k}\right).$$

So,

$$\mathbf{f} = \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_{N-1} \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 \\ \omega^2 \\ \vdots \\ \omega^{2(N-1)} \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 1 \\ \omega^{N-2} \\ \vdots \\ \omega^{(N-2)(N-1)} \end{bmatrix} = \frac{\sqrt{N}}{2}\left(\mathbf{F}^{(2)} + \mathbf{F}^{(N-2)}\right), \tag{9.57}$$

where we denote the columns of the matrix F by $F = \begin{bmatrix}\mathbf{F}^{(0)} & \mathbf{F}^{(1)} & \cdots & \mathbf{F}^{(N-1)}\end{bmatrix}$

Denote the columns of the $N \times N$ identity matrix by $I = \begin{bmatrix}\mathbf{d}^{(0)} & \mathbf{d}^{(1)} & \cdots & \mathbf{d}^{(N-1)}\end{bmatrix}$. The usual notations for the columns of the identity matrix are $\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \ldots, \mathbf{e}^{(N)}$, so $\mathbf{d}^{(k)} = \mathbf{e}^{(k+1)}$.

By orthonormality of the columns of F, the discrete Fourier transform of $f(t)$ is

$$\mathbf{F} = DFT[\,\mathbf{f}\,] = F^*\left(\frac{\sqrt{N}}{2}\,\mathbf{F}^{(2)} + \frac{\sqrt{N}}{2}\,\mathbf{F}^{(N-2)}\right) = \frac{\sqrt{N}}{2}\left(\mathbf{d}^{(2)} + \mathbf{d}^{(N-2)}\right) = \frac{\sqrt{N}}{2}\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \end{bmatrix},$$

that is, $F_2 = F_{N-2} = \frac{\sqrt{N}}{2}$ and $F_\ell = 0$ for all other indices ℓ. ◯

The fact that a sinusoidal of frequency 2, in units of $\frac{2\pi}{N}$, has *DFT* with nonzero second and $(N-2)$th components is an example of **aliasing**.

Other commonly used "wrap around" or circularity notations define $F_{-\ell} \triangleq F_{N-\ell}$, $\mathbf{F}^{(-\ell)} \triangleq \mathbf{F}^{(N-\ell)}$, and $\mathbf{d}^{(-\ell)} \triangleq \mathbf{d}^{(N-\ell)}$, for $\ell = 0, 1, \ldots, N$. Using the latter notation, we see that

$$\mathbf{F} = F^* \left(\frac{\sqrt{N}}{2} \mathbf{F}^{(2)} + \frac{\sqrt{N}}{2} \mathbf{F}^{(-2)} \right) = \frac{\sqrt{N}}{2} \left(\mathbf{d}^{(2)} + \mathbf{d}^{(-2)} \right)$$

when $\mathbf{f}$ is the discrete vector of values of $f(t) = \frac{1}{2} \left(e^{2(i2\pi t/N)} + e^{-2(i2\pi t/N)} \right)$.

As a last comment about Example 9.25, finding the *DFT* of a sinusoidal whose frequency is a multiple of the basic frequency is very similar to the problem of finding the Fourier series of a sinusoidal whose frequency is a multiple of the basic frequency. In that special case we were able to find the Fourier coefficients without calculating integrals. In Example 9.25 we were able to find the *DFT* components without calculating the summations in the definition in the analysis equation, (9.50).

We caution that there are many slightly different definitions of the *DFT* in current use. For example, instead of defining it as in the analysis equation (9.50), some authors leave out the factor of $\frac{1}{\sqrt{N}}$ in $F_\ell \triangleq \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \ldots$ but then have a factor of $\frac{1}{N}$ instead of $\frac{1}{\sqrt{N}}$ in the synthesis equation, $f_k = \cdots$. Yet other authors consider sequences as being $\{ \quad \}_{k=1}^{N}$ rather than $\{ \quad \}_{k=0}^{N-1}$, so they define the *DFT* by $F_\ell \triangleq \frac{1}{\sqrt{N}} \sum_{k=1}^{N} f_k \, e^{-i2\pi(k-1)(\ell-1)/N}$, for $\ell = 1, \ldots, N$; again, they may also leave out the factor of $\frac{1}{\sqrt{N}}$.

Of course, electrical engineers may replace $i \triangleq \sqrt{-1}$ by $j \triangleq \sqrt{-1}$. This is one reason why we do not use j as an index in our definitions in this section.

Example 9.26

Let $f(t)$ be the square wave of period N given by

$$f(t) = \begin{Bmatrix} 1, & 0 \le t < \frac{N}{2} \\ \\ 0, & \frac{N}{2} \le t < N \end{Bmatrix}. \tag{9.58}$$

If N is an even positive integer, find the analysis and synthesis of $f(t)$.

Method: Unfortunately, there is no advantage in using the orthonormality of the columns of the $N \times N$ matrix F as we did in Example 9.25. From (9.58),

$$f_k = f(k) = \begin{Bmatrix} 1, & k = 0, 1, \ldots, \frac{N}{2} - 1 \\ \\ 0, & k = \frac{N}{2}, \ldots, N \end{Bmatrix},$$

so the analysis equation gives

$$F_\ell = \frac{1}{\sqrt{N}} \sum_{k=0}^{\frac{N}{2}-1} 1 \cdot e^{-i2\pi k\ell/N}, \text{ for } \ell = 0, 1, \ldots, N-1.$$

For $\ell = 0$, $e^{-i2\pi\ell/N} = 1$, so $e^{-i2\pi k\ell/N} = \left(e^{-i2\pi\ell/N} \right)^k = 1$ for $k = 0, \ldots, \frac{N}{2} - 1$. We calculate

$$F_0 = \frac{1}{\sqrt{N}} \sum_{k=0}^{\frac{N}{2}-1} 1 \cdot 1 = \frac{1}{\sqrt{N}} \cdot \frac{N}{2} = \frac{\sqrt{N}}{2}.$$

For $\ell = 1, \ldots, N-1$, $e^{-i2\pi\ell/N} \neq 1$, but we can still use the fact that $e^{-i2\pi k\ell/N} = \left(e^{-i2\pi\ell/N}\right)^k$ for $k = 0, \ldots, \frac{N}{2} - 1$. We calculate

$$F_\ell = \frac{1}{\sqrt{N}} \sum_{k=0}^{\frac{N}{2}-1} 1 \cdot \left(e^{-i2\pi\ell/N}\right)^k = \frac{1}{\sqrt{N}} \cdot \frac{1 - \left(e^{-i2\pi\ell/N}\right)^{N/2}}{1 - e^{-i2\pi\ell/N}} = \frac{1}{\sqrt{N}} \cdot \frac{1 - e^{-i\pi\ell}}{1 - e^{-i2\pi\ell/N}}$$

$$= \frac{1}{\sqrt{N}} \cdot \frac{1 - (-1)^\ell}{1 - e^{-i2\pi\ell/N}} = \frac{2}{\sqrt{N}(1 - e^{-i2\pi\ell/N})} \cdot \begin{Bmatrix} 1, & \ell = 1, 3, \ldots, N-1 \\ 0, & \ell = 2, 4, \ldots, N-2 \end{Bmatrix} \cdot \bigcirc$$

Example 9.27

A square wave voltage $v(t)$ is periodic with period 8 ms and is given by

$$v(t) = \begin{Bmatrix} 1, & 0 \leq t < 0.004 \\ \\ 0, & 0.004 \leq t < 0.008 \end{Bmatrix} \textit{volts}. \tag{9.59}$$

Find $\mathbf{V}$=$DFT[\mathbf{v}] = \begin{bmatrix} V_0 & V_1 \cdots V_7 \end{bmatrix}^T$ and its **power spectrum** $\begin{bmatrix} |V_0|^2 & |V_1|^2 \cdots |V_7|^2 \end{bmatrix}^T$. Give graphs of (a) $\mathcal{R}e(DFT[\,\mathbf{v}\,]) = \{\mathcal{R}e(V_\ell)\}_{\ell=0}^7$, (b) $\mathcal{I}m(DFT[\,\mathbf{v}\,]) = \{\mathcal{I}m(V_\ell)\}_{\ell=0}^7$, and (c) the power spectrum.

Method: Measure time, t, in milliseconds, so $v(t)$ is periodic with period $N = 8$. Example 9.26 with $N = 8$ gives

$$\mathbf{V} = DFT[\,\mathbf{v}\,]$$

$$= \left[\sqrt{2} \quad \frac{1}{\sqrt{2}(1-\omega^{-1})} \quad 0 \quad \frac{1}{\sqrt{2}(1-\omega^{-3})} \quad 0 \frac{1}{\sqrt{2}(1-\omega^{-5})} \quad 0 \quad \frac{1}{\sqrt{2}(1-\omega^{-7})}\right]^T,$$

where $\omega^{-1} = e^{-i2\pi/8} = e^{-i\pi/4} = \frac{1}{\sqrt{2}}(1 - i)$. To be more explicit,

$$DFT[\mathbf{v}] \approx$$

$$[1.414 \quad .3536 - i0.8536 \quad 0 \quad .3536 - i0.1464 \quad 0 \quad .3536 + i0.1464 \quad 0 \quad .3536 + i0.8536]^T$$

and its real and imaginary parts are graphed in Figure 9.26a and b.

The power spectrum is the vector

$$\left[2 \quad \frac{1}{2(2-\sqrt{2})} \quad 0 \quad \frac{1}{2(2+\sqrt{2})} \quad 0 \frac{1}{2(2+\sqrt{2})} \quad 0 \quad \frac{1}{2(2-\sqrt{2})}\right]^T$$

$$\approx [2 \quad 0.8536 \quad 0 \quad 0.1464 \quad 0 \quad 0.1464 \quad 0 \quad 0.8536]^T$$

and is graphed in Figure 9.27. $\bigcirc$

Mathematica and **MATLAB®**

Mathematica's command

$$\mathsf{W = Fourier[v,\ FourierParameters \rightarrow \{0, -1\}]}$$

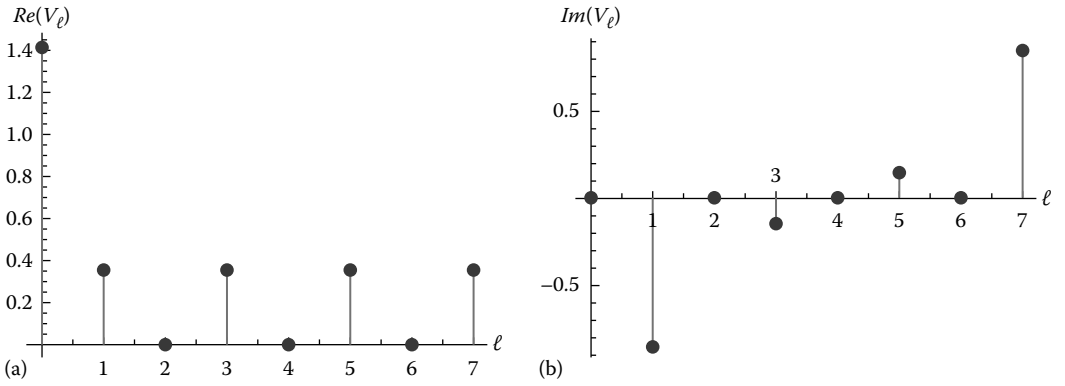


FIGURE 9.26
Example 9.27 (a) and (b).

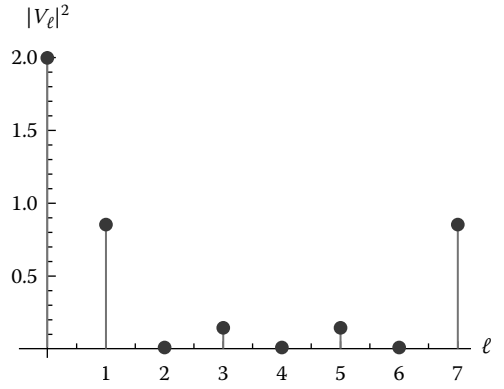


FIGURE 9.27
Example 9.27 (c) Power spectrum.

implements our definition of the *DFT* except that its *DFT* vector is indexed by $1, 2, \ldots, 8$, that is, is

$$\mathrm{W} = [\,\text{first number second number } \ldots \text{ eighth number}\,]$$

instead of our vector

$$\mathbf{V} = DFT[\,\mathbf{v}\,] = [\,V_0 \ldots V_7\,].$$

Basically, we used the *Mathematica* command

$$\text{ListPlot[Abs[V}^2\text{], PlotStyle} \rightarrow \text{\{Blue, PointSize[Large]\}, Filling} \rightarrow \text{Axis, FillingStyle} \rightarrow \text{Red,}$$

$$\text{AxesLabel} \rightarrow \text{\{Text[Style["}\ell\text{", Italic, 18]],Text[Style["|V}_\ell|^2\text{", Italic, 18]]\}]}$$

to get the plot in Figure 9.26, although we did make some small drawing changes for clarity.

Example 9.28

In MATLAB, with $N = 4$, input the vector

$$a = [1, -0.6 - 0.8i, -0.28 + 0.96i, 0.936 - 0.352i]$$

The MATLAB command fft(a) produced

$$[\,1.0560 - 0.1920i, \quad 0.8320 + 0.5760i, \quad 0.3840 + 2.1120i, \quad 1.7280 - 2.4960i\,],$$

which is $\sqrt{N} = 2$ times the *DFT* as we have defined it. MATLAB leaves out the factor $\frac{1}{\sqrt{N}}$ that is in our definition of *DFT*. ◯

9.5.1 Convolution and Auto-Correlation

Given two sequences $\mathbf{x} = \{x_k\}_{k=0}^{N-1}$ and $\mathbf{y} = \{y_k\}_{k=0}^{N-1}$ we can form their **discrete convolution**, defined by

$$\mathbf{x} * \mathbf{y} = \{(\mathbf{x} * \mathbf{y})_k\}_{k=0}^{N-1} \triangleq \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} x_m y_{k-m}. \tag{9.60}$$

We use the wrap-around convention that $y_{-n} \triangleq y_{N-n}$ if $n = 0, \ldots, N$.

Theorem 9.13

(a) $\mathbf{x} * \mathbf{y} = \mathbf{y} * \mathbf{x}$ (b) If $\mathbf{X} = DFT[\,\mathbf{x}\,]$ and $\mathbf{Y} = DFT[\,\mathbf{y}\,]$ then the *DFT* of $\mathbf{x} * \mathbf{y}$ is given by

$$DFT[\,\mathbf{x} * \mathbf{y}\,] = \{X_\ell Y_\ell\}_{\ell=0}^{N-1}.$$

Why? You will explain (a) in Problem 9.5.3.8. For (b), let $\mathbf{z} \triangleq \mathbf{x} * \mathbf{y}$ and, as usual, $\omega \triangleq e^{i2\pi/N}$. By (9.50) we calculate that for $\ell = 0, 1, \ldots, N-1$, the ℓth component of $DFT[\,\mathbf{z}\,]$ is

$$Z_\ell \triangleq \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} z_k\, \omega^{-k\ell} = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \left(\frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} x_m y_{k-m} \right) \omega^{-k\ell}.$$

Exchanging the order of finite summations and writing $\omega^{-k\ell} = \omega^{-m\ell}\, \omega^{-(k-m)\ell}$ gives that

$$Z_\ell = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} x_m \omega^{-m\ell} \cdot \left(\frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} y_{k-m} \omega^{-(k-m)\ell} \right) = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} x_m \omega^{-m\ell} \cdot Y_\ell$$

$$= Y_\ell \cdot \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} x_m \omega^{-m\ell} = Y_\ell X_\ell. \; \square$$

Above, we used the change of index $K = k - m$ and wrap-around to see that

$$\frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} y_{k-m} \omega^{-(k-m)\ell} = \frac{1}{\sqrt{N}} \sum_{K=0}^{N-1} y_K \omega^{-K\ell} = Y_\ell.$$

The **auto-correlation** of a sequence $\mathbf{x} = \{x_k\}_{k=0}^{N-1}$ is defined to be the sequence $\mathbf{z} = \{z_k\}_{k=0}^{N-1}$ where

$$z_k \triangleq \sum_{m=0}^{N-1} x_m \overline{x}_{m-k}. \tag{9.61}$$

This is the discrete convolution of $\mathbf{x}$ with the complex conjugate of a reversed order version of $\mathbf{x}$.

Theorem 9.14

The *DFT* of the auto-correlation is the power spectrum.

Why? In terms of the sequence $\mathbf{x}$, the *DFT* of its auto-correlation, $\mathbf{z}$ is given by

$$Z_\ell = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} z_k \omega^{-k\ell} = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \left(\frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} x_m \overline{x}_{m-k}\, \omega^{-k\ell} \right)$$

$$= \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} x_m\, \omega^{-m\ell} \cdot \left(\frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \overline{x}_{m-k}\, \omega^{-(k-m)\ell} \right) = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} x_m\, \omega^{-m\ell} \cdot \left(\frac{1}{\sqrt{N}} \sum_{K=0}^{N-1} \overline{x}_K\, \omega^{K\ell} \right),$$

where we defined $K \triangleq m - k$. But

$$\overline{X_\ell} = \overline{\frac{1}{\sqrt{N}} \sum_{K=0}^{N-1} x_K\, \omega^{-K\ell}} = \frac{1}{\sqrt{N}} \sum_{K=0}^{N-1} \overline{x}_K\, \omega^{K\ell},$$

so the *DFT* of its auto-correlation, $\mathbf{z}$, is given by

$$Z_\ell = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} x_m\, \omega^{-m\ell} \cdot \left(\overline{X_\ell} \right) = \overline{X_\ell} \cdot \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} x_m\, \omega^{-m\ell} = \overline{X_\ell} \cdot X_\ell = |X_\ell|^2. \ \square$$

9.5.2 Fast Fourier Transform

Mathematicians found tricks for calculating the discrete Fourier transform. There are many clever algorithms that improve the speed and memory storage requirements of the *DFT*.

Collectively these algorithms are called the **Fast Fourier transform (*FFT*)**. The improvements were so dramatic as to make feasible the solution of problems that otherwise would have required so much computer processing time as to be effectively insoluble.

Nevertheless, it should be emphasized that the *FFT* produces the same end result as the *DFT* would theoretically produce if it used sufficiently high precision arithmetic and were allowed to run for a sufficiently long time. Of course, *DFT* would be useless if it takes years to produce results that our technology needs to know in milliseconds.

Let N be a positive, *even* integer and denote $\omega = e^{i2\pi/N}$. We choose N to be even because the *fast* part of the phrase *FFT* can come from breaking down calculations into even versus odd parts.

Let a sequence $\mathbf{x} = \{x_k\}_{k=0}^{N-1}$ have *DFT*

$$\mathbf{X} = DFT_N[\mathbf{x}] = \{X_\ell\}_{\ell=0}^{N-1}, \text{ where } X_\ell = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} x_k \, \omega^{-k\ell},$$

which is a *DFT* of size N. Define the even part of $\mathbf{x}$ to be

$$\mathbf{y} = \{y_m\}_{m=0}^{\frac{N}{2}-1}, \text{ where } y_m = x_{2m},$$

and define the odd part of $\mathbf{x}$ to be

$$\mathbf{z} = \{z_m\}_{m=0}^{\frac{N}{2}-1}, \text{ where } z_m = x_{2m+1}.$$

So, $\mathbf{x}$ consists of the interleaving of $\mathbf{y}$ and $\mathbf{z}$:

$$\mathbf{x} = \begin{bmatrix} y_0 & z_0 & y_1 & z_1 & \cdots & y_{\frac{N}{2}-1} & z_{\frac{N}{2}-1} \end{bmatrix}^T.$$

Correspondingly, each of the even and odd parts has a *DFT* of size $\frac{N}{2}$: Because

$$e^{-i2\pi/(N/2)} = \omega^{-2}.$$

we see that $\mathbf{Y} = DFT_{N/2}[\,\mathbf{y}\,]$ is given by

$$Y_n \triangleq \frac{1}{\sqrt{N/2}} \sum_{m=0}^{\frac{N}{2}-1} y_m \, \omega^{-2mn}$$

and $\mathbf{Z} = DFT_{N/2}[\,\mathbf{z}\,]$ is given by

$$Z_n \triangleq \frac{1}{\sqrt{N/2}} \sum_{m=0}^{\frac{N}{2}-1} z_m \, \omega^{-2mn}.$$

The $DFT_N[\,\mathbf{x}\,] = \{X_\ell\}_{\ell=0}^{N-1}$ of the original sequence can be constructed from the $DFT_{N/2}$ of the even and odd parts: For $\ell = 1, \ldots, \frac{N}{2} - 1$,

$$X_\ell = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} x_k\, \omega^{-k\ell} = \frac{1}{\sqrt{2}} \left(\frac{1}{\sqrt{N/2}} \sum_{m=0}^{\frac{N}{2}-1} x_{2m}\, \omega^{-2m\ell} + \frac{1}{\sqrt{N/2}} \sum_{m=0}^{\frac{N}{2}-1} x_{2m+1}\, \omega^{-(2m+1)\ell} \right)$$

$$= \frac{1}{\sqrt{2}} \left(\frac{1}{\sqrt{N/2}} \sum_{m=0}^{\frac{N}{2}-1} y_m\, \omega^{-2m\ell} + \omega^{-\ell} \cdot \frac{1}{\sqrt{N/2}} \sum_{m=0}^{\frac{N}{2}-1} z_m\, \omega^{-2m\ell} \right) = \frac{1}{\sqrt{2}} \left(Y_\ell + \omega^{-\ell} \cdot Z_\ell \right).$$

For $\ell = \frac{N}{2}, \ldots, N-1$,

$$X_\ell = \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{N/2}} \left(\sum_{m=0}^{\frac{N}{2}-1} x_{2m}\, \omega^{-2m\left(\ell - \frac{N}{2}\right)} \cdot \omega^{-2m \cdot \frac{N}{2}} + \omega^{-\ell} \cdot \sum_{m=0}^{\frac{N}{2}-1} x_{2m+1}\, \omega^{-2m\left(\ell - \frac{N}{2}\right)} \cdot \omega^{-2m \cdot \frac{N}{2}} \right).$$

But, $\omega^{-2m \cdot \frac{N}{2}} = \omega^{-mN} = \left(\omega^N\right)^{-m} = 1$ for all integers m, so for $\ell = \frac{N}{2}, \ldots, N-1$,

$$X_\ell = \frac{1}{\sqrt{2}} \cdot \left(\frac{1}{\sqrt{N/2}} \sum_{m=0}^{\frac{N}{2}-1} x_{2m}\, \omega^{-2m\left(\ell - \frac{N}{2}\right)} + \omega^{-\ell} \cdot \frac{1}{\sqrt{N/2}} \sum_{m=0}^{\frac{N}{2}-1} x_{2m+1}\, \omega^{-2m\left(\ell - \frac{N}{2}\right)} \right).$$

Using the fact that $\omega^{\frac{N}{2}} = e^{i\pi} = -1$, so $\omega^{-\ell} = -\omega^{-\ell - \frac{N}{2}}$, we have that for $\ell = \frac{N}{2}, \ldots, N-1$,

$$X_\ell = \frac{1}{\sqrt{2}} \left(Y_{\ell - \frac{N}{2}} - \omega^{-\left(\ell - \frac{N}{2}\right)} \cdot Z_{\ell - \frac{N}{2}} \right).$$

Further, if $\frac{N}{2}$ is even, then we can break down $\mathbf{y}$ and $\mathbf{z}$ into *their* even and odd parts, etc. This is an example of a "divide and conquer" strategy.

In summary,

$$Z_\ell = \frac{1}{\sqrt{2}} \begin{cases} Y_\ell + \omega^{-\ell} \cdot Z_\ell, & \ell = 0, \ldots, \frac{N}{2} - 1 \\ Y_{\ell - \frac{N}{2}} - \omega^{-\left(\ell - \frac{N}{2}\right)} \cdot Z_{\ell - \frac{N}{2}}, & \ell = \frac{N}{2}, \ldots, N-1 \end{cases}. \tag{9.62}$$

So, why is this *fast*? Most of the computer processing time done by the *DFT* or *FFT* is in multiplications, and it turns out that when N is "large," the *FFT* defined above requires significantly fewer multiplications than does *DFT*.

First, we note that the *DFT* requires N^2 multiplications.* For *each* of $\ell = 0, 1, \ldots, N-1$, we need N multiplications to calculate Z_ℓ by the *DFT* defined in (9.50), so the *DFT* requires N^2 multiplications.

On the other hand, when N is a power of two we will use an *inductive argument*† to see why calculating the *FFT* requires only $2N\,logN$ multiplications and is thus much faster than the *DFT* when N is "large."

Assume that it takes only $2 \cdot \frac{N}{2} \cdot log\frac{N}{2}$ multiplications to calculate each of $\mathbf{Y} = DFT_{N/2}[\,\mathbf{y}\,]$ and $\mathbf{Z} = DFT_{N/2}[\,\mathbf{z}\,]$. Using (9.62), calculating $\mathbf{X} = DFT_{N/2}[\,\mathbf{x}\,]$ requires $2 \cdot \left(2 \cdot \frac{N}{2} \cdot log\frac{N}{2}\right)$ multiplications to get the Y_m's and the Z_m's and also $\frac{N}{2}$ multiplications to calculate the products $\omega^{-\ell} \cdot Z_\ell$, for $\ell = 0, \ldots, \frac{N}{2} - 1$. The latter are used to calculate X_ℓ for $\ell = 0, \ldots, \frac{N}{2} - 1$ but can be stored to calculate X_ℓ for $\ell = \frac{N}{2}, \ldots, N-1$.

So, the number of multiplications needed to calculate the *FFT* is

$$2 \cdot \left(2 \cdot \frac{N}{2}\, log\frac{N}{2}\right) + \frac{N}{2} = 2N \cdot (log\, N - log\, 2) + \frac{N}{2} = 2N log\, N + (0.5 - 2log2)N < 2N log\, N.$$

By inductive reasoning, when N is a power of two we have an upper bound on the number of multiplications needed, although we do need to verify the beginning step that when $N = 2$, the number of multiplications needed is two: when $\mathbf{x} = [x_0 \quad x_1]^T$, its *DFT* is $[X_0 \quad X_1]^T$ where $X_0 = \frac{1}{\sqrt{2}}(x_0 - x_1)$ and $X_1 = \frac{1}{\sqrt{2}}(x_0 + x_1)$. So, when $N = 2$ the number of multiplications needed is two, which is less than $2N \cdot log\, N = 4log\, 2$.

If N is not a power of two then some algorithms "pad" the vector $\mathbf{x}$ with zeros to create a vector whose length is a power of two. At first it might appear paradoxical that increasing the size of the problem of calculating the *DFT* makes it faster to do, but we saw how efficient it is to calculate the *FFT* when the size is a power of two.

One particularly useful application of the FFT is evaluating a discrete convolution $\mathbf{x} * \mathbf{y}$ by an indirect method: By Theorem 9.13(a),

$$\mathbf{x} * \mathbf{y} = DFT^{-1}\big[\,\{X_\ell Y_\ell\}_{\ell=0}^{N-1}\,\big].$$

The numbers of multiplications needed to calculate this are (a) $2Nlog\, N$ to get $\mathbf{X} = DFT\,[\,\mathbf{x}\,]$, (b) $2Nlog\, N$ to get $\mathbf{y} = DFT\,[\,\mathbf{y}\,]$, (c) N to get $\{X_\ell Y_\ell\}_{\ell=0}^{N-1}$, and (d) $2Nlog\, N$ to get $DFT^{-1}\big[\,\{X_\ell Y_\ell\}_{\ell=0}^{N-1}\,\big]$, for a total of $N + 6N \cdot log\, N$ rather than the N^2 to directly calculate the discrete convolution $\mathbf{x} * \mathbf{y}$.

---

* We neglect in the count the $(N-1)$ multiplications needed to create the multipliers $\omega^{-k\ell}$. First, as we saw in (9.51), we only need the powers $\omega^{-1}, \ldots, \omega^{-(N-1)}$, not $\omega^{-1}, \ldots, \omega^{-(N-1)(N-1)}$. Second of all, we can calculate $\omega^{-1}, \ldots, \omega^{-(N-1)}$ by storing the successive products $\omega^{-2} = \omega^{-1} \cdot \omega^{-1}$, $\omega^{-3} = \omega^{-1} \cdot \omega^{-2}$, …, $\omega^{-(N-1)} = \omega^{-1} \cdot \omega^{-(N-2)}$. Finally, we can calculate and store "offline" $\omega^{-1}, \ldots, \omega^{-(N-1)}$ just once at the beginning of solving a problem.

† We are following the argument of (Cochran et al. 1967).

Learn More About It

Perhaps the most cited mathematics paper is "An algorithm for the machine calculation of complex Fourier series," J. W. Cooley and J. W. Tukey, *Math. Comput*. **19** (1965), 297–301. In addition to the paper by Cochran, et al. cited earlier, two of the many good references to the details of *FFT* algorithms are *Computational Frameworks for the Fast Fourier Transform*, by Charles Van Loan, *S. I. A. M. Frontiers Appl. Mat*, Volume 10, © 1992, and *A First Course in Fourier Analysis*, by David W. Kammler, Cambridge University Press, © 2007. The former often gives "pseudo-code" for *FFT* algorithms. Many other interesting properties of the matrix F are found in Phillip J. Davis, *Circulant Matrices*, John Wiley & Sons, Inc., © 1979.

9.5.3 Problems

In problems 1–3 assume $f(t)$ is periodic with period N. Find its *DFT*.

1. $f(t) = \sin\left(\frac{4\pi t}{N}\right)$.
2. $f(t) = \cos\left(\frac{6\pi t}{N}\right)$.
3. $f(t) = a^t$, for $0 \le t \le N$ and is extended to be periodic with period N; assume the constant a satisfies $a = e^{-i2\pi k_0/N}$ and k_0 is a positive integer with $1 < k_0 < N$.

In problems 4–7 a graph or table of values of $f(t)$ is given for $0 \le t \le 4\,ms$. For both (a) $N = 4$ (b) $N = 8$, (i) write down the vector of function values $\mathbf{f} \triangleq [f_0 \quad \dots \quad f_{N-1}]^T$ and (ii) $\mathbf{F} = DFT_N[f(t)]$. You may use technology but document what technology you used by stating what command you used and in what software package.

4. $f(t)$ graphed in Figure 9.28.
5. $f(t)$ graphed in Figure 9.29.
6. $f(t)$ whose data is in Table 9.3.
7. $f(t)$ whose data is in Table 9.4.
8. Explain why Theorem 9.13(a) is true.
9. Use the *FFT* algorithm *by hand* for the sequence

$$a = [1, -0.6 - 0.8i, -0.28 + 0.96i, 0.936 - 0.352i]$$

of Example 9.28.

10. Abstraction of Examples 9.27 and 9.28 suggest that if $\mathbf{f} = \{f_k\}_{k=0}^{N-1}$ is a real sequence then

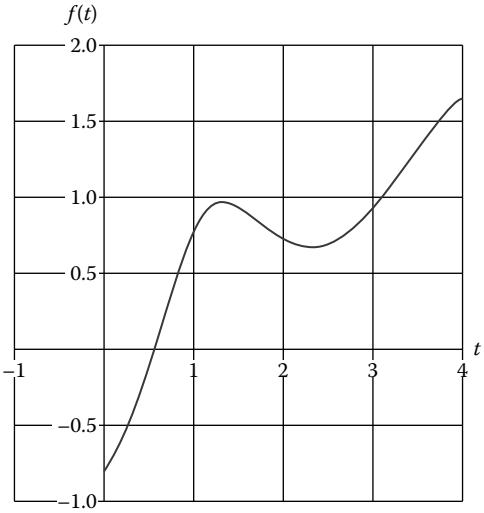


FIGURE 9.28
Problem 9.5.3.4.

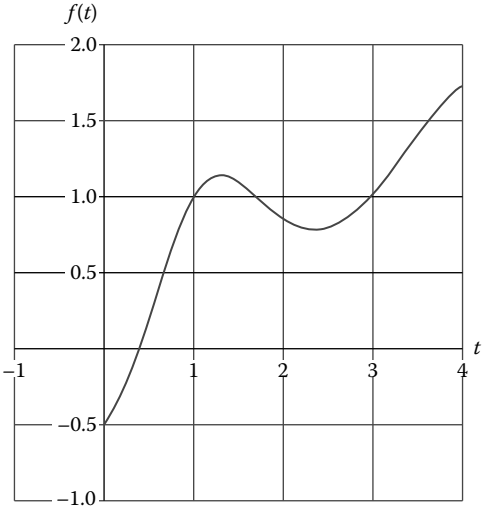


FIGURE 9.29
Problem 9.5.3.5.

TABLE 9.3

Problem 9.5.3.6

t	0	0.5	1.0	1.5	2.0	2.5	3.0	3.6	4.0
$f(t)$	8	7	7	6	6	6	7	8	9

TABLE 9.4

Problem 9.5.3.7

t	0	0.5	1.0	1.5	2.0	2.5	3.0	3.6	4.0
$f(t)$	2	3	2	2	0	−1	−2	−1	0

(a) Its $DFT_N[\mathbf{f}] = \mathbf{F} = \{F_\ell\}_{\ell=0}^{N-1}$ symmetric in the sense that $F_{N-\ell} = \overline{F_\ell}$, and

(b) Its power spectrum is symmetric, in the sense that $|F_{N-\ell}| = |\overline{F_\ell}| = |F_\ell|$, for $\ell = 0, \ldots, N$. [We used the wrap around convention that $F_{-\ell} \triangleq F_{N-\ell}$ if $\ell = 0, \ldots, N$.]

11. Establish the converse of Problem 9.5.3.10(a), that is, that if $F_{N-\ell} = \overline{F_\ell}$, for $\ell = 0, \ldots, N$, where $DFT_N[\mathbf{f}] = \mathbf{F} = \{F_\ell\}_{\ell=0}^{N-1}$, then $\mathbf{f}$ is a real sequence. [We used the "wrap around" convention that $F_{-\ell} \triangleq F_{N-\ell}$ if $\ell = 0, \ldots, N$.]

9.6 Sturm–Liouville Problems

The ODE in (9.20) in Section 9.3 can be generalized in many ways that are useful for engineering problems. We start with

Definition 9.5

(a) A **Sturm-Liouville problem** is given by an ODE in the form

$$\big(p(x)X'(x)\big)' + (\lambda s(x) + q(x))X(x) = 0,\ a < x < b, \tag{9.63}$$

along with the familiar BCs

$$\left\{\begin{array}{l} \epsilon_0 X(a) - \epsilon_1 X'(a) = 0, \\ \\ \gamma_0 X(b) + \gamma_1 X'(b) = 0 \end{array}\right\}, \tag{9.64}$$

where $\epsilon_0, \epsilon_1, \gamma_0, \gamma_1$ are given constants (scalars). We will assume always that either $\epsilon_0 \neq 0$ or $\epsilon_1 \neq 0$, and either $\gamma_0 \neq 0$ or $\gamma_1 \neq 0$.

(b) A Sturm-Liouville problem is **regular** if all of the functions p, p', q, s are defined, continuous, and real-valued on the interval $[a, b]$, $p(x) > 0$ on $[a, b]$, and $s(x) \geq 0$ and is not identically zero on $[a, b]$.

The results for a regular Sturm–Liouville problem are pretty much the same as for the ODE-BVP eigenvalue problem with a much simpler ODE. It is unusual in any subject to have such a successful generalization.

Theorem 9.15

For a regular Sturm-Liouville eigenvalue problem (9.63) and (9.64), there are infinitely many eigenvalues

$$\lambda_1 < \lambda_2 < \lambda_3 < \cdots \to \infty,$$

and for each eigenvalue λ_n the only eigenfunctions are the nonzero multiples of a single eigenfunction

$$X_n(x).$$

The eigenvalues are found by solving a characteristic equation and the eigenfunctions satisfy an orthogonality relation

$$0 = \int_a^b s(x)X_n(x)X_m(x)dx, \text{ for } n \neq m. \tag{9.65}$$

Further, for nice functions we have a generalized Fourier expansion

$$f(x) \doteq \sum_{n=1}^{\infty} A_n X_n(x), \ a < x < b,$$

where

$$A_n = \frac{\int_a^b s(x)X_n(x)f(x)dx}{\int_a^b s(x)|X_n(x)|^2 dx}. \tag{9.66}$$

Finally, at most finitely many eigenvalues can be negative, and, if $\epsilon_0, \epsilon_1 > 0$, $\gamma_0, \gamma_1 \geq 0$ and $q(x) \leq 0$ on the interval $[\,a, b\,]$, then all eigenvalues are positive.

Except for Example 9.29, we will not give most of the explanations for Theorem 9.15.

Example 9.29

Explain the orthogonality relation (9.65).

The orthogonality relation also helps explain the formula for the coefficients A_n in (9.66).

Method: Suppose λ_n and λ_m are two distinct eigenvalues, with corresponding eigenfunctions $X_n(x)$ and $X_m(x)$. From the ODE, with λ replaced by λ_n and $X(x)$ replaced by $X_n(x)$, we get

$$\big(p(x)X_n'(x)\big)' + (\lambda_n s(x) + q(x))X_n(x) = 0, \ a < x < b, \tag{9.67}$$

and similarly we get

$$\big(p(x)X_m'(x)\big)' + (\lambda_m s(x) + q(x))X_m(x) = 0,\ a < x < b.$$

Multiply the first ODE, that is, (9.67), by $X_m(x)$, and subtract from that $X_n(x)$ times the second ODE to get

$$0 = X_m(x)\left(p(x)X_n'(x)\right)' + \lambda_n X_m(x)s(x)X_n(x) - X_n(x)\left(p(x)X_m'(x)\right)' - \lambda_m X_n(x)s(x)X_m(x).$$

Integrate from a to b to get

$$\begin{aligned} 0 = \int_a^b 0\,dx &= \int_a^b \left(X_m(x)\left(p(x)X_n'(x)\right)' - X_n(x)\left(p(x)X_m'(x)\right)'\right)dx \\ &\quad + (\lambda_n - \lambda_m)\int_a^b s(x)X_n(x)X_m(x)\,dx. \end{aligned} \tag{9.68}$$

For the first term, integration by parts gives

$$\begin{aligned} \int_a^b X_m(x)\big(p(x)X_n'(x)\big)'\,dx &= \Big[X_m(x)\big(p(x)X_n'(x)\big)\Big]_a^b - \int_a^b X_m'(x)\big(p(x)X_n'(x)\big)dx \\ &= X_m(b)p(b)X_n'(b) - X_m(a)p(a)X_n'(a) - \int_a^b p(x)X_m'(x)X_n'(x)\,dx, \end{aligned}$$

and similarly

$$-\int_a^b X_n(x)\big(p(x)X_m'(x)\big)'\,dx = -X_n(b)p(b)X_m'(b) + X_n(a)p(a)X_m'(a) + \int_a^b p(x)X_n'(x)X_m'(x)\,dx.$$

So, in (9.68), after canceling two terms of $\int_a^b p(x)X_n'(x)X_m'(x)dx$ and combining the terms evaluated at $x = b$ and $x = a$, we have

$$\begin{aligned} 0 &= p(b)\big(X_m(b)X_n'(b) - X_n(b)X_m'(b)\big) - p(a)\big(X_m(a)X_n'(a) - X_n(a)X_m'(a)\big) \\ &\quad + (\lambda_n - \lambda_m)\int_a^b s(x)X_n(x)X_m(x)\,dx. \end{aligned} \tag{9.69}$$

We claim that each of the two terms involving evaluations at $x = b$ or $x = a$ cancel. Consider two cases: If in the boundary condition at $x = b$ we have $\gamma_1 = 0$, then to satisfy the BC there we get $0 = \gamma_0 X_n(b) + 0 \cdot X_n'(b) = \gamma_0 X_n(b)$, hence $X_n(b) = 0$ and likewise $X_m(b) = 0$, so $p(b)\big(X_m(b)X_n'(b) - X_n(b)X_m'(b)\big) = p(b)(0 \cdot X_n'(b) - X_n'(b) \cdot 0 = 0$. On the other hand, if $\gamma_1 \neq 0$, then the BC that X_n satisfies at $x = b$ yields $X_n'(b) = -\dfrac{\gamma_0}{\gamma_1}X_n(b)$, and similarly $X_m'(b) = -\dfrac{\gamma_0}{\gamma_1}X_m(b)$, so

$$X_m(b)X_n'(b) - X_n(b)X_m'(b) = \cancel{X_m(b)\left(-\frac{\gamma_0}{\gamma_1}X_n(b)\right)} - \cancel{X_n(b)\left(-\frac{\gamma_0}{\gamma_1}X_m(b)\right)} = 0.$$

Similarly,

$$X_m(a)X_n'(a) - X_n(a)X_m'(a) = 0.$$

So our claim is correct. So, (9.69) reduces to

$$0 = (\lambda_n - \lambda_m) \int_a^b s(x) X_n(x) X_m(x)\, dx.$$

Because $\lambda_n \neq \lambda_m$, we can divide through by $(\lambda_n - \lambda_m)$ to get the orthogonality relation (9.65). ◯

9.6.1 Other Sturm–Liouville Problems

There are other kinds of Sturm–Liouville problems. For example, if the continuous function $s(x)$ is not strictly positive on the interval $[a, b]$ then the Sturm–Liouville problem is not regular. Also, the problem is not regular if one of the functions p, p', s, q fails to be continuous on the interval $[a, b]$.

Suppose the Sturm–Liouville problem satisfies all of the requirements to be regular *except* that the nonnegative function $s(x)$ fails to be positive on $[a, b]$ and/or $q(x)$ fails to satisfy $|q(a^+)| < \infty$ or $|q(b^-)| < \infty$. We will refer to such a situation as a **singular Sturm Liouville problem**. We will see in Chapter 11 that such singular ODE-BVPs *naturally* occur when solving partial differential equation models in certain physical geometries, for example, problems in a disk or cylinder, and we will understand why we call such ODE-BVPs "singular."

In Sections 11.5 and 11.6 we will often use results concerning the eigenvalue problem

$$\left\{ \begin{array}{c} \frac{d}{dr}\left(r \frac{dR}{dr}\right) + (\lambda r - \frac{n^2}{r}) R(r) = 0,\ 0 < r < a, \\ \\ |\, R(0^+)\,| < \infty, \\ \\ \gamma_0 R(a) + \gamma_1 \frac{dR}{dr}(a) = 0 \end{array} \right\}, \tag{9.70}$$

where n is a nonnegative integer. The ODE here is Bessel's equation of order n. The singular Sturm-Liouville problem (9.70) has infinitely many eigenvalues λ_{nm} satisfying

$$0 < \lambda_{n0} < \lambda_{n1} < \lambda_{n2} < \cdots \to \infty.$$

The corresponding eigenfunctions are $R_m(r) = J_n(\lambda_{nm} r)$ and satisfy the orthogonality relation

$$0 = \int_0^a r R_m(r) R_{m'}(r)\, dr, \text{ for } m \neq m'. \tag{9.71}$$

Notice that even though this is a *singular* Sturm–Liouville problem, the orthogonality relation is the same as found in (9.65): the functions p, s, q here are $p(r) = r, s(r) = r$, $q(r) = -\frac{n^2}{r}$.

Notice also that we can rewrite the orthogonality relation as $0 = \int_0^a R_m(r) R_{m'}(r) r\, dr$, for $m \neq m'$, and we recall that $r dr$ is part of the element of area in polar coordinates, $dA = r dr d\theta$, that we saw in Sections 6.6 and 7.3. As we will see in Section 11.5, it is not an accident that the $r\, dr$ of polar coordinates shows up in the study of Bessel functions, because Bessel function originated in the solution of partial differential equations in polar coordinates.

Example 9.30

Show the positivity of all eigenvalues of

$$\left\{\begin{array}{c} r\big(rR'\big)' + \lambda r^2 R = 0,\ 0 < r < a \\ \\ |\, R(0^+)\,| < \infty,\ R(a) = 0 \end{array}\right\},$$

which is an ODE-BVP for Bessel's equation of order zero that we will use in Section 11.5.

Method: Multiply through the ODE by $\frac{1}{r}\ R(r)$ and integrate on the interval $0 < r < a$ to get

$$0 = \int_0^a R(r)\big(rR'(r)\big)'\,dr + \lambda \int_0^a r\big(R(r)\big)^2\,dr.$$

Integrate by parts to get

$$0 = \Big[R(r)\big(rR'(r)\big)\Big]_0^a - \int_0^a R'(r)\big(rR'(r)\big)\,dr + \lambda \int_0^a r\big(R(r)\big)^2\,dr,$$

hence

$$\lambda \int_0^a r\big(R(r)\big)^2\,dr = -aR(a)R'(a) + \lim_{r\to 0^+} rR(r)R'(r) + \int_0^a R'(r)\big(rR'(r)\big)\,dr,$$

and the BCs imply

$$\lambda \int_0^a r\big(R(r)\big)^2\,dr = \int_0^a r\big(R'(r)\big)^2\,dr.$$

Note that an eigenfunction $R(r)$ cannot have $R'(r) \equiv 0$ on $(0, a]$ because of the boundary condition $R(a) = 0$.

Because both $\int_0^a r\big(R(r)\big)^2 dr > 0$ and $\int_0^a r\big(R'(r)\big)^2 dr \geq 0$ for a nontrivial differentiable eigenfunction $R(r)$, it follows that $\lambda \geq 0$. As to whether $\lambda = 0$ can be an eigenvalue, if $\lambda = 0$ then $R(r)$ satisfies the Cauchy-Euler ODE $r^2R'' + rR' = 0$. As we discussed in Section 3.5, substituting in $R = r^m$ gives characteristic equation $m(m-1) + m = 0$, hence the roots are $m = 0, m = 0$ and the ODE has general solution $R(r) = c_1 + c_2 \ln(r)$. Because of the BC $|\, R(0^+)\,| < \infty$, we must have $c_2 = 0$, so $R(r) = c_1$. The only way for $R(r)$ to satisfy the BC $R(a) = 0$ is for $c_1 = 0$, hence $R(r) \equiv 0$ and there is no eigenfunction. So, all of the eigenvalues have $\lambda \geq 0$ and $\lambda \neq 0$, so all eigenvalues are positive. ◯

Example 9.31

In Section 11.5 we will use the solutions of the eigenvalue problem

$$\left\{\begin{array}{c} \big((1-x^2)\Psi'(x)\big)' + \lambda\Psi(x) = 0,\ -1 < x < 1 \\ \\ |\Psi(-1^+)| < \infty,\ |\Psi(1^-)| < \infty, \end{array}\right\}. \tag{9.72}$$

The ODE here is Legendre's equation, as in Appendix C. There are infinitely many eigenvalues $\lambda_n = n(n+1)$, and the corresponding eigenfunctions are the Legendre polynomials $P_n(x)$. They satisfy the orthogonality relation

$$0 = \int_{-1}^{1} P_n(x)P_{n'}(x)\,dx, \text{ for } n \neq n'.$$

The orthogonality relation is the same as for the results for regular Sturm-Liouville problems, even though (9.72) is singular.

As we will see in Section 11.6, the eigenvalue problem (9.72) and hence the Legendre polynomials are used in solving Laplace's equation in spherical coordinates. When angle ϕ lies in the interval $(0, \pi)$ and we substitute $x = \cos\phi$, we have correspondingly $-1 < x < 1$. The orthogonality relation involves

$$0 = \int_{-1}^{1} P_n(x)P_{n'}(x)\,dx = \int_{\pi}^{0} P_n(\cos\phi)P_{n'}(\cos\phi)\big(-\sin\phi\,d\phi\big) = \int_{0}^{\pi} P_n(\cos\phi)P_{n'}(\cos\phi)\sin\phi\,d\phi.$$

Again, it is no accident that $\sin\phi\,d\phi$ appears in the orthogonality relation, because in spherical coordinates the element of volume is $dV = \rho^2\sin\phi\,d\rho\,d\phi\,d\theta$. ◯

9.6.2 A Composite Media Problem

Example 9.32

Suppose $X(x)$ satisfies ODE

$$\big(\kappa(x)X'(x)\big)' + \lambda X(x) = 0, \quad 0 < x < 1 \text{ and } 1 < x < 2, \tag{9.73}$$

where the function $\kappa(x)$ is defined piecewise by

$$\kappa(x) = \begin{cases} \frac{1}{2}, & 0 < x < 1 \\ \frac{1}{3}, & 1 < x < 2 \end{cases}.$$

Suppose $X(x)$ satisfies boundary conditions

$$X'(0) = X'(2) = 0, \tag{9.74}$$

and **interface conditions**

$$\lim_{x\to 1^-} X(x) = \lim_{x\to 1^+} X(x) \quad \text{and} \quad \lim_{x\to 1^-} \big(\kappa(x)X'(x)\big) = \lim_{x\to 1^+} \big(\kappa(x)X'(x)\big). \tag{9.75}$$

Find all of the eigenvalues and corresponding eigenfunctions.

As we will see in Section 10.2 and Example 11.5 in Section 11.1, the jump in the coefficient κ in the ODE, as well as these new, interface conditions at $x = 1$, correspond physically to joining two types of materials at $x = 1$. For example, the two materials could be two metals and κ could be a coefficient of thermal conductivity. Physically, the two interface conditions require the continuity of both the temperature and the heat flux at $x = 1$.

Method: We will find a characteristic equation for this problem: to do this, we will solve the ODE separately on the two subintervals $[0, 1]$ and $[1, 2]$ and then **match** the two solutions using the interface boundary conditions. Define $X^{(1)} = X(x)$ for $0 < x < 1$ and $X^{(2)} = X(x)$ for $1 < x < 2$.

First consider the case of $\lambda = 0$: On subinterval $[0, 1], \kappa(x) = \frac{1}{2}$, so in the ODE there is $\frac{1}{2}X''(x) + 0 \cdot X(x) = 0$, that is, $X''(x) = 0$, so the solutions are $X^{(1)}(x) = c_1 + c_2x$. The BC $X'(0) = 0$ yields $c_2 = 0$, so $X^{(1)}(x) = c_1$, $0 < x < 1$.

On $[1, 2], \kappa(x) = \frac{1}{3}$, so in the ODE there is $\frac{1}{3}X''(x) + 0 \cdot X(x) = 0$. The solutions are $X^{(2)}(x) = d_1 + d_2x$. The BC $X'(2) = 0$ yields that $d_2 = 0$, so $X^{(2)}(x) = d_1$, $1 < x < 2$.

Matching the solutions $X^{(1)}(x)$ and $X^{(2)}(x)$ requires $c_1 = X^{(1)}(1^-) = X^{(2)}(1^+) = d_1$, hence $c_1 = d_1$, as well as $0 = \lim_{x\to 1^-} \kappa(x)(X^{(1)})'(x) = \lim_{x\to 1^+} \kappa(x)(X^{(2)})'(x) = 0$, which

effectively imposes no further requirement on c_1 or d_1. So, $\lambda = 0$ is an eigenvalue, with corresponding eigenfunctions

$$X(x) = \begin{Bmatrix} c_1, & 0 < x < 1 \\ c_1, & 1 < x < 2 \end{Bmatrix},$$

and we can define $X(1) = \lim_{x\to 1^-} c_1 = \lim_{x\to 1^+} c_1 = c_1$, so $X(x) = c_1 \cdot 1$. So, an eigenfunction $X_0(x) \equiv 1$ corresponds to eigenvalue $\lambda_0 = 0$.

For $\lambda > 0$, on $[0,1], \kappa(x) = \frac{1}{2}$, so in the ODE there is $\frac{1}{2}X''(x) + \lambda X(x) = 0$, so the solutions are $X^{(1)}(x) = c_1 \cos(\sqrt{2\lambda}\, x) + c_2 \sin(\sqrt{2\lambda}\, x)$. The BC $X'(0) = 0$ yields $c_2 = 0$, so

$$X^{(1)}(x) = c_1 \cos(\sqrt{2\lambda}\, x), \quad (X^{(1)})'(x) = -c_1\sqrt{2\lambda} \sin(\sqrt{2\lambda}\, x),\ 0 < x < 1.$$

On $[1,2], \kappa(x) = \frac{1}{3}$, so the ODE there is $\frac{1}{3}X''(x) + \lambda X(x) = 0$. We could take the solutions there, $X^{(2)}(x)$, to be a linear combination of $\cos(\sqrt{3\lambda}\, x)$ and $\sin(\sqrt{3\lambda}\, x)$, but instead we can use another set of basic solutions that will help us solve the BC $X'(2) = 0$. Let

$$X^{(2)}(x) = d_1 \cos\left(\sqrt{3\lambda}\,(2-x)\right) + d_2 \sin\left(\sqrt{3\lambda}\,(2-x)\right).$$

The chain rule yields

$$(X^{(2)})'(x) = d_1\sqrt{3\lambda} \sin\left(\sqrt{3\lambda}\,(2-x)\right) - d_2\sqrt{3\lambda} \cos\left(\sqrt{3\lambda}\,(2-x)\right).$$

The BC $X'(2) = 0$ yields

$$0 = d_1\sqrt{3\lambda} \sin\left(\sqrt{3\lambda}\,(2-2)\right) - d_2\sqrt{3\lambda} \cos\left(\sqrt{3\lambda}\,(2-2)\right) = \sqrt{3\lambda} d_1 \cdot 0 + d_2 \cdot 1,$$

hence $d_2 = 0$. So*

$$X^{(2)}(x) = d_1 \cos\left(\sqrt{3\lambda}\,(2-x)\right), \quad \left(X^{(2)}\right)'(x) = d_1\sqrt{3\lambda} \sin\left(\sqrt{3\lambda}\,(2-x)\right),\ 1 < x < 2.$$

Matching the solutions $X^{(1)}(x)$ and $X^{(2)}(x)$ requires, by the interface conditions (9.75),

$$c_1 \cos(\sqrt{2\lambda}) = X^{(1)}(1^-) = X^{(2)}(1^+) = d_1 \cos(\sqrt{3\lambda}),$$

that is,

$$\cos(\sqrt{2\lambda})c_1 - \cos(\sqrt{3\lambda})d_1 = 0, \tag{9.76}$$

as well as

$$-\frac{1}{2}\sqrt{2\lambda}\, c_1 \sin(\sqrt{2\lambda}) = \lim_{x\to 1^-} \kappa(x)(X^{(1)})'(x) = \lim_{x\to 1^+} \kappa(x)(X^{(2)})'(x) = \frac{1}{3}\sqrt{3\lambda}\, d_1 \sin(\sqrt{3\lambda}),$$

that is,

$$-\frac{1}{2}\sqrt{2\lambda}\, \sin(\sqrt{2\lambda})c_1 - \frac{1}{3}\sqrt{3\lambda}\, \sin(\sqrt{3\lambda})d_1 = 0. \tag{9.77}$$

Equations (9.76) and (9.77) form a homogeneous system in two unknowns, c_1, d_1. After dividing the last equation through by $\sqrt{\frac{\lambda}{6}}$ (recall that we assumed $\lambda > 0$ in this case), we can write the system as

$$A \begin{bmatrix} c_1 \\ d_1 \end{bmatrix} \triangleq \begin{bmatrix} \cos(\sqrt{2\lambda}) & -\cos(\sqrt{3\lambda}) \\ -\sqrt{3}\sin(\sqrt{2\lambda}) & -\sqrt{2}\sin(\sqrt{3\lambda}) \end{bmatrix} \begin{bmatrix} c_1 \\ d_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

* In Section 11.3 we will call "clairvoyance" the method of using $\left\{\cos\left(\sqrt{3\lambda}\,(2-x)\right), \sin\left(\sqrt{3\lambda}\,(2-x)\right)\right\}$ instead of $\left\{\cos(\sqrt{3\lambda}\, x), \sin(\sqrt{3\lambda}\, x)\right\}$. If we use the latter we would arrive at the same solution for $X^{(2)}(x)$ but it would take more work, involving trigonometric identities, to satisfy the boundary conditions.

There is a nontrivial solution for the eigenfunction

$$X(x) = \begin{cases} X^{(1)}(x), & 0 < x < 1 \\ \\ X^{(2)}(x), & 1 < x < 2 \end{cases}$$

if and only if the system of two equations has a nontrivial solution for c_1, d_1, that is, if and only if

$$|A| = \begin{vmatrix} \cos(\sqrt{2\lambda}) & -\cos(\sqrt{3\lambda}) \\ \\ -\sqrt{3}\sin(\sqrt{2\lambda}) & -\sqrt{2}\sin(\sqrt{3\lambda}) \end{vmatrix} = 0.$$

This reduces to the characteristic equation

$$0 = -\sqrt{2}\cos(\sqrt{2\lambda})\sin(\sqrt{3\lambda}) - \sqrt{3}\sin(\sqrt{2\lambda})\cos(\sqrt{3\lambda}). \tag{9.78}$$

We can use a graphical method in Problem 9.6.4.9 to see why there are infinitely many eigenvalues and we can use Newton's Method to get good approximations to a finite number of them. (See Problem 8.1.8.13 for finding some eigenvalues.)

As to the corresponding eigenfunctions, we can use the "adjugate matrix method," of Theorem 2.2 in Section 2.1, as in Example 2.5 in Section 2.1, to find a nontrivial solution for the vector $\begin{bmatrix} c_1 \\ d_1 \end{bmatrix}$: We calculate that

$$adj(A) = \begin{bmatrix} -\sqrt{2}\sin(\sqrt{3\lambda}) & \cos(\sqrt{3\lambda}) \\ \\ \sqrt{3}\sin(\sqrt{2\lambda}) & \cos(\sqrt{2\lambda}) \end{bmatrix}.$$

Any nonzero column of $adj(A)$ is a useful solution for $\begin{bmatrix} c_1 \\ d_1 \end{bmatrix}$.

Denote $adj(A) = [\,\mathbf{v}_1 \quad \mathbf{v}_2\,]$. The first column is nonzero, because*

$$||\mathbf{v}_1||^2 = \left(-\sqrt{2}\sin(\sqrt{3\lambda})\right)^2 + \left(\sqrt{3}\sin(\sqrt{2\lambda})\right)^2 \neq 0.$$

By the way, using the fact that λ being an eigenvalue yields

$$-\sqrt{2}\cos(\sqrt{2\lambda})\sin(\sqrt{3\lambda}) = \sqrt{3}\sin(\sqrt{2\lambda})\cos(\sqrt{3\lambda}),$$

we calculate that

$$\cos(\sqrt{2\lambda})\mathbf{v}_1 = \sqrt{3}\sin(\sqrt{2\lambda})\mathbf{v}_2,$$

confirming that† $rank(adj(A)) = 1$, as predicted by Theorem 1.33 in Section 1.6.

So, we can take $\begin{bmatrix} c_1 \\ \\ d_1 \end{bmatrix} = \mathbf{v}_1 = \begin{bmatrix} -\sqrt{2}\sin(\sqrt{3\lambda}) \\ \\ \sqrt{3}\sin(\sqrt{2\lambda}) \end{bmatrix}$. Correspondingly, our eigenfunctions are

$$X_n(x) = \begin{cases} -\sqrt{2}\sin(\sqrt{3\lambda})\cos(\sqrt{2\lambda}\,x), & 0 < x < 1 \\ \\ \sqrt{3}\sin(\sqrt{2\lambda})\cos(\sqrt{3\lambda}\,(2-x)), & 1 < x < 2 \end{cases},$$

where λ is any of the eigenvalues λ_n, $n = 1, 2, \ldots$.

We leave it to the reader to explain why there are no eigenvalues with $\lambda < 0$. ◯

---

* This is by a somewhat pure mathematical explanation: $\sin(\sqrt{2\lambda}) = 0$ and $\sin(\sqrt{3\lambda}) = 0$ can both be true only if both $\sqrt{2\lambda} = n\pi$ and $\sqrt{3\lambda} = m\pi$ for some integers m, n, which implies that $\frac{n}{m} = \frac{\sqrt{2\lambda}}{\sqrt{3\lambda}} = \sqrt{\frac{2}{3}}$, hence $\sqrt{\frac{2}{3}}$ is a rational number. We'll spare you how that leads to a contradiction.

† Because we cannot have both $\sin(\sqrt{2\lambda}) = 0$ and $\cos(\sqrt{2\lambda}) = 0$.

9.6.3 Fourth-Order ODE-BVP

Example 9.33

Follow the steps below to find all eigenvalues and eigenfunctions of the fourth order ODE-BVP

$$\left\{\begin{array}{c} X^{(iv)}(x)+\lambda X=0,\ 0<x<L \\ \\ X(0)=X''(0)=X(L)=X'(L)=0 \end{array}\right\}. \tag{9.79}$$

The boundary conditions correspond to a beam **clamped** at $x=0$ and **fixed** at $x=L$, and the eigenvalue parameter $\lambda=-EI$, where E is the Young's modulus and I is the moment of inertia.

(a) First, explain why $\lambda=0$ is not an eigenvalue.

(b) Second, explain why $\lambda>0$ cannot be an eigenvalue.

(c) Third, find all eigenvalues and eigenfunctions for the case $\lambda<0$.

Method:

(a) For $\lambda=0$ the ODE in (9.79) becomes $X''''(x)=0$, so $X(x)=c_0+c_1x+\frac{1}{2}c_2x^2+\frac{1}{6}c_3x^3$, where the c_i's are arbitrary constants. Substitute $X(x)$ and $X''(x)=c_2+c_3x$ into the first two BCs to get

$$\left\{\begin{array}{ll} 0= & X(0)=c_0 \\ 0= & X''(0)=c_2 \end{array}\right\}.$$

Substitute these into $X(x)$ to reduce it to $X(x)=c_1x+\frac{1}{6}c_3x^3$. Substitute $X(x)$ and $X'(x)=c_1+\frac{1}{2}c_3x^2$ into the last two BCs to get the system of equations

$$\left\{\begin{array}{lll} 0= & X(L)=c_1L & +\frac{1}{6}c_3L^3 \\ \\ 0= & X'(L)=c_1 & +\frac{1}{2}c_3L^2 \end{array}\right\},$$

that is,

$$\begin{bmatrix} L & \frac{1}{6}L^3 \\ \\ 1 & \frac{1}{2}L^2 \end{bmatrix}\begin{bmatrix} c_1 \\ \\ c_3 \end{bmatrix}=\begin{bmatrix} 0 \\ \\ 0 \end{bmatrix}.$$

Because

$$\begin{vmatrix} L & \frac{1}{6}L^3 \\ \\ 1 & \frac{1}{2}L^2 \end{vmatrix}=\frac{1}{3}L^3\neq 0,$$

it follows that $c_1=c_3=0$. This yields $X(x)\equiv 0$, so $\lambda=0$ is not an eigenvalue of this fourth order ODE-BVP.

(b) Multiply through ODE $-\lambda X=X''''$ by $X(x)$ and integrate from $x=0$ to $x=L$ to get

$$-\lambda\int_0^L (X(x))^2\,dx=\int_0^L X(x)X''''(x)\,dx\,.$$

Similar to work in Example 9.30, use integration by parts and the BCs $X(0) = X(L) = 0$ to get

$$-\lambda \int_0^L (X(x))^2\,dx = \big[\, X(x)X'''(x) \,\big]_0^L - \int_0^L X'(x)X'''(x)\,dx$$

$$= \cancel{X(L)}X'''(L) - \cancel{X(0)}X'''(0) - \int_0^L X'(x)X'''(x)\,dx = -\int_0^L X'(x)X'''(x)\,dx.$$

Use integration by parts again to get

$$-\lambda \int_0^L (X(x))^2\,dx = -\big[\, X'(x)X''(x) \,\big]_0^L + \int_0^L X''(x)X''(x)\,dx,$$

$$= -X'(L)X''(L) + X'(0)X''(0) + \int_0^L \big(X''(x)\big)^2\,dx.$$

The BCs $X''(0) = X'(L) = 0$ imply

$$-\lambda \int_0^L (X(x))^2\,dx = \int_0^L \big(X''(x)\big)^2\,dx \geq 0. \tag{9.80}$$

Because* $\int_0^L (X(x))^2\,dx > 0$, (9.80) yields $\lambda \leq 0$. So, $\lambda > 0$ cannot be an eigenvalue.

(c) Define $\omega = (-\lambda)^{1/4}$ for convenience. The ODE $X'''' - \omega^4 X = 0$ has characteristic polynomial $s^4 - \omega^4$, which has roots $s = \pm\omega, \pm i\omega$. The solutions of the ODE are

$$X(x) = c_1 \cosh(\omega x) + c_2 \sinh(\omega x) + c_3 \cos(\omega x) + c_4 \sin(\omega x).$$

Substitute that and $X''(x) = \omega^2\big(c_1 \cosh(\omega x) + c_2 \sinh(\omega x) - c_3 \cos(\omega x) - c_4 \sin(\omega x)\big)$ into the first two BCs to get

$$\left\{\begin{array}{rl} 0 = & X(0) = c_1 + c_3 \\ 0 = & X''(0) = c_1 - c_3 \end{array}\right\},$$

which implies $c_1 = c_3 = 0$. Substitute these into $X(x)$ to reduce it to $X(x) = c_2 \sinh(\omega x) + c_4 \sin(\omega x)$. Substitute $X(x)$ and $X'(x) = \omega\big(c_2 \cosh(\omega x) + c_4 \cos(\omega x)\big)$ into the last two BCs to get the system of equations

$$\left\{\begin{array}{rlcl} 0 = & X(L) & = c_2 \sinh(\omega L) & + \quad c_4 \sin(\omega L) \\ 0 = & X'(L) & = \omega\big(c_2 \cosh(\omega L) & + \quad c_4 \cos(\omega L)\big) \end{array}\right\},$$

that is,

$$\begin{bmatrix} \sinh(\omega L) & \sin(\omega L) \\ \cosh(\omega L) & \cos(\omega L) \end{bmatrix} \begin{bmatrix} c_2 \\ c_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{9.81}$$

There is a nontrivial solution for $X(x)$ if, and only if,

$$0 = \begin{vmatrix} \sinh(\omega L) & \sin(\omega L) \\ \cosh(\omega L) & \cos(\omega L) \end{vmatrix} = \sinh(\omega L)\cos(\omega L) - \cosh(\omega L)\sin(\omega L)\,,$$

* If not, $X(x) \equiv 0$ on $[\,0, L\,]$.

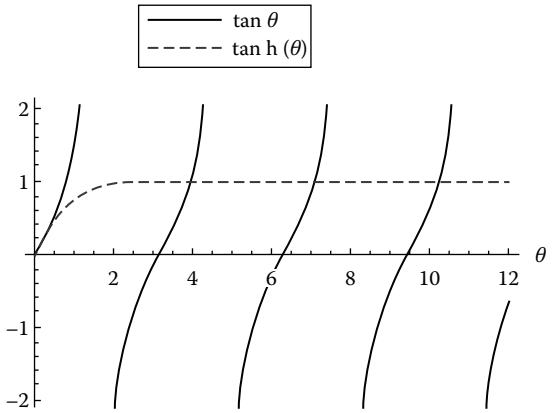


FIGURE 9.30
Graphical method for finding eigenvalues.

that is, if, and only if, $\sinh(\omega L)\cos(\omega L) = \cosh(\omega L)\sin(\omega L)$, which is true only if $\dfrac{\sinh(\omega L)}{\cosh(\omega L)} = \dfrac{\sin(\omega L)}{\cos(\omega L)}$, that is,

$$\tanh(\theta) = \tan(\theta), \tag{9.82}$$

where $\theta = \omega L$ and $\tanh(\theta) \triangleq \dfrac{\sinh(\theta)}{\cosh(\theta)}$.

Figure 9.30's "graphical solution" of (9.82) shows that there are infinitely many eigenvalues

$$\lambda_n = -\left(\frac{\theta_n}{L}\right)^4, \; n = 1, 2, \ldots .$$

Note that $\theta_n \approx \frac{\pi}{4} + (n-1)\pi$, as $n \to \infty$. For the corresponding eigenfunctions, again we can use the "adjugate matrix method" of Theorem 2.2 in Section 2.1, as in Example 9.32, to find a solution of (9.81):

$$\begin{bmatrix} c_2 \\ c_4 \end{bmatrix} = \begin{bmatrix} \cos(\theta_n) \\ -\cosh(\theta_n) \end{bmatrix},$$

The corresponding eigenfunctions are

$$X_n(x) = \cos(\theta_n)\sinh\left(\frac{\theta_n x}{L}\right) - \cosh(\theta_n)\sin\left(\frac{\theta_n x}{L}\right), \; n = 1, 2, \ldots . \bigcirc$$

Eigenfunctions, also known as *mode shapes*, for ODE $X''''(x) + \lambda X = 0$ with various combinations of boundary conditions, are given on page 6.97 of Merritt et al. (1996).

9.6.4 Problems

1. Explain the very last result mentioned in Theorem 9.15, that is, explain why $\epsilon_0, \epsilon_1 > 0$, $\gamma_0, \gamma_1 \geq 0$ and $q(x) \leq 0$ on the interval $[a, b]$ together imply that all eigenvalues for problems (9.63) and (9.64) are positive.

For the ODE-BVP in problems 2–4, (a) find all the eigenvalues and (b) the orthogonality relation. Note that the ODE is a Cauchy-Euler equation.

2. $\left\{\begin{array}{c}(r^2R')' + \lambda R(r) = 0,\ a < r < b \\ \\ R'(a) = R'(b) = 0\end{array}\right\}$
3. $\left\{\begin{array}{c}(rR')' + \lambda \frac{1}{r} R(r) = 0,\ a < r < b \\ \\ R'(a) = R'(b) = 0\end{array}\right\}$
4. $\left\{\begin{array}{c}r^2R'' + rR + \lambda R = 0,\ 1 < r < 3 \\ \\ R(1) = R(3) = 0\end{array}\right\}$

For the ODE-BVP in problems 5–7, (a) find the characteristic equation satisfied by all eigenvalues, and, if possible, their values, and (b) the orthogonality relation satisfied by the eigenfunctions.

5. $\left\{\begin{array}{c}X'' + \lambda X = 0,\ \frac{\pi}{4} < x < \frac{3\pi}{4} \\ \\ X\left(\frac{\pi}{4}\right) - 3X'\left(\frac{\pi}{4}\right) = 0 \\ \\ X\left(\frac{3\pi}{4}\right) + X'\left(\frac{3\pi}{4}\right) = 0\end{array}\right\}$
6. $\left\{\begin{array}{c}X'' + 2\beta X' + \lambda X(x) = 0, 0 < x < L, \\ \\ X(0) = X(L) = 0\end{array}\right\}$, where $\lambda > \beta^2$.
7. $\left\{\begin{array}{c}\frac{d^4X}{dx^4} + \lambda X = 0,\ 0 < x < L \\ \\ X(0) = X'(0) = X(L) = X'(L) = 0\end{array}\right\}$

The boundary conditions correspond to a beam "fixed" at both $x = 0$ and $x = L$.

8. Suppose ODE-BVP

$$\left\{\begin{array}{c}\psi'' + q(x)\psi(x) = 0,\ a < x < b \\ \\ \psi(a) = \psi(b) = 0\end{array}\right\}$$

has a solution $\psi(x)$ and nonhomogeneous ODE-BVP

$$\left\{\begin{array}{c}y'' + q(x)y(x) = f(x),\ a < x < b \\ \\ y(a) = y(b) = 0\end{array}\right\}$$

has a solution $y(x)$. Show that

$$\int_a^b \psi(x)y(x) = 0.$$

[This is an example of a "Fredholm Alternative" result, similar to Problem 1.4.1.7.]

9. For Example 9.32, we derived the characteristic equation (9.78), that is,

$$0 = -\sqrt{2}\cos(\sqrt{2\lambda})\sin(\sqrt{3\lambda}) - \sqrt{3}\sin(\sqrt{2\lambda})\cos(\sqrt{3\lambda}).$$

Use a graphical method to analyze this. For example, you could move the second term to the other side and then divide through by $\sqrt{3}\sin(\sqrt{3\lambda})\cos(\sqrt{2\lambda})$ to get

$$0 = \sqrt{\frac{2}{3}} + \frac{\tan(\sqrt{2\lambda})}{\tan(\sqrt{3\lambda})}.$$

[To use Newton's Method, it's probably better to let $\theta \triangleq \sqrt{\lambda}$ and define a function by

$$f(\theta) = -\sqrt{2}\cos(\sqrt{2}\,\theta)\sin(\sqrt{3}\,\theta) - \sqrt{3}\sin(\sqrt{2}\,\theta)\cos(\sqrt{3}\,\theta).]$$

10. Assume that $\lambda > 0$. For the ODE-BVP

$$(\star)\left\{\begin{array}{c} X'' + \lambda X(x) = 0,\ 0 < x < L \\ X(0) - hX'(0) = 0,\ X(L) + hX'(L) = 0 \end{array}\right\},$$

(a) Find an equation in λ whose solutions give eigenvalues for $(\star)$. Do not solve the equation, that is, do not find the eigenvalues.

(b) For the eigenvalues of $(\star)$ use the adjugate matrix method of Theorem 2.2 in Section 2.1, as in Examples 9.32 and 9.33, to find a formula for the corresponding eigenfunctions.

11. Theorem 9.15 guarantees that the problem

$$\left\{\begin{array}{c} X'' + \lambda X(x) = 0, a < x < b, \\ \epsilon_0 X(a) - \epsilon_1 X'(a) = 0, \\ \gamma_0 X(b) + \gamma_1 X'(b) = 0 \end{array}\right\}$$

has infinitely many eigenvalues λ_n and corresponding eigenfunctions $X_n(x)$. For those n for which $\omega_n^2 = \lambda_n > 0$, $X_n(x)$ satisfies the undamped oscillator ODE, so $X_n(x) = c_{1,n}\cos(\omega_n x) + c_{2,n}\sin(\omega_n x)$. Substitute this into the BCs and use the adjugate matrix method of Theorem 2.2 in Section 2.1 to find $c_{2,n}$ in terms of $c_{1,n}$ and thus find a way to express $X_n(x)$ in terms of $\omega_n, \cos(\omega_n x), \sin(\omega_n x), \epsilon_0, \epsilon_1, \gamma_0, \gamma_1, a, b$.

12. Define the function $\kappa(x)$ piecewise by $\kappa(x) = \begin{cases} 1, & 0 < x < 1 \\ \frac{1}{4}, & 1 < x < 2 \end{cases}$. Find all eigenvalues and corresponding eigenfunctions for problem $(\star)$ below.

$$(\star) \begin{cases} \big(\kappa(x)X'(x)\big)' + \lambda X(x) = 0, \\ 0 < x < 1 \text{ and } 1 < x < 2, \\ \\ X'(0) = X'(2) = 0, \\ \\ X(1^-) = X(1^+), \quad \lim_{x\to 1^-} \kappa(x)X'(x) \\ = \lim_{x\to 1^+} \kappa(x)X'(x) \end{cases}.$$

13. [Small project suggested by Professor Allen Hunt] Consider Example 9.32 with instead

$$\kappa(x) = \begin{cases} \kappa_1, & 0 < x < 1 \\ \\ \kappa_2, & 1 < x < 2 \end{cases}.$$

If $\sqrt{\kappa_2/\kappa_1}$ is a rational number, that is, the ratio of integers, find all the eigenvalues and eigenfunctions. For a smaller project, make a specific choice of κ_1, κ_2 for which $\sqrt{\kappa_2/\kappa_1}$ is rational.

9.7 Rayleigh Quotient

Consider a Sturm–Liouville eigenvalue problem with special boundary conditions

$$\begin{cases} \big(p(x)X'(x)\big)' + (\lambda s(x) + q(x))X(x) = 0, \ a < x < b, \\ \\ X(a) = X(b) = 0 \end{cases}.$$

Multiply the ODE by $X(x)$ and then integrate from a to b to get

$$0 = \int_a^b X(x) \cdot \big(p(x)X'(x)\big)'\, dx + \int_a^b X(x) \cdot (\lambda s(x) + q(x))X(x)\, dx.$$

Similar to how we established the orthogonality relation (9.65) in Section 9.6, use integration by parts on the first integral to get

$$0 = \big[X(x) \cdot p(x)X'(x)\big]_a^b - \int_a^b X'(x) \cdot p(x)X'(x)\, dx + \int_a^b X(x) \cdot (\lambda s(x) + q(x))X(x)\, dx.$$

Because $X(a) = X(b) = 0$, the terms in the brackets are zero. Put the first integral on the other side of the equation to get

$$\int_a^b p(x)\big(X'(x)\big)^2\,dx = \lambda \int_a^b s(x)\big(X(x)\big)^2\,dx + \int_a^b q(x)\big(X(x)\big)^2\,dx,$$

and then move the second integral on the right-hand side to the left-side of the equation to get

$$\int_a^b \big(-q(x)\big(X(x)\big)^2 + p(x)\big(X'(x)\big)^2\big)dx = \lambda \int_a^b s(x)\big(X(x)\big)^2\,dx.$$

If we assume that $s(x)$ satisfies one of the assumptions of a regular Sturm-Liouville problem, namely, that

$$s(x) \geq 0 \text{ and is not identically zero on } [a, b],$$

it is reasonable to expect that $X(x)$ being an eigenfunction implies that there is no division by zero problem in rewriting the last identity involving integrals as

$$\lambda = \frac{\int_a^b \big(p(x)\big(X'(x)\big)^2 - q(x)\big(X(x)\big)^2\big)\,dx}{\int_a^b s(x)\big(X(x)\big)^2 dx}.$$

This explains why the eigenvalue λ can be expressed in terms of definite integrals involving the eigenfunction.

This looks more interesting than it really is, because it seems to say that in order to evaluate λ we should find the eigenfunction $X(x)$ first in order to evaluate the integrals. But, to find $X(x)$ we would seem to need to solve the ODE which involves the unknown value λ! That is circular logic. It is therefore pretty amazing that the quotient can be used to find an eigenvalue(s).

Definition 9.6

For a *regular* Sturm–Liouville (SL) problem (9.63) in Section 9.6, with its general BCs (9.64) in Section 9.6, the **Rayleigh quotient** is given by

$$\mathcal{R}_{SL}[\,X(x)\,] \triangleq \frac{-\big[p(x)X(x)X'(x)\big]_a^b + \int_a^b \big(p(x)\big(X'(x)\big)^2 - q(x)\big(X(x)\big)^2\big)\,dx}{\int_a^b s(x)\big(X(x)\big)^2 dx}. \tag{9.83}$$

Remarks:

(a) As we will see in Section 11.1, the minimum eigenvalue λ_1 is used in calculating the "time constant" which indicates how quickly the temperature distribution of an object settles down to be close to the equilibrium temperature distribution. In fact, the Rayleigh Quotient estimate can be very useful in making approximations, or "back of the envelope" calculations, for physical problems.

(b) The Rayleigh Quotient for a Sturm–Liouville problem is "homogeneous" in the sense that if X is a nice, nontrivial function on the interval (a, b) and β is a nonzero scalar, then

$$\mathcal{R}_{SL}[\,\beta X(x)\,] = \mathcal{R}_{SL}[\,X(x)\,]. \tag{9.84}$$

We'll see the application of this remark in Example 9.35.

Theorem 9.16

For a regular Sturm-Liouville problem (9.63) and (9.64) in Section 9.6, the minimum eigenvalue, λ_1, is given by

$$\lambda_1 = \min\{\mathcal{R}_{SL}[X(x)] : \text{ nice, differentiable, nontrivial } X(x)\}. \tag{9.85}$$

In the above, "$X(x) \not\equiv 0$" means "$X(x)$ is not identically zero on the interval (a, b)."

Theorem 9.16 is quite a deep result because its explanation can use the advanced mathematical subject of "Functional Analysis." In fact, the origins of that subject were exactly in the symbolic calculations we have just done, and expressions such as

$$-\big[p(x)X(x)X'(x)\big]_a^b + \int_a^b \big(p(x)\big(X'(x)\big)^2 - q(x)\big(X(x)\big)^2\big)\,dx$$

and

$$\int_a^b s(x)\big(X(x)\big)^2\,dx$$

are called "quadratic functionals." While the Theorem is true, I am sweeping under the carpet all sorts of abstract mathematical details about what such a minimization entails, including why there is such a minimum value. Also, in Section 10.6 and Chapter 14 we will see examples of quadratic functionals and their interpretation as energy, either kinetic and potential.

Example 9.34

Use Theorem 9.16 and a single **trial function** $X(x)$ to get an estimate for the minimum eigenvalue for the ODE-BVP eigenvalue problem

$$\left\{\begin{array}{c} X''(x)+\lambda X(x)=0,\ 0<x<L, \\ hX(0)-X'(0)=0, \\ X'(L)=0 \end{array}\right\}, \tag{9.86}$$

where $h>0$ is a given heat transfer coefficient.

Method: According to Definition 9.6, the Rayleigh Quotient for this problem is

$$\mathcal{R}_{SL}[X(x)]=\frac{-\big(X(L)X'(L)-X(0)X'(0)\big)+\int_0^L \big(X'(x)\big)^2\,dx}{\int_0^L \big(X(x)\big)^2\,dx}=\frac{h\big(X(0)\big)^2+\int_0^L \big(X'(x)\big)^2\,dx}{\int_0^L \big(X(x)\big)^2\,dx}.$$

For a specific function $X(x)$, the single number $\mathcal{R}_{SL}[\,X(x)\,]$ will be an estimate for the minimum eigenvalue λ_1.

If $X(x)$ were an eigenfunction of (9.86) then the Rayleigh Quotient might exactly equal the minimum eigenvalue. So, it makes sense to choose $X(x)$ that satisfies the boundary conditions $hX(0)-X'(0)=X'(L)=0$.

A quadratic function

$$X(x)=\alpha+\beta x+\frac{1}{2}\,\xi x^2$$

will satisfy the BCs if we correctly choose values for the parameters α,β,ξ. Because $X'(x)=\beta+\xi x$, the two BCs require $hX(0)-X'(0)=h\cdot\alpha-\beta=0$ and $X(L)=\beta+\xi L=0$. This yields $\alpha=\frac{\beta}{h}$ and $\xi=-\frac{\beta}{L}$. So, for all β, the function $X(x)=\frac{\beta}{h}+\beta x-\frac{\beta}{2L}x^2=\beta\left(\frac{1}{h}+x-\frac{1}{2L}x^2\right)$ satisfies the BCs. But by (9.84), $\mathcal{R}_{SL}\left[\,\beta\left(\frac{1}{h}+x-\frac{1}{2L}x^2\right)\right]=\mathcal{R}_{SL}\left[\left(\frac{1}{h}+x-\frac{1}{2L}x^2\right)\right]$, so it will be enough to get an estimate using the single trial function

$$X(x)\triangleq\frac{1}{h}+x-\frac{1}{2L}x^2.$$

We calculate that $X(0)=\frac{1}{h}$ and $X'(x)=1-\frac{1}{L}x$, so

$$\mathcal{R}_{SL}\left[\frac{1}{h}+x-\frac{1}{2L}x^2\right]=\frac{\frac{1}{h}+\int_0^L\left(1-\frac{1}{L}x\right)^2dx}{\int_0^L\left(\frac{1}{h}+x-\frac{1}{2L}x^2\right)^2dx}=\ldots=\frac{\frac{1}{h}+\frac{1}{3}L}{\frac{L}{h^2}+\frac{2L^2}{3h}+\frac{2L^3}{15}}=\frac{15h+5h^2L}{15L+10hL^2+2h^2L^3}.$$

So, we get an estimate for the minimum eigenvalue of problem (9.86):

$$\lambda_1\approx\frac{15h+5h^2L}{15L+10hL^2+2h^2L^3}.$$

For the example where $L=2$ and $h=\frac{1}{4}$, we get

$$\lambda_1\approx\frac{35}{328}\approx 0.1067073171.$$

This compares quite well with $\lambda_1\approx 0.10669081097$, which we got in Example 9.16 in Section 9.3 using Newton's Method to find an approximate solution of the characteristic equation for (9.86). Also, because the function $X(x)=4+x-\frac{1}{4}\,x^2$ satisfies the boundary

conditions and gives a good estimate for λ_1, we can use $X(x)$ as an approximation for an eigenfunction. ◯

Example 9.35

Use Theorem 9.16 and a family of functions to get an estimate for the minimum eigenvalue for (9.86), the same ODE-BVP eigenvalue problem as in Example 9.34.

Method: Again, the Rayleigh Quotient for this problem is

$$\mathcal{R}_{SL}[\,X(x)\,] = \frac{h\big(X(0)\big)^2 + \int_0^L \big(X'(x)\big)^2\,dx}{\int_0^L \big(X(x)\big)^2\,dx}.$$

We can substitute into the Rayleigh Quotient a family of **trial functions** and get an estimate for the minimum eigenvalue by minimizing $\mathcal{R}_{SL}[\,X(x)\,]$ for that family. Let's try $X(x) = \alpha + \beta x$, that is, all "linear" functions, as the trial functions. By (9.84), we don't need two adjustable parameters α, β in our family of trial functions, since

$$\mathcal{R}_{SL}[\,\alpha + \beta x)\,] = \mathcal{R}_{SL}\left[\,\alpha\Big(1 + \frac{\beta}{\alpha}\,x\Big)\,\right] = \mathcal{R}_{SL}\Big[\,1 + \frac{\beta}{\alpha}\,x\,\Big]$$

for all $\alpha \neq 0$. Denoting $\mu = \frac{\beta}{\alpha}$, we have

$$\mathcal{R}_{SL}[1 + \mu x] = \frac{h + \int_0^L \mu^2 dx}{\int_0^L (1 + \mu x)^2\,dx} = \frac{h + \mu^2 L}{L + \mu L^2 + \frac{1}{3}\mu^2 L^3} = \frac{h + \mu^2 L}{L(1 + \mu L + \frac{1}{3}\mu^2 L^2)} \triangleq F(\mu).$$

The minimum value of $F(\mu)$ will give us an estimate for the minimum eigenvalue λ_1.

For the example with $h = \frac{1}{4}$ and $L = 2$ and $F(\mu) = \dfrac{0.25 + 2\mu^2}{2(1 + 2\mu + \frac{4}{3}\mu^2)}$,

the *Mathematica* command* NMinimize [F[μ], μ] has output {0.11064, {μ− > 0.129787}}.

So, using a family of all linear functions as trial functions, we estimate the minimum eigenvalue of $\lambda \approx 0.11064$ for the ODE-BVP eigenvalue problem (9.86). We have no interest in the value of μ that gives an estimate for the minimum of the Rayleigh Quotient. (If we had chosen the family of trial functions to be **admissible**, that is, to satisfy the BCs, then the value of μ and the corresponding trial function $X(x) = 1 + \mu x$ would be of interest as an approximation for an eigenfunction.) This compares pretty well with the approximate minimum eigenvalue $\lambda \approx 0.10669084$ that we found in Example 9.16 in Section 9.3. ◯

9.7.1 Problems

1. For $\mathcal{R}_{SL}[\,X(x)\,]$ given in (9.83), explain why (9.84) is true for all constants $\beta \neq 0$ and nice, differentiable nontrivial functions $X(x)$.

* Or, we could have used the analytical, exact methods of minimization found in Calculus I.

2. Use a single "trial" function $X(x)$ to get an estimate for the minimum eigenvalue for ODE-BVP eigenvalue problem

$$\left\{\begin{array}{c} X''(x) + \lambda X(x) = 0,\ 0 < x < L \\ X'(0) = 0, \\ hX(L) + X'(L) = 0 \end{array}\right\},$$

where $h > 0$ is a given heat transfer coefficient. Write down what the Rayleigh quotient is here.

3. Use a family of "trial" functions to get an estimate for the minimum eigenvalue for ODE-BVP eigenvalue problem

$$\left\{\begin{array}{c} X''(x) + \lambda X(x) = 0,\ 0 < x < L \\ X'(0) = 0, \\ hX(L) + X'(L) = 0 \end{array}\right\},$$

where $h > 0$ is a given heat transfer coefficient. Write down what the Rayleigh quotient is here.

4. Find a nontrivial second degree polynomial function $X(x)$ that satisfies the boundary conditions $X(0) = X(L) = 0$. Use $X(x)$ as a single "trial" function to get an estimate for the minimum eigenvalue for ODE-BVP eigenvalue problem

$$\left\{\begin{array}{c} X''(x) + \lambda x X(x) = 0,\ 0 < x < L \\ X(0) = X(L) = 0 \end{array}\right\}.$$

Write down what the Rayleigh quotient is here.

5. Find a function(s) $X(x)$ of the form $\cos \omega x$ that satisfies the boundary conditions $X'(0) = X'(L) = 0$. Use $X(x)$ as a "trial" function to get an estimate for the minimum eigenvalue for the ODE-BVP eigenvalue problem

$$\left\{\begin{array}{c} X''(x) + \lambda x X(x) = 0,\ 0 < x < L \\ X'(0) = X'(L) = 0 \end{array}\right\}.$$

Write down what the Rayleigh quotient is here.

6. Use a single "trial" function $X(x)$ to get an estimate for the minimum eigenvalue for the ODE-BVP eigenvalue problem

$$\left\{\begin{array}{c} \big((1+x)X'(x)\big)' + \lambda X(x) = 0,\ 0 < x < L \\ X'(0) = X(L) = 0 \end{array}\right\}.$$

Write down what the Rayleigh quotient is here.

7. A light cantilever beam is clamped at the end $x = L$ with vertical yawing axis and loaded at the free end $x = 0$ with load P acting on the longitudinal, that is, vertical, axis. As in Example 2.4 of (Temple and Bickley 1956), assume the beam's vertical displacement $X(x)$ satisfies (ODE) $ACX'' + P^2x^2X = 0$ for $0 < x < L$, where A is the

torsional rigidity and C is the flexural rigidity of the beam section. The geometrical assumptions at the ends yields the boundary conditions $X'(0) = X(L) = 0$.

This problem will demonstrate, by example, a method for improving accuracy by using higher order boundary conditions to refine a trial function in the Rayleigh Quotient method.

(a) Write down the corresponding Rayleigh Quotient $\mathcal{R}_{SL}$. The eigenvalue parameter is $\lambda = P^2$, so the minimum load is $P_1 = \sqrt{\lambda_1}$. For $P < P_1$, the beam is undeflected, that is, has $X(x) \equiv 0$.

(b) Because $X(x)$ satisfies the ODE, take the limit, as $x \to 0^+$, to explain why ODE yields the boundary condition $X''(0) = 0$.

(c) Because $X(x)$ satisfies the ODE, take the limit, as $x \to L^-$, to explain why ODE yields the boundary condition $X''(L) = 0$.

(d) Take the derivative with respect to x of both sides of the ODE to conclude that $ACX''' + P^2(2xX + x^2X') = 0$. After that, take the limit, as $x \to 0^+$, to explain why ODE yields the boundary condition $X''''(0) = 0$.

(e) Find a polynomial $q(x)$ of degree 5 that satisfies all five boundary conditions $q'(0) = q''(0) = q'''(0) = q(L) = q''(L) = 0$. [Note that q is not required to satisfy the ODE itself.] Use $q(x)$ as a trial function in the Rayleigh Quotient to get an estimate on the load in terms of the other physical parameters, specifically

$$P_1 \approx \sqrt{\frac{858}{53}}\,\frac{\sqrt{AC}}{L^2} \approx 4.023\frac{\sqrt{AC}}{L^2}.$$

By the way, from a table of zeros of the Bessel function of order $-\frac{1}{4}$ we get $P_1 \approx 4.0126\dfrac{\sqrt{AC}}{L^2}$.

9.8 Parseval's Theorems and Applications

As we will see in Chapter 10, the square of a function, integrated on an interval, can measure the total amount of some physical quantity. For example, if the mass density of a string is 100 gm/m and the vertical position of points on the string are given by $y = u(x,t)$ for $0 < x < L$, then the total kinetic energy of the string in *joules* is

$$(0.1) \cdot \int_0^L \left|\frac{\partial u}{\partial t}(x,t)\right|^2 dx.$$

Using orthogonality we can express that in terms of Fourier sine series coefficients.

In Section 2.4, we saw Parseval identities for vectors in $\mathbb{R}^n$, specifically (2.19)(b) and (2.19)(c) in Section 2.4.

First, let's derive the first Parseval's Identity for a Fourier series: Suppose $f(x)$ is given on the interval $[\,-L, L\,]$. Then

$$f(x) \doteq f_s(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right).$$

Let us now integrate the square of f on the interval $[\,-L, L\,]$. We will not worry about the pure mathematical technicalities that are needed to justify the convergence of the infinite sums and why we can move the $\sum_{n=1}^{\infty}$ operation in or out of the integral that defines the inner product. So, in a non-rigorous way we have

$$\begin{aligned}
&\int_{-L}^{L} |f(x)|^2 dx = \langle f(x), f(x)\rangle \\
&= \left\langle \frac{a_0}{2} + \sum_{m=1}^{\infty} a_m \cos\left(\frac{m\pi x}{L}\right) + \sum_{m=1}^{\infty} b_m \sin\left(\frac{m\pi x}{L}\right),\ \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right) \right\rangle \\
&= \left(\frac{a_0}{2}\right)^2 \langle 1,\ 1\rangle + a_0 \sum_{n=1}^{\infty} a_n \left\langle 1, \cos\left(\frac{n\pi x}{L}\right)\right\rangle + a_0 \sum_{n=1}^{\infty} b_n \left\langle 1, \sin\left(\frac{n\pi x}{L}\right)\right\rangle \\
&\quad + 2\sum_{n=1}^{\infty}\sum_{m=1}^{\infty} a_m b_n \left\langle \cos\left(\frac{m\pi x}{L}\right),\ \sin\left(\frac{n\pi x}{L}\right)\right\rangle + \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} a_m a_n \left\langle \cos\left(\frac{m\pi x}{L}\right),\ \cos\left(\frac{n\pi x}{L}\right)\right\rangle \\
&\quad + \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} b_m b_n \left\langle \sin\left(\frac{m\pi x}{L}\right),\ \sin\left(\frac{n\pi x}{L}\right)\right\rangle \\
&= \left(\frac{a_0}{2}\right)^2 2L + a_0 \cdot 0 + a_0 \cdot 0 + 2\sum_{n=1}^{\infty}\sum_{m=1}^{\infty} a_m b_n \cdot 0 + \sum_{n=1}^{\infty} a_n a_n L + \sum_{n=1}^{\infty} b_n b_n L \\
&= 2L\left(\left(\frac{a_0}{2}\right)^2 + \frac{1}{2}\sum_{n=1}^{\infty}\left(a_n^2 + b_n^2\right)\right).
\end{aligned}$$

Theorem 9.17

Assuming both sides exist,

$$\frac{1}{2L}\int_{-L}^{L} |f(x)|^2\, dx = \left(\frac{a_0}{2}\right)^2 + \frac{1}{2}\sum_{n=1}^{\infty}\left(a_n^2 + b_n^2\right),$$

where the Fourier series of f on the interval $[\,-L, L\,]$ is given by

$$f(x) \doteq f_s(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right).$$

In a similar way we can use orthogonality to establish Parseval's Theorems for a Fourier sine series or a Fourier cosine series:

Theorem 9.18

Assuming both sides exist,

$$\frac{1}{L}\int_0^L |f(x)|^2 dx = \frac{1}{2}\sum_{n=1}^{\infty} b_n^2,$$

where the Fourier sine series of f on the interval $[\,0, L\,]$ is given by

$$f(x) \doteq f_{sin}(x) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right),$$

where $b_n = \dfrac{2}{L}\int_0^L f(x)\sin\left(\dfrac{n\pi x}{L}\right) dx$.

Theorem 9.19

Assuming both sides exist,

$$\frac{1}{L}\int_0^L |f(x)|^2 dx = \left(\frac{a_0}{2}\right)^2 + \frac{1}{2}\sum_{n=1}^{\infty} a_n^2,$$

where the Fourier cosine series of f on the interval $[\,0, L\,]$ is given by

$$f(x) \doteq f_{cos}(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right),$$

where $a_0 = \dfrac{2}{L}\int_0^L f(x)\,dx$ and $a_n = \dfrac{2}{L}\int_0^L f(x)\cos\left(\dfrac{n\pi x}{L}\right) dx$.

Example 9.36

Suppose that the current through a resistor of $R\,\Omega$ is

$$I(t) = \sum_{n=1}^{N} b_n \sin\left(\frac{n\pi t}{L}\right).$$

Find the total energy used by that resistor over one period of time.

Method: For $n = 1, 2, \ldots, N$, the function $\sin\left(\frac{n\pi t}{L}\right)$ is periodic with period $\frac{2\pi}{\frac{n\pi}{L}} = \frac{2L}{n}$. For $n = 1, \ldots, N$, each $\sin\left(\frac{n\pi t}{L}\right)$ is periodic with period $2L$, so $I(t)$ is periodic with period $2L$.

To expand $I(t)$ in a Fourier series with period $2L$, we see that it is already given to us as

$$I(t) = \frac{0}{2} + \sum_{n=1}^{\infty} 0 \cdot \cos\left(\frac{n\pi t}{L}\right) + \sum_{n=1}^{N} b_n \sin\left(\frac{n\pi t}{L}\right) + \sum_{n=N+1}^{\infty} 0 \cdot \sin\left(\frac{n\pi t}{L}\right).$$

Across a resistor of $R\ \Omega$, the voltage drop is $V = RI$, so the power used by the resistor is $P = IV = RI^2$. The total energy used by the resistor over one period is $\int_0^{2L} R(I(t))^2\,dt$, by Problem 9.1.8.22. Applying Parseval's Theorem 9.17 and problem 9.1.8.22, we get that

$$E = \int_0^{2L} R\big(I(t)\big)^2 dt = RL \sum_{n=1}^{N} b_n^2$$

is the total energy used by the resistor over one period of time. ○

We can also use the various forms of Parseval's Theorem to establish exact values for interesting infinite series.

Example 9.37

Use Parseval's Theorem for a Fourier cosine series for $f(x) = x$ to establish the identity

$$\frac{\pi^4}{96} = \sum_{k=1}^{\infty} \frac{1}{(2k-1)^4} = \frac{1}{1^4} + \frac{1}{3^4} + \frac{1}{5^4} + \cdots.$$

Method: From Example 9.9 in Section 9.2, the Fourier cosine series of $f(x)$ for $0 < x < \pi$ is

$$x \doteq f_{cos}(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right) = \frac{\pi}{2} + \sum_{k=1}^{\infty} \frac{4}{\pi(2k-1)^2} \cos\big((2k-1)x\big).$$

Applying Theorem 9.19 gives us

$$\frac{\pi^2}{3} = \frac{1}{\pi}\int_0^{\pi} x^2 dx = \frac{1}{\pi}\int_0^{\pi} |f(x)|^2 dx = \left(\frac{a_0}{2}\right)^2 + \frac{1}{2}\sum_{n=1}^{\infty} a_n^2 = \left(\frac{\pi}{2}\right)^2 + \frac{1}{2}\sum_{k=1}^{\infty}\left(\frac{4}{\pi(2k-1)^2}\right)^2$$

$$= \frac{\pi^2}{4} + \frac{8}{\pi^2}\sum_{k=1}^{\infty}\frac{1}{(2k-1)^4},$$

hence, as desired,

$$\sum_{k=1}^{\infty}\frac{1}{(2k-1)^4} = \frac{\pi^2}{8}\left(\frac{\pi^2}{3} - \frac{\pi^2}{4}\right) = \frac{\pi^4}{96}. \bigcirc$$

9.8.1 Best Approximation by a Partial Sum of a Fourier Series

A $f(x)$ defined and continuous on the interval $(-L, L)$ is **square integrable** if

$$\int_{-L}^{L} |f(x)|^2\,dx < \infty.$$

The set of all such functions can be made larger by considering sequences $\{f_N(x)\}_{N=1}^{\infty}$ of such functions and saying $f_N \to f$ if the sequence of real numbers

$$\int_{-L}^{L} |f(x) - f_N(x)|^2 dx \to 0, \text{ as } N \to \infty.$$

We can consider the set $L^2[-L, L]$ to be the set of all such f, although to be honest, elements of $L^2[-L, L]$ are not exactly functions but are "represented" by functions; in addition, such functions have to be "measurable" and the concept of integration is actually "Lebesgue integration." We're "sweeping under the rug" a tremendous amount of beautiful and useful mathematics, but for our purpose of applying mathematics we will think of $L^2[-L, L]$ as a Hilbert space of "functions" when given the real inner product

$$\langle f, g \rangle \triangleq \int_{-L}^{L} f(x)g(x)\, dx$$

or the complex inner product

$$\langle f, g \rangle \triangleq \int_{-L}^{L} f(x)\overline{g(x)}\, dx.$$

Given a function in $L^2[-L, L]$, it can be expanded in a Fourier series

$$f(x) \doteq f_s(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos\left(\frac{n\pi x}{L}\right) + b_n \sin\left(\frac{n\pi x}{L}\right)\right)$$

in the sense that the sequence $\{s_N(x)\}_{N=1}^{\infty}$ of partial sums of the Fourier series has

$$\int_{-L}^{L} |f(x) - s_N(x)|^2 dx \to 0, \quad \text{as } N \to \infty.$$

Here,

$$s_N(x) = \frac{a_0}{2} + \sum_{n=1}^{N} \left(a_n \cos\left(\frac{n\pi x}{L}\right) + b_n \sin\left(\frac{n\pi x}{L}\right)\right),$$

where the formulas for $a_0; a_1, a_2, \ldots, b_1, b_2, \ldots$ were given in (9.6) through (9.8) in Section 9.1.

Theorem 9.20

If f is in $L^2[-L, L]$ then the function g_N in

$$\mathcal{V}_N \triangleq Span\left\{1, \cos\left(\frac{\pi x}{L}\right), \sin\left(\frac{\pi x}{L}\right), \ldots, \cos\left(\frac{N\pi x}{L}\right), \sin\left(\frac{N\pi x}{L}\right)\right\}$$

that is closest to f, that is, which minimizes

$$||f - g_N||^2 \triangleq \int_{-L}^{L} |f(x) - g_N(x)|^2 dx,$$

is $g_N = s_N$. In other words, the **best approximation** by a linear combination of the first $2N + 1$ Fourier terms is a partial sum of the Fourier series of f.

Why? It is easy to normalize to get an o.n. set

$$\left\{\frac{1}{2L} \cdot 1, \frac{1}{L}\cos\left(\frac{\pi x}{L}\right), \frac{1}{L}\sin\left(\frac{\pi x}{L}\right), \ldots, \frac{1}{L}\cos\left(\frac{N\pi x}{L}\right), \frac{1}{L}\sin\left(\frac{N\pi x}{L}\right)\right\}.$$

After that, use Theorem 2.54(c) in Section 2.10. ◯

Theorem 9.21

(**Bessel's Inequality**) If f is in $L^2[-L, L]$ then s_N, a partial sum of its Fourier series, satisfies

$$L \cdot \left(\frac{a_0^2}{2} + \sum_{n=1}^{N}(|a_n|^2 + |b_n|^2)\right) = ||\, s_N\,||^2 \leq ||f||^2|| \;\triangleq \int_{-L}^{L} |f|^2\, dx.$$

9.8.2 Complex Fourier Series

Recall that we defined an inner product on complex valued functions by (9.30) in Section 9.4, that is,

$$\langle f(x), g(x)\rangle \triangleq \int_{-L}^{L} f(x)\overline{g(x)}\, dx,$$

where $\bar{z}$ denotes the complex conjugate of z.

We note first that

$$\int_{-L}^{L} |f(x)|^2\, dx = \int_{-L}^{L} f(x)\,\overline{f(x)}\, dx = \langle f(x), f(x)\rangle$$

$$= \langle f_{\mathbb{C}}(x), f_{\mathbb{C}}(x) \rangle = \left\langle \sum_{n=-\infty}^{\infty} c_n e^{in\pi x/L}, \sum_{m=-\infty}^{\infty} c_m e^{im\pi x/L} \right\rangle$$

$$= \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} c_n \overline{c_m} \cdot \int_{-L}^{L} e^{i(n-m)\pi x/L} dx = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} c_n \overline{c_m} \cdot \begin{Bmatrix} 2L, & n = m \\ 0, & n \neq m \end{Bmatrix}.$$

The above calculations are not mathematically rigorous proof because we do not study the convergence of the infinite series and how that interacts with taking integrals.

We have the **Parseval identity**

$$\int_{-L}^{L} |f(x)|^2 \, dx = 2L \sum_{n=-\infty}^{\infty} |c_n|^2. \tag{9.87}$$

It follows that the **root mean square** (**RMS**) value of f is

$$f_{RMS} \triangleq \sqrt{\frac{1}{2L} \int_{-L}^{L} |f(x)|^2 \, dx} = \sqrt{\sum_{n=-\infty}^{\infty} |c_n|^2}. \tag{9.88}$$

9.8.3 Fourier Transforms

Theorem 9.22

(Parseval's Theorem) If $F(\omega) = \mathcal{F}[\, f(t) \,](\omega)$ and $G(\omega) = \mathcal{F}[\, g(t) \,](\omega)$ exist then

$$(a) \qquad \int_{-\infty}^{\infty} f(t) \, \overline{g(t)} \, dt = \int_{-\infty}^{\infty} F(\omega) \, \overline{G(\omega)} \, d\omega \tag{9.89}$$

and, in particular with $g = f$,

$$(b) \qquad \int_{-\infty}^{\infty} |f(t)|^2 \, dt = \int_{-\infty}^{\infty} |F(\omega)|^2 \, d\omega \tag{9.90}$$

Learn More About It

A derivation of Theorem 9.22 is in Section 1.3 of *Fourier Transforms*, by Ian N. Sneddon, Dover Publications., Inc., © 1995.

9.8.4 Problems

1. Explain why Theorem 9.18 is true, using work similar to what justified Theorem 9.17, or by applying Theorem 9.17 to the special case of a function that is odd on the interval $-L < x < L$.
2. Explain why Theorem 9.19 is true, using work similar to what justified Theorem 9.17, or by applying Theorem 9.17 to the special case of a function that is odd on the interval $-L < x < L$.
3. Find the power used by a resistor of $R\ \Omega$ by the periodic current

$$I(t) = \frac{1}{2} + \sum_{n=1}^{\infty} \left(e^{-n} \cos\left(\frac{n\pi t}{L}\right) + e^{-2n} \sin\left(\frac{n\pi t}{L}\right) \right) \text{ amps}$$

   over one period of time. Simplify your conclusion using, for example, $e^{-2n} = (e^{-2})^n$, and the sum of a geometric series,

$$\sum_{n=1}^{\infty} r^n = \frac{1}{1-r}, \text{ if } |r| < 1.$$

4. For $f(x) = x$,
   (a) use Parseval's Theorem for a Fourier sine series to establish the identity

$$\frac{\pi^2}{6} = \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \cdots.$$

   (b) If you use a different interval than the one you used in part (a), do you get the same identity? Why, or why not?
5. Establish an identity for an infinite series by using one of Parseval's Theorems applied to an example of a Fourier series, Fourier cosine series, or Fourier sine series found in an earlier section, either in the narrative or the problems. (Your example should not duplicate Example 9.37 or Problem 9.8.4.4.)
6. Establish a Parseval identity for a generalized Fourier series.
7. If you have read Section 2.10, "Hilbert Space," use Paresval's Identity and Definition 2.39 to explain why the sequence of functions $\{\frac{1}{\sqrt{\pi}} \sin(nx)\}_{n=1}^{\infty}$ converges weakly to the zero function in $L^2(-\pi, \pi)$, even though $||\frac{1}{\sqrt{\pi}} \sin(nx)|| = 1$ for all $n = 1, 2, \ldots$.

Key Terms

admissible: Example 9.35 in Section 9.7
aliasing: after (9.57) in Section 9.5
analysis equation: before (9.55) in Section 9.5
auto-correlation: (9.61) in Section 9.5
average value: after (9.8) in Section 9.1

Bessel's Inequality: Theorem 9.21 in Section 9.8
best approximation: Theorem 9.20 in Section 9.8
characteristic equation: before (9.19) in Section 9.3
clamped (boundary conditions): Example 9.33 in Section 9.6
convolution on the real line: before Theorem 9.10 in Section 9.4, Problem 17.1.2.8
digital filter: before (9.51) in Section 9.5
discrete convolution: (9.60) in Section 9.5
Discrete Fourier Transform (*DFT*): (9.50) in Section 9.5
eigenfunction, **eigenvalue**: after (9.17) in Section 9.3
even: before Theorem 9.2 in Section 9.1
extend periodically, **periodic extension**: before Example 9.7 in Section 9.1
Fast Fourier transform (*FFT***)**: after (9.61) in Section 9.5
finite Fourier transform: (9.49) in Section 9.5
finite jump discontinuity: Definition 9.1 in Section 9.1
finite pulse: Example 9.19 in Section 9.4
finite wave train: Example 9.21 in Section 9.4
fixed (boundary conditions): Example 9.33 in Section 9.6
Fourier coefficients: (9.1) in Section 9.1
Fourier cosine series, **Fourier cosine coefficients**: (9.12) in Section 9.2
Fourier series: (9.1) in Section 9.1
Fourier sine series, **Fourier sine coefficients**: (9.13) in Section 9.2
generalized Fourier series expansion: before (9.20) in Section 9.3
graphical method: after (9.27) in Section 9.3
Hadamard matrix: Example 9.24 in Section 9.5
half-range expansion: before (9.12) in Section 9.2
Hermitian conjugate: before (9.51) in Section 9.5
interface conditions: (9.75) in Section 9.6
match: Example 9.32 in Section 9.6
nice function: Definition 9.2 in Section 9.1
odd: before Theorem 9.2 in Section 9.1
Parseval identity: Theorems 9.17 through 9.19 in Section 9.8 and Theorem 9.22 in Section 9.8, and (9.87) in Section 9.8
partial sums: (9.3) in Section 9.1
piecewise smooth: after Definition 9.2 in Section 9.1
periodic boundary conditions: (9.28) in Section 9.3
power spectrum: Example 9.27 in Section 9.5
Rayleigh quotient: Definition 9.6 in Section 9.7
regular: Definition 9.5 in Section 9.6
root mean square (**RMS**): (9.88) in Section 9.8
singular Sturm Liouville problem: before (9.70) in Section 9.6
square integrable: before Theorem 9.20 in Section 9.8
square integrability condition: (9.19) in Section 9.3
square wave: Example 9.5 in Section 9.1
Sturm-Liouville problem: Definition 9.5 in Section 9.6
synthesis equation: (9.55) in Section 9.5
time series: before (9.50) in Section 9.5
trial function: Example 9.34 in Section 9.7
unitary matrix: Example 9.23 in Section 9.5

Mathematica Command

NMinimize[F[μ], μ]: Example 9.35 in Section 9.7

References

Cochran, W.T, Cooley, J.W et al. What is the fast Fourier transform? *IEEE Transactions on Audio Electroacousting* AU-15, 45–55, 1967.

Davis, P.J. *Circulant Matrices*. John Wiley & Sons, Inc., New York, 1979.

Friedman, B. *Lectures on Applications-Oriented Mathematics*, ed. by V. Twersky, Wiley Classics Library, 1991.

Merritt, F.S. et al. *Standard Book for Civil Engineers*. McGraw-Hill, New York, 1996.

Temple, G. and Bickley, W.G. *Rayleigh's Principle and Its Applications to Engineering*, Dover Publications Inc., Mineola, NY, 1956.

10

Partial Differential Equations Models

10.1 Integral and Partial Differential Equations

In this chapter, we will discuss a variety of physical problems that can be modeled by integral equations, typically over planar regions or spatial regions. We will see that Chapter 7's divergence theorem, Stokes' theorem, and change of variable formulas enable us to turn integral equations into partial differential equations (PDEs).

In Chapters 11, 16, and 17, we will learn methods for getting solution formulas for PDEs, and in Chapters 12 and 14, we will learn methods for numerical approximation of solutions of PDEs.

Our first physical applications are to flow of mass or heat energy.

Recall from Example 7.36 in Section 7.5 that

$$\oiint_{\mathcal{S}} \varrho\, \mathbf{v} \bullet \widehat{\mathbf{n}}\, dS$$

is the rate of mass flow out of $\mathcal{S}$. Here ϱ is the mass density, in units of mass per volume, $\mathbf{v}$ is the velocity vector field of the fluid; and $\mathcal{S}$ is a constant, that is, time-independent, closed, piecewise smooth, oriented surface enclosing a solid $\mathcal{V}$ having outward unit normal vector $\widehat{\mathbf{n}}$. We may refer briefly to such a situation as $\mathcal{S}$ **encloses a "control" volume** $\mathcal{V}$.

In a physical problem in which a substance can flow through the boundary of a control volume, the divergence theorem can relate that flow, which is an integral over the surface bounding the control volume, to an integral over the control volume.

Define $M = \iiint_{\mathcal{V}} \varrho\, dV$ to be the total mass of the fluid in the solid $\mathcal{V}$. In Example 7.39 in Section 7.6, we saw that

$$\oiint_{\mathcal{S}} \varrho\, \mathbf{v} \bullet \widehat{\mathbf{n}}\, dS = -\frac{dM}{dt} = -\iiint_{\mathcal{V}} \frac{\partial \varrho}{\partial t}\, dV.$$

The integral equation

$$\iiint_{\mathcal{V}} \frac{\partial \varrho}{\partial t}\, dV + \oiint_{\mathcal{S}} \varrho\, \mathbf{v} \bullet \widehat{\mathbf{n}}\, dS = 0 \tag{10.1}$$

expresses the physical law of conservation of mass.

In Example 7.39 in Section 7.6, we used the divergence theorem, Theorem 7.14 in Section 7.6, along with reasoning using continuity and the arbitrariness of $\mathcal{V}$, to see why (10.1) implies the differential equation (7.41) in Section 7.6, that is,

$$\frac{\partial \varrho}{\partial t} + \nabla \bullet (\varrho \mathbf{v}) = 0, \tag{10.2}$$

which likewise expresses the physical law of conservation of mass.

In this and later sections, we will see other physical situations where a physical law expressed as an integral equation can also be expressed as a PDE. In Chapter 11, we will learn a technique for solving many PDEs.

Example 10.1

Find an integral equation that expresses conservation of heat energy in a fixed control volume, and use that to establish a PDE that expresses the same law.

Method: Let $e = e(x, y, z, t)$ be the **density of heat energy** in a substance, in units of *energy/volume*, and let $\mathcal{S}$ enclose a control volume $\mathcal{V}$. Then, analogously to total mass, the total heat energy in $\mathcal{V}$ is

$$E \triangleq \iiint_{\mathcal{V}} e \, dV.$$

The **heat flux vector** $\mathbf{q} = \mathbf{q}(x, y, z, t)$ measures the time rate of heat flow, in units of *energy/(time·area)*. Figure 10.1 depicts heat flowing in or out of parts of a solid $\mathcal{V}$. Suppose $\mathcal{S}$ has outward unit normal vector $\widehat{\mathbf{n}}$. Conservation of heat energy says that

$$\frac{dE}{dt} = -\oiint_{\mathcal{S}} \mathbf{q}(\mathbf{r}, t) \bullet \widehat{\mathbf{n}} \, dS + \iiint_{\mathcal{V}} Q(\mathbf{r}, t) \, dV,$$

where $Q = Q(\mathbf{r}, t)$ is the rate of heat production, in units of *energy/(time · volume)*. It follows that

$$\iiint_{\mathcal{V}} \frac{\partial e}{\partial t} \, dV = -\oiint_{\mathcal{S}} \mathbf{q}(\mathbf{r}, t) \bullet \widehat{\mathbf{n}} \, dS + \iiint_{\mathcal{V}} Q(\mathbf{r}, t) \, dV. \tag{10.3}$$

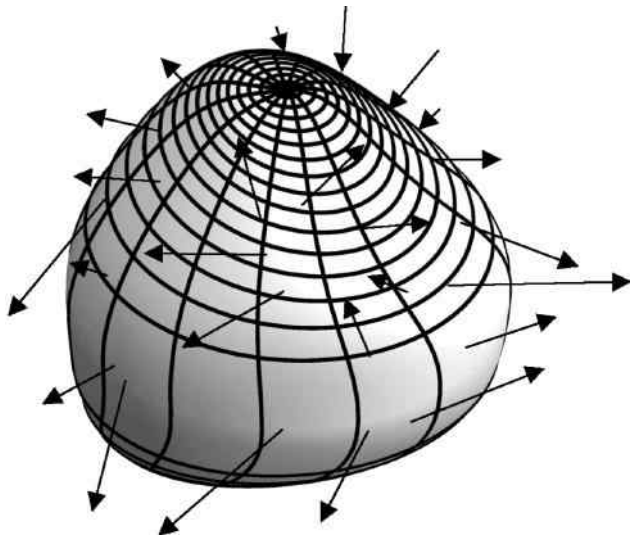


FIGURE 10.1
Heat flux vectors.

The divergence theorem enables us to rewrite this as the integral equation

$$\iiint_{\mathcal{V}} \frac{\partial e}{\partial t}\, dV = -\iiint_{\mathcal{V}} \nabla \bullet \mathbf{q}(\mathbf{r},t)\, dV + \iiint_{\mathcal{V}} Q(\mathbf{r},t)\, dV. \tag{10.4}$$

As in Example 7.39 in Section 7.6, reasoning using continuity and the arbitrariness of $\mathcal{V}$ explains why (10.4) implies the PDE

$$\frac{\partial e}{\partial t} = -\nabla \bullet \mathbf{q} + Q. \tag{10.5}$$

We have omitted the dependence on $(\mathbf{r},t)$ only so as to write the PDE in a compact form. ◯

Example 10.2

Derive the one space dimensional version of the heat balance equation (10.5) in a thin, homogeneous rod $0 \le x \le L$, $0 \le y^2 + z^2 \le R^2$, where R, L are constants.

Method: For an arbitrary control volume $\mathcal{V}$, we derived (10.3), that is,

$$\iiint_{\mathcal{V}} \frac{\partial e}{\partial t}\, dV = -\oiint_{\mathcal{S}} \mathbf{q}(\mathbf{r},t) \bullet \widehat{\mathbf{n}}\, dS + \iiint_{\mathcal{V}} Q(\mathbf{r},t)\, dV.$$

It is natural to rewrite an arbitrary cylindrical slice $\mathcal{V} = \{(x,y,z) : a < x < b, 0 \le y^2 + z^2 \le R^2\}$ in an alternative cylindrical coordinates as $\mathcal{V} = \{(x,\mathrm{r},\vartheta) : a < x < b, 0 \le \mathrm{r} < R, -\pi < \vartheta \le \pi\}$. The surface $\mathcal{S}$ that bounds $\mathcal{V}$ consists of three parts, similarly to Example 7.36 in Section 7.5: (1) $\mathcal{S}_-$, the disk $x = a$, $0 \le y^2 + z^2 \le R^2$, on which $\widehat{\mathbf{n}} = -\hat{\boldsymbol{\imath}}$; (2) $\mathcal{S}_+$, the disk $x = b$, $0 \le y^2 + z^2 \le R^2$, on which $\widehat{\mathbf{n}} = \hat{\boldsymbol{\imath}}$ and (3) $\tilde{\mathcal{S}}$, the lateral surface $a < x < b$, $\mathrm{r} = R$, $-\pi < \vartheta \le \pi$, on which $\widehat{\mathbf{n}} = \hat{\mathbf{e}}_\mathrm{r} = \cos\vartheta\, \hat{\boldsymbol{\imath}} + \sin\vartheta\, \hat{\boldsymbol{\jmath}}$. In these cylindrical coordinates, $\mathbf{q} = q_x\, \hat{\boldsymbol{\imath}} + q_\mathrm{r}\, \hat{\mathbf{e}}_\mathrm{r} + q_\vartheta\, \hat{\mathbf{e}}_\vartheta$.

So,

$$-\oiint_{\mathcal{S}} \mathbf{q}(\mathbf{r},t) \bullet \widehat{\mathbf{n}}\, dS = -\iint_{\mathcal{S}_-} \mathbf{q}(\mathbf{r},t) \bullet \widehat{\mathbf{n}}\, dS - \iint_{\mathcal{S}_+} \mathbf{q}(\mathbf{r},t) \bullet \widehat{\mathbf{n}}\, dS - \iint_{\tilde{\mathcal{S}}} \mathbf{q}(\mathbf{r},t) \bullet \widehat{\mathbf{n}}\, dS$$

$$= -\int_{-\pi}^{\pi}\int_0^R \mathbf{q}(a,\mathrm{r},\vartheta,t) \bullet (-\hat{\boldsymbol{\imath}})\mathrm{r}\, d\mathrm{r}\, d\vartheta - \int_{-\pi}^{\pi}\int_0^R \mathbf{q}(b,\mathrm{r},\vartheta,t) \bullet (\hat{\boldsymbol{\imath}})\mathrm{r}\, d\mathrm{r}\, d\vartheta - \int_a^b\int_{-\pi}^{\pi} \mathbf{q}(x,R,\vartheta,t) \bullet \hat{\mathbf{e}}_\mathrm{r}\, d\vartheta\, dx$$

So (10.3) is

$$\int_a^b\int_{-\pi}^{\pi}\int_0^R \frac{\partial e}{\partial t}(x,\mathrm{r},\vartheta,t)\mathrm{r}\, d\mathrm{r}\, d\vartheta\, dx = \int_{-\pi}^{\pi}\int_0^R \mathbf{q}(a,\mathrm{r},\vartheta,t) \bullet \hat{\boldsymbol{\imath}}\, \mathrm{r}\, d\mathrm{r}\, d\vartheta + \int_{-\pi}^{\pi}\int_0^R \mathbf{q}(b,\mathrm{r},\vartheta,t) \bullet (-\hat{\boldsymbol{\imath}})\mathrm{r}\, d\mathrm{r}\, d\vartheta$$

$$-\int_a^b\int_{-\pi}^{\pi} \mathbf{q}(x,R,\vartheta,t) \bullet \hat{\mathbf{e}}_\mathrm{r}\, d\vartheta\, dx + \int_a^b\int_{-\pi}^{\pi}\int_0^R Q(x,\mathrm{r},\vartheta,t)\mathrm{r}\, d\mathrm{r}\, d\vartheta\, dx$$

$$= \int_{-\pi}^{\pi}\int_0^R q_x(a,\mathrm{r},\vartheta,t)\mathrm{r}\, d\mathrm{r}\, d\vartheta - \int_{-\pi}^{\pi}\int_0^R q_x(b,\mathrm{r},\vartheta,t)\mathrm{r}\, d\mathrm{r}\, d\vartheta - \int_a^b\int_{-\pi}^{\pi} q_\mathrm{r}(x,R,\vartheta,t) d\vartheta\, dx$$

$$+ \int_a^b\int_{-\pi}^{\pi}\int_0^R Q(x,\mathrm{r},\vartheta,t)\mathrm{r}\, d\mathrm{r}\, d\vartheta\, dx.$$

If we assume that e, $\mathbf{q}$, and Q do not depend on r or ϑ and we assume that $q_r \equiv 0$, that is, there is no heat flux out of the rod at any point on the lateral surface of the rod, then we can integrate to get

$$\pi R^2 \int_a^b \frac{\partial e}{\partial t}(x,t)dx = -\pi R^2 \left(q_x(b,t) - q_x(a,t)\right) - 0 + \pi R^2 \int_a^b Q(x,t)\,dx.$$

The first two terms can be rewritten using

$$q_x(b,t) - q_x(a,t) = \int_a^b \frac{\partial q_x}{\partial x}(x,t)\,dx.$$

After dividing through by the constant πR^2 and moving terms to the left-hand side, we get

$$\int_a^b \left(\frac{\partial e}{\partial t}(x,t) + \frac{\partial q_x}{\partial x}(x,t) - Q(x,t)\right) dx = 0.$$

This being true for all a, b satisfying $0 < a < b < L$ implies the PDE

$$\frac{\partial e}{\partial t}(x,t) = -\frac{\partial q_x}{\partial x}(x,t) + Q(x,t),$$

the one-space-dimensional version of the heat balance equation. ○

Effectively, we used a "control interval" $a < x < b$ instead of a control volume $\mathcal{V}$.

10.1.1 Maxwell's Equations of Electromagnetism

The electromotive force $\mathcal{E}$ produced by an electric field $\mathbf{E}$ along a simple, closed, piecewise smooth curve $\mathcal{C}$ is defined by

$$\mathcal{E} = \oint_{\mathcal{C}} \mathbf{E} \bullet d\mathbf{r}. \tag{10.6}$$

In a physical problem in which a vector field is integrated over a closed, simple curve that is the boundary curve of a surface, Stokes' theorem can relate that line integral to an integral over the surface it bounds.

Michael Faraday observed that if an electric current is changing in time, it can induce a current in another conductor. Define c to be the speed of electromagnetic waves in a medium; for example, the speed of light in a vacuum is* $c = 2.99792458 \times 10^8$ m/s. Assume that, as in Stokes' theorem, $\mathcal{C}$ is a simple, closed, piecewise smooth boundary curve for $\mathcal{S}$, an oriented, piecewise smooth, parametrized surface, and assume that $\mathcal{C}$ and $\mathcal{S}$ are fixed, that is, independent of time. The Faraday–Maxwell's law of Induction states that $\mathcal{E}$ is related to the magnetic flux density $\mathbf{B}$ by

* From the National Institute of Standards and Technology, http://www.nist.gov/pml/div823/museum-timeline.cfm, by William B. Penzes, paraphrased, "On October 20, 1983, the meter was redefined to be the distance traveled by light in vacuum during a time interval of 1/299,792,458 of a second... The second is determined to an uncertainty, $U = 1$ part in 10^{14} by the Cesium clock."

$$\mathcal{E} = -\frac{1}{c}\frac{d}{dt}\left(\iint_{\mathcal{S}} \mathbf{B} \bullet d\mathbf{S}\right). \tag{10.7}$$

Combining (10.6) and (10.7) yields

$$-\frac{1}{c}\frac{d}{dt}\left(\iint_{\mathcal{S}} \mathbf{B} \bullet d\mathbf{S}\right) = \oint_{\mathcal{C}} \mathbf{E} \bullet d\mathbf{r}. \tag{10.8}$$

By Stokes' theorem

$$\oint_{\mathcal{C}} \mathbf{E} \bullet d\mathbf{r} = \iint_{\mathcal{S}} (\nabla \times \mathbf{E}) \bullet d\mathbf{S}. \tag{10.9}$$

Because $\mathcal{S}$ is fixed,

$$\frac{d}{dt}\left(\iint_{\mathcal{S}} \mathbf{B} \bullet d\mathbf{S}\right) = \iint_{\mathcal{S}} \frac{\partial \mathbf{B}}{\partial t} \bullet d\mathbf{S}$$

Putting the latter together with (10.8) and (10.9) gives one of Maxwell's integral equations of electromagnetism:

$$-\frac{1}{c}\iint_{\mathcal{S}} \frac{\partial \mathbf{B}}{\partial t} \bullet d\mathbf{S} = \iint_{\mathcal{S}} (\nabla \times \mathbf{E}) \bullet d\mathbf{S}. \tag{10.10}$$

To express the PDE version of the physical law in (10.10), rewrite it as

$$\iint_{\mathcal{S}} \left(\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} + \nabla \times \mathbf{E}\right) \bullet d\mathbf{S} = 0$$

and then reason that $\mathcal{S}$ is an *arbitrary*-oriented, piecewise smooth, parametrized surface; hence,

$$\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} + \nabla \times \mathbf{E} = 0,$$

that is,

$$\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} = -\nabla \times \mathbf{E}. \tag{10.11}$$

The units of $||\mathbf{E}||$ are V/m, and the units of $||\mathbf{B}||$ are $\mathrm{Wb/m^2}$, so the units of the two sides of (10.11) are

$$\frac{1}{\mathrm{s}} \cdot \frac{\mathrm{Wb}}{\mathrm{m}^2} \cdot \mathrm{m}^2 \text{ and } \frac{\mathrm{V}}{\mathrm{m}} \cdot \mathrm{m},$$

respectively. The equality in (10.11) thus implies that

$$1\,\text{Wb} = \frac{1\,\text{V}}{1\,\text{s}}.$$

The magnetomotive force $\mathcal{M}$ produced by a magnetic field $\mathbf{H}$ along a simple, closed, piecewise smooth curve $\mathcal{C}$ is defined by

$$\mathcal{M} = \oint_{\mathcal{C}} \mathbf{H} \bullet d\mathbf{r}. \tag{10.12}$$

Note that $\mathbf{H} = \mu\,\mathbf{B}$, where μ is the magnetic permeability of the medium, in units of henrys/m. Conceivably, μ could be a 3×3 matrix if the medium is anisotropic.

Ampère—Maxwell's Law states that $\mathcal{M}$ is related to the electric field density $\mathbf{D}$ and the electric current density $\mathbf{J}$ by

$$\mathcal{M} = \iint_{\mathcal{S}} \mathbf{J} \bullet d\mathbf{S} + \frac{d}{dt}\left(\iint_{\mathcal{S}} \mathbf{D} \bullet d\mathbf{S}\right), \tag{10.13}$$

where, as in Stokes' theorem, $\mathcal{C}$ is a simple, closed, piecewise smooth boundary curve for $\mathcal{S}$, an oriented, piecewise smooth, parametrized surface.

Note that $\mathbf{D} = \epsilon\,\mathbf{E}$, where ϵ is the electric permittivity* of the medium, in units of farads/m. Conceivably, ϵ could be a 3×3 matrix if the medium is anisotropic.

Combining (10.12) and (10.13), Ampère—Maxwell's law can be restated as

$$\iint_{\mathcal{S}} \mathbf{J} \bullet d\mathbf{S} + \frac{d}{dt}\left(\iint_{\mathcal{S}} \mathbf{D} \bullet d\mathbf{S}\right) = \oint_{\mathcal{C}} \mathbf{H} \bullet d\mathbf{r}. \tag{10.14}$$

In Problem 10.1.3.4, you will explain how to use (10.14) to get

$$\frac{\partial \mathbf{D}}{\partial t} = -\mathbf{J} + \boldsymbol{\nabla} \times \mathbf{H}. \tag{10.15}$$

Equations (10.11) and (10.15) are two of the four Maxwell's equations of electromagnetism, a crowning achievement of nineteenth century physics.

10.1.2 Continuum Mechanics

Here we will discuss the deformation of a "body" of matter. We will derive an integral equation of continuum mechanics by using both the divergence theorem and the change of variables formula for integration over a spatial region.

A **reference state** consists of particles at "material points" $\mathbf{X}$ in some initial, control volume $\mathcal{V}_0 = \mathcal{V}(0)$ in $\mathbb{R}^3$. The **Lagrangian description** says to follow the motion of each particle overtime. This motion is described by a function $\mathbf{x} = \boldsymbol{\phi}(\mathbf{X}, t)$ satisfying $\boldsymbol{\phi}(\mathbf{X}, 0) = \mathbf{X}$.

* From http://www.scienceworld.Wolfram.com, the permittivity of a vacuum is $\epsilon_0 \approx 8.8542 \times 10^{-12}$ F/m.

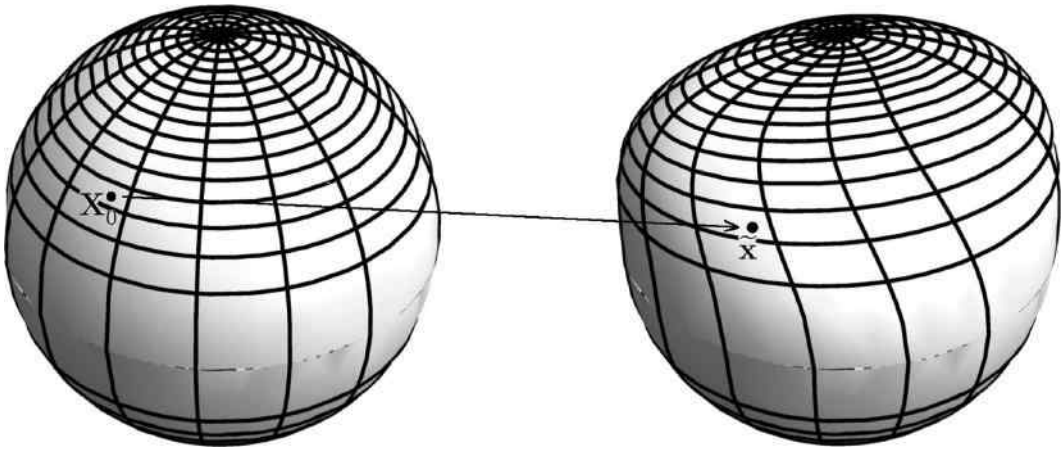


FIGURE 10.2
Deformation of a ball.

This notation means that the particle that begins at $\mathbf{X}$ is located at position $\boldsymbol{\phi}(\mathbf{X}, t)$ at time t. In this way, the material points in a control volume $\mathcal{V}_0$ move around to become the region

$$\mathcal{V}(t) \triangleq \{\boldsymbol{\phi}(\mathbf{X}, t) : \mathbf{X} \text{ in } \mathcal{V}_0\}$$

at time t. To understand the notation, it helps to have Figure 10.2's picture illustrating the concept involved.

We expect that the function $\boldsymbol{\phi}(\mathbf{X}, t)$ should depend smoothly on both $\mathbf{X}$ and t. For any fixed time $t > 0$, the map

$$\mathbf{X} \mapsto \boldsymbol{\phi}(\mathbf{X}, t)$$

is a transformation, possibly not linear. But, we expect that the map should be invertible, that is, for any point $\mathbf{x}$ in $\mathcal{V}(t)$, there should be *exactly one* initial point $\mathbf{X}$ such that

$$\mathbf{x} = \boldsymbol{\phi}(\mathbf{X}, t).$$

Also we expect that the inverse map

$$\mathbf{x} \mapsto \mathbf{X},$$

defines an inverse function

$$\mathbf{X} = \boldsymbol{\phi}^{-1}(\mathbf{x}, t)$$

that depends smoothly on both $\mathbf{x}$ and t.

We have already seen an example of such a situation in Section 6.8 where we "rigidly rotated" the points on or inside the planet Earth around an axis $\boldsymbol{\omega}$. By "rigidly," we mean that the distance between points is unchanged by the rotation. A particle that begins in the

planet at location $\mathbf{X}$ at time $t = 0$ will be at position $\boldsymbol{\phi}(\mathbf{X}, t) = e^{t\Omega}\,\mathbf{X}$ at time t. In this simple example, we can explicitly find the inverse map: it is given by

$$\boldsymbol{\phi}^{-1}(\mathbf{x}, t) = e^{-t\Omega}\,\mathbf{x}$$

because $\mathbf{x} = \boldsymbol{\phi}(\mathbf{X}, t) = e^{t\Omega}\,\mathbf{X}$ implies that

$$e^{-t\Omega}\,\mathbf{x} = e^{-t\Omega}(e^{t\Omega}\mathbf{X}) = (e^{-t\Omega}\,e^{t\Omega})\mathbf{X} = I\mathbf{X} = \mathbf{X}.$$

Define $\varrho_0 = \varrho_0(\mathbf{X})$ to be the mass density at point $\mathbf{X}$, at time $t = 0$, and define $\varrho(\mathbf{x}, t)$ to be the mass density at point $\mathbf{x}$, at time $t \geq 0$. Note that $\varrho(\mathbf{X}, 0) = \varrho_0(\mathbf{X})$.

Conservation of mass tells us that for $t \geq 0$,

$$\iiint_{\mathcal{V}(t)} \varrho(\mathbf{x}, t)\, d\mathbf{x} = \iiint_{\mathcal{V}_0} \varrho_0(\mathbf{X})\, d\mathbf{X}. \tag{10.16}$$

Here, $d\mathbf{x}$ is short for $dx_1 dx_2 dx_3$, and $d\mathbf{X}$ is short for $dX_1 dX_2 dX_3$.

Define the Jacobian determinant by

$$J(\mathbf{X}, t) = \det\left(\frac{\partial \boldsymbol{\phi}(\mathbf{X}, t)}{\partial \mathbf{X}}\right).$$

We assume $J(\mathbf{X}, t) \neq 0$ at all $(\mathbf{X}, t)$.

By the change of variables formula for the transformation $\mathbf{X} \mapsto \boldsymbol{\phi}(\mathbf{X}, t)$,

$$\iiint_{\mathcal{V}(t)} \varrho(\mathbf{x}, t)\, d\mathbf{x} = \iiint_{\mathcal{V}_0} \varrho(\boldsymbol{\phi}(\mathbf{X}, t), t)\, J(\mathbf{X}, t)\, d\mathbf{X}, \tag{10.17}$$

Why? Recall from Section 6.6 that the Jacobian measures how an element of volume, $d\mathbf{X}$, is transformed by a change of variables. The assumption that $J(\mathbf{X}, t) \neq 0$ at all $(\mathbf{X}, t)$ both guarantees that the element of volume $d\mathbf{x} = |J(\mathbf{X}, t)|\, d\mathbf{X}$ never vanishes and guarantees the smoothness and invertibility of the transformation $\mathbf{X} \mapsto \boldsymbol{\phi}(\mathbf{X}, t)$. Further,

$$J(\mathbf{X}, 0) = \det\left(\frac{\partial \boldsymbol{\phi}}{\partial \mathbf{X}}(\mathbf{X}, 0)\right) > 0$$

because for small $\triangle t$, $\boldsymbol{\phi}(\mathbf{X}, 0) \approx \mathbf{X} + \triangle t\, \dfrac{\partial \boldsymbol{\phi}}{\partial t}(\mathbf{X}, 0)$ implies $J(\mathbf{X}, 0) = \det(I) = 1 > 0$ and continuity in t implies $J(\mathbf{X}, t) > 0$ at all $(\mathbf{X}, t)$; hence, $|J(\mathbf{X}, t)| \equiv J(\mathbf{X}, t)$.

For example, for the rigid rotation $\boldsymbol{\phi}(\mathbf{X}, t) = e^{t\Omega}\,\mathbf{X}$, the Jacobian determinant is identically 1 or -1 because $\dfrac{\partial \boldsymbol{\phi}(\mathbf{X}, t)}{\partial \mathbf{X}} = e^{t\Omega}$ is a real, orthogonal matrix for all t.

Combining (10.16) and (10.17) gives

$$\iiint_{\mathcal{V}_0} \varrho_0(\mathbf{X})\, d\mathbf{X} = \iiint_{\mathcal{V}_0} \varrho(\boldsymbol{\phi}(\mathbf{X}, t), t)\, J(\mathbf{X}, t)\, d\mathbf{X}.$$

This being true for all control volumes $\mathcal{V}_0$ implies a PDE expression of conservation of mass that is given by

$$\varrho_0(\mathbf{X}) = \varrho(\boldsymbol{\phi}(\mathbf{X}, t), t)\, J(\mathbf{X}, t). \tag{10.18}$$

Suppose further that a nice surface $\mathcal{S}$ is part of the boundary of a solid $\mathcal{V}$ and $\widehat{\mathbf{n}}$ is the outward unit normal at a point $\mathbf{r}$ on $\mathcal{S}$. At a point $\mathbf{r}$ on $\mathcal{S}$, a **contact force** acting on the solid at a point on the bounding surface $\mathcal{S}$ will exert an element of force proportional to the surface area, that is,

$$d\mathbf{F} = \boldsymbol{\sigma}\, dS,$$

for some vector $\boldsymbol{\sigma}$ known as the **stress**, or **traction**, **vector**.

We expect stress to depend on the direction of the unit normal $\widehat{\mathbf{n}}$. The simplest such dependence is by a linear transformation, that is,

$$\boldsymbol{\sigma} = [\tau]\,\widehat{\mathbf{n}},$$

where the 3×3 matrix $[\tau]$ is called the **stress tensor**.

Recalling from Chapter 7 the notation $d\mathbf{S} = \widehat{\mathbf{n}}\, dS$, we see that

$$d\mathbf{F} = [\tau]\, d\mathbf{S}. \tag{10.19}$$

The **displacement**, or **deformation**, of the solid is defined by

$$\mathbf{U}(\mathbf{X}, t) \triangleq \mathbf{x} - \mathbf{X} = \boldsymbol{\phi}(\mathbf{X}, t) - \mathbf{X}.$$

Note that $\mathbf{U}(\mathbf{X}, 0) \equiv \mathbf{0}$. Define

$$\mathbf{u} = \mathbf{u}(\mathbf{x}, t) = \mathbf{U}(\boldsymbol{\phi}^{-1}(\mathbf{x}, t), t).$$

The **total linear momentum** in the volume $\mathcal{V}(t)$ is

$$\iiint\limits_{\mathcal{V}(t)} \varrho(\mathbf{x}, t)\, \frac{\partial \mathbf{u}(\mathbf{x}, t)}{\partial t}\, d\mathbf{x}.$$

Using the change of variables $\mathbf{X} \mapsto \boldsymbol{\phi}(\mathbf{X}, t)$ and conservation of mass,

$$\iiint\limits_{\mathcal{V}(t)} \varrho(\mathbf{x}, t)\, \frac{\partial \mathbf{u}(\mathbf{x}, t)}{\partial t}\, d\mathbf{x} = \iiint\limits_{\mathcal{V}_0} \varrho(\boldsymbol{\phi}(\mathbf{X}, t)\, \frac{\partial \mathbf{U}(\mathbf{X}, t)}{\partial t}\, J(\mathbf{X}, t)\, d\mathbf{X} = \iiint\limits_{\mathcal{V}_0} \varrho_0(\mathbf{X})\, \frac{\partial \mathbf{U}(\mathbf{X}, t)}{\partial t}\, d\mathbf{X}. \tag{10.20}$$

Newton's second law of motion states that

$$\begin{aligned} &\frac{d}{dt}[\text{total linear momentum in the volume } \mathcal{V}(t)] \\ &\quad = \text{sum of the forces acting on points in } \mathcal{V}(t) \\ &\qquad \text{and the forces in contact with the boundary of } \mathcal{V}(t). \end{aligned}$$

The **body forces**, $\mathbf{f}(\mathbf{x}, t)$, are the forces such as gravity that act on points inside $\mathcal{V}(t)$. So, using (10.20),

$$\frac{d}{dt}\left[\iiint_{\mathcal{V}_0} \varrho_0(\mathbf{X})\,\frac{\partial \mathbf{U}(\mathbf{X},t)}{\partial t}\,d\mathbf{X}\right] = \iiint_{\mathcal{V}(t)} \mathbf{f}(\mathbf{x},t)d\mathbf{x} + \oiint_{\partial\mathcal{V}(t)} d\mathbf{F}, \tag{10.21}$$

where $\partial\mathcal{V}(t)$ is the nice surface that encloses $\mathcal{V}(t)$. We want the integral on the left-hand side to be over $\mathcal{V}_0$ rather than over $\mathcal{V}(t)$ because it will make it easier later to take the derivative with respect to t.

The contact forces are given by (10.19), so the divergence theorem implies

$$\oiint_{\partial\mathcal{V}(t)} d\mathbf{F} = \oiint_{\partial\mathcal{V}(t)} [\tau]\widehat{\mathbf{n}}\,dS = \iiint_{\mathcal{V}(t)} \nabla \bullet [\tau]\,d\mathbf{x}, \tag{10.22}$$

where we write $[\tau]$ in terms of its rows

$$[\tau] \triangleq \begin{bmatrix} \boldsymbol{\tau}_1^T \\ \boldsymbol{\tau}_2^T \\ \boldsymbol{\tau}_3^T \end{bmatrix}$$

and then define

$$\nabla \bullet [\tau] \triangleq \begin{bmatrix} \nabla \bullet \boldsymbol{\tau}_1^T \\ \nabla \bullet \boldsymbol{\tau}_2^T \\ \nabla \bullet \boldsymbol{\tau}_3^T \end{bmatrix}.$$

Combining (10.21) and (10.22), we get

$$\frac{d}{dt}\left[\iiint_{\mathcal{V}_0} \varrho_0(\mathbf{X})\,\frac{\partial \mathbf{U}(\mathbf{X},t)}{\partial t}\,d\mathbf{X}\right] = \iiint_{\mathcal{V}(t)} \mathbf{f}(\mathbf{x},t)d\mathbf{x} + \iiint_{\mathcal{V}(t)} \nabla \bullet [\tau]\,d\mathbf{x}. \tag{10.23}$$

The left-hand side of (10.23) can be rewritten as

$$\frac{d}{dt}\left[\iiint_{\mathcal{V}_0} \varrho_0(\mathbf{X})\,\frac{\partial \mathbf{U}(\mathbf{X},t)}{\partial t}\,d\mathbf{X}\right] = \iiint_{\mathcal{V}_0} \frac{\partial}{\partial t}\left[\varrho_0(\mathbf{X})\,\frac{\partial \mathbf{U}(\mathbf{X},t)}{\partial t}\right] d\mathbf{X} = \iiint_{\mathcal{V}_0} \varrho_0(\mathbf{X})\,\frac{\partial^2 \mathbf{U}(\mathbf{X},t)}{\partial t^2}\,d\mathbf{X}$$

$$= \iiint_{\mathcal{V}_0} \varrho(\boldsymbol{\phi}(\mathbf{X},t),t)\,J(\mathbf{X},t)\,\frac{\partial^2 \mathbf{U}(\mathbf{X},t)}{\partial t^2}\,d\mathbf{X} = \iiint_{\mathcal{V}(t)} \varrho(\mathbf{x},t)\,\frac{\partial^2 \mathbf{u}(\mathbf{x},t)}{\partial t^2}\,d\mathbf{x}.$$

So, (10.23) is equivalent to

$$\iiint_{\mathcal{V}(t)} \varrho(\mathbf{x},t)\,\frac{\partial^2 \mathbf{u}(\mathbf{x},t)}{\partial t^2}\,d\mathbf{x} = \iiint_{\mathcal{V}(t)} \mathbf{f}(\mathbf{x},t)\,d\mathbf{x} + \iiint_{\mathcal{V}(t)} \nabla \bullet [\tau]\,d\mathbf{x}.$$

This being true for all $\mathcal{V}(t)$, the PDE version of Newton's second law of motion is the **Cauchy equation**

$$\varrho(\mathbf{x},t)\,\frac{\partial^2 \mathbf{u}(\mathbf{x},t)}{\partial t^2} = \mathbf{f}(\mathbf{x},t) + \nabla \bullet [\tau]. \tag{10.24}$$

By the way, conservation of angular momentum implies that the stress tensor is a symmetric matrix, that is, $[\tau]^T = [\tau]$.

Learn More About It

Our derivation of the PDEs of continuum mechanics follows the exposition in *Applied Solid Mechanics*, Peter Howell, Gregory Kozyreff, and Jim Ockendon, Cambridge University Press, 2009. There is much more in that book, including the derivation of the symmetry of the stress tensor.

10.1.3 Problems

In problems 1–3, as in Example 10.2, use integral equation (10.3) in $\mathbb{R}^3$ to

1. Derive the two-space-dimensional version of the heat balance equation (10.5) in a short, rectangular slab $\{(x,y,z) : 0 \le z \le H,\ 0 \le x \le a,\ 0 \le y \le b\}$, assuming the temperature and boundary conditions (BCs) in the usual rectangular coordinates do not depend on z.
2. Derive the one-space-dimensional version of the heat balance equation (10.5) in a short, cylindrical slab $\{(x,y,z) : 0 \le z \le H,\ 0 \le x^2 + y^2 \le R^2\}$, assuming the temperature and BCs in the usual cylindrical coordinates do not depend on z or θ. [Hint: In Example 10.2, we used an arbitrary cylindrical slice over $a < x < b$ because we assumed no dependence on r or ϑ. In this problem, the analogous arbitrary solid is a cylindrical shell $\{(r,\theta,z) : a < r < b, 0 \le \theta \le 2\pi, 0 < z < H\}$.]
3. Derive the one-space-dimensional version of the heat balance equation (10.5) in a thin, homogeneous rod assuming that the *lateral* side of the rod is not insulated but instead is losing heat according to Newton's law of cooling.
4. Use the Ampere–Maxwell's law, (10.14), and Stokes' theorem to derive (10.15).

10.2 Heat Equations

Recall from Section 10.1, specifically (10.5), that

$$\frac{\partial e}{\partial t} = -\nabla \bullet \mathbf{q} + Q,$$

where

$e = e(\mathbf{r}, t)$ is the density of heat energy in a substance, in units of *energy/volume*

the heat flux vector $\mathbf{q} = \mathbf{q}(\mathbf{r}, t)$ measures the time rate of heat flow, in units of *energy/(time · area)*

$Q = Q(\mathbf{r}, t)$ is the rate of heat production, in units of *energy/time*

Before we can get a solvable mathematical problem, we need further concepts. In particular, we will use Fourier's law of heat conduction, which relates the heat flux vector to the gradient of the temperature distribution. Such a relationship is an example of a "constitutive law."

Also in this section, we will see a special case when the physical situation involving heat energy is in equilibrium, that is, the temperature does not depend on time. We will have what is known as the "steady-state problem" whose solution is the "steady state temperature distribution."

Let $u = u(\mathbf{r}, t)$ be the **temperature**, in units of °C and let $c = c(\mathbf{r}, u)$ be the **specific heat**, in units of *energy/(mass* $\cdot$ °C). As usual, $\varrho = \varrho(\mathbf{r}, u)$ denotes the mass density, in units of *mass/volume*. If the specific heat does not depend* on the temperature, then

$$e = c\varrho u, \tag{10.25}$$

that is, the temperature is proportional to the amount of heat energy, and (10.4) in Section 10.1 becomes

$$\frac{\partial}{\partial t}[c\varrho u] = -\nabla \bullet \mathbf{q} + Q. \tag{10.26}$$

We need one more assumption in order to get PDEs we can solve in Chapter 11: We need a constitutive relation between the heat flux, $\mathbf{q}$, and temperature, u. The simplest form of such a relation is **Fourier's law of heat conduction**

$$\mathbf{q} = -\kappa \nabla u, \tag{10.27}$$

where the **thermal conductivity** $\kappa = \kappa(\mathbf{r}, u)$ either is a positive scalar or is a positive definite matrix all of whose entries are nonnegative. These sign conventions on κ guarantee that (10.27) implies that heat flows from hot spatial points to cool points. Note that if heat flows preferentially in some directions then κ would be a matrix other than a multiple of the identity matrix.

---

* If the specific heat does depend on the temperature, then $e(\mathbf{r}, u) = \int_0^u \varrho(\mathbf{r}, \xi)\, c(\mathbf{r}, \xi)\, d\xi$.

Substituting (10.27) into (10.26) gives

$$\frac{\partial}{\partial t}[c\varrho u] = \nabla \bullet (\kappa \nabla u) + Q. \tag{10.28}$$

A commonly used special case is when there is heat conduction in a homogeneous material; the latter means that c, ϱ, and κ are constants and κ is a scalar. In this case, (10.28) becomes

$$\frac{\partial u}{\partial t} = \alpha \nabla^2 u + \frac{Q}{c\varrho}, \tag{10.29}$$

where

$$\alpha = \frac{\kappa}{c\varrho}$$

is called the **thermal diffusivity**. If, further, there is no heat source or sink, then $Q \equiv 0$. We call

$$\frac{\partial u}{\partial t} = \alpha \nabla^2 u \tag{10.30}$$

the **heat equation***.

Recall that for a first-order ODE $\frac{dy}{dt} = f(t, y)$, there is an existence and uniqueness theorem that states that if an initial condition (IC) $y(t_0) = y_0$ is given, then there will be exactly one solution of the IVP (3.19) in Section 3.2, that is,

$$\left\{ \begin{array}{l} \frac{dy}{dt} = f(t, y) \\ \\ y(t_0) = y_0 \end{array} \right\},$$

as long as $f(t, y)$ satisfies a technical hypothesis such as its dependence on y being differentiable or satisfying a Lipschitz condition. (The latter will be discussed in Chapter 18.)

Similarly, in order to predict the temperature distribution $u = u(\mathbf{r}, t)$ later in time, we must know the initial temperature distribution, that is, the function

$$u(\mathbf{r}, t_0),$$

at least for $\mathbf{r}$ in some spatial region in which we want to solve the PDEs (10.29) or (10.30).

In addition, in order to solve the PDE in a spatial region $\mathcal{V}$, we need to know something about the temperature's or the heat flux's behavior on $\mathcal{V}$'s boundary surface $\mathcal{S}$. This makes sense because certainly those behaviors can affect the temperature inside $\mathcal{V}$. It turns out that, mathematically, it suffices to know either the temperature or the heat flux. We can specify the temperature on some part(s) of $\mathcal{S}$ and the heat flux on another part(s).

* Similarly, if c is the concentration of a substance, it may satisfy the heat equation, called also the **diffusion equation** in this context.

Here are some commonly used BCs than can be specified at a point $\mathbf{r}$ on the surface $\mathcal{S}$:

$$\begin{aligned} &\text{Dirichlet BC :} && \text{Specified temperature, } u(\mathbf{r},t) \\ &\text{Neumann BC :} && \text{Specified heat flux in normal direction, that is, } \mathbf{q}(\mathbf{r},t)\bullet\widehat{\mathbf{n}} \\ &\text{Robin BC :} && \text{Specified linear combination of } u(\mathbf{r},t) \text{ and } \mathbf{q}(\mathbf{r},t)\bullet\widehat{\mathbf{n}}. \end{aligned} \tag{10.31}$$

The Neumann BC specifies the heat flux in the normal direction because that is the component of flux that passes through the boundary; the tangential component of the flux just moves along, rather than through, the boundary, as we can see in (10.3) in Section 10.1.

Fourier's law of heat conduction, $\mathbf{q}=-\kappa\nabla u$, says that the component of flux in the normal direction is

$$\widehat{\mathbf{n}}\bullet\mathbf{q} = -\widehat{\mathbf{n}}\bullet(\kappa\nabla u) = -\kappa\frac{\partial u}{\partial n},$$

assuming κ is a scalar. So, a Neumann BC in that case can be stated as

$$\frac{\partial u}{\partial n} = g(\mathbf{r},t) \tag{10.32}$$

for some given function g.

The Robin BC comes from combining Newton's law of cooling, which states that the rate of heat loss is proportional to the difference between a hot object and a cool medium, with a Neumann BC. Suppose that a medium's temperature, $u_0(t)$, is given outside of $\mathcal{V}$ but near a point $\mathbf{r}$ on the boundary of $\mathcal{V}$. Then

$$\widehat{\mathbf{n}}\bullet\mathbf{q} = k(\mathbf{r})\,(u(\mathbf{r},t) - u_0(t)) \tag{10.33}$$

expresses the proportionality, where k is a constant. Similarly to what was said for a Neumann BC, we can rewrite (10.33) as

$$u(\mathbf{r},t) + h(\mathbf{r})\,\frac{\partial u}{\partial n} = g(\mathbf{r},t). \tag{10.34}$$

On different parts of the boundary, different BCs can be given. For example, if $\mathcal{V}$ is the box

$$\mathcal{V} = \{\,(x,y,z) :\, 0\le x\le a,\quad 0\le y\le b,\quad 0\le z\le c\,\},$$

then on the six boundary faces of $\mathcal{V}$, we might specify the BCs, for $0<t<\infty$,

$$\left\{\begin{array}{c} u(0,y,z,t) = u(a,y,z,t) = 0, \quad \text{for } 0<y<b,\ 0<z<c \\[2ex] \dfrac{\partial u}{\partial y}(x,0,z,t) = u(x,b,z,t) = 0, \quad \text{for } 0<x<a,\ 0<z<c \\[2ex] u(x,y,0,t) - h\dfrac{\partial u}{\partial z}(x,y,0,t) = 1,\ \dfrac{\partial u}{\partial z}(x,y,c,t) = 0, \quad \text{for } 0<x<a,\ 0<y<b \end{array}\right\}. \tag{10.35}$$

10.2.1 Steady-State Temperature

A **steady-state temperature**, also known as an **equilibrium temperature**, is one that is not changing in time. For example, if u satisfies the heat equation (10.30) in a solid $\mathcal{V}$ *and* has $\dfrac{\partial u}{\partial t} \equiv 0$, then

$$0 = \nabla^2 u \text{ in } \mathcal{V}, \tag{10.36}$$

after substituting $\dfrac{\partial u}{\partial t} \equiv 0$ and dividing through by the positive scalar α.

Equation (10.36) is called **Laplace's equation** or the **potential equation**. For both physical and mathematical reasons, closely related to (10.36) is **Poisson's equation**,

$$-\frac{Q(\mathbf{r})}{c\varrho} = \alpha\, \nabla^2 u \text{ in } \mathcal{V}.$$

This equation is satisfied by a steady-state temperature that satisfies (10.29) if $Q = Q(\mathbf{r})$ does not depend on time, t. It makes sense that we can't expect to have a steady-state temperature if there is a time varying source of heat.

In the context of heat equations, the concept of "steady state" does *not* allow for oscillation in time. This is different from the "steady-state solutions" we had in the study of ODEs. We are using the commonly used terminologies that, unfortunately, are not consistent with each other. Referring to "equilibrium temperature" rather than to "steady-state temperature" may help us avoid confusion.

10.2.2 Lower Dimensional Problems

In practical problems, it is often true that we can ignore one or more spatial variables and thus consider heat flow in one or two spatial dimensions instead of three.

Example 10.3

Suppose $\mathcal{V}$ is a thin cylindrical rod

$$\{(x, r, \vartheta) : 0 \le r < R, 0 \le \vartheta \le 2\pi, 0 \le x \le L\}$$

whose longitudinal axis is the x-axis. If the material properties c, ϱ, κ do not vary with r or ϑ *and* the longitudinal side

$$\{(x, r, \vartheta) : r = R, 0 \le \vartheta \le 2\pi, 0 \le x \le L\}$$

is kept perfectly insulated, then we may look for a solution of the heat equation (10.30), that is, $\dfrac{\partial u}{\partial t} = \alpha\, \nabla^2 u$, satisfying $\dfrac{\partial u}{\partial r} \equiv 0$ and $\dfrac{\partial u}{\partial \vartheta} \equiv 0$. Using the result and method of Example 10.2 in Section 10.1, (10.30) becomes

$$\frac{\partial u}{\partial t} = \alpha\, \frac{\partial^2 u}{\partial x^2},\ 0 < x < L,\ 0 < t < \infty. \tag{10.37}$$

Along with PDE (10.37), we would need to specify BCs on the rest of the boundary, namely, the disks $0 \le r < R$ and $-\pi < \vartheta \le \pi$ at $x = 0$ and $x = L$, again with no dependence on r or ϑ.

For example, if the ends $x=0$ and $x=L$ are perfectly insulated, $u = u(x,t)$ needs to satisfy the BCs

$$\frac{\partial u}{\partial x}(0,t) = \frac{\partial u}{\partial x}(L,t) = 0,\ 0 < t < \infty. \tag{10.38}$$

Along with the perfect insulation on the longitudinal side, (10.38) guarantees that no heat is flowing into or out of the rod, because at the left of the rod, the heat flux *out* of the rod is

$$\mathbf{q}(0,t) \bullet (-\hat{\boldsymbol{\imath}}) = \kappa \nabla u(0,t) = \kappa \frac{\partial u}{\partial x}(0,t) = 0,$$

and, similarly, at the right end the heat flux out of the rod is zero.

If the BCs on either of the disks $0 \le r < R$ and $-\pi < \vartheta \le \pi$ at $x=0$ and $x=L$ involve either r or ϑ dependence by the temperature, for example, a Dirichlet BC specifying a nonconstant temperature dependence

$$u|_{x=0} = f(r,\vartheta),$$

then we should not expect the solution to be independent of r and ϑ, even if inside the thin rod, the material properties c, ϱ, κ do not depend on r or ϑ.

Example 10.4

Suppose a thin, longitudinally insulated homogeneous rod has its left end kept at 20°C and heat is leaving the right end at a rate of $100\,\mathrm{W/m^2}$. Assume the mass density of the material in the rod is $\varrho\,\mathrm{kg/m^3}$, the specific heat is c J/(kg ·°C), and the thermal conductivity is κ W/(m ·°C). Also, assume that heat is being generated inside the rod at a constant rate of $Q\,\mathrm{W/m^3}$. Find the equilibrium temperature distribution.

Method: Because the longitudinal side is insulated, the material is homogeneous, and the source term and the BCs do not depend on r, ϑ, or t, we need to look for an equilibrium temperature $u = u(x)$ that solves

$$0 = \alpha \frac{d^2u}{dx^2} + \frac{Q}{c\varrho}$$

and $u(0) = 20$, as well as a BC at $x = L$. The latter has

$$\mathbf{q}(L) \bullet \hat{\boldsymbol{\imath}} = 100\,\mathrm{W/m^2}.$$

As is often the case in physical problems, it helps to keep track of the units.

At $x = L$, Fourier's law of heat conduction gives

$$\mathbf{q}(L) = -\left(\kappa\ \frac{\mathrm{W}}{\mathrm{m}\cdot{}^\circ\mathrm{C}}\right) \cdot \left(\frac{du}{dx}(L)\ \frac{{}^\circ\mathrm{C}}{\mathrm{m}}\right)\hat{\boldsymbol{\imath}};$$

hence,

$$100\,\mathrm{W/m^2} = -\left(\kappa\ \frac{\mathrm{W}}{\mathrm{m}\cdot{}^\circ\mathrm{C}}\right) \cdot \left(\frac{du}{dx}(L)\ \frac{{}^\circ\mathrm{C}}{\mathrm{m}}\right).$$

So, the appropriate BC at the right end is $\dfrac{du}{dx}(L) = -\dfrac{100}{\kappa}$, in units of °C/m. Note that κ is constant because we assumed that the rod's material is homogeneous.

Because $\alpha = \dfrac{\kappa}{c\varrho}$, the ODE to be satisfied by $u(x)$ can be rewritten as

$$u'' = -\frac{Q}{\kappa}.$$

The right-hand side is a constant, so indefinite integration with respect to x of both sides gives

$$u' = -\frac{Qx}{\kappa} + c_1,$$

where c_1 is an arbitrary constant. Using the BC at $x = L$, we get

$$-\frac{100}{\kappa} = u'(L) = -\frac{QL}{\kappa} + c_1;$$

hence, $c_1 = (QL - 100)/\kappa$, and so

$$u' = \frac{1}{\kappa}(-Qx + QL - 100).$$

A second indefinite integration with respect to x of both sides gives

$$u = \frac{1}{\kappa}\left(-\frac{Q}{2}x^2 + (QL - 100)x\right) + c_2,$$

where c_2 is an arbitrary constant. The BC at the left end requires

$$20 = u(0) = 0 + c_2.$$

So, the equilibrium temperature distribution is

$$u(x) = 20 + \frac{1}{\kappa}\left(-\frac{Q}{2}x^2 + (QL - 100)x\right). \bigcirc$$

It turns out that $u'(0) > 0$; hence, there is a positive heat flux *out* of the left end of the rod, not just the right end. The BC at $x = L$ *required* positive heat flux out of the right end of the rod. Positive heat flux out of the left end of the rod is a result of the combined effects of the ODE and BCs.

The units of both $\alpha\dfrac{d^2u}{dx^2}$ and $\dfrac{Q}{c\varrho}$ are °C/s. Also, note that the radius of the thin rod is not needed to solve this problem.

For a thin solid rod of copper, the physical parameters include $\kappa \approx 390\,\mathrm{W/(m\cdot{}^\circ C)}$. Supposing also that we have more information, specifically that $Q = 60\,\mathrm{W/m^3}$ and $L = 2\,\mathrm{m}$, then the equilibrium temperature distribution is

$$u(x) \approx 20 + 0.002564\left(-30x^2 + 20x\right),\ 0 < x < 2,$$

which is graphed in Figure 10.3. Note that $u'(2) \approx -0.2564$ at the right end.

Example 10.5

The region between concentric spheres of radii 3 and 6 cm is filled with a homogeneous material whose thermal conductivity is constant $\kappa = 50\,\mathrm{W/(m\cdot{}^\circ C)}$ and is generating heat at a rate of $960\,\mathrm{w/m^3}$. Suppose the inside of the shell, that is, the ball of radius 3 cm,

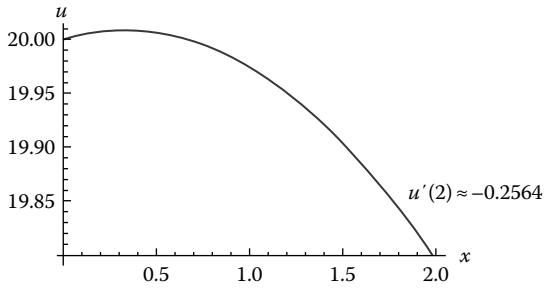


FIGURE 10.3
Temperature distribution.

consists of a perfect thermal insulator and the outer boundary of the shell is kept at a constant temperature of 10°C. Find the steady-state temperature distribution between the spheres.

Method: Let $u = u(\rho)$ be the steady-state temperature; it does not depend on ϕ, θ because the material is homogeneous and the BCs $\dfrac{\partial u}{\partial n}(0.03) = 0$ and $u(0.06) = 10$, do not depend on ϕ, θ, or t. On the inner sphere, the outward unit normal vector is $\widehat{\mathbf{n}} = -\hat{\mathbf{e}}_\rho$, so $\dfrac{\partial u}{\partial n} = -\dfrac{\partial u}{\partial \rho}$ there.

The steady-state temperature satisfies the PDE $0 = \kappa\, \nabla^2 u + Q$. Using the Laplacian in spherical coordinates given at the end of Section 6.7, the ODE BVP we need to solve is

$$0 = 50\, \frac{1}{\rho^2}\, \frac{d}{d\rho}\left[\rho^2 \frac{du}{d\rho}\right] + 960, \quad \frac{du}{d\rho}(0.03) = 0, \quad u(0.06) = 10. \tag{10.39}$$

Rewrite the ODE in (10.39) as

$$\frac{d}{d\rho}\left[\rho^2 \frac{du}{d\rho}\right] = -19.2\, \rho^2,$$

and then integrate once with respect to ρ to get

$$\rho^2 \frac{du}{d\rho} = -6.4\rho^3 + c, \tag{10.40}$$

where c is an arbitrary constant. At this point, we may substitute in the inner BC to get

$$0 = 0.03^2\, \frac{du}{d\rho}(0.03) = -6.4(0.03^3) + c;$$

hence, $c = 1.728 \times 10^{-4}$. Substitute this into (10.40) to get $\rho^2 \dfrac{du}{d\rho} = -6.4\rho^3 + 1.728 \times 10^{-4}$. Divide through by ρ^2 to get

$$\frac{du}{d\rho} = -6.4\rho + 1.728 \times 10^{-4} \rho^{-2},$$

and then integrate with respect to ρ to get

$$u(\rho) = -3.2\rho^2 - 1.728 \times 10^{-4} \rho^{-1} + c_2,$$

where c_2 is an arbitrary constant. We substitute in the outer BC to get

$$10 = u(0.06) = -3.2(0.06^2) - 1.728 \times 10^{-4}(0.06^{-1}) + c_2.$$

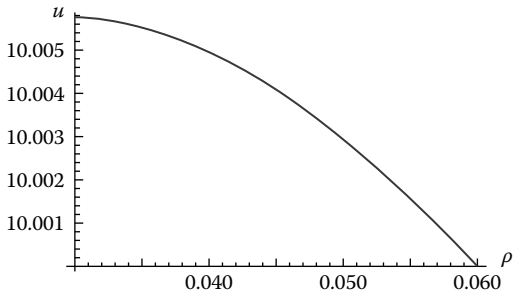


FIGURE 10.4
Temperature distribution between concentric spheres.

This gives

$$c_2 = 10 + 1.152 \times 10^{-2} + 2.88 \times 10^{-3} = 10.0144.$$

The equilibrium temperature distribution is

$$u(\rho) = 10.0144 - 3.2\rho^2 - 1.728 \times 10^{-4}\rho^{-1}$$

and is graphed in Figure 10.4. ◯

Example 10.6

Set up, but do not solve, a PDE and BCs for the steady-state temperature distribution in a thin homogeneous plate $0 \le x \le a, 0 \le y \le b$. Assume that there is no heat source or sink inside the plate, and the top and bottom flat surfaces of the plate are insulated. Assume that in the xy-plane, the "vertical" sides $x = 0$ and $x = a$ are insulated, the bottom horizontal side is kept at $0°C$, and the top horizontal side has temperature distribution $u = ax - x^2$.

Method: Let $u = u(x, y)$ be the steady-state temperature; it does not depend on z because the material is homogeneous, the plate is thin, and the top and bottom surfaces are insulated. The steady-state temperature satisfies Laplace's equation; because of the lack of z dependence, this becomes $0 = \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2}$. For two of the BCs, we are given that the sides $x = 0, 0 < y < b$, and $x = a, 0 < y < b$, are insulated. There, $\hat{\mathbf{n}} = \pm\hat{\boldsymbol{\imath}}$; hence, those BCs are $\dfrac{\partial u}{\partial x}(0, y) = \dfrac{\partial u}{\partial x}(a, y) = 0$. On the other two sides, the temperature is specified. Putting everything together, the steady-state temperature should satisfy the problem consisting of PDE, BC on each of the vertical sides of a rectangle, and BC on each of the horizontal sides of a rectangle:

$$\left\{ \begin{array}{ll} 0 = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, & 0 < x < a,\ 0 < y < b \\ \frac{\partial u}{\partial x}(0, y) = \frac{\partial u}{\partial x}(a, y) = 0, & 0 < y < b \\ u(x, 0) = 0,\ u(x, b) = ax - x^2, & 0 < x < a. \end{array} \right\} \cdot \bigcirc$$

Figure 10.5 summarizes the conclusions of Example 10.6.

We will learn how to use Fourier series to solve this problem in Section 11.3.

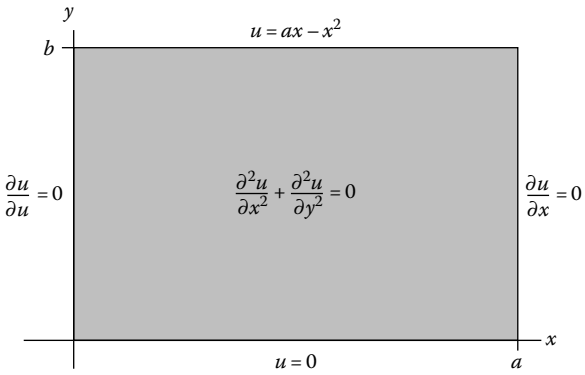


FIGURE 10.5
Boundary value problem.

10.2.3 Composite Rod

Suppose a thin homogeneous rod for $0 \le x \le L$, as in Example 10.3, consists of two different materials joined at the point $x = 1$, where $1 < L$. Then each of the material properties c, ϱ, κ may be two different constants depending on whether $x < 1$ or $x > 1$. Generally we will assume that the temperature and the heat flux vector should be continuous at $x = 1$, that is,

$$u(1^-, t) = u(1^+, t) \quad \text{and} \quad \mathbf{q}(1^-, t) = \mathbf{q}(1^+, t).$$

Because of the thin rod assumption, we assume that $\mathbf{q}(x, t) = q_x \hat{\boldsymbol{\imath}}$. Further, by using Fourier's law, we have

$$\mathbf{q}(x, t) = -\kappa(x) \frac{\partial u}{\partial x}(x, t) \hat{\boldsymbol{\imath}}.$$

So, continuity of heat flux requires

$$\lim_{x \to 1^-} \left(\kappa(x) \frac{\partial u}{\partial x}(x, t) \right) = \lim_{x \to 1^+} \left(\kappa(x) \frac{\partial u}{\partial x}(x, t) \right),$$

that is,

$$\kappa(1^-) \frac{\partial u}{\partial x}(1^-, t) = \kappa(1^+) \frac{\partial u}{\partial x}(1^+, t).$$

Yet another interesting situation is the case of **contact resistance** when two pieces of the *same* material are not perfectly joined at $x = 1$. In this case, the BCs usually assumed are continuity of heat flux and the dependence of the heat flux on the temperature jump at the interface:

$$\frac{\partial u}{\partial x}(1^-, t) = \frac{\partial u}{\partial x}(1^+, t) \quad \text{and} \quad u(1^+, t) - u(1^-, t) = h \frac{\partial u}{\partial x}(1, t).$$

10.2.4 Problems

1. Solve the ODE BVP

$$\left\{\begin{array}{c} \dfrac{1}{r}\dfrac{d}{dr}\left[r\dfrac{du}{dr}\right]=-1,\ a<r<b \\ \\ u(a)=T_0,\ u'(b)=0 \end{array}\right\},$$

   where $0 < a < b$ and T_0 are constants.

2. Find the equilibrium temperature for

$$\left\{\begin{array}{c} \dfrac{\partial u}{\partial t}=\dfrac{\partial^2 u}{\partial x^2}+x,\ 0<x<L, 0<t<\infty \\ \\ \dfrac{\partial u}{\partial x}(0,t)-2u(0,t)=0, u(L,t)=T_1, 0<t<\infty \\ \\ u(x,0)=f(x),\ 0<x<L \end{array}\right\}$$

   in terms of the unspecified constants T_1 and L.

3. Find the equilibrium temperature for the problem

$$\left\{\begin{array}{c} \dfrac{\partial u}{\partial t}=\dfrac{\partial^2 u}{\partial x^2}-\eta(u(x,t)-\bar{T}), 0<x<L, 0<t<\infty \\ \\ u(0,t)=T_0, u(L,t)=T_1, 0<t<\infty \\ \\ u(x,0)=f(x),\ 0<x<L \end{array}\right\}$$

   in terms of the unspecified positive constants $\eta, \bar{T}, T_0, T_1$, and L.

4. A thin copper rod is 0.3 m long. Assume the lateral surface of the rod is perfectly insulated. Assuming there is no internal heat source or sink, what is the equilibrium temperature distribution in the rod if the left end is kept at temperature T_0 and the right end is kept at temperature $2T_0$?

5. Consider the problem of finding the equilibrium temperature distribution in a thin, homogeneous rod with internal heat source proportional to the temperature modeled by $\dfrac{\partial u}{\partial t}=\dfrac{\partial^2 u}{\partial x^2}+4u,\ 0<x<L$. Assume the lateral surface of the rod is perfectly insulated and the rod is insulated at the right end and is kept at a constant nonzero temperature at the left end. For which value(s), if any, of L, is there (a) exactly one solution? (b) infinitely many solutions, and (c) no solution?

6. Suppose the equilibrium temperature distribution $u = u(x)$ in a thin, homogeneous rod of length 1 satisfies

$$\left\{\begin{array}{c} 0=u''+\pi^2 u \\ u'(0)=A,\ u'(1)=B \end{array}\right\},$$

where A, B are unspecified constants. Find a condition(s) on A and B that guarantees that there is an equilibrium temperature distribution, and then find it.

7. A thin copper rod has length L. Assume the lateral surface of the rod is perfectly insulated. What is the equilibrium temperature distribution in the rod if the left end is kept insulated, the right end is kept at temperature T_1, and there is an internal heat source $Q(x) = \cos\left(\frac{\pi x}{L}\right)$? Assume that the thermal conductivity, κ, is constant.
8. Suppose a thin rod made of a homogeneous material has constant cross section and length L, and suppose there is a heat source $Q(x) = g''(x)$ for some function g. [Do not specify g or try to solve for it.] Assume the lateral surface of the rod is perfectly insulated. Suppose the rod is in thermal equilibrium and the temperatures at the left and right ends are constants T_0, T_1, respectively. (a) Find the equilibrium temperature distribution in the rod, in terms of T_0, T_1, L, and $g(x)$, and (b) find the rate of heat flow out of the rod at the left end, in terms of T_0, T_1, L, and $g'(x)$.
9. Describe the physical situation modeled by the problem

$$\frac{\partial u}{\partial t} = \frac{1}{\rho^2}\frac{\partial}{\partial \rho}\left[\rho^2 \frac{\partial u}{\partial \rho}\right] + \frac{1}{\rho^2 \sin\phi}\frac{\partial}{\partial \phi}\left[\sin\phi \frac{\partial u}{\partial \phi}\right] - Q_0,$$

$$\frac{\pi}{2} \le \phi \le \pi, 0 \le \rho < a, 0 < t < \infty$$

$$\frac{\partial u}{\partial \rho}(a, \phi, t) = 0, \ \frac{\pi}{2} \le \phi \le \pi, 0 < t < \infty$$

$$\frac{1}{\rho}\frac{\partial u}{\partial \phi}\left(\rho, \frac{\pi}{2}, t\right) = -\left(u\left(\rho, \frac{\pi}{2}, t\right) - 20\right), \quad 0 \le \rho < a, 0 < t < \infty$$

$$u(\rho, \phi, 0) = 40, \ \tfrac{\pi}{2} \le \phi \le \pi, 0 \le \rho < a.$$

In particular, in what physical region is this taking place?

10. Write down mathematical equations to model the physical situation of a double-walled cylindrical rod of length L whose cross section is shown in Figure 10.6. The inner, unshaded region contains a solid metal in which heat is being conducted. The outer, shaded region contains a different metal whose inner wall dissipates

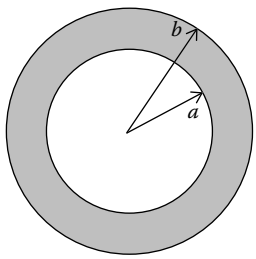


FIGURE 10.6
Problem 10.2.4.10.

heat from the inner solid and whose outer wall dissipates heat to the surrounding medium, both according to Newton's law of cooling. The $z = 0$ end of the rod is kept at a constant temperature T_0, and the end $z = L$ is insulated. Assume that the material properties of the two metals are different.

11. Regarding remarks after Example 10.4, recall that the component of flux in the normal direction is $\hat{\mathbf{n}} \bullet \mathbf{q} = -\kappa \dfrac{\partial u}{\partial n}$, if κ is a scalar. Suppose that the object is a thin, homogeneous rod. Note that at the left end of the rod, $\hat{\mathbf{n}} = -\hat{\boldsymbol{\imath}}$; at the right end of the rod, $\hat{\mathbf{n}} = \hat{\boldsymbol{\imath}}$. So, the normal derivative at the *left* end of the rod is $\dfrac{\partial u}{\partial n} = -\dfrac{\partial u}{\partial x}$. Explain why (a) heat flux out of the rod at $x = 0$ is positive if, and only if, $\dfrac{\partial u}{\partial x}\Big|_{at\ x=0} > 0$, and (b) heat flux out of the right end of the rod at $x = L$ is positive if, and only if, $\dfrac{\partial u}{\partial x}\Big|_{at\ x=L} < 0$.
12. Suppose the temperature is in equilibrium in a thin annular plate $a < r < b$ made of a homogeneous material in which there is no heat generation or absorption inside. Suppose the annulus outer circle is kept at temperature 0°C. and the inner circle is kept at temperature $T_0 \triangleq u(a)$. If the temperature distribution in the plate does not depend on θ, find a formula for the total heat flow out of the plate through the inner circle, in terms of the unspecified constants a, b, T_0 and the unspecified constant thermal conductivity κ.
13. Suppose the temperature is in equilibrium in a thin annular plate $a < r < b$ made of a homogeneous material in which there is no heat generation or absorption inside. Suppose the annulus inner circle is kept at temperature 0°C, and β is the total rate of heat flow out of the plate through the outer circle. If the temperature distribution in the plate does not depend on θ, find a formula for the constant temperature $T \triangleq u(b)$, in terms of the unspecified constants a, b, β and the unspecified constant thermal conductivity κ.
14. Assume a homogeneous ball has radius a in m, the spherical surface is kept at 20°C, and inside the ball heat is generated at a rate of $1000\,\mathrm{W/m^3}$. Find (a) the steady-state temperature distribution, and (b) the rate at which heat is flowing out of the ball. Your conclusions should be in terms of the unspecified parameter a.
15. Three pieces of wire, with thermal conductivities $\kappa_1, \kappa_2, \kappa_3$, respectively, are joined in perfect thermal contact at the point $(x, y) = (0, 0)$, as shown in Figure 10.7.

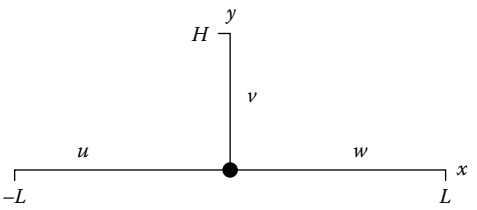


FIGURE 10.7
Three wires in perfect contact.

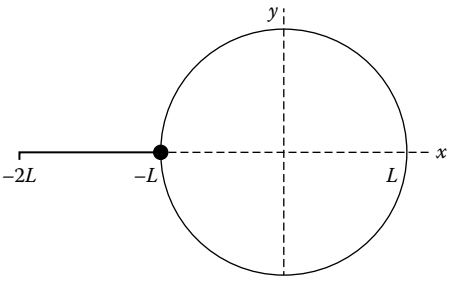


FIGURE 10.8
Line and circle in perfect contact.

Let $u(x,t)$, $-L < x < 0$, $v(y,t)$, $0 < y < H$, and $w(x,t)$, $0 < x < L$ be the temperature distributions in the three pieces of wire. Write down all of the BCs needed to model perfect thermal contact at $(0,0)$.

16. Two pieces of wire, with thermal conductivities κ_1, κ_2, respectively, are joined in perfect thermal contact at the point $(x,y) = (-L,0)$, as shown in Figure 10.8. The first piece of wire is a line segment, and the second piece is a circle. Let $u(x,t)$, $-2L < x < 2L$, and $v(\theta,t)$, $-\pi < \theta < \pi$ be the temperature distributions in the two pieces of wire. Write down all of the BCs needed to model perfect thermal contact at $(-L,0)$.
17. Suppose the sectors,

$$\mathcal{D}_1 = \left\{(r,\theta) : 0 < r < a, 0 < \theta < \frac{\pi}{4}\right\},$$

$$\mathcal{D}_2 = \left\{(r,\theta) : 0 < r < a, \frac{\pi}{4} < \theta < \frac{\pi}{2}\right\},$$

with thermal conductivities κ_1, κ_2, respectively, are joined in perfect thermal contact. Write down all of the BCs needed to model perfect thermal contact, assuming the temperature in $\mathcal{D}_i$ is $u_i = u_i(r,\theta,t)$, for $i = 1,2$.

In problems 18–20, the temperature of a thin, homogeneous rod of constant cross-sectional area A satisfies the problem

$$\left\{\begin{array}{c} \dfrac{\partial u}{\partial t} = \alpha \dfrac{\partial^2 u}{\partial x^2} + \dfrac{1}{c\varrho} Q(x),\ 0 < x < L,\ 0 < t \\ \\ \dfrac{\partial u}{\partial x}(0,t) = \dfrac{\partial u}{\partial x}(L,t) = 0 \end{array}\right\},$$

where α is a constant. Define the total energy in the rod by $E = c\varrho A \int_0^L u(x,t)\,dx$. Then $E = E(t)$, that is, E may be a function of t.

18. If $Q(x) \equiv 0$, explain why $E(t)$ is constant. [Hint: What ODE does $E(t)$ satisfy?]

19. If $Q(x) \equiv Q_0$ is constant and $E(t) = E(0) + \beta t$ for some constant β, find Q_0 in terms of the other constants mentioned in the problem. [Hint: What ODE does $E(t)$ satisfy?]
20. If $\bar{Q} \triangleq \frac{1}{L}\int_0^L Q(x)\,dx$ is the average value of Q and $E(t) = E(0)+\beta t$ for some constant β, find $\bar{Q}$ in terms of the other constants mentioned in the problem. [Hint: What ODE does $E(t)$ satisfy?]
21. Suppose that $U(x,t)$ solves the problem

$$\left\{\begin{array}{l} \dfrac{\partial U}{\partial t} = \dfrac{\partial^2 U}{\partial x^2} - U,\ 0 < x < \pi,\ t > 0, \\ \\ U(0,t) = U(\pi,t) = 0,\ t > 0, \\ \\ U(x,0) = e^x f(x),\ 0 < x < \pi, \end{array}\right\}$$

and $T(x,t) \triangleq e^{-x}U(x,t)$.

(a) Explain why $T(x,t)$ will solve the PDE BVP–IVP

$$\left\{\begin{array}{l} \dfrac{\partial T}{\partial t} = \dfrac{\partial^2 T}{\partial x^2} + 2\dfrac{\partial T}{\partial x}, 0 < x < \pi, t > 0 \\ \\ T(0,t) = T(\pi,t) = 0,\ t > 0, \\ \\ T(x,0) = f(x),\ 0 < x < \pi. \end{array}\right\}. \tag{10.41}$$

(b) Use this to solve (10.41) for the initial condition $T(x,0) = \sin\left(\dfrac{\pi x}{L}\right),\ 0 < x < \pi$.

22. Suppose $u = u(x,t)$ satisfies the PDE-BVP

$$\left\{\begin{array}{l} \dfrac{\partial u}{\partial t} = \alpha \dfrac{\partial^2 u}{\partial x^2},\ 0 < x < L, 0 < t \\ \\ \dfrac{\partial u}{\partial x}(0,t) = u(L,t) = 0,\ 0 < t \end{array}\right\}.$$

Define $G(t) \triangleq \int_0^L \frac{1}{2}((u(x,t))^2\,dx$. Explain why $\dot{G}(t) \leq 0$ for $0 < t$.

23. Suppose the temperature is in equilibrium in a spherical shell $a < \rho < b$ made of a homogeneous material. Suppose the shell's inner sphere is kept at temperature 20°C. and β is the total rate of heat flow out of the shell through the outer sphere $\rho = b$. If the temperature distribution in the shell is spherically symmetric, that is, does not depend on angles ϕ or θ, find a formula for the constant temperature $T \triangleq u(b)$, in terms of the unspecified constants a, b, β and the unspecified constant thermal conductivity κ.

10.3 Potential Equations

As we saw, a steady-state temperature may satisfy Laplace's equation

$$\nabla^2 u = 0 \tag{10.42}$$

or Poisson's equation

$$\nabla^2 u = -\frac{Q}{\kappa}. \tag{10.43}$$

In fact, many static or steady physical problems are solvable using Laplace's or Poisson's equation.

For example, suppose that the fluid is flowing in a region $\mathcal{V}$ and the fluid velocity is

$$\mathbf{v} = u\,\hat{\boldsymbol{\imath}} + v\,\hat{\boldsymbol{\jmath}} + w\,\hat{\boldsymbol{k}}.$$

Potential flow assumes that the vector field $\mathbf{v}$ is exact, that is, that

$$\mathbf{v} = \nabla\Phi \tag{10.44}$$

for some scalar potential function Φ. If the fluid is incompressible, then

$$\nabla \bullet \mathbf{v} = 0. \tag{10.45}$$

In that case, substitute (10.44) into (10.45) to get

$$0 = \nabla \bullet \mathbf{v} = \nabla \bullet (\nabla\Phi) = \nabla^2\Phi,$$

so Φ would satisfy Laplace's equation. This example of fluid flow is one of many reasons why Laplace's equation is referred to as a potential equation.

By the way, if we had assumed that $\mathbf{v}$ is **irrotational**, that is, that the **vorticity**

$$\boldsymbol{\omega} \triangleq \nabla \times \mathbf{v}$$

is the zero vector *and* that $\mathbf{v}$ is continuously differentiable in $\mathcal{V}$, then Theorem 6.3 in Section 6.4 would guarantee that $\mathbf{v}$ is exact and thus that there is a potential function Φ.

In addition, in this section, we will learn about some physically significant mathematical properties of solutions of Laplace's and Poisson's equation.

10.3.1 Magnetostatics

Another source of potential equation models is magnetostatics. If the magnetic flux density, $\mathbf{B}$, does not depend on time, t, then a consequence of Faraday–Maxwell's law, specifically (10.11) in Section 10.1, implies that

$$\nabla \times \mathbf{E} = \mathbf{0}. \tag{10.46}$$

Analogously to the case of irrotational fluid flow, if $\mathbf{E}$ is continuously differentiable in a region $\mathcal{V}$, then (10.46) would imply that there exists a "scalar" potential function Φ, that is, a function satisfying

$$\nabla\Phi = -\mathbf{E}. \tag{10.47}$$

In addition, another of Maxwell's laws relates the electric field to the electric charge density, ϱ:

$$\nabla \bullet \mathbf{E} = 4\pi\varrho. \tag{10.48}$$

If we substitute (10.47) into (10.48), then we see that Φ satisfies Poisson's equation because

$$-4\pi\varrho = \nabla \bullet (-\mathbf{E}) = \nabla \bullet (\nabla\Phi),$$

which yields

$$\nabla^2\Phi = -4\pi\varrho.$$

10.3.2 Boundary Conditions

As for the steady-state heat equation, it makes sense that what happens on the boundary of a region can affect what happens inside. This is true whether the region is in $\mathbb{R}^3$, or $\mathbb{R}^2$, or even $\mathbb{R}^1$ if symmetry, or homogeneity and thinness, allows us to look for a solution in a lower number of spatial dimensions.* As for heat problems, for Laplace's or Poisson's equation, everywhere on the boundary, we should specify a BC. Again, we may specify different types of BCs on different parts of the boundary, as in Example 10.5 in Section 10.2.

10.3.3 Properties of Solutions

Example 10.7

Suppose $u = u(\mathbf{r})$ satisfies Laplace's equation at all $\mathbf{r}$ in a region $\mathcal{V}$ whose boundary, denoted by $\partial\mathcal{V}$, is a simple, closed, piecewise smooth, positively oriented surface $\mathcal{S}$. Explain why

$$\oiint_{\mathcal{S}} \frac{\partial u}{\partial n}\, dS = 0. \tag{10.49}$$

Method: By the divergence theorem,

$$\oiint_{\mathcal{S}} \frac{\partial u}{\partial n}\, dS = \oiint_{\mathcal{S}} (\nabla u) \bullet \hat{\mathbf{n}}\, dS = \iiint_{\mathcal{V}} \nabla \bullet (\nabla u)\, dV = \iiint_{\mathcal{V}} \nabla^2 u\, dV = \iiint_{\mathcal{V}} 0\, dV = 0. \bigcirc$$

* Historically, such lower dimensional problems were often amenable to special mathematical techniques. Even today, computational methods that would take too much processing time in three space dimensions may give conclusions for lower dimensional versions of a physical problem.

Example 10.8

Suppose $u = u(\mathbf{r})$ satisfies the Neumann problem for Laplace's equation, that is,

$$\left\{ \begin{array}{ll} \nabla^2 u(\mathbf{r}) = 0, & \mathbf{r} \text{ in } \mathcal{V} \\ \frac{\partial u}{\partial n} = g(\mathbf{r}), & \mathbf{r} \text{ in } \mathcal{S} \end{array} \right\},$$

where the nice surface $\mathcal{S}$ encloses $\mathcal{V}$. Then g must satisfy the **solvability condition**

$$\oiint_{\mathcal{S}} g(\mathbf{r}) dS = 0, \tag{10.50}$$

that is, the average value of g on $\mathcal{S}$ must be zero.

The results of both Examples 10.7 and 10.8 make sense in the case where u, the solution of Laplace's equation, is the steady-state temperature distribution. Laplace's equation models a steady-state temperature when (a) the material is homogeneous and thus the thermal conductivity κ is constant in $\mathcal{V}$ and (b) there is no heat sink or source in $\mathcal{V}$. In this case,

$$\kappa \oiint_{\mathcal{S}} \frac{\partial u}{\partial n} dS$$

measures the total heat flux out of $\mathcal{V}$ through $\mathcal{S}$. To have a steady-state temperature distribution and no heat sink or source, the total heat flux out of $\mathcal{V}$ must be zero.

The total heat flux can be zero if heat flux out of one part of $\mathcal{S}$ is balanced by heat flux into another part of $\mathcal{S}$. So, the solvability condition does *not* imply that g is identically zero.

Example 10.9

Suppose $u = u(\mathbf{r})$ satisfies the problem

$$\left\{ \begin{array}{ll} \nabla^2 u(\mathbf{r}) = 0, & \mathbf{r} \text{ in the half disk } \mathcal{D} = \{(r,\theta) : 0 \le r < a, 0 < \theta < \pi\} \\ \frac{\partial u}{\partial \theta}(r,0) = g_1(r), & 0 < r < a \\ \frac{\partial u}{\partial r}(a,\theta) = g_2(\theta), & 0 < \theta < \pi \\ \frac{\partial u}{\partial \theta}(r,\pi) = g_3(r), & 0 < r < a \end{array} \right\}.$$

Find the solvability condition in terms of the functions g_1, g_2, g_3.

Method: The positively oriented boundary of $\mathcal{D}$ consists of three curves, as depicted in Figure 10.9. Parametrize the curves by

$$\mathcal{C}_1 : \mathbf{r}_1 = r\hat{\mathbf{e}}_r\big|_{\theta=0},\ 0 \le r \le a\,.$$

$$\mathcal{C}_2 : \mathbf{r}_2 = a\hat{\mathbf{e}}_r,\ 0 \le \theta \le \pi\,.$$

$$\mathcal{C}_3 : \mathbf{r}_3 = (a-t)\hat{\mathbf{e}}_r\big|_{\theta=\pi},\ 0 \le t \le a.$$

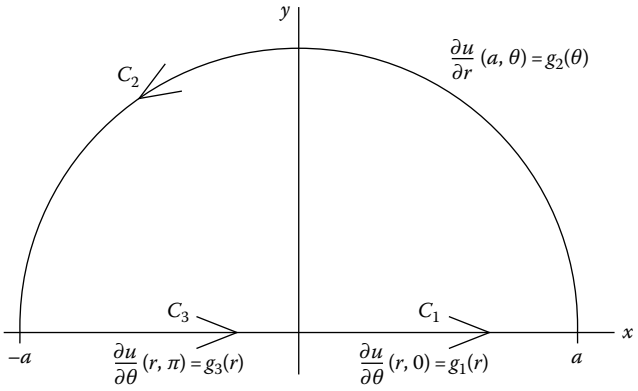


FIGURE 10.9
Example 10.9.

Similar to Example 10.8's solvability condition (10.49), a solution of a Neumann problem for Laplace's equation in two space dimensions must satisfy the solvability condition

$$0 = \oint_{\mathcal{C}} \frac{\partial u}{\partial n}\, ds.$$

We calculate

$$0 = \oint_{\mathcal{C}} \frac{\partial u}{\partial n}\, ds = \int_{\mathcal{C}_1} \frac{\partial u}{\partial n}(r,0)\, ds + \int_{\mathcal{C}_2} \frac{\partial u}{\partial n}(a,\theta)\, ds + \int_{\mathcal{C}_3} \frac{\partial u}{\partial n}(a-t,\pi)\, ds.$$

In polar coordinates, that is, cylindrical coordinates without z, $\nabla u = \dfrac{\partial u}{\partial r}\,\hat{\mathbf{e}}_r + \dfrac{1}{r}\dfrac{\partial u}{\partial \theta}\,\hat{\mathbf{e}}_\theta$.

On $\mathcal{C}_1$, the *outward* unit normal vector is $\widehat{\mathbf{n}} = -\hat{\mathbf{e}}_\theta$, so $\dfrac{\partial u}{\partial n} = -\dfrac{1}{r}\dfrac{\partial u}{\partial \theta}$. On $\mathcal{C}_1$, $ds = \left|\left|\dfrac{d\mathbf{r}}{dr}\right|\right| dr = \left|\left|\dfrac{d}{dr}[r\hat{\mathbf{e}}_r]\right|\right| dr = \left|\left|\hat{\mathbf{e}}_r\right|\right| dr = dr$.

On $\mathcal{C}_2$, $\widehat{\mathbf{n}} = \hat{\mathbf{e}}_r$, so $\dfrac{\partial u}{\partial n} = \dfrac{\partial u}{\partial r}$ and $ds = \left|\left|\dfrac{d\mathbf{r}}{d\theta}\right|\right| dr = \left|\left|\dfrac{d}{d\theta}\left[a\hat{\mathbf{e}}_r\right]\right|\right| d\theta = \left|\left|a\hat{\mathbf{e}}_\theta\right|\right| d\theta = a\, d\theta$.

On $\mathcal{C}_3$, $\widehat{\mathbf{n}} = \hat{\mathbf{e}}_\theta$, so $\dfrac{\partial u}{\partial n} = \dfrac{1}{r}\dfrac{\partial u}{\partial \theta}$ and $ds = \left|\left|\dfrac{d\mathbf{r}}{dt}\right|\right| dt = \left|\left|\dfrac{d}{dt}\left[(a-t)\hat{\mathbf{e}}_r\right]\right|\right| dt = ||-\hat{\mathbf{e}}_r\Big|\Big| dt = dt$.

The solvability condition is

$$0 = -\int_0^a \frac{1}{r}\frac{\partial u}{\partial \theta}(r,0)\, dr + a\int_0^\pi \frac{\partial u}{\partial r}(a,\theta)\, d\theta + \int_0^a \frac{1}{a-t}\frac{\partial u}{\partial \theta}(a-t,\pi)\, dt.$$

In the latter term, make the change of variables $r = a - t$ to get $\int_0^a \dfrac{1}{a-t}\dfrac{\partial u}{\partial \theta}(a-t,\pi)\, dt = \int_a^0 \dfrac{1}{r}\dfrac{\partial u}{\partial \theta}(r,\pi)(-dr) = \int_0^a \dfrac{1}{r}\dfrac{\partial u}{\partial \theta}(r,\pi)\, dr$.

The solvability condition is

$$0 = -\int_0^a \frac{1}{r} g_1(r)\, dr + a\int_0^\pi g_2(\theta)\, d\theta + \int_0^a \frac{1}{r} g_3(r)\, dr,$$

that is, the solvability condition is

$$\int_0^a \frac{1}{r}\, g_1(r)\, dr - \int_0^a \frac{1}{r}\, g_3(r)\, dr = a \int_0^\pi g_2(\theta)\, d\theta. \bigcirc$$

Theorem 10.1

(Minimum and maximum principles) Suppose u satisfies the Dirichlet problem for Laplace's equation in an open region $\mathcal{V}$ in $\mathbb{R}^3$ enclosed by a nice surface $\partial\mathcal{V}$, that is,

$$\left\{ \begin{array}{ll} \nabla^2 u(\mathbf{r}) = 0, & \mathbf{r} \text{ in } \mathcal{V} \\ \\ u = g(\mathbf{r}), & \mathbf{r} \text{ in } \partial\mathcal{V} \end{array} \right\}.$$

Suppose that g is continuous on $\partial\mathcal{V}$, and denote by g_{min} and g_{max} the minimum and maximum values of g on $\partial\mathcal{V}$. Then

(a) $g_{min} \leq u(\mathbf{r}) \leq g_{max}$ for all $\mathbf{r}$ inside $\mathcal{V}$.
(b) If $g(\mathbf{r}_0) > g_{min}$ for at least one $\mathbf{r}_0$ in $\partial\mathcal{V}$, then $u(\mathbf{r}) > g_{min}$ for all $\mathbf{r}$ in $\mathcal{V}$.
(c) If $g(\mathbf{r}_0) < g_{max}$ for at least one $\mathbf{r}_0$ in $\partial\mathcal{V}$, then $u(\mathbf{r}) < g_{max}$ for all $\mathbf{r}$ in $\mathcal{V}$.

Results in $\mathbb{R}^2$ are analogous, with $\mathcal{V}$ replaced by $\mathcal{D}$ and nice surface $\partial\mathcal{V}$ replaced by nice curve $\mathcal{C}$.

Theorem 10.2

(Poisson's formula for a disk) Suppose u satisfies the Dirichlet problem for Laplace's equation in the disk $\mathcal{D} = \{(r,\theta) : 0 \leq r < a, -\pi < \theta \leq \pi\}$, that is,

$$\left\{ \begin{array}{ll} \nabla^2 u(\mathbf{r}) = 0, & \mathbf{r} \text{ in disk } \mathcal{D} \\ \\ u(a,\theta) = g(\theta), & -\pi < \theta \leq \pi \end{array} \right\}.$$

Suppose that $g(\theta)$ is continuous and periodic with period 2π. Then

$$u(r,\theta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{a^2 - r^2}{a^2 + r^2 - 2ar\cos(\theta - \vartheta)}\, g(\vartheta)\, d\vartheta = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{||\mathbf{R}||^2 - ||\mathbf{r}||^2}{||\mathbf{R} - \mathbf{r}||^2}\, g(\vartheta)\, d\vartheta,$$

where $\mathbf{R} = \mathbf{R}(\vartheta) = a\left(\cos\vartheta\, \hat{\boldsymbol{\imath}} + \sin\vartheta\, \hat{\boldsymbol{\jmath}}\right)$, $-\pi \leq \vartheta \leq \pi$, parametrizes the circle $||\mathbf{R}|| = a$ that bounds the disk and let $\mathbf{r} = \mathbf{r}(r,\vartheta) = r\left(\cos\theta\, \hat{\boldsymbol{\imath}} + \sin\theta\, \hat{\boldsymbol{\jmath}}\right)$ be any point inside the disk, that is, has $||\mathbf{r}|| < a$.

So, Poisson's formula gives a formula for the values of u inside the disk in terms of the boundary values of u on the circle.

In particular, at the center of the disk, that is, at $\mathbf{r} = \mathbf{0}$,

$$u|_{origin} = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(\vartheta)\, d\vartheta,$$

that is, $u(\mathbf{0})$ equals the average of the boundary value of u.

This result generalizes to $\mathbb{R}^3$.

Theorem 10.3

Suppose u satisfies the Dirichlet problem for Laplace's equation in a ball $\mathcal{V} = \{\mathbf{r}: ||\mathbf{r} - \mathbf{r}_0|| < a\}$, that is,

$$\left\{ \begin{array}{ll} \nabla^2 u(\mathbf{r}) = 0, & \mathbf{r} \text{ in ball } \mathcal{V} \\ u(\mathbf{r}) = g(\mathbf{r}), & \text{on } ||\mathbf{r} - \mathbf{r}_0|| = a \end{array} \right\}.$$

Suppose that $g(\mathbf{r})$ is continuous on the sphere $\{\mathbf{r} : ||\mathbf{r} - \mathbf{r}_0|| = a\}$. Then

$$u(\mathbf{r}_0) = \frac{1}{4\pi a^2} \oiint_{\{\mathbf{r}: ||\mathbf{r}-\mathbf{r}_0||=a\}} g(\mathbf{r})\, dS,$$

that is, the value of u at the center of the ball is the average of the boundary values of u on the sphere.

It is also true* that if u satisfies the Laplace's equation in the ball $\{\mathbf{r} : ||\mathbf{r} - \mathbf{r}_0|| < a\}$ and is continuous on the closed ball $\{\mathbf{r} : ||\mathbf{r} - \mathbf{r}_0|| \le a\}$, then

$$u(\mathbf{r}_0) = \frac{1}{\pi a^3} \iiint_{\{\mathbf{r}: ||\mathbf{r}-\mathbf{r}_0||\le a\}} u(\mathbf{r})\, dV.$$

10.3.4 Problems

1. Suppose $u = u(x)$ solves the 1D Laplace's equation $u''(x) = 0$ for $0 < x < L$. Explain why $u'(L) - u'(0) = 0$. [The latter is the 1D version of the solvability condition for the Neumann problem.]
2. Suppose $u = u(x)$ solves a 1D Poisson's equation $u''(x) = -Q(x)$ for $0 < x < L$. Find an equation satisfied by $u'(0), u'(L)$, and $\int_0^L Q(x)\, dx$.
3. A coaxial cable has inner radius 1.3 cm and outer radius 2.5 cm. Suppose the electric potential is a constant V_0 on the surface $r = 1.3$, is a constant V_1 on the surface

* Nice explanations for this and for Theorem 10.3 are in Section 2.2.2 of Evans (1988).

$r = 2.5$, and satisfies Laplace's equation and is circularly symmetric between those two surfaces. Find the potential between the two surfaces.

4. For the problem

$$\left\{ \begin{array}{c} \nabla^2 u = 0,\ 0 < x < L,\ 0 < y < H \\ \\ \dfrac{\partial u}{\partial x}(0,y) = 0,\ \dfrac{\partial u}{\partial x}(L,y) = g(y),\ 0 < y < H \\ \\ \dfrac{\partial u}{\partial y}(x,0) = f(x),\ \dfrac{\partial u}{\partial y}(x,H) = 0,\ 0 < x < L \end{array} \right\},$$

find the solvability condition in terms of the unspecified functions f, g. Express your final conclusion in terms of $\int_0^L f(x)\,dx$ and $\int_0^H g(y)\,dy$.

5. For what value of the constant k does the problem

$$\left\{ \begin{array}{c} \nabla^2 u = 0,\ 0 < x < L,\ 0 < y < H \\ \\ \dfrac{\partial u}{\partial x}(0,y) = y,\ \dfrac{\partial u}{\partial x}(L,y) = 0,\ 0 < y < H \\ \\ \dfrac{\partial u}{\partial y}(x,0) = 0,\ \dfrac{\partial u}{\partial y}(x,H) = k,\ 0 < x < L \end{array} \right\}$$

have a solution? Express your final conclusion for k in terms of L and H.

6. Suppose u satisfies the Neumann problem for Laplace's equation in the half ball $\{(\rho,\phi,\theta) : 0 \le \rho < a, 0 \le \phi < \frac{\pi}{2}, -\pi < \theta \le \pi\}$. Write down the solvability condition in as much detail as possible, including *use of* specific formulas for the outward unit normal vector and the gradient in spherical coordinates.

7. Suppose $\nabla^2 u = 0$ in the region $\mathcal{D}$ that lies between the rectangle and the circle $r = a$ drawn in Figure 10.10 and suppose the boundary of the rectangle is insulated except on the upper horizontal line segment $\{(x,y) : -L < x < L, y = H\}$. Find the

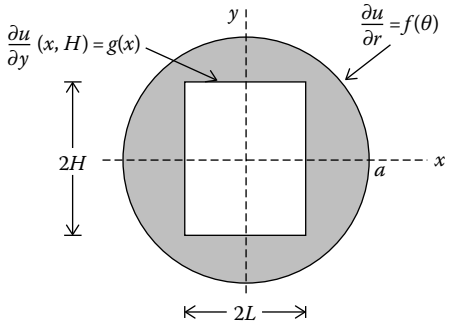


FIGURE 10.10
Problem 10.3.4.7.

solvability condition in terms of the unspecified functions f, g. Express your final conclusion in terms of $\int_{-\pi}^{\pi} f(\theta)\,d\theta$ and $\int_{-L}^{L} g(x)\,dx$.

8. Suppose the temperature is in equilibrium in a thin annular plate $a < r < b$ made of a homogeneous material with no internal heat source or sink. If the temperature $u = u(r, \theta)$ satisfies

$$b\,\frac{\partial u}{\partial r}(b, \theta) \equiv a\,\frac{\partial u}{\partial r}(a, \theta),$$

does it satisfy the solvability condition?

9. For

$$\left\{ \begin{array}{c} \dfrac{\partial u}{\partial t} = \dfrac{1}{r}\dfrac{\partial}{\partial r}\left[r\,\dfrac{\partial u}{\partial r}\right] - 4,\ 0 \le r < a \\ \\ |\,u(0^+, t)\,| < \infty,\ \dfrac{\partial u}{\partial r}(a, t) = 5 \end{array} \right\},$$

for what value of "a" is there an equilibrium temperature distribution? (b) for all such value(s) of a, find the equilibrium solution(s) and confirm that the solution satisfies the solvability condition.

10. For what value(s) of "a" is there an equilibrium temperature distribution for

$$\left\{ \begin{array}{c} \dfrac{\partial u}{\partial t} = \dfrac{1}{r}\dfrac{\partial}{\partial r}\left[r\,\dfrac{\partial u}{\partial r}\right] - \dfrac{4u}{r^2},\ 0 \le r < a \\ \\ |\,u(0^+, t)\,| < \infty,\ \dfrac{\partial u}{\partial r}(a, t) = 5 \end{array} \right\}?$$

For all such value(s) of a, find the equilibrium solution and confirm that the solution satisfies the solvability condition.

11. For the Poisson's equation in Problem 10.2.4.7, (a) find the solvability condition, and (b) use your result from part (a) to find the rate of heat flow out of the right end of the homogeneous rod.

12. Suppose $u = u(\mathbf{r})$ satisfies Poisson's equation $\kappa \nabla^2 u = -Q$ in a solid $\mathcal{V}$ for some unspecified function $Q(\mathbf{r})$. Explain why

$$\iint_{\mathcal{S}} \frac{\partial u}{\partial n}\,dS + \iiint_{\mathcal{V}} \frac{Q}{\kappa}\,dV = 0,$$

where κ is constant and the positively oriented, piecewise smooth, parametrized surface $\mathcal{S}$ encloses $\mathcal{V}$.

13. Suppose $u = u(\mathbf{r})$ satisfies the Neumann problem for an anisotropic variant of Laplace's equation, that is,

$$\left\{ \begin{array}{ll} \nabla \bullet (\kappa(\mathbf{r})\nabla u) = 0, & \mathbf{r} \text{ in } \mathcal{V} \\ \\ \widehat{\mathbf{n}} \bullet (\kappa(\mathbf{r})\nabla u) = g(\mathbf{r}), & \mathbf{r} \text{ in } \mathcal{S} \end{array} \right\},$$

where the solid $\mathcal{V}$ is enclosed by a positively oriented, piecewise smooth, parametrized surface $\mathcal{S}$. Find the solvability condition to be satisfied by u.

14. Suppose $u = u(x, y, z)$ satisfies Poisson's equation $\kappa \nabla^2 u = -Q(x, y, z)$ in a solid $\mathcal{V}$ for some unspecified function Q, where κ is constant. Is

$$\iiint_{\mathcal{V}} \nabla u \bullet \nabla u \, dV = \iint_{\mathcal{S}} u \frac{\partial u}{\partial n} dS + \iiint_{\mathcal{V}} \frac{uQ}{\kappa} dV,$$

where the positively oriented, piecewise smooth, parametrized surface $\mathcal{S}$ encloses $\mathcal{V}$? If so, why? If not, why not? [Hints: Multiply the PDE through by u and then use Corollary 6.1 in Section 6.7.]

15. Suppose $u = u(r, \theta)$ satisfies Laplace's equation in the *exterior* of a disk, specifically for $a < r < \infty$. Define a change of variables by $p = \frac{1}{r}$ and let $U(p, \theta) = u\left(\frac{1}{p}, \theta\right)$. Explain why U satisfies the PDE

$$\frac{\partial^2 U}{\partial p^2} + \frac{1}{p}\frac{\partial U}{\partial p} + \frac{1}{p^2}\frac{\partial^2 U}{\partial \theta^2} = 0.$$

10.4 Wave Equations

The simplest **wave equation** is, in one space dimension,

$$\frac{1}{c^2}\frac{\partial^2 y}{\partial t^2} = \frac{\partial^2 y}{\partial x^2}, \tag{10.51}$$

where c is called the wave speed. The formula for the wave speed in a string is $c = \sqrt{T_0/\varrho_0}$, where T_0 is the tension in the string and ϱ_0 is the linear mass density. As we will see in Section 11.2, (10.51) has "product solutions" whose dependence on time, t, is given by $c_1 \cos\left(\frac{n\pi ct}{L}\right) + c_2 \sin\left(\frac{n\pi ct}{L}\right)$. These are periodic functions having frequencies of vibration $\omega_n = \frac{n\pi c}{L}$, that is, frequencies $f_n = \frac{\omega}{2\pi} = \frac{nc}{2L}$ Hz.

10.4.1 Guitar String

Example 10.10

Suppose a guitar string has mass density of 4.466×10^{-3} kg/m and length 640.0 mm and is under tension equal to the weight of a mass of 9.03 kg at sea level. Find the basic frequency of vibration of the string.

Method: The $n = 1$ mode has the basic frequency, that is, the basic frequency is $f_1 = \frac{c}{2L}$ Hz. The tension is found by multiplying 9.03 kg by 9.80665 m/s^2, the acceleration due to gravity at sea level, so $T_0 \approx (9.03) \cdot (9.80665)\ N \approx 88.554\ N$. The wave speed is $c \approx \sqrt{T_0/\varrho_0} = \sqrt{88.554\,\text{N}/(4.466 \times 10^{-3}\,\text{kg/m})} \approx 140.81$ m/s. The basic frequency is

$$f_1 \approx \frac{140.81\,\text{m/s}}{2(0.640\,\text{m})} \approx 110\,\text{Hz}. \bigcirc$$

This frequency of 110 Hz gives a "pitch" of $A2$, in "scientific notation," that is, A in "Helmholtz notation," and is a minor tenth below "middle C." The fifth lowest of the six guitar strings is tuned to this pitch.

Guitar players also sometimes finger a string so as to excite a higher harmonic, that is, not the basic frequency, because that way, they can play very high notes that also have a "mystical" or ethereal "tone color."

By the way, the true response of a guitar string is nonlinear, so even if only a few modes are excited initially, many other modes can be involved, although beyond a certain frequency, the intensity of the modes will be insignificant and human beings won't hear them, anyway. The relative magnitudes of the amplitudes of the higher harmonics are responsible for the tone colors of various musical instruments, for example, why bagpipes don't sound like trumpets.

The PDE modeling vibrations of a string generalize in two space dimensions to

$$\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}. \tag{10.52}$$

This can be used to model vibrations of a membrane.

If the material is not homogeneous, then we will see an even more general linear PDE; for example, in one space dimension,

$$\varrho(x)\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x}\left[T(x)\frac{\partial}{\partial x}\right]. \tag{10.53}$$

Even more general is the nonlinear system of PDEs

$$\varrho(x)\frac{\partial^2 \mathbf{r}}{\partial t^2} = \frac{\partial}{\partial x}\left[\mathbf{n}\left(\frac{\partial \mathbf{r}}{\partial x}\right)\right].$$

In addition, there are models for vibrations of plates and other solids and models for electromagnetic waves in transmission lines and waveguides.

10.4.2 Vibrating String

Suppose a string, when undisturbed and unstretched, occupies the interval $0 \leq X \leq L$ on the X-axis. This is called the **reference configuration** or **reference state**, as we referred to it in Section 10.1's discussion of continuum mechanics. At a later time, the string forms a curve

$$\mathbf{r} = \mathbf{r}(X, t).$$

In effect, a material particle whose initial position vector was $X\hat{\boldsymbol{\imath}}$ will be on the string at a point whose position vector is $\mathbf{r}(X, t)$. So, for each $t \geq 0$, we have a map of material particles

$$X \mapsto \mathbf{r}(X, t).$$

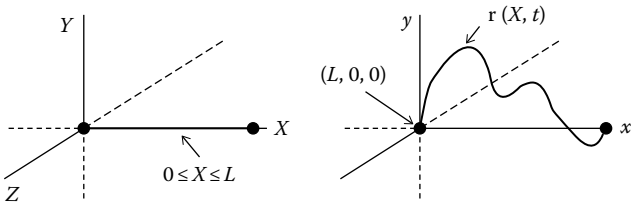


FIGURE 10.11
String.

For example, in Figure 10.11, we see that the material particle at the origin is mapped to the origin and the material particle at $X = L$ is mapped to the point $(x, y, z) = (L, 0, 0)$. So,

$$\mathbf{r}(0, t) = \mathbf{0} \quad \text{and} \quad \mathbf{r}(L\,\hat{\boldsymbol{\imath}}, t) = L\,\hat{\boldsymbol{\imath}}. \tag{10.54}$$

For each fixed t, the map $X \mapsto \mathbf{r}(X, t)$, $0 \le X \le L$, gives a parametric curve.

We say that the string is **tied at the ends** if BC (10.54) are satisfied. Figure 10.11 depicts the string at some time after it has moved, possibly by an out-of-the-xy-plane motion.

The **local elongation** of the string is

$$\left\| \frac{\partial \mathbf{r}}{\partial X}(X, t) \right\|,$$

which we will assume is positive at all X in $0 \le X \le L$ and all $t > 0$. (Technically, we should talk about $\frac{\partial \mathbf{r}}{\partial X}(0^+, t)$ rather than about $\frac{\partial \mathbf{r}}{\partial X}(0, t)$, but we will avoid being fussy.)

Let $\mathbf{T} = \mathbf{T}(X, t)$ be the **contact force** or **tension force** exerted *on* the material segment $[0, X)$ *by* the material segment $(X, L]$. One definition of a **string** is that it is a zero thickness idealization of material that has the property that $\mathbf{T}(X, t)$ is tangent to the string at $\mathbf{r}(X, t)$, that is,

$$\mathbf{T}(X, t) = \beta \cdot \frac{\partial \mathbf{r}}{\partial X}(X, t)$$

for some scalar function β.* A string is said to be **elastic** if β is a function of $\left|\left|\frac{\partial \mathbf{r}}{\partial X}(X, t)\right|\right|$. In this case, the tension force is

$$\mathbf{T}(X, t) = \beta\left(\left|\left|\frac{\partial \mathbf{r}}{\partial X}(X, t)\right|\right|\right) \frac{\partial \mathbf{r}}{\partial X}(X, t). \tag{10.55}$$

Consider any interval $I \triangleq (a, b)$ with $0 < a < b < L$. Denote by $\mathcal{P}$ the piece of the string that came from the material in I, that is,

$$\mathcal{P} = \{\mathbf{r}(X, t) : a \le X \le b\} = \mathbf{r}(I, t).$$

* The dependence of β on $\mathbf{r}(X, t)$ may be quite complicated. For example, a **viscoelastic** string would have β depending upon the "history" of $\frac{\partial \mathbf{r}}{\partial X}$, that is, values of $\frac{\partial \mathbf{r}}{\partial X}(X, t - s)$ for $s > 0$.

Newton's second law of motion says that

$$\frac{d}{dt}\big[\text{ linear momentum in } \mathcal{P}\,\big] = \text{sum of forces on } \mathcal{P}.$$

On the left end of $\mathcal{P}$, the material from $[0, a)$ exerts a contact force

$$-\mathbf{T}(a, t)$$

on $\mathcal{P}$; similarly, the material from $(b, 1]$ exerts a contact force

$$\mathbf{T}(b, t)$$

on $\mathcal{P}$. In addition, there may be "body" forces $\mathbf{f}(X, t)$ acting on each point $\mathbf{r}(X, t)$ in $\mathcal{P}$. Body forces may include gravity or frictional forces.

The total linear momentum in $\mathcal{P}$ is

$$\int_a^b \varrho(X)\,\frac{\partial \mathbf{r}}{\partial t}(X, t)\,dX,$$

where $\varrho(X)$ is the linear mass density, and the total body force acting on $\mathcal{P}$ is

$$\int_a^b \mathbf{f}(X, t)\,dX.$$

So, Newton's second law of motion says that

$$\int_a^b \varrho(X)\,\frac{\partial^2 \mathbf{r}}{\partial t^2}(X, t)\,dX = \frac{d}{dt}\left[\int_a^b \varrho(X)\frac{\partial \mathbf{r}}{\partial t}(X, t)\,dX\right] = -\mathbf{T}(a, t) + \mathbf{T}(b, t) + \int_a^b \mathbf{f}(X, t)\,dX. \quad (10.56)$$

By the fundamental theorem of calculus,

$$-\mathbf{T}(a, t) + \mathbf{T}(b, t) = \int_a^b \frac{\partial \mathbf{T}}{\partial X}(X, t)\,dX,$$

so, we can rewrite (10.56) as

$$\int_a^b \left(\varrho(X)\,\frac{\partial^2 \mathbf{r}}{\partial t^2}(X, t) - \frac{\partial \mathbf{T}}{\partial X}(X, t) - \mathbf{f}(X, t)\right) dX.$$

This being true for all a, b with $0 < a < b < L$, if we assume that $\mathbf{r}(X, t)$ is smooth enough, then this implies that $\mathbf{r}$ satisfies the PDE

$$\varrho(X)\,\frac{\partial^2 \mathbf{r}}{\partial t^2}(X, t) = \frac{\partial \mathbf{T}}{\partial X}(X, t) + \mathbf{f}(X, t),\ 0 < X < L,\ 0 < t < \infty. \quad (10.57)$$

For an elastic string, the constitutive law (10.55) holds. Substitute that into (10.57) to get the PDE

$$\varrho(X)\,\frac{\partial^2 \mathbf{r}}{\partial t^2} = \frac{\partial}{\partial X}\left[\beta\left(\left|\left|\frac{\partial \mathbf{r}}{\partial X}(X,t)\right|\right|\right)\frac{\partial \mathbf{r}}{\partial X}\right] + \mathbf{f},\ 0 < X < L,\ 0 < t < \infty. \tag{10.58}$$

A special case is when

$$\beta\left(\left|\left|\frac{\partial \mathbf{r}}{\partial X}(X,t)\right|\right|\right) \equiv T_0 \text{ and } \varrho(X) \equiv \varrho_0,$$

that is, the tension coefficient is constant and the mass density is constant. If we assume also that the motions of the string have small amplitude and come close to remaining in the xy-plane, we can approximate $\mathbf{r}(X,t) \approx X\hat{\boldsymbol{\imath}} + y(X,t)\hat{\boldsymbol{\jmath}} \approx x\hat{\boldsymbol{\imath}} + y(x,t)\hat{\boldsymbol{\jmath}}$, so the PDE becomes (10.51), that is,

$$\frac{1}{c^2}\,\frac{\partial^2 y}{\partial t^2} = \frac{\partial^2 y}{\partial x^2},$$

where $c = \sqrt{T_0/\varrho_0}$.

Along with a PDE such as (10.51), it makes sense to specify BC at material points $X = 0$ and $X = L$. Also, analogous to our study of vibrations of a spring in Section 3.3, we should specify initial conditions, usually giving as data the functions of x given by

$$y(x,0),\ 0 < x < L$$

and

$$\frac{\partial y}{\partial t}(x,0),\ 0 < x < L.$$

10.4.3 Speed of Sound

Gases *are* compressible, so the density $\varrho = \varrho(x,t)$ is not necessarily constant. Let $u = u(x,t)$ be the gas velocity in one space dimension, P the pressure, and f be the sum of the body forces. The basic model for vibrations of the gas the

$$\text{Continuity equation}: \quad \frac{\partial \varrho}{\partial t} + \frac{\partial}{\partial x}\big[\varrho u\big] = 0, \tag{10.59}$$

that is, the 1D version of Euler's conservation of mass PDE, (7.41) in Section 7.6 or (10.2) in Section 10.1, and the

$$\text{Momentum equation}: \quad \varrho\,\frac{\partial u}{\partial t} + \varrho\, u\,\frac{\partial u}{\partial x} + \frac{\partial P}{\partial x} = \varrho f, \tag{10.60}$$

which follows from Newton's second law of motion and neglecting viscosity. Sometimes (10.60) is referred to as "Euler's equation of motion for an 'ideal' fluid." The first two terms of (10.60) combined can be rewritten in terms of the "material derivative" as $\varrho\frac{Du}{dt}$.

By "still air," we mean the constant physical state $\varrho \equiv \varrho_0, P = P_0$, and $u \equiv 0$. This is the analogue of the reference state we mentioned in the study of continuum mechanics.

We will consider "small disturbances," that is,

$$\left\{\begin{array}{l} \varrho = \varrho_0 + \varepsilon\, \varrho_1(x,t) \\ \\ u = \varepsilon\, u_1(x,t) \end{array}\right\}, \tag{10.61}$$

where we expect ε to be very small, which we denote by $|\varepsilon| \ll 1$. We will also assume that $f \equiv 0$, that is, that the body forces are so small that we can neglect them.

Noting that ϱ_0 is constant, substitute (10.61) into (10.59) to get

$$0 = \frac{\partial}{\partial t}\Big[\varrho_0 + \varepsilon\, \varrho_1\Big] + \frac{\partial}{\partial x}\Big[\,(\varrho_0 + \varepsilon\, \varrho_1)\varepsilon\, u_1\Big] = \frac{\partial \varrho_0}{\partial t} + \varepsilon\, \frac{\partial \varrho_1}{\partial t} + \varepsilon\, \varrho_0 \frac{\partial u_1}{\partial x} + \varepsilon^2\, \frac{\partial}{\partial x}\Big[\varrho_1 u_1\Big],$$

that is,

$$0 = 0 + \varepsilon\, \frac{\partial \varrho_1}{\partial t} + \varepsilon\, \varrho_0\, \frac{\partial u_1}{\partial x} + \varepsilon^2\, \frac{\partial}{\partial x}\Big[\varrho_1 u_1\Big]. \tag{10.62}$$

Because $|\varepsilon| \ll 1$, terms involving ε^2 should be significantly smaller than terms involving ε^0 or ε^1.

Neglecting the ε^2 term in (10.62) and then dividing by ε gives

$$\frac{\partial \varrho_1}{\partial t} + \varrho_0 \frac{\partial u_1}{\partial x} = 0. \tag{10.63}$$

Similarly, substitute (10.61) into (10.60) and recall that we assumed $f \equiv 0$; after that, neglect terms involving ε^2 or ε^3. This process gives

$$\begin{aligned} 0 &= \big(\varrho_0 + \varepsilon\, \varrho_1\big) \frac{\partial \big(\varepsilon\, u_1\big)}{\partial t} + \big(\varrho_0 + \varepsilon\, \varrho_1\big)\big(\varepsilon\, u_1\big)\, \frac{\partial \big(\varepsilon\, u_1\big)}{\partial x} + \frac{\partial P}{\partial x} \\ &= \varepsilon\, \varrho_0 \frac{\partial u_1}{\partial t} + \varepsilon^2\, \varrho_1 \frac{\partial u_1}{\partial t} + \varepsilon^2 \varrho_0\, u_1\, \frac{\partial u_1}{\partial x} + \varepsilon^3 \varrho_1\, u_1\, \frac{\partial u_1}{\partial x} + \frac{\partial P}{\partial x}, \end{aligned}$$

hence,

$$\varepsilon\, \varrho_0 \frac{\partial u_1}{\partial t} + \frac{\partial P}{\partial x} = 0. \tag{10.64}$$

To have a system of PDEs that is solvable, we will need a constitutive relationship. For a gas, let's assume that the pressure is a function of the mass density:

$$P = P(\varrho).$$

Then the chain rule and Taylor series together give

$$\frac{\partial P}{\partial x} = \frac{\partial}{\partial x}\Big[P\left(\varrho(x,t)\right)\Big] = \frac{dP}{d\varrho}\cdot\frac{\partial \varrho}{\partial x} = \frac{dP}{d\varrho}\left(\varrho(x,t)\right)\cdot\frac{\partial \varrho}{\partial x}(x,t) = \frac{dP}{d\varrho}(\varrho_0+\varepsilon\,\varrho_1)\cdot\frac{\partial}{\partial x}\Big[\varrho_0+\varepsilon\,\varrho_1\Big].$$

$$= \left(\frac{dP}{d\varrho}(\varrho_0) + \varepsilon\,\varrho_1\,\frac{d^2P}{d\varrho^2}(\varrho_0)+\cdots\right)\cdot\left(0+\varepsilon\,\frac{\partial \varrho_1}{\partial x}\right),$$

where "$\cdots$" is short for terms involving ε^2. This gives

$$\frac{\partial P}{\partial x} = \varepsilon\,\frac{dP}{d\varrho}(\varrho_0)\cdot\frac{\partial \varrho_1}{\partial x}+\cdots.$$

Substitute this into (10.64), neglect terms involving ε^2, and then divide by ε to get

$$\varrho_0\frac{\partial u_1}{\partial t} + \frac{dP}{d\varrho}(\varrho_0)\cdot\frac{\partial \varrho_1}{\partial x} = 0. \tag{10.65}$$

The system of equations (10.63) and (10.65), that is,

$$\left\{\begin{array}{c} \dfrac{\partial \varrho_1}{\partial t} = -\varrho_0\dfrac{\partial u_1}{\partial x} \\[2ex] \varrho_0\dfrac{\partial u_1}{\partial t} = -P'(\varrho_0)\cdot\dfrac{\partial \varrho_1}{\partial x} \end{array}\right\}, \tag{10.66}$$

is called the "linearization" of (10.59) and (10.60).

The linearized system can be reduced to a single PDE, specifically the wave equation: Partially differentiate (10.63) with respect to t to get

$$\frac{\partial^2 \varrho_1}{\partial t^2} + \varrho_0\frac{\partial^2 u_1}{\partial x\partial t} = 0 \tag{10.67}$$

and take the partial derivative of (10.65) with respect to x to get

$$\varrho_0\frac{\partial^2 u_1}{\partial t\partial x} + P'(\varrho_0)\cdot\frac{\partial^2 \varrho_1}{\partial x^2} = 0. \tag{10.68}$$

Solve (10.68) for $\varrho_0\frac{\partial^2 u_1}{\partial t\partial x}$ and substitute into (10.67) to get

$$\frac{\partial^2 \varrho_1}{\partial t^2} = P'(\varrho_0)\cdot\frac{\partial^2 \varrho_1}{\partial x^2}.$$

This is the wave equation in one space dimension and wave speed $\boxed{c=\sqrt{P'(\varrho_0)}}$, which is the **speed of sound**.

The analysis we used, including the use of a small parameter ε, is an example of a "perturbation" method.

To find the value of the speed of sound, it is necessary to use a definite expression for the constitutive law $P = P(\varrho)$. For an "ideal" gas, in **adiabatic conditions**, that is, when there is no or negligible heat exchange with surroundings, $PV^\gamma = k$ for some constants γ, k. It follows that

$$P = kV^{-\gamma}.$$

Further, there is a relationship between volume, V, and mass density, ϱ: Let M be the total mass of a gas in a fixed volume V. Then $M = \varrho V$, hence $V^{-1} = \varrho/M$. It follows that

$$P = kV^{-\gamma} = k\left(V^{-1}\right)^\gamma = k\left(\frac{\varrho}{M}\right)^\gamma;$$

hence,

$$P = kM^{-\gamma}\varrho^\gamma,$$

and thus,

$$P'(\varrho_0) = k \cdot \gamma\, M^{-\gamma}\varrho_0^{\gamma-1}. \tag{10.69}$$

But, in still air, that is, when $\varrho = \varrho_0, P = P_0$, and $V = V_0 \triangleq M/\varrho_0$,

$$k = P_0\, V^\gamma = P_0\left(\frac{M}{\varrho_0}\right)^\gamma = P_0\,\varrho_0^{-\gamma}\,M^\gamma.$$

Substitute this into (10.69) to get

$$P'(\varrho_0) = \left(P_0\,\varrho_0^{-\gamma}M^\gamma\right)\left(\gamma\, M^{-\gamma}\varrho_0^{\gamma-1}\right);$$

hence,

$$P'(\varrho_0) = \frac{\gamma\, P_0}{\varrho_0}. \tag{10.70}$$

So, according to this model, the speed of sound is

$$c = \sqrt{\frac{\gamma\, P_0}{\varrho_0}}.$$

Assuming that "standard conditions" have temperature 273.15 K, that is, 0°C, atmospheric pressure at sea level being $P_0 = (1\,\text{atm}) \approx 101,323\,\text{N/m}^2$, and density of air being

$\varrho_0 \approx 1.293\,\text{kg/m}^3$. Assuming the adiabatic gas constant is $\gamma \approx 1.4$,

$$c = \sqrt{P'(\varrho_0)} = \sqrt{\frac{\gamma\, P_0}{\varrho_0}} \approx \sqrt{\frac{(1.4)(101,323\,\text{N/m}^2)}{1.293\ \text{kg/m}^3}} \approx 331.22\,\text{m/s} \approx 1086.7\,\text{ft/s}.$$

This is quite close to the experimental value of 1088 ft/s for the speed of sound.

10.4.4 Linear Elasticity

In Section 10.1, we derived the continuum mechanics PDE implied by Newton's second law of motion (10.24) in Section 10.1, that is,

$$\varrho(\mathbf{x}, t)\, \frac{\partial^2 \mathbf{u}(\mathbf{x}, t)}{\partial t^2} = \mathbf{f}(\mathbf{x}, t) \ + \ \boldsymbol{\nabla} \bullet [\tau], \tag{10.71}$$

where
the displacement vector is $\mathbf{u} \triangleq \mathbf{x}(\mathbf{X}, t) - \mathbf{X}$,
material particles $\mathbf{X}$ are in a region $\mathcal{V}_0 = \mathcal{V}(0)$, $\mathbf{f}$ is the sum of the body forces,
$[\tau]$ is the stress tensor due to contact forces.

The 3×3 matrix is symmetric because of conservation of angular momentum. Also, as the system evolves, the material particles may move, so the material occupies the region $\mathcal{V}(t) \triangleq \{\mathbf{x}(\mathbf{X}, t) : \mathbf{X} \text{ in } \mathcal{V}_0\}$. Writing the displacement vector as

$$\mathbf{u} = u_1(\mathbf{X}, t)\hat{\boldsymbol{\imath}} + u_2(\mathbf{X}, t)\,\hat{\boldsymbol{\jmath}} + u_3(\mathbf{X}, t)\hat{\boldsymbol{k}},$$

we define the 3×3 **linearized strain tensor** $[\varepsilon]$ in terms of its entries:

$$\varepsilon_{ij} \triangleq \frac{1}{2}\left(\frac{\partial u_i}{\partial X_j} + \frac{\partial u_j}{\partial X_i}\right).$$

The exact strain tensor has other terms in its ij-th entry, namely, $\sum_{k=1}^{3} \left(\frac{\partial u_k}{\partial X_i}\,\frac{\partial u_k}{\partial X_j}\right)$, which we neglect based on the assumption that displacements are small.

Note that

$$[\varepsilon] = \frac{1}{2}\left(\left[\frac{\partial \mathbf{u}}{\partial \mathbf{X}}\right] + \left[\frac{\partial \mathbf{u}}{\partial \mathbf{X}}\right]^T\right)$$

is a symmetric matrix.

Example 10.11

In Sections 1.2 and 6.6, we called $\begin{bmatrix} 1 & 0 \\ h & 1 \end{bmatrix}$ the matrix of a linear shear transformation on $\mathbb{R}^2$. Assuming h is a scalar constant, find the linear strain tensor for the corresponding

mapping of material particles given by

$$\mathbf{x}(\mathbf{X}) \triangleq \begin{bmatrix} 1 & 0 & 0 \\ h & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{X}.$$

Method: The displacement is

$$\mathbf{u}(\mathbf{X}) \triangleq \mathbf{x}(\mathbf{X}) - \mathbf{X} = \begin{bmatrix} 0 & 0 & 0 \\ h & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{X} = \begin{bmatrix} 0 \\ hX_1 \\ 0 \end{bmatrix},$$

so the strain tensor is

$$[\varepsilon] = \frac{h}{2} \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \bigcirc$$

Definition 10.1

The **principal strains** are the eigenvalues of the symmetric matrix $[\varepsilon]$.

By Theorem 2.23 in Section 2.6, the principal strains are real numbers because $[\varepsilon]$ is real and symmetric. In Example 10.11, the principal strains are $0, \pm\frac{h}{2}$.

Definition 10.2

Plane strain is when $[\varepsilon]$ has one of the forms

$$\left[\begin{array}{cc:c} \varepsilon_{11} & \varepsilon_{12} & 0 \\ \varepsilon_{21} & \varepsilon_{22} & 0 \\ \hdashline 0 & 0 & 0 \end{array}\right], \left[\begin{array}{c:c:c} \varepsilon_{11} & 0 & \varepsilon_{13} \\ \hdashline 0 & 0 & 0 \\ \hdashline \varepsilon_{31} & 0 & \varepsilon_{33} \end{array}\right], \left[\begin{array}{c:cc} 0 & 0 & 0 \\ \hdashline 0 & \varepsilon_{22} & \varepsilon_{23} \\ 0 & \varepsilon_{32} & \varepsilon_{33} \end{array}\right], \tag{10.72}$$

or, more generally,

$$\mathcal{O} \left[\begin{array}{cc:c} \varepsilon_{11} & \varepsilon_{12} & 0 \\ \varepsilon_{21} & \varepsilon_{22} & 0 \\ \hdashline 0 & 0 & 0 \end{array}\right] \mathcal{O}^T, \tag{10.73}$$

for some 3×3 orthogonal matrix $\mathcal{O}$.

In the latter case of (10.73), $\{\mathcal{O}\hat{\boldsymbol{\imath}}, \mathcal{O}\hat{\boldsymbol{\jmath}}\}$ is an orthonormal basis for the plane in which there is strain and $\mathcal{O}\hat{\boldsymbol{k}}$ is normal to that plane.

In any case, we have plane strain if, and only if, at least one of the principal strains is zero.

By itself, the PDE (10.71) cannot be solved, even if the body forces are zero. To have a solvable problem, we need to have a constitutive law that gives the stress tensor in terms of the displacement vector. To that end, we say that we have **linear elasticity** if Hooke's law holds in the sense that the stress tensor is a linear transformation of the linearized strain tensor, that is,

$$[\tau] = C[\varepsilon],$$

where C is a linear transformation from the space of all 3×3 symmetric matrices to the space of 3×3 symmetric matrices. In terms of indices, for each $1 \leq i \leq 3$ and each $1 \leq j \leq 3$,

$$\tau_{ij} = \sum_{k=1}^{3} \sum_{\ell=1}^{3} c_{ijkl} \varepsilon_{k\ell}.$$

Fortunately, the 81 possible entries in C can be reduced to 36 due to the fact that both $[\tau]$ and $[\varepsilon]$ must be symmetric matrices. If we further assume that the material is **isotropic**, that is, it is homogeneous in the sense that it "bends in the same way in all directions," then it turns out that C is specified by exactly two numbers, λ, μ, and

$$\tau_{ij} = \lambda \cdot (\varepsilon_{11} + \varepsilon_{22} + \varepsilon_{33})\delta_{ij} + 2\mu\, \varepsilon_{ij}, \tag{10.74}$$

where δ_{ij} is the Kronecker delta, that is,

$$\delta_{ij} \triangleq \begin{Bmatrix} 1 & i = j \\ 0 & i \neq j \end{Bmatrix}.$$

We can rewrite (10.74) more compactly as

$$[\tau] = \lambda \cdot (\varepsilon_{11} + \varepsilon_{22} + \varepsilon_{33})I + 2\mu[\varepsilon].$$

Lamé constants are the constants λ and μ. Another name for μ is the **shear modulus**.

We will substitute the isotropic constitutive law (10.74) into the PDE (10.71). Further, for small amplitude motions we can make the approximations $\mathbf{x}(\mathbf{X}, t) \approx \mathbf{X}$ and $\varrho(\mathbf{x}, t) \approx \varrho_0(\mathbf{X})$ to get the "Navier" or "Lamé" equation,

$$\varrho_0(\mathbf{X})\, \frac{\partial^2 \mathbf{u}(\mathbf{X}, t)}{\partial t^2} = \mathbf{f}(\mathbf{X}, t) \; + \; (\lambda + \mu)\nabla\big[\nabla \bullet \mathbf{u}\big] + \mu\, \nabla^2 \mathbf{u}.$$

Alternatively, using the identity

$$\nabla^2 \mathbf{u} = \nabla\big[\nabla \bullet \mathbf{u}\big] - \nabla \times (\nabla \times \mathbf{u}),$$

we can rewrite the Navier–Lamé equation as

$$\varrho_0(\mathbf{X})\, \frac{\partial^2 \mathbf{u}(\mathbf{X}, t)}{\partial t^2} = \mathbf{f}(\mathbf{X}, t) \; + \; (\lambda + 2\mu)\nabla\big[\nabla \bullet \mathbf{u}\big] - \mu\, \nabla \times (\nabla \times \mathbf{u}).$$

10.4.5 Linear Elastostatics

For a steady state, that is, for the elastostatic problem, $\dfrac{\partial \mathbf{u}(\mathbf{X}, t)}{\partial t} \equiv \mathbf{0}$ implies

$$\mathbf{0} = \mathbf{f}(\mathbf{X}, t) + (\lambda + 2\mu)\nabla\,[\nabla \bullet \mathbf{u}] - \mu\,\nabla \times (\nabla \times \mathbf{u}).$$

Uniaxial stretching, say along the x-axis, is when displacements are given by

$$\mathbf{u} = \alpha \begin{bmatrix} x \\ -\nu\, y \\ -\nu\, z \end{bmatrix},$$

for some ν. In Problem 10.4.6.6, you will explain why for these displacements, the strain tensor is the diagonal matrix

$$[\varepsilon] = \begin{bmatrix} \alpha & 0 & 0 \\ 0 & -\alpha\nu & 0 \\ 0 & 0 & -\alpha\nu \end{bmatrix}. \tag{10.75}$$

Assuming isotropy, Hooke's law says that the stress tensor is

$$[\tau] = \begin{bmatrix} \alpha\,(\lambda(1-2\nu)+2\mu) & 0 & 0 \\ 0 & \alpha\,(\lambda(1-2\nu)-2\mu\nu) & 0 \\ 0 & 0 & \alpha\,(\lambda(1-2\nu)-2\mu\nu) \end{bmatrix}, \tag{10.76}$$

as you will explain in Problem 10.4.6.7. If ν is chosen to equal **Poisson's ratio**, that is,

$$\nu = \frac{\lambda}{2(\lambda+\mu)},$$

then, as you will explain in Problem 10.4.6.8,

$$[\tau] = \begin{bmatrix} \alpha \cdot 2\mu\,(1+\nu) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{10.77}$$

In this case, $\tau_{11} = E\alpha$, where the **Young's modulus** is defined by

$$E \triangleq \frac{\mu \cdot (3\lambda + 2\mu)}{\lambda + \mu},$$

in terms of which the Lamé constants are given by

$$\mu = \frac{E}{2(1+\nu)};\ \text{hence},\ \lambda = \frac{\nu\, E}{(1+\nu)(1-2\nu)}, \tag{10.78}$$

as you will explain in Problem 10.4.6.9.

Finally, we briefly mention that the theory of finite thickness beams, rods, and plates takes into account "transverse shear" forces and that each piece of such a 3D object

exerts a bending moment(s) on the rest of the object. Thus, the equation of motion takes into account a spatial derivative(s) of the shear force, and the latter may be modeled as being proportional to a spatial derivative(s) of the bending moment. Further, the bending moment may be modeled as being approximately proportional to a second derivative(s) of the deformation in the longitudinal direction(s). This leads to linear PDE models, for example, of the forms

$$\varrho\ddot{w} = -EIw^{(iv)} + Tw'' - \varrho g,\ 0 < x < L.$$

$$\varrho\ddot{\mathbf{w}} = -E[\mathcal{I}]\mathbf{w}^{(iv)} + T\mathbf{w}'' + \varrho\mathbf{f},\ 0 < x < L.$$

$$\varrho\ddot{w} = -D\nabla^4 w + T\nabla^2 w - \varrho g,\ (x, y) \text{ in } \mathcal{D}.$$

for beams, rods, and plates, respectively. The symbol $\mathcal{I}$ stands for a 2×2 inertia matrix, and $\mathbf{w} = [v(x,t) \quad w(x,t)]^T$ are the displacements in the directions transverse to the rod. In the third PDE, we assumed that the stress tensor is isotropic. There are even more sophisticated, nonlinear, approximate models such as von Kármán's plate theory.

Learn More About It

Our derivation of PDEs for vibrating strings is based on part of "The equations for large vibrations of strings," by Stuart S. Antman, *Am. Math. Monthly* **87** (1980) pp. 359–370. As in Section 10.1, our discussion of the PDEs of elasticity follows Chapter 1 of *Applied Solid Mechanics*, Peter Howell, Gregory Kozyreff, and Jim Ockendon, Cambridge University Press, 2009. In addition, the latter book's Chapter 4 is a good reference for the linear theory of finite thickness beams, rods, and plates.

Example 10.11 is based on *Continuum Mechanics*, by Segel, p. 156, Example 3. Problem 10.4.6.10 is a version of Exercise 2.2.2 on page 40 of *Partial Differential Equations: An Introduction*, by Walter A. Strauss, John Wiley & Sons, Inc., 1992.

10.4.6 Problems

1. In deriving the wave equation for a vibrating string, we assumed that the mass density, ϱ, and the tension force, T_0, were constants. Instead, let the mass density and the tension force be functions of x and derive a PDE for the string.
2. In the transmission line equation,

$$\frac{\partial^2 i}{\partial x^2} = LC\frac{\partial^2 i}{\partial t^2} + (LG + RC)\frac{\partial i}{\partial t} + RGi,$$

   suppose there is no current leakage, that is, $G = 0$, and let $i(x,t) = e^{-pt}u(x,t)$. Choose the constant p so that the transmission line equation reduces to a PDE of the form $\dfrac{\partial^2 u}{\partial x^2} = c^2\dfrac{\partial^2 u}{\partial t^2} - ru$ and state the values of the constants c, r in terms of R, L, C.
3. Suppose a $13\frac{7}{8}$ in.($\approx$352 mm) long mandolin string is under 21.3 lb of tension and is tuned to a pitch of 440 Hz, that is, $A4$ in the scientific notation for pitches. What

is the mass density of the string, in kg/m? Assume that at sea level, an object whose mass is 1 kg weighs about 9.80665 N, that is, about 2.2 lb.

4. Find a real, orthogonal matrix $\mathcal{O}$ such that

$$\left[\begin{array}{c|cc} 0 & 0 & 0 \\ \hline 0 & d & c \\ & & \\ 0 & b & a \end{array}\right] = \mathcal{O}\left[\begin{array}{cc|c} a & b & 0 \\ & & \\ c & d & 0 \\ \hline 0 & 0 & 0 \end{array}\right]\mathcal{O}^T.$$

[Hint: What matrix multiplication has the effect of swapping the variables x and z?]

5. Find a real, orthogonal matrix $\mathcal{O}$ such that

$$\left[\begin{array}{ccc} a & 0 & b \\ & & \\ 0 & 0 & 0 \\ & & \\ c & 0 & d \end{array}\right] = \mathcal{O}\left[\begin{array}{cc|c} a & b & 0 \\ & & \\ c & d & 0 \\ \hline 0 & 0 & 0 \end{array}\right]\mathcal{O}^T.$$

[Hint: What matrix multiplication has the effect of swapping the variables y and z?]

6. Explain why in the strain tensor is the diagonal matrix (10.75) in the example of isotropic, uniaxial stretching.
7. Explain why Hooke's law for isotropic, uniaxial stretching gives (10.76).
8. Explain why the stress tensor $[\tau]$ is given by (10.77) for the uniaxial stretching example, assuming ν is chosen to equal Poisson's ratio.
9. Explain why the Lamé constants for isotropic, uniaxial stretching are given by (10.78).
10. Suppose that $u(x,t)$ satisfies the 1D wave equation with wave speed $\sqrt{c} = 1$, that is, $\dfrac{\partial^2 u}{\partial t^2} = \dfrac{\partial^2 u}{\partial x^2}$. Define the **energy density** function $e(x,t)$ and the **momentum density** function by

$$e(x,t) = \frac{1}{2}\left(\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial t}\right)^2\right),\ p(x,t) = \frac{\partial u}{\partial x}\,\frac{\partial u}{\partial t},$$

respectively. Explain why both $e(x,t)$ and $p(x,t)$ satisfy that very same 1D wave equation.

11. The **principle of virtual work** for the string equation (10.57) can be obtained from it by taking the dot product of both sides of that PDE with an unspecified "test function" $\boldsymbol{\eta} = \boldsymbol{\eta}(X,t)$, integrating both sides with respect to X over the interval $[\,0, L\,]$, using integration by parts (with respect to X) on the term involving $\boldsymbol{\eta}(X,t) \bullet \dfrac{\partial \mathbf{T}}{\partial X}(X,t)$, and then integrating both sides of the equation with respect to t over the interval $[0,\infty)$. Finally, use integration by parts (with respect to t) on the term

involving $\boldsymbol{\eta}(X,t) \bullet \varrho(X)\dfrac{\partial^2 \mathbf{r}}{\partial t^2}(X,t)$. Use this process to explain why we have the principle of virtual work equation

$$\int_0^{\infty}\int_0^{L}\left(\frac{\partial \boldsymbol{\eta}}{\partial X} \bullet \mathbf{T} - \mathbf{f} \bullet \boldsymbol{\eta} + \varrho(X)\frac{\partial \boldsymbol{\eta}}{\partial t} \bullet \frac{\partial \mathbf{r}}{\partial t}\right) dX dt = 0.$$

We also say (10.57) has the latter integral equation as its "weak form," whose meaning is related to the weak convergence idea discussed in Section 2.10.

10.5 D'Alembert Wave Solutions

Solutions of the wave equation (10.51) in Section 10.4 in one space dimension, that is,

$$\frac{\partial^2 y}{\partial t^2} = c^2 \frac{\partial^2 y}{\partial x^2}, \tag{10.79}$$

along with initial conditions defined on the real line, that is,

$$\left\{ \begin{array}{l} y(x,0) = \Phi(x) \\ \\ \frac{\partial y}{\partial t}(x,0) = \Psi(x) \end{array} \right\}, -\infty < x < \infty, \tag{10.80}$$

can be expressed as a sum of **d'Alembert wave solutions**:

$$y(x,t) = f(x+ct) + g(x-ct), \tag{10.81}$$

assuming that $\Phi(x)$ and $\Psi(x)$ are piecewise continuous on the real line. In the process of justifying (10.81), we will be specific about the relationship between the initial conditions, (10.80), and the solution (10.81).

First, we should understand the physical behavior of a solution such as $y(x,t) = f(x+ct)$: As t increases, it gives a wave traveling to the left with speed $c > 0$. Equivalently, as t decreases, it gives a wave traveling to the right. It depends on whether you are watching a movie run forward or backward! An example is depicted in Figure 10.12.

Example 10.12

(a) Find the wave speed from the graphs in Figure 10.12, assuming time t is measured in seconds and spatial position x is measured in meters.

(b) If the graphs in Figure 10.12 correspond to a solution of a vibrating string whose linear mass density is 4 g/m, what is the magnitude of the tension force in the string?

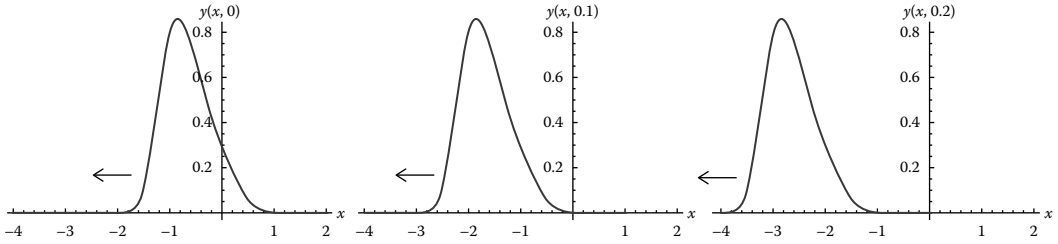


FIGURE 10.12
Wave moving left.

Method:

(a) Define $y(x,0) = f(x)$. It appears that the solution is a wave moving to the left and $y(x,0.1) \approx f(x+1)$. We want $f(x + c \cdot (0.1)) = y(x,0.1) \approx f(x+1)$, so $c \cdot (0.1) = 1$. The wave speed is approximately $c = 10\,\text{m/s}$.

(b) Recall that for a vibrating string, the wave speed is given by $c = \sqrt{T_0/\varrho_0}$. So, using the result of part (a), we have $10\,\text{m/s} = \sqrt{T_0/(0.004\,\text{kg/m})}$; hence, the tension force in the string has magnitude

$$T_0 = (100\,\text{m}^2/\text{s}^2) \cdot (0.004\,\text{kg/m}) = 0.4\,\text{kg} \cdot \text{m/s} = 0.4\,\text{N}. \bigcirc$$

10.5.1 Zero Initial Velocity

Example 10.13

Use (10.81) to solve the IVP consisting of the wave equation PDE (10.79) and a special case of the ICs: $y(x,0) = \Phi(x)$, $\frac{\partial y}{\partial t}(x,0) \equiv 0, -\infty < x < \infty$.

Method: In (10.81), f and g are functions of a single variable, for example, $f(x+ct) = f(\xi)|_{\xi=x+ct}$. Before substituting (10.81) into the ICs, it helps to calculate that, using the chain rule,

$$\frac{\partial y}{\partial t}(x,t) = \frac{\partial}{\partial t}\left[f(x+ct) + g(x-ct)\right] = f'(x+ct)\frac{\partial}{\partial t}\left[x+ct\right] + g'(x+ct)\frac{\partial}{\partial t}\left[x-ct\right];$$

hence,

$$\frac{\partial y}{\partial t}(x,t) = \frac{\partial}{\partial t}\left[f(x+ct) + g(x-ct)\right] = cf'(x+ct) - cg'(x-ct). \tag{10.82}$$

So, the ICs require that f and g satisfy the system of equations

$$\left\{\begin{array}{l} \Phi(x) = y(x,0) = f(x) + g(x) \\ \\ 0 \equiv \frac{\partial y}{\partial t}(x,0) = c\,\left(f'(x) - g'(x)\right) \end{array}\right\}, -\infty < x < \infty. \tag{10.83}$$

The second equation in (10.83) implies that $f'(x) \equiv g'(x)$. The fundamental theorem of calculus for functions of a single variable implies that

$$f(x) \equiv g(x) + c_1,$$

for some constant c_1. Substituting this into the first equation in (10.83) implies that

$$\Phi(x) \equiv f(x) + g(x) = g(x) + c_1 + g(x);$$

hence,

$$g(x) = \frac{1}{2}\,\Phi(x) - \frac{1}{2}\,c_1 \text{ and } f(x) = \frac{1}{2}\,\Phi(x) + \frac{1}{2}\,c_1.$$

So, the solution to the IVP is

$$y(x,t) = f(x+ct) + g(x-ct) = \frac{1}{2}\,\Phi(x+ct) + \cancel{\frac{1}{2}c_1} + \frac{1}{2}\,\Phi(x-ct) - \cancel{\frac{1}{2}c_1},$$

that is,

$$y(x,t) = \frac{1}{2}\,\left(\Phi(x+ct) + \Phi(x-ct)\right). \bigcirc \tag{10.84}$$

Example 10.14

If $y(x,0) \triangleq \begin{Bmatrix} 1, & 1 \le x < 2 \\ 0, & \text{all other x} \end{Bmatrix}$, $c = 10\,\text{m/s}$, and $y(x,t)$ is the solution of the IVP of Example 10.13, (a) sketch the solution $y(x,t)$ at times $t = 0, 0.05, 0.1, 0.2$ s, and (b) find as explicit a formula for the solution $y(x,t)$ as possible.

Method: (a) (10.84) gives us a "graphical method" of sketching the solution $y(x,t)$ at time $t > 0$: First, we can rewrite the IC as

$$y(x,0) = \frac{1}{2}\,\Phi(x) + \frac{1}{2}\,\Phi(x), \quad \text{where} \quad \Phi(x) = \begin{Bmatrix} 1, & 1 \le x < 2 \\ 0, & \text{all other x} \end{Bmatrix}$$

and sketch this decomposition in Figure 10.13. Because the wave speed is $c = 10\,\text{m/s}$, it is convenient to measure x in meters and t in seconds.

(i) At time $t = 0.05$, we have $y(x, 0.05) = \frac{1}{2}\,\Phi(x+0.5) + \frac{1}{2}\,\Phi(x-0.5)$. Figure 10.14 shows on the left the graph of $y(x, 0.05)$ versus x for $-1 < x < 4$. It is constructed as the sum of the graph of $\frac{1}{2}\,\Phi(x)$, shifted left by 0.5 m, and the graph of $\frac{1}{2}\,\Phi(x)$, shifted right by 0.5 m.

(ii) At time $t = 0.1$, we have $y(x, 0.1) = \frac{1}{2}\,\Phi(x+1) + \frac{1}{2}\,\Phi(x-1)$. Figure 10.15 shows on the left the graph of $y(x, 0.1)$ as the sum of the graph of $\frac{1}{2}\,\Phi(x)$, shifted left by 1 m, and the graph of $\frac{1}{2}\,\Phi(x)$, shifted right by 1 m.

(iii) At time $t = 0.2$, we have $y(x, 0.2) = \frac{1}{2}\,\Phi(x+2) + \frac{1}{2}\,\Phi(x-2)$. Figure 10.16 shows on the left the graph of $y(x, 0.2)$ as the sum of the graph of $\frac{1}{2}\,\Phi(x)$, shifted left by 2 m, and the graph of $\frac{1}{2}\,\Phi(x)$, shifted right by 2 m.

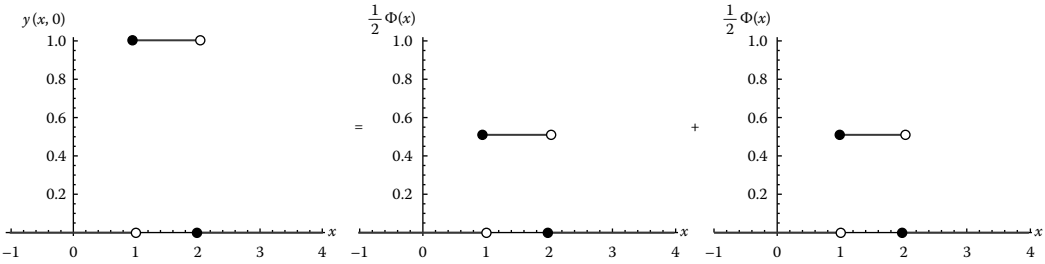


FIGURE 10.13
Decomposing the initial condition.

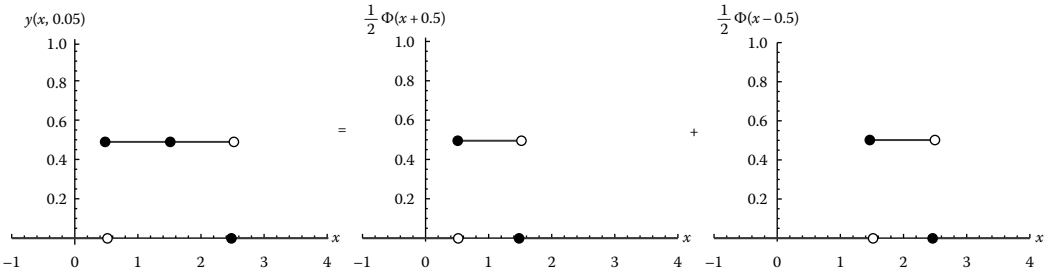


FIGURE 10.14
$y(x, 0.05)$.

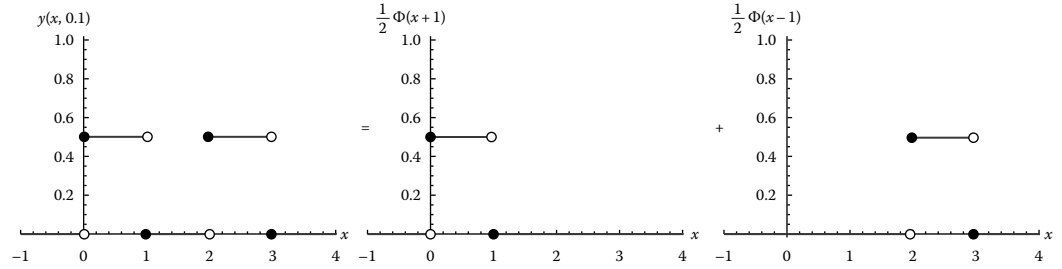


FIGURE 10.15
$y(x, 0.1)$.

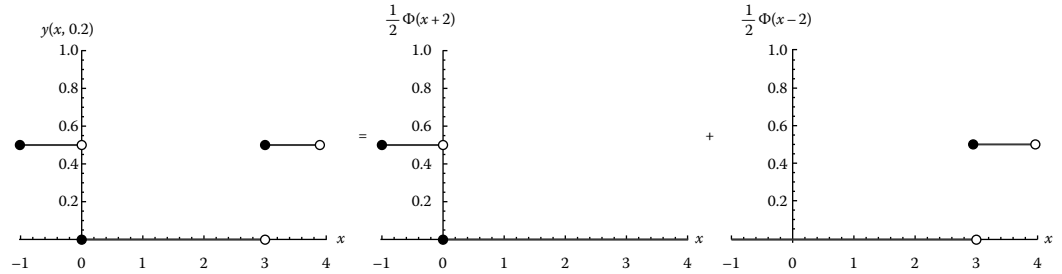


FIGURE 10.16
$y(x, 0.2)$.

(b) To write as explicit a *formula* for $y(x, t)$ as possible, we recall that

$$\Phi(x) \text{ is nonzero only for } 1 \le x < 2.$$

By replacing x by $\xi \triangleq x + ct$, we see that

$$\Phi(x + ct) = \Phi(\xi) \text{ is nonzero only for } 1 \le \xi = x + ct < 2,$$

that is,

$$\Phi(x + ct) \text{ is nonzero only for } 1 - ct \le x < 2 - ct.$$

It helps to create a new notation for a "moving interval": Define

$$\mathcal{I}_t^- \triangleq [1 - ct, 2 - ct),$$

that is, x is in $\mathcal{I}_t^-$ if, and only if, $1 - ct \le x < 2 - ct$. Similarly, we define

$$\mathcal{I}_t^+ \triangleq [1 + ct, 2 + ct).$$

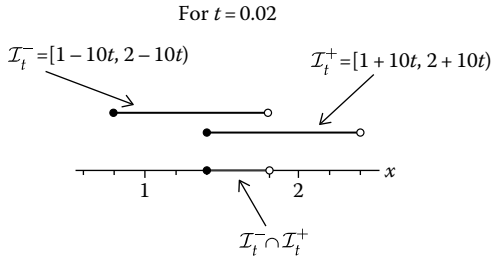


FIGURE 10.17
$\mathcal{I}_t^-$ and $\mathcal{I}_t^+$ *overlap.*

In terms of these notations, we have that

$$\Phi(x + ct) \text{ is nonzero only for } x \text{ in } \mathcal{I}_t^-$$
$$\text{and}$$
$$\Phi(x - ct) \text{ is nonzero only for } x \text{ in } \mathcal{I}_t^+.$$

For $t \geq 0$, $\mathcal{I}_t^-$, and $\mathcal{I}_t^+$ *overlap* if, and only if, $1 + ct < 2 - ct$, that is, $2ct < 1$, that is, $0 \leq t < \frac{1}{2c}$. In this particular example, $c = 10$, so

$$\mathcal{I}_t^- \text{ and } \mathcal{I}_t^+ \textit{ overlap} \iff 0 \leq t < 0.05.$$

Figure 10.17 illustrates this with the example of $t = 0.02$.

For $0 \leq t < 0.05$,

$$y(x,t) = \frac{1}{2}\,\Phi(x + ct) + \frac{1}{2}\,\Phi(x - ct)$$

$$= \begin{Bmatrix} 0.5, & 1 - 10t \leq x < 2 - 10t \\ 0, & x < 1 - 10t \text{ and } x \geq 2 - 10t \end{Bmatrix} + \begin{Bmatrix} 0.5, & 1 + 10t \leq x < 2 + 10t \\ 0, & x < 1 + 10t \text{ and } x \geq 2 + 10t \end{Bmatrix}$$

$$= \begin{Bmatrix} 0, & x < 1 - 10t \\ 0.5, & 1 - 10t \leq x < 1 + 10t \\ 0.5, & 1 + 10t \leq x < 2 - 10t \\ 0, & 2 - 10t \leq x < 2 + 10t \\ 0, & x \geq 2 + 10t \end{Bmatrix} + \begin{Bmatrix} 0, & x < 1 - 10t \\ 0, & 1 - 10t \leq x < 1 + 10t \\ 0.5, & 1 + 10t \leq x < 2 - 10t \\ 0.5, & 2 - 10t \leq x < 2 + 10t \\ 0, & x \geq 2 + 10t \end{Bmatrix}$$

$$= \begin{Bmatrix} 0 + 0, & x < 1 - 10t \\ 0.5 + 0, & 1 - 10t \leq x < 1 + 10t \\ 0.5 + 0.5, & 1 + 10t \leq x < 2 - 10t \\ 0 + 0.5, & 2 - 10t \leq x < 2 + 10t \\ 0 + 0, & x \geq 2 + 10t \end{Bmatrix} = \begin{Bmatrix} 0, & x < 1 - 10t \\ 0.5, & 1 - 10t \leq x < 1 + 10t \\ 1, & 1 + 10t \leq x < 2 - 10t \\ 0.5, & 2 - 10t \leq x < 2 + 10t \\ 0, & x \geq 2 + 10t \end{Bmatrix}.$$

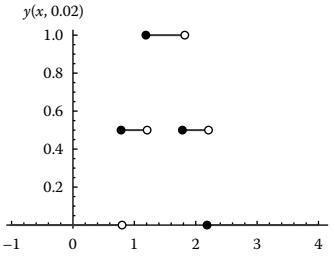


FIGURE 10.18
$y(x, 0.02)$.

For example, for $t = 0.02$, the solution is given by

$$y(x, 0.02) = \begin{Bmatrix} 0, & x < 0.8 \\ 0.5, & 0.8 \le x < 1.2 \\ 1, & 1.2 \le x < 1.8 \\ 0.5, & 1.8 \le x < 2.2 \\ 0, & x \ge 2.2 \end{Bmatrix},$$

which is graphed in Figure 10.18.

For $t \ge 0.05$, the intervals $\mathcal{I}_t^-$ and $\mathcal{I}_t^+$ do not overlap, so

$$\begin{aligned} y(x,t) &= \frac{1}{2}\,\Phi(x+ct) + \frac{1}{2}\,\Phi(x-ct) \\ &= \begin{Bmatrix} 0, & x < 1-10t \\ 0.5, & 1-10t \le x < 2-10t \\ 0, & x \ge 2-10t \end{Bmatrix} + \begin{Bmatrix} 0, & x < 1+10t \\ 0.5, & 1+10t \le x < 2+10t \\ 0, & x \ge 2+10t \end{Bmatrix} \\ &= \begin{Bmatrix} 0, & x < 1-10t \\ 0.5, & 1-10t \le x < 2-10t \\ 0, & 2-10t \le x < 1+10t \\ 0.5, & 1+10t \le x < 2+10t \\ 0, & x \ge 2+10t \end{Bmatrix}. \bigcirc \end{aligned}$$

The jumps in the initial condition $y(x, 0) = \Phi(x)$ remain in the solution at time $t \neq 0$. This means that our concept of solution should include the possibility that the PDE is not satisfied at some isolated x in the interval $-\infty < x < \infty$.

Also, unlike what happens with Fourier series, the d'Alembert solution does not use an average at a jump.

10.5.2 Writing the Solution Using Step Functions

For the purpose of using a calculator or computer, it can be useful to use step functions to express the solution of Example 10.14, that is, $y(x,t) = \frac{1}{2}(\Phi(x+10t) + \Phi(x-10t))$ in (10.84). To do that, we write the "pulse" $\Phi(x)$ in terms of step functions:

$$\Phi(x) = \begin{Bmatrix} 0, & x < 1 \\ 1, & 1 \le x < 2 \\ 0, & 2 \le x \end{Bmatrix} = \begin{Bmatrix} 0, & x < 1 \\ 1, & 1 \le x < 2 \\ 1-1, & 2 \le x \end{Bmatrix} = step(x-1) - step(x-2),$$

where we define the function $step(x)$ on the whole real line by

$$step(x) = \begin{Bmatrix} 0, & x < 0 \\ & \\ 1, & x \ge 0 \end{Bmatrix}.$$

Using this notation, Example 10.14's solution (10.84) can be expressed as

$$y(x,t) = \frac{1}{2}\left(step(x+10t-1) - step(x+10t-2) - step(x-10t-1) + step(x-10t-2)\right).$$

10.5.3 Support

There is another interesting way to look at Example 10.14's solution: $y(x,t)$ is the sum of $\frac{1}{2}\Phi(x+10t)$, a wave traveling to the left, and $\frac{1}{2}\Phi(x-10t)$, a wave traveling to the right. In Example 10.14, $\Phi(x)$ is zero outside of the interval $[1,2)$. The **support** of $\Phi(x)$ is defined to be the set of x for which $\Phi(x)$ is nonzero. The interval $[1,2)$ is the support of $\Phi(x)$. Similarly, the support of $\frac{1}{2}\Phi(x+10t)$ is the interval $\mathcal{I}_t^- = [1-10t, 2-10t)$, and the support of $\frac{1}{2}\Phi(x-10t)$ is the interval $\mathcal{I}_t^+ = [1+10t, 2+10t)$. In Figure 10.19, we see how the intervals $\mathcal{I}_t^{\pm}$ change as t changes. Each horizontal line $t = constant$ intersects Figure 10.19a

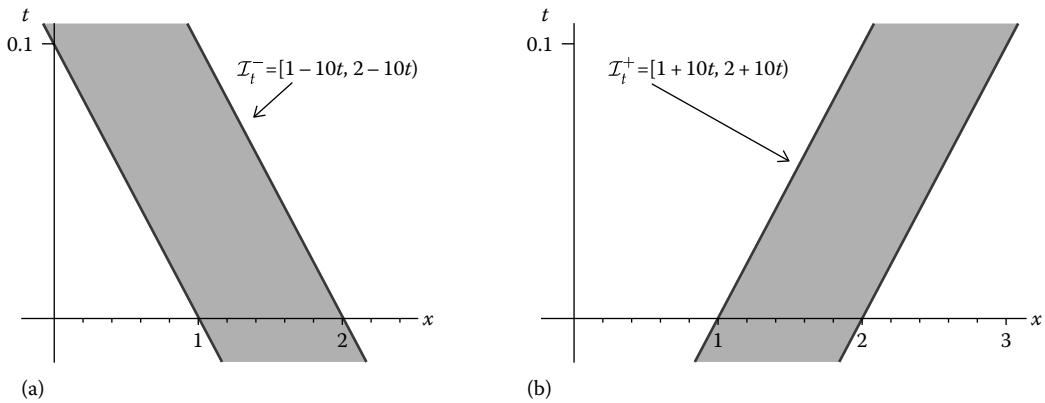


FIGURE 10.19
(a) $\mathcal{I}_t^-$ and (b) $\mathcal{I}_t^+$.

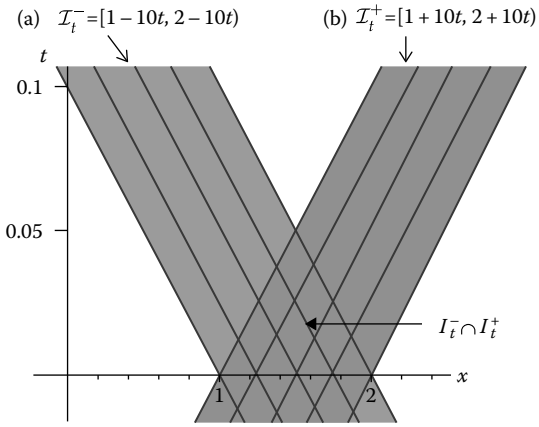


FIGURE 10.20
(a) $\mathcal{I}_t^-$ and (b) $\mathcal{I}_t^+$.

in the interval of x values in $\mathcal{I}_t^-$ and intersects Figure 10.19b in the interval of x values in $\mathcal{I}_t^-$.

We can combine Figure 10.19a and b to give Figure 10.20's picture of the support of $y(x,t) = \frac{1}{2}\,\Phi(x + 10t) + \frac{1}{2}\,\Phi(x - 10t)$.

10.5.4 Zero Initial Displacement

Example 10.15

Use (10.81) to solve the IVP consisting of the wave PDE (10.79) and the special case of the ICs being $y(x,0) \equiv 0$ and $\frac{\partial y}{\partial t}(x,0) = \Psi(x)$, both for $-\infty < x < \infty$.

Method: Using (10.81), that is, solutions in the form $y(x,t) = f(x + ct) + g(x - ct)$, we calculated in Example 10.13 that $\frac{\partial y}{\partial t}(x,t) = c\,f'(x + ct) - c\,g'(x - ct)$.

So, the ICs require that f and g satisfy the system of equations

$$\left\{ \begin{array}{c} 0 \equiv y(x,0) = f(x) + g(x) \\ \\ \Psi(x) = \frac{\partial y}{\partial t}(x,0) = c\,\left(f'(x) - g'(x)\right) \end{array} \right\}, -\infty < x < \infty. \tag{10.85}$$

The first equation in (10.85) implies that $g(x) \equiv -f(x)$; hence, $g'(x) \equiv -f'(x)$. Substituting that into the second equation in (10.85) implies that we need

$$\Psi(x) = 2c\,f'(x)$$

The fundamental theorem of calculus for functions of a single variable implies that

$$f(x) = c_1 + \frac{1}{2c}\int_0^x \Psi(\xi)\,d\xi,$$

for some constant c_1, and this implies that

$$g(x) = -c_1 - \frac{1}{2c}\int_0^x \Psi(\xi)\,d\xi.$$

So, the solution to the IVP is

$$y(x,t) = f(x+ct) + g(x-ct) = \cancel{c_1} + \frac{1}{2c}\int_0^{x+ct} \Psi(\xi)\,d\xi - \cancel{c_1} - \frac{1}{2c}\int_0^{x-ct} \Psi(\xi)\,d\xi,$$

that is,

$$y(x,t) = \frac{1}{2c}\int_0^{x+ct} \Psi(\xi)\,d\xi - \frac{1}{2c}\int_0^{x-ct} \Psi(\xi)\,d\xi = \frac{1}{2c}\int_0^{x+ct} \Psi(\xi)\,d\xi + \frac{1}{2c}\int_{x-ct}^{0} \Psi(\xi)\,d\xi.$$

The solution to this IVP in the special case that $y(x,0) \equiv 0$ is

$$y(x,t) = \frac{1}{2c}\int_{x-ct}^{x+ct} \Psi(\xi)\,d\xi. \bigcirc \tag{10.86}$$

We can combine the results of Examples 10.13 and 10.15 to get the next result.

Theorem 10.4

If $\Phi(x)$ and $\Psi(x)$ are piecewise continuous functions of x for $-\infty < x < \infty$, then d'Alembert solution of the IVP (10.79) and (10.80), that is,

$$\left\{ \begin{array}{l} \dfrac{\partial^2 y}{\partial t^2} = c^2 \dfrac{\partial^2 y}{\partial x^2} \\ \\ y(x,0) = \Phi(x) \\ \\ \dfrac{\partial y}{\partial t}(x,0) = \Psi(x) \end{array} \right\}, -\infty < x < \infty, \tag{10.87}$$

is given by

$$y(x,t) = \frac{1}{2}\,\Phi(x+ct) - \frac{1}{2}\,\Phi(x-ct) + \frac{1}{2c}\int_{x-ct}^{x+ct} \Psi(\xi)\,d\xi. \tag{10.88}$$

Why? Denote by $y_1(x,t) = \frac{1}{2}\,\Phi(x+ct) - \frac{1}{2}\,\Phi(x-ct)$ and $y_2(x,t) = \frac{1}{2c}\int_{x-ct}^{x+ct} \Psi(\xi)\,d\xi$, the solutions of Examples 10.13 and 10.15, respectively. Because PDE (10.79) is linear, $y(x,t) = y_1(x,t) + y_2(x,t)$ also satisfies (10.79). Also, $y(x,t)$ satisfies the ICs (10.80) because

$$y(x,0) = y_1(x,0) + y_2(x,0) = \Phi(x) + 0 = \Phi(x)$$

and

$$\frac{\partial y}{\partial t}(x,0) = \frac{\partial y_1}{\partial t}(x,0) + \frac{\partial y_2}{\partial t}(x,0) = 0 + \Psi(x) = \Psi(x). \square$$

Example 10.16

Solve, as explicitly as possible, the IVP consisting of ICs $y(x,0) \equiv 0$ and $\frac{\partial y}{\partial t}(x,0) \triangleq \begin{Bmatrix} \sin x, & 0 \le x < \pi \\ 0, & \text{all other } x \end{Bmatrix}$ for $t > 0$. Assume the wave speed is $c = 3\pi$ m/s. Sketch the solution $y(x,t)$ versus x at three interesting, positive times t of your choosing.

Method: By Theorem 10.4, the solution is

$$y(x,t) = \frac{1}{2c} \int_{x-ct}^{x+ct} \Psi(\xi)\, d\xi,$$

where $\Psi(x) = \frac{\partial y}{\partial t}(x,0)$. Unfortunately, it is not true that $y(x,t) = \frac{1}{2c} \int_{x-ct}^{x+ct} 1\, d\xi$.

Instead, using $c = 3\pi$, we can write

$$\begin{aligned} y(x,t) &= \int_{x-3\pi t \le \xi \le x+3\pi t} \Psi(\xi)\, d\xi \\ &= \int_{x-3\pi t \le \xi \le x+3\pi t \text{ and } 0 \le \xi < \pi} 1\, d\xi + \int_{x-3\pi t \le \xi \le x+3\pi t \text{ and } (\xi < 0 \text{ or } \xi \ge \pi)} 0\, d\xi, \end{aligned}$$

that is,

$$y(x,t) = \int_{[x-3\pi t, x+3\pi t] \cap [0,\pi)} \sin \xi\, d\xi. \tag{10.89}$$

The **intersection** symbol, $\cap$, says that the integral in (10.89) is performed over only those ξ that are in the *overlap* of the interval $[x - 3\pi t, x + 3\pi t]$ and the interval $[0, \pi)$.

Formula (10.89) looks nice and compact but to use it requires further analysis! We need to study the overlap of the intervals $[x - 3\pi t, x + 3\pi t]$ and the interval $[0, \pi)$. For any fixed $t > 0$, define the interval

$$J_x \triangleq [x - 3\pi\, t, x + 3\pi\, t] = \{\xi : x - 3\pi t \le \xi \le x + 3\pi\, t\}.$$

For any fixed $t > 0$, the interval J_x "slides" along the ξ real line as x varies. In studying the intersection $[x - 3\pi t, x + 3\pi t] \cap [0, \pi)$, ξ must also be located in the interval $[0, \pi)$, which does not vary as x varies.

We will study that intersection by letting x increase and thus sliding J_x along until it overlaps the interval $[0, \pi)$ and then slides past the interval $[0, \pi)$. It's as if J_x is a pulse of material that moves in relation to the fixed interval $[0, \pi)$. The latter interval is the support of the initial velocity datum $\Psi(x)$.

Illustrated in Figure 10.21 are the six and only possible relationships the interval J_x can have with the interval $[0, \pi)$:

1. J_x is to the left of $[0, \pi)$.
2. Part of J_x is to the left of $[0, \pi)$, and part of J_x is contained in, but unequal to, $[0, \pi)$.

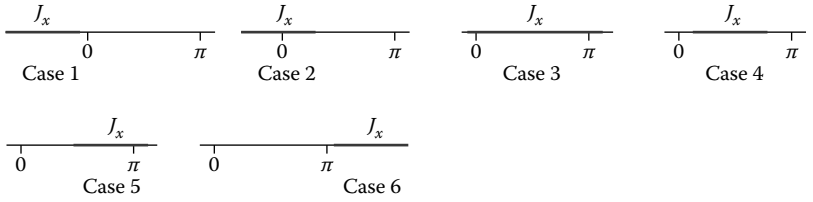


FIGURE 10.21
Interval cases.

3. Part of J_x is to the left of $[0, \pi)$, all of $[0, \pi)$ is contained in J_x, and part of J_x is to the right of $[0, \pi)$.
4. All of J_x is contained in, but unequal to, $[0, \pi)$.
5. Part of J_x is contained in $[0, \pi)$ and part of J_x is to the right of $[0, \pi)$.
6. J_x is to the right of $[0, \pi)$.

Let us analyze if and when each of those six cases occurs, *for* $t > 0$: cases (3) and (4) have a "transitional" nature and will turn out to be useful to study first.

Case (3) can occur *only* if the length of the interval J_x is greater than or equal to π, and, conversely, case (4) can occur *only* if the length of J_x is less than π. The length of $J_x = [x - 3\pi t, x + 3\pi t]$ is $6\pi t$. So, case (3) can occur only if $t \geq \frac{1}{6}$, and case (4) can occur only if $0 < t < \frac{1}{6}$. So, it makes sense to consider separately the cases (a) $t \geq \frac{1}{6}$ and (b) $0 < t < \frac{1}{6}$.

(a) First, suppose $\mathbf{0 < t < \frac{1}{6}}$. Case (1) occurs when the right endpoint of the interval J_x is to the left of the interval $[0, \pi)$, that is, when $x + 3\pi\, t < 0$, that is, when $x < -3\pi t$.

Case (2) occurs when J_x "straddles" $[0, \pi)$ "on the left," specifically the right endpoint of J_x is at or to the right of the left endpoint of the interval $[0, \pi)$ *and* the left endpoint of J_x is to the left of $[0, \pi)$. This happens when both $x + 3\pi t \geq 0$ and $x - 3\pi t < 0$, that is, when $x \geq -3\pi t$ and $x < 3\pi t$, that is, when $-3\pi t \leq x < 3\pi t$.

Case (3) cannot occur because we supposed that $0 < t < \frac{1}{6}$.

Case (4) occurs when the left endpoint of J_x is at, or to the right of, 0 *and* the right endpoint of J_x is to the left of π, that is, when $x - 3\pi t \geq 0$ and $x + 3\pi t < \pi$, that is, when $x \geq 3\pi t$ and $x < (1 - 3t)\pi$, that is, when $3\pi t \leq x < (1 - 3t)\pi$.

Case (5) occurs when J_x straddles $[0, \pi)$ on the right, specifically when (i) the right endpoint of J_x is to the right of the interval $[0, \pi)$ *and* (ii) the left endpoint of J_x is in the interval $[0, \pi)$. This happens when both (i) $x + 3\pi t \geq \pi$ and (ii) $0 \leq x - 3\pi t < \pi$, that is, when (i) $x \geq \pi - 3\pi t$ and (ii)(A) $x \geq 3\pi t$ and (ii)(B) $x < \pi + 3\pi t$. Of the inequalities (i) $x \geq (1 - 3t)\pi$ and (ii)(A) $x \geq 3\pi t$, (i) is more demanding because $(1 - 3t)\pi > 3t\pi$ follows from the assumption in part (a) that $0 < t < \frac{1}{6}$. So, case (5) occurs when (i) $x \geq (1 - 3t)\pi$ and (ii)(B) $x < (1 + 3t)\pi$, that is, when $(1 - 3t)\pi \leq x < (1 + 3t)\pi$.

Case (6) occurs when the left endpoint of J_x is to the right of the interval $[0, \pi)$, that is, when $x - 3\pi t \geq \pi$, that is, when $x \geq (1 + 3t)\pi$.

Integrating

Now that we understand the cases for $0 < t < \frac{1}{6}$, we can evaluate $y(x, t)$, as given in (10.89):

1. For $x < -3\pi t$, $J_x \cap [0, \pi) = \varnothing$, the **empty set**, so $y(x, t) = 0$. [The integral over nothing is nothing.]

2. For $-3\pi t \le x < 3\pi t$, the part of the interval J_x that overlaps $[0,\pi)$ is $[0, x+3\pi t]$, so

$$y(x,t) = \int\limits_{J_x \cap [0,\pi)} \sin\xi \, d\xi = \int\limits_{0}^{x+3\pi t} \sin\xi \, d\xi = \Big[-\cos\xi \Big]_0^{x+3\pi t} = 1 - \cos(x+3\pi t).$$

3. Case 3 cannot occur if $0 < t < \frac{1}{6}$.
4. For $3\pi t \le x < (1-3t)\pi$, all of J_x lies inside $[0,\pi)$. So, $J_x \cap [0,\pi) = J_x$ and thus using the difference of cosines formula, which we saw in Section 4.2 when discussing the beats phenomenon,

$$\begin{aligned} y(x,t) &= \int\limits_{J_x \cap [0,\pi)} \sin\xi \, d\xi = \int\limits_{x-3\pi t}^{x+3\pi t} \sin\xi \, d\xi = \Big[-\cos\xi \Big]_{x-3\pi t}^{x+3\pi t} \\ &= -\cos(x+3\pi t) + \cos(x-3\pi t) \\ &= 2\sin\left(\frac{x+3\pi t-(x-3\pi t)}{2}\right)\sin\left(\frac{x+3\pi t+(x-3\pi t)}{2}\right) = 2\sin(3\pi t)\sin x. \end{aligned}$$

5. For $(1-3t)\pi \le x < (1+3t)\pi$, the part of the interval J_x that overlaps $[0,\pi)$ is $[x-3\pi t, \pi]$, so

$$y(x,t) = \int\limits_{J_x \cap [0,\pi)} \sin\xi \, d\xi = \int\limits_{x-3\pi t}^{\pi} \sin\xi \, d\xi = \Big[-\cos\xi \Big]_{x-3\pi t}^{\pi} = 1 + \cos(x-3\pi t).$$

6. For $x \ge (1+3t)\pi$, again $J_x \cap [0,\pi) = \varnothing$, so $y(x,t) = 0$.

Putting all of the cases that can occur together, for $0 < t < \frac{1}{6}$,

$$y(x,t) = \begin{cases} 0, & x < -3\pi t \\ 1-\cos(x+3\pi t), & -3\pi t \le x < 3\pi t \\ 2\sin(3\pi t)\sin x, & 3\pi t \le x < (1-3t)\pi \\ 1+\cos(x-3\pi t), & (1-3t)\pi \le x < (1+3t)\pi \\ 0, & x \ge (1+3t)\pi \end{cases}. \tag{10.90}$$

(b) Second, suppose $\mathbf{t \ge \frac{1}{6}}$. It makes sense to begin by studying the "transitional" case that can occur that couldn't occur in part (a).

Case (3) occurs when part of J_x is to the left of $[0,\pi)$, all of $[0,\pi)$ is contained in J_x, *and* part of J_x is to the right of $[0,\pi)$. This occurs when the left endpoint of J_x, that is, $x-3\pi t$, is to the left of 0 and the right endpoint of J_x, that is, $x+3\pi t$, is to the right of π. This case occurs when both $x-3\pi t < 0$ and $x+3\pi t \ge \pi$, that is, when $x < 3\pi t$ and $x \ge \pi - 3\pi t$. So, case (3) occurs when $(1-3t)\pi \le x < 3\pi t$.

Case (1) occurs when J_x is to the left of $[0,\pi)$, that is, when $x+3\pi t<0$. So, in part (b) as in part (a), case (1) occurs when $x<-3\pi t$.

Case (2) occurs when J_x straddles $[0,\pi)$ on the left, specifically when (i) the left endpoint of J_x is to the left of $[0,\pi)$ *and* (ii) the right endpoint of J_x is in the interval $[0,\pi)$. This happens when (i) $x-3\pi t<0$ and (ii) $0\le x+3\pi t<\pi$, that is, (i) $x<3\pi t$ and (ii)(A) $-3\pi t\le x$ and (ii)(B) $x<\pi-3\pi t$. Of these three inequalities, (i) requires $x<3\pi t$, and (ii)(B) requires $x<(1-3t)\pi$; but because we assumed that $t\ge\frac{1}{6}$, $(1-3t)<0$, so (ii)(B) implies $x<0$, which is more demanding than the requirement of (i) that $x<3\pi t$. So, case (2) occurs when both $-3\pi t\le x$ and $x<(1-3t)\pi$, that is, when $-3\pi t\le x<(1-3t)\pi$.

Case (4) cannot occur because we supposed that $t\ge\frac{1}{6}$.

Case (5) occurs when the left endpoint of J_x is in $[0,\pi)$ and the right endpoint of J_x is to the right of π. So, this requires (i) $0\le x-3\pi t<\pi$ and (ii) $x+3\pi t\ge\pi$, that is, when (i)(A) $x\ge 3\pi t$, and (i)(B) $x<(1+3t)\pi$, and (ii) $x\ge(1-3t)\pi t$. We are assuming in part (b) that $t\ge\frac{1}{6}$, so of the two inequalities (i)(A) $x\ge 3\pi t$ and (ii) $x\ge(1-3t)\pi t$, the more demanding requirement is that $x\ge 3\pi t$. So, case (5) occurs when $3\pi t\le x<(1+3t)\pi$.

The analysis of where case (6) occurs is the same as for part (a): case (6) occurs when $x\ge(1+3t)\pi$.

Integrating

Now that we understand the cases for $t\ge\frac{1}{6}$, we can evaluate $y(x,t)$, as given in (10.89):

1. For $x<-3\pi t$, $J_x\cap[0,\pi)=\varnothing$, the **empty set**, so $y(x,t)=0$.
2. For $-3\pi t\le x<(1-3t)\pi$, the part of the interval J_x that overlaps $[0,\pi)$ is $[0,x+3\pi t]$, so

$$y(x,t)=\int\limits_{J_x\cap[0,\pi)}\sin\xi\,d\xi=\int\limits_0^{x+3\pi t}\sin\xi\,d\xi=\Big[-\cos\xi\Big]_0^{x+3\pi t}=1-\cos(x+3\pi t).$$

3. For $(1-3t)\pi\le x<3\pi t$, all of $[0,\pi)$ lies inside J_x, so $J_x\cap[0,\pi)=[0,\pi)$, and thus,

$$y(x,t)=\int\limits_{J_x\cap[0,\pi)}\sin\xi\,d\xi=\int\limits_0^{\pi}\sin\xi\,d\xi=\Big[-\cos\xi\Big]_0^{\pi}=2.$$

4. Case 4 cannot occur if $t\ge\frac{1}{6}$.
5. For $3\pi t\le x<(1+3t)\pi$, the part of the interval J_x that overlaps $[0,\pi)$ is $[x-3\pi t,\pi)$, so

$$y(x,t)=\int\limits_{J_x\cap[0,\pi)}\sin\xi\,d\xi=\int\limits_{x-3\pi t}^{\pi}\sin\xi\,d\xi=\Big[-\cos\xi\Big]_{x-3\pi t}^{\pi}=1+\cos(x-3\pi t).$$

6. For $x\ge(1+3t)\pi$, again $J_x\cap[0,\pi)=\varnothing$ so $y(x,t)=0$.

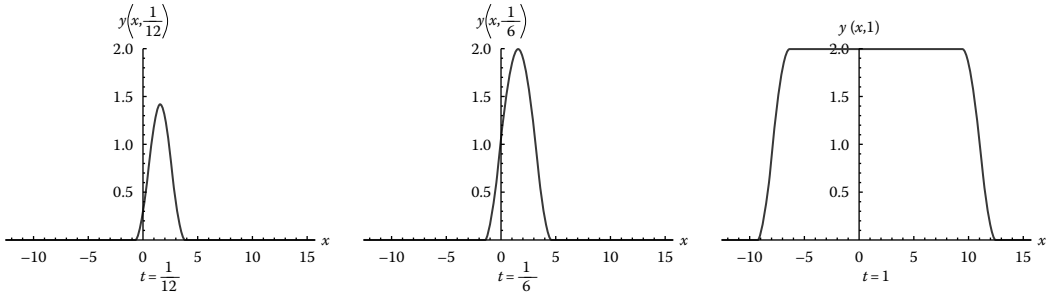


FIGURE 10.22
$y\left(x, \frac{1}{12}\right)$, $y\left(x, \frac{1}{6}\right)$, and $y(x, 1)$.

Putting all of the cases that can occur together, for $t \geq \frac{1}{6}$,

$$y(x,t) = \begin{cases} 0, & x < -3\pi t \\ 1 - \cos(x + 3\pi t), & -3\pi t \leq x < (1 - 3t)\pi \\ 2, & (1 - 3t)\pi \leq x < 3\pi t \\ 1 + \cos(x - 3\pi t), & 3\pi t \leq x < (1 + 3t)\pi \\ 0, & x \geq (1 + 3t)\pi \end{cases} . \bigcirc \tag{10.91}$$

The graphs of the solution $y(x,t)$ versus x are given in Figure 10.22 for times $t = \frac{1}{12}, \frac{1}{6}, 1$. The graphs for those values of t come from, respectively, (10.90), (10.91), and (10.91).

If we ask for the graph of $y(x,t)$ versus x for only one particular value of t, then we wouldn't have to do all of the work as in Example 10.16; instead, we would first decide which of case (a) or (b) is appropriate for that particular value of t.

10.5.5 Problems

1. An infinite string with linear mass density 0.01 kg/m has a tension of 4 N. It has an initial displacement of $\Phi(x)$ m, which is sketched in Figure 10.23 for $-2 \leq x < 2$ and which is zero for all other x. The string is released from rest with that

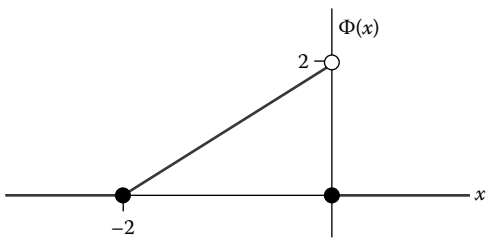


FIGURE 10.23
Problem 10.5.5.1.

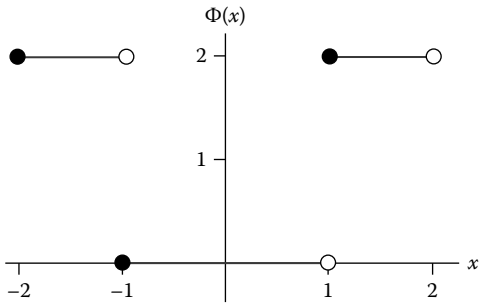


FIGURE 10.24
Problem 10.5.5.2.

displacement. (a) Determine a formula for the subsequent motion of the string in as explicit a form as possible, (b) sketch the graph of $u(x, 0.025)$, and (c) sketch the graph of $u(x, 0.05)$.

2. An infinite string with linear mass density 0.01 kg/m has a tension of 4 N. It has an initial displacement of $\Phi(x)$ m, which is sketched in Figure 10.24 for $-2 \le x < 2$ and which is zero for all other x. The string is released from rest with that displacement. (a) Determine a formula for the subsequent motion of the string in as explicit a form as possible, (b) sketch the graph of $u(x, 0.025)$, and (c) sketch the graph of $u(x, 0.05)$.

3. An infinite string satisfies $\dfrac{\partial^2 u}{\partial t^2} = 100\dfrac{\partial^2 u}{\partial x^2}$, $\dfrac{\partial u}{\partial t}(x, 0) \equiv 0$, and

$$u(x,0) = \begin{Bmatrix} 0, & x < -1 \\ x+1, & -1 \le x < 0 \\ 1, & 0 \le x < 1 \\ 0, & 1 \le x \end{Bmatrix}.$$

Sketch the graph of $u(x, 0.07)$.

4. An infinite string with linear mass density 0.1 kg/m is held taut by a force of 10 N. Its initial displacement is

$$u(x,0) = \begin{Bmatrix} 0, & x < -1 \\ 0.2, & -1 \le x < 0 \\ 0.2(1-x), & 0 \le x < 1 \\ 0, & 1 \le x \end{Bmatrix}$$

and is released from rest. Find a formula for the subsequent motion of the string in as explicit a form as possible, and sketch the solution at $t = 0.05$.

5. An infinite string with wave speed $c = 20\,\text{m/s}$ is given initial velocity

$$\Psi(x) = \begin{cases} 20\,\text{m/s}, & 0 \le x < 2 \\ 0, & \text{all other } x \end{cases}$$

and has zero initial displacement. Find (a) $u(x, 0.025)$ and (b) $u(x,t)$ in as explicit a form as possible.

6. An infinite string with wave speed $c = 20\,\text{m/s}$ is given initial velocity

$$\Psi(x) = \begin{cases} 4\,\text{m/s}, & 0 \le x < 1 \\ 0, & \text{all other } x \end{cases}$$

and has zero initial displacement. Find (a) $u(x, 0.025)$ and (b) $u(x,t)$ in as explicit a form as possible.

7. An infinite string with wave speed $c = 20\,\text{m/s}$ is given initial velocity

$$\Psi(x) = \begin{cases} 20\,\text{m/s}, & 2 \le x < 3 \\ 0, & \text{all other } x \end{cases}$$

and has zero initial displacement. Find (a) $u(x, 0.025)$ and (b) $u(x,t)$ in as explicit a form as possible.

8. Look for **plane wave** solutions of the form $u(x,y,t) = f(k_1 x + k_2 y - ct)$ for the wave equation in two space dimensions

$$\frac{\partial^2 u}{\partial t^2} = c^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right).$$

Find an equation that the pair (k_1, k_2) must satisfy. By the way, $\sqrt{k_1^2 + k_2^2}$ is called the **wave number**.

10.6 Short Take: Conservation of Energy in a Finite String

If $u(x,t)$ satisfies the PDE of a vibrating string

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2},\ 0 < x < L,\ t > 0, \tag{10.92}$$

where $c^2 = T_0/\varrho_0$, on the interval $0 < x < L$, *and* the BCs

$$u(0,t) = u(L,t) = 0,\ t > 0, \tag{10.93}$$

then we will explain why the quantity $E(t)$ defined by

$$E(t) \triangleq \int_0^L \frac{1}{2}\left(T_0\left(\frac{\partial u}{\partial x}\right)^2 + \varrho_0\left(\frac{\partial u}{\partial t}\right)^2\right) dx \tag{10.94}$$

is, in fact, constant. The natural physical interpretation of $\int_0^L \frac{1}{2}\varrho_0\left(\frac{\partial u}{\partial t}\right)^2 dx$ is that of total kinetic energy in the string, and a somewhat (but not entirely) natural physical interpretation of $\int_0^L \frac{1}{2}T_0\left(\frac{\partial u}{\partial x}\right)^2 dx$ is that of the total potential energy in the string. So, if we can explain why $E(t)$ is constant, we will have explained why the principle of conservation of energy holds true for this mathematical model.

We calculate that

$$\frac{dE}{dt} = \int_0^L \frac{1}{2}\frac{\partial}{\partial t}\left(T_0\left(\frac{\partial u}{\partial x}\right)^2 + \varrho_0\left(\frac{\partial u}{\partial t}\right)^2\right) dx = \int_0^L T_0\,\frac{\partial u}{\partial x}\,\frac{\partial^2 u}{\partial x \partial t}\,dx + \int_0^L \varrho_0\,\frac{\partial u}{\partial t}\,\frac{\partial^2 u}{\partial t^2}\,dx.$$

Using integration by parts in the first term and using the PDE in the second term gives

$$\frac{dE}{dt} = \left[T_0\,\frac{\partial u}{\partial x}\,\frac{\partial u}{\partial t}\right]_0^L - \int_0^L T_0\frac{\partial^2 u}{\partial x^2}\,\frac{\partial u}{\partial t}\,dx + \int_0^L \varrho_0\frac{\partial u}{\partial t}\,c^2\,\frac{\partial^2 u}{\partial x^2}\,dx.$$

Regarding the first terms, that is, $\frac{\partial u}{\partial x}(L,t)\frac{\partial u}{\partial t}(L,t) - \frac{\partial u}{\partial x}(0,t)\frac{\partial u}{\partial t}(0,t)$, the BC $u(0,t) = 0$ for all $t > 0$ implies that $\frac{\partial u}{\partial t}(0,t) = 0$ for all $t > 0$.* Similarly $u(L,t) = 0$ for all $t > 0$ implies that $\frac{\partial u}{\partial t}(L,t) = 0$. It follows that the first term is zero, so

$$\frac{dE}{dt} = -\int_0^L T_0\frac{\partial^2 u}{\partial x^2}\,\frac{\partial u}{\partial t}\,dx + \int_0^L \varrho_0\frac{\partial u}{\partial t}\,c^2\,\frac{\partial^2 u}{\partial x^2}\,dx,$$

which equals zero because $c^2 = T_0/\varrho_0$.

So, the principle of conservation of energy holds true for this PDE BVP. Notice that the initial conditions were never used in establishing the conservation of energy. In fact, using the initial conditions $u(x,0) = f(x),\ \frac{\partial u}{\partial t}(x,0) = g(x),\ 0 < x < L$, we see that

* This makes sense but requires some advanced mathematical justification concerning functions of two variables $u(x,t)$, specifically why $\lim_{x\to 0+}\left(\frac{\partial u}{\partial t}(x,t)\right) = \frac{\partial}{\partial t}\left(\lim_{x\to 0+} u(x,t)\right)$.

$$E(t) = E(0) = \int_0^L \frac{1}{2}\left(T_0\left(\frac{df}{dx}\right)^2 + \varrho_0\left(g(x)\right)^2\right)dx$$

for all $t > 0$.

While the total energy, E, is constant, the same cannot be said for either the kinetic energy, KE, or the potential energy, PE, in the string. But, we will see that both KE and PE are periodic functions of time, t.

In Section 11.2, specifically (11.25), we will see that the general solution of the PDE BVP is of the form

$$u(x,t) = \sum_{n=1}^{\infty}\left(a_n \cos\left(\frac{n\pi ct}{L}\right) + b_n \sin\left(\frac{n\pi ct}{L}\right)\right)\sin\left(\frac{n\pi x}{L}\right).$$

We assume we can move partial derivatives inside infinite sums to get

$$\frac{\partial u}{\partial t}(x,t) = \sum_{n=1}^{\infty}\left(\left(\frac{n\pi c}{L}\right)\left(-a_n \sin\left(\frac{n\pi ct}{L}\right) + b_n \cos\left(\frac{n\pi ct}{L}\right)\right)\right)\sin\left(\frac{n\pi x}{L}\right)$$

and

$$\frac{\partial u}{\partial x}(x,t) = \sum_{n=1}^{\infty}\left(\left(\frac{n\pi}{L}\right)\left(a_n \cos\left(\frac{n\pi ct}{L}\right) + b_n \sin\left(\frac{n\pi ct}{L}\right)\right)\right)\cos\left(\frac{n\pi x}{L}\right).$$

So, by Parseval's Theorem 9.20 in Section 9.8, denoting

$$B_n(t) \triangleq \frac{n\pi c}{L}\left(-a_n \sin\left(\frac{n\pi ct}{L}\right) + b_n \cos\left(\frac{n\pi ct}{L}\right)\right),$$

$$KE(t) \triangleq \int_0^L \frac{1}{2}\varrho_0\left(\frac{\partial u}{\partial t}\right)^2 dx = \frac{\varrho_0}{2}\cdot\frac{L}{2}\sum_{n=1}(B_n(t))^2$$

$$= \frac{\varrho_0 L}{4}\sum_{n=1}^{\infty}\left(\frac{n\pi c}{L}\right)^2\left(-a_n \sin\left(\frac{n\pi ct}{L}\right) + b_n \cos\left(\frac{n\pi ct}{L}\right)\right)^2.$$

Similarly, by Parseval's Theorem 9.21 in Section 9.8,

$$PE(t) \triangleq \int_0^L \frac{1}{2}T_0\left(\frac{\partial u}{\partial x}\right)^2 dx = \frac{T_0 L}{4}\sum_{n=1}^{\infty}\left(\frac{n\pi}{L}\right)^2\left(a_n \cos\left(\frac{n\pi ct}{L}\right) + b_n \sin\left(\frac{n\pi ct}{L}\right)\right)^2.$$

Incidentally, these calculations offer an alternative way to verify that conservation of energy holds for this PDE BVP, by explaining why the sum

$$\varrho_0 c^2 \left(-a_n \sin\left(\frac{n\pi ct}{L}\right) + b_n \cos\left(\frac{n\pi ct}{L}\right)\right)^2 + T_0 \left(a_n \cos\left(\frac{n\pi ct}{L}\right) + b_n \sin\left(\frac{n\pi ct}{L}\right)\right)^2$$

is constant.

By the way, at URL

⟨http://proxima.math.wright.edu/People/Larry_Turyn/strings.html⟩

you will find a brief movie showing a vibration of a plucked string and the potential energy in the string as a function of time. (It's an old web page and those are the only functioning parts of the page.)

10.6.1 Problems

1. Let T be the period of $PE(t), KE(t)$ as a function of t. Use Parseval's theorem from Section 9.8 to explain why the average value of the potential energy, that is, $\overline{PE} = \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} PE(t)dt$, and the average value of the kinetic energy, $\overline{KE} = \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} KE(t)dt$, are equal, and find those average values.
2. [R. Haberman] Here's a model for the vibrations of a string whose right end is fixed but whose left end is attached to a mass on a spring and is thus free to move up or down: The same PDE (10.92) holds along with the BC

$$ku(0,t) - T_0 \frac{\partial u}{\partial x}(0,t) = 0,\ u(L,t) = 0,\ 0 < t < \infty,$$

   where k is the spring constant. [We assume that the mass of the spring itself is negligible.] Show there is conservation of total energy with the potential energy formula modified to include the potential energy of the left-end spring, that is, $KE(t) \triangleq \int_0^L \frac{1}{2}\varrho_0 \left(\frac{\partial u}{\partial t}\right)^2 dx$ and $PE(t) \triangleq \int_0^L \frac{1}{2}\varrho_0 \left(\frac{\partial u}{\partial x}\right)^2 dx + \frac{1}{2}k\,(u(0,t))^2$.
3. Use the trigonometric identity $\cos^2\theta + \sin^2\theta = 1$ to explain why the function of time given by (10.94) is constant.

Key Terms

adiabatic conditions: before (10.69) in Section 10.4
body forces: after (10.20) in Section 10.1
Cauchy equation: (10.24) in Section 10.1
composite rod: after Example 10.6 in Section 10.2
contact force: after (10.18) in Section 10.1, before (10.55) in Section 10.4
contact resistance: end of Section 10.2
continuum mechanics: after (10.15) in Section 10.1
d'Alembert wave solutions: (10.81) in Section 10.5

deformation, displacement: after (10.19) in Section 10.1
density of heat energy: Example 10.1 in Section 10.1
elastic: before (10.55) in Section 10.4
empty set: after Figure 10.21
energy density: Problem 10.4.6.10
equilibrium temperature: before (10.36) in Section 10.2
Fourier's law of heat conduction: (10.27) in Section 10.2
heat equation: (10.30) in Section 10.2
heat flux vector: Example 10.1 in Section 10.1
intersection symbol: after (10.89) in Section 10.5
irrotational: after (10.44) in Section 10.3
isotropic: before (10.74) in Section 10.4
Lagrangian description: after (10.15) in Section 10.1
Lamé constants: after (10.74) in Section 10.4
Laplace's equation: (10.36) in Section 10.2
linear elasticity: after (10.73) in Section 10.4
linear elastostatics: before (10.75) in Section 10.4
linearized strain tensor: after (10.71) in Section 10.4
local elongation: after (10.54) in Section 10.4
magnetostatics: before (10.46) in Section 10.3
Maxwell's equations of electromagnetism: before (10.6) in Section 10.1
momentum density: Problem 10.4.6.10
plane strain: Definition 10.2 in Section 10.4
plane wave: Problem 10.5.5.8
Poisson's equation: after (10.36) in Section 10.2
Poisson's ratio: after (10.76) in Section 10.4
potential equation: (10.36) in Section 10.2
potential flow: (10.44) in Section 10.3
principle of virtual work: Problem 10.4.6.11
principal strains: Definition 10.1 in Section 10.4
reference configuration: before (10.54) in Section 10.4
reference state: after (10.15) in Section 10.1, before (10.54) in Section 10.4
shear modulus: after (10.74) in Section 10.4
solvability condition: (10.50) in Section 10.3
specific heat: beginning of Section 10.2
speed of sound: after (10.68) in Section 10.4
steady-state temperature: before (10.36) in Section 10.2
stress tensor: before (10.19) in Section 10.1
stress vector: after (10.18) in Section 10.1
string: before (10.55) in Section 10.4
support: before Figure 10.19
temperature: beginning of Section 10.2
tension force: (10.55) in Section 10.4
thermal conductivity: (10.27) in Section 10.2
thermal diffusivity: after (10.29) in Section 10.2
tied at the ends: after (10.54) in Section 10.4
total linear momentum: after (10.19) in Section 10.1
traction vector: after (10.18) in Section 10.1
uniaxial stretching: before (10.75) in Section 10.4

vibrating string: before (10.54) in Section 10.4
vorticity: after (10.44) in Section 10.3
wave number: Problem 10.5.5.8
Young's modulus: before (10.77) in Section 10.4

Reference

Evans, L.C. *Partial Differential Equations*. American Mathematical Society, Providence, RI, 1998.

11

Separation of Variables for PDEs

11.1 Heat Equation in One Space Dimension

Here we begin using the "method of separation of variables" to solve models with partial differential equations (PDEs). Our presentation, based on eigenfunction expansions, is initially a little different from most books but uses the engineering style of standardization: Our method uses over and over again off-the-shelf components whose properties are well understood and then extended in their usefulness.

In Section 10.2, we studied models for heat flux and temperature in heat conduction, diffusion of substances, and propagation of long electromagnetic waves in a good conductor. We saw that integral forms of heat balance led to the heat equation in one or more dimensions, saw how the concept of heat flux is involved in some boundary conditions, and calculated equilibrium solutions for 1D problems.

The PDE–boundary value problem–initial value problem (PDE-BVP-IVP)

$$\frac{\partial T}{\partial t}(x,t) = \alpha \frac{\partial^2 T}{\partial x^2}(x,t),\ 0 < x < L,\ t > 0, \tag{11.1}$$

$$T(0,t) = T(L,t) = 0,\ t > 0, \tag{11.2}$$

$$T(x,0) = f(x),\ 0 < x < L, \tag{11.3}$$

will be solved by the separation of variables method discussed in the following. The region in the xt-plane, that is, $0 < x < L,\ t > 0$, is depicted in Figure 11.1. The two boundary conditions (11.2) are on the parallel sides $x = 0$ and $x = L$. First, we need some background knowledge.

The ordinary differential equation–boundary value problem (ODE-BVP)

$$\left\{ \begin{array}{l} X''(x) + \lambda X(x) = 0, \\ \\ X(0) = X(L) = 0 \end{array} \right\} \tag{11.4}$$

has infinitely many solutions for the *eigenvalues* $\lambda_n = \left(\frac{n\pi}{L}\right)^2,\ n = 1, 2, \ldots$, and corresponding *eigenfunctions* $X_n(x) = \sin\left(\frac{n\pi x}{L}\right),\ n = 1, 2, ...$; for all other values of λ, (11.4) has only the trivial solution, that is, $X(x) \equiv 0$. We justified these conclusions in Example 9.14 in Section 9.3.

Notice that the boundary conditions in (11.4) are the same as the boundary conditions (11.2) applied to a function $X(x)$ rather than to a function of two variables such as $T(x,t)$.

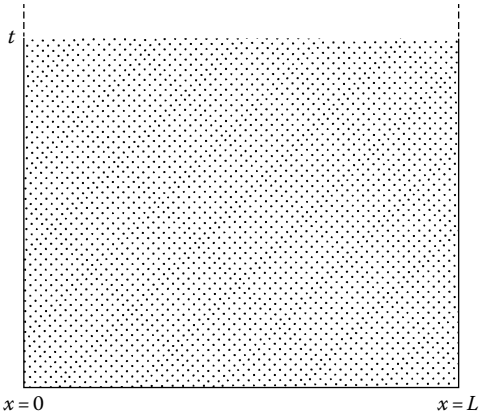


FIGURE 11.1
The region $0 < x < L,\ t > 0$ in the xt-plane.

With that background knowledge, we can solve (11.1) through (11.3): Because of the "homogeneous boundary conditions on parallel sides," that is, $T(0,t) = T(L,t) = 0, t > 0$, we can imagine there are "product solutions" of the PDE in the *assumed* form $T(x,t) = X(x)G(t)$, where X solves the ODE-BVP (11.4). Because $X(x)$ satisfies the boundary conditions $X(0) = X(L) = 0$, such a product solution $T(x,t)$ automatically satisfies the boundary conditions $T(0,t) = T(L,t) = 0$. By the **method of separation of variables**, we mean this technique of using product solutions in which one (or more) of the function factors satisfies a pair of boundary conditions on parallel sides.

Let $X = X_n(x) = \sin\left(\frac{n\pi x}{L}\right)$ and substitute $T(x,t) = \sin\left(\frac{n\pi x}{L}\right) G(t)$ into the PDE $\frac{\partial}{\partial t}T(x,t) = \alpha \frac{\partial^2}{\partial x^2}T(x,t)$ to get

$$\sin\left(\frac{n\pi x}{L}\right)\frac{dG}{dt} = -\alpha\left(\frac{n\pi}{L}\right)^2 \sin\left(\frac{n\pi x}{L}\right) G(t),\ n = 1,2,\ldots;$$

hence,

$$\frac{dG}{dt} = -\alpha\left(\frac{n\pi}{L}\right)^2 G(t),\ n = 1,2,\ldots\ .$$

This is the simplest kind of first-order constant coefficients ODE, the solutions being

$$G_n(t) = b_n e^{-\alpha\left(\frac{n\pi}{L}\right)^2 t},\ n = 1,2,\ldots,$$

where b_n is an arbitrary constant. The subscript n in the formulas indicates that there is dependence on n, for example, for each value of n, one might have a varying constant b_n. The product solutions are

$$T_n(x,t) = b_n \sin\left(\frac{n\pi x}{L}\right) e^{-\alpha\left(\frac{n\pi}{L}\right)^2 t},\ n = 1,2,\ldots.$$

What's left in solving the PDE-BVP-IVP? The initial condition is $T(x,0) = f(x),\ 0 < x < L$, that is, (11.3). The method is to use *all* of the product solutions by using the infinite series

$$T(x,t) = \sum_{n=1}^{\infty} T_n(x,t) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right) e^{-\alpha\left(\frac{n\pi}{L}\right)^2 t}. \tag{11.5}$$

This is similar to our method of using Fourier analysis to solve a forced oscillator problem in Example 9.13 in Section 9.2.

Substitute (11.5) into the initial condition (11.3), that is, $T(x,0) = f(x),\ 0 < x < L$, to get

$$f(x) = T(x,0) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right). \tag{11.6}$$

Fourier analysis of this sine series yields

$$b_n = \frac{2}{L}\int_0^L f(x) \sin\left(\frac{n\pi x}{L}\right) dx.$$

So, we can say that the solution is given by

$$T(x,t) = \sum_{n=1}^{\infty} \left(\frac{2}{L}\int_0^L f(x) \sin\left(\frac{n\pi x}{L}\right) dx \right) \sin\left(\frac{n\pi x}{L}\right) e^{-\alpha\left(\frac{n\pi}{L}\right)^2 t}.$$

Example 11.1

Solve the PDE-BVP-IVP

$$\left\{ \begin{array}{c} \frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2},\ 0 < x < 5,\ t > 0, \\ \\ T(0,t) = T(5,t) = 0,\ t > 0, \\ \\ T(x,0) = f(x) = \left\{ \begin{array}{ll} 56x, & 0 < x < \frac{5}{2} \\ \\ 280 - 56x, & \frac{5}{2} < x < 5 \end{array} \right\} \end{array} \right\}. \tag{11.7}$$

Method: This problem fits directly into the form of problem (11.1) through (11.3). Note especially that there are two homogeneous boundary conditions $T(0,t) = T(5,t) = 0$, so $L = 5$ in this problem and the solution of the PDE and BVP is

$$T(x,t) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{5}\right) e^{-\alpha\left(\frac{n\pi}{5}\right)^2 t}.$$

The initial condition is satisfied by solving $f(x) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{5}\right)$, $0 < x < 5$. Fourier sine series analysis implies that this is done by finding the coefficients:

$$b_n = \frac{2}{5}\int_0^5 f(x)\sin\left(\frac{n\pi x}{5}\right)dx = \frac{2}{5}\int_0^{5/2}(56x)\sin\left(\frac{n\pi x}{5}\right)dx + \frac{2}{5}\int_{5/2}^5 (280-56x)\sin\left(\frac{n\pi x}{5}\right)dx$$

$$= \frac{112}{5}\left(\left[\frac{-5x}{n\pi}\cos\left(\frac{n\pi x}{5}\right) + \frac{25}{(n\pi)^2}\sin\left(\frac{n\pi x}{5}\right)\right]_0^{\frac{5}{2}}\right.$$

$$\left. + \left[\frac{-5(5-x)}{n\pi}\cos\left(\frac{n\pi x}{5}\right) - \frac{25}{(n\pi)^2}\sin\left(\frac{n\pi x}{5}\right)\right]_{\frac{5}{2}}^5\right)$$

$$= \frac{112}{5}\left(\left(\frac{-25}{2n\pi}\cos\left(\frac{n\pi}{2}\right) + \frac{25}{(n\pi)^2}\sin\left(\frac{n\pi}{2}\right) + 0 - 0\right)\right.$$

$$\left. + \left(-0 - 0 + \frac{25}{2n\pi}\cos\left(\frac{n\pi}{2}\right) + \frac{25}{(n\pi)^2}\sin\left(\frac{n\pi}{2}\right)\right)\right)$$

$$= \frac{1120}{(n\pi)^2}\sin\left(\frac{n\pi}{2}\right).$$

So the solution is

$$T(x,t) = \frac{1120}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2}\sin\left(\frac{n\pi}{2}\right)\sin\left(\frac{n\pi x}{5}\right)e^{-\alpha\left(\frac{n\pi}{5}\right)^2 t}. \bigcirc \tag{11.8}$$

Notice that we *did* bother to evaluate the definite integrals and we *did* state a final conclusion giving the solution to the whole PDE-BVP-IVP.

For purposes of computation, one would only take a finite number of terms in the series (11.8). Also, half of the terms in the series are zero, because when n is even, $\sin\left(\frac{n\pi}{2}\right) = 0$. So, we need only to include the odd-indexed terms, that is, when $n = 2k-1$, $k =, 1, 2, \ldots$. We can rewrite (11.8), after noting that $\sin\left(\frac{(2k-1)\pi}{2}\right) = \sin(k\pi - \frac{\pi}{2}) = -\cos(k\pi) = -(-1)^k$:

$$T(x,t) = -\frac{1120}{\pi^2}\sum_{k=1}^{\infty}\frac{(-1)^k}{(2k-1)^2}\sin\left(\frac{(2k-1)\pi x}{5}\right)e^{-\alpha\left(\frac{(2k-1)\pi}{5}\right)^2 t}.$$

Of physical interest is the decay, as $t \to \infty$, of the solution $T(x,t)$. Each of the functions $e^{-\alpha\left(\frac{n\pi}{5}\right)^2 t} \to 0$ as $t \to \infty$. Recall that for an ODE, the time constant indicates how long it takes for a solution to decay to $\frac{1}{e}$ of its initial value. The time constant for a heat equation PDE is defined to be the reciprocal of the slowest decay rate r_{min}. Because $T(x,t)$ includes many different decaying exponential functions, "weighted" by $\left(\frac{2}{5}\left(\int_0^5 T(x,0)\sin\left(\frac{n\pi x}{5}\right)dx\right)\sin\left(\frac{n\pi x}{5}\right)\right)$, we *can't* guarantee that $T(x,\tau) = \frac{1}{e}T(x,0)$. Nevertheless, for physical intuition, it is still useful to think of the time constant as being how long it takes for the solution to decay in a standard way.

The function with $n=1$ has the slowest **rate of decay**, $r_{min} = \alpha \left(\frac{\pi}{5}\right)^2 = \frac{\alpha \pi^2}{25}$. The corresponding **time constant** is

$$\tau \triangleq \frac{1}{r_{min}} = \frac{25}{\alpha \pi^2}.$$

Example 11.2

A 2 m long steel rod, with thermal diffusivity $\alpha \approx 1.24 \times 10^{-5} \mathrm{m}^2/\mathrm{s}$ with insulated lateral surface, has its left end maintained at a temperature of 300°C and its right end maintained at 20°C. Determine the temperature as a function of x and t if the initial condition is given by

$$T(x,0) = f(x) = \begin{Bmatrix} 300, & 0 < x < \frac{5}{2} \\ 580 - 112x, & \frac{5}{2} < x < 5 \end{Bmatrix}. \tag{11.9}$$

Method: The physical problem translates to a mathematical problem, specifically the PDE-BVP-IVP

$$\left\{ \begin{array}{c} \frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2},\ 0 < x < 5,\ t > 0, \\ \\ T(0,t) = 300,\ T(5,t) = 20,\ t > 0, \\ \\ T(x,0) = \begin{Bmatrix} 300, & 0 < x < \frac{5}{2} \\ 580 - 112x, & \frac{5}{2} < x < 5 \end{Bmatrix} \end{array} \right\}. \tag{11.10}$$

Unfortunately, this problem does *not* directly fit into the form of problem (11.1) through (11.3) because of the *inhomogeneous* boundary conditions $T(0,t) = 300$ and $T(5,t) = 20$. But we will see that we can find the solution of this problem by a two-step process, and the second step will turn out to be Example 11.1.

It turns out that the correct form of the solution is not simply the infinite series (11.5) but instead the sum of a *equilibrium solution* plus the infinite series (11.5). To be explicit, the solution will be in the form

$$T(x,t) = v(x) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{5}\right) e^{-\alpha\left(\frac{n\pi}{5}\right)^2 t}. \tag{11.11}$$

As we saw in Chapter 10, the problem in (11.10), the equilibrium solution, $v(x)$, satisfies $\frac{\partial v}{\partial t} \equiv 0$, and we want it to *also* satisfy the PDE in (11.10); hence, $0 = \alpha \frac{\partial^2 v}{\partial x^2}$. Because v is a function only of x, the partial derivatives are really ordinary derivatives, so $v(x)$ should satisfy ODE $0 = \alpha\, v''(x)$. The equilibrium solution of (11.10) should satisfy also the boundary conditions (BCs) $v(0) = 300,\ v(5) = 20$. So, altogether, $v(x)$ should satisfy

$$\left\{ \begin{array}{l} 0 = \alpha\, v''(x),\ 0 < x < 5, \\ \\ v(0) = 300,\ v(5) = 20. \end{array} \right\}. \tag{11.12}$$

Before solving the inhomogeneous ODE-BVP (11.12), let's see what it will accomplish for us: If $v(x)$ solves (11.12) and $T(x,t) = v(x) + w(x,t)$, that is, $w(x,t) = T(x,t) - v(x)$, is to solve the original PDE in (11.10), then applying the BCs to $T(x,t)$ gives

$$w(0,t) = T(0,t) - v(0) = 300 - 300 = 0 \quad \text{and}$$

$$w(5,t) = T(5,t) - v(5) = 20 - 20 = 0.$$

Because $v(x)$ will solve the linear homogeneous PDE (11.1) and $T(x,t)$ should solve the same linear homogeneous PDE, $w = T - v$ must solve it, too. So, putting this all together implies that $w(x,t)$ should satisfy the PDE-BVP

$$\left\{ \begin{array}{l} \frac{\partial w}{\partial t} = \alpha \frac{\partial^2 w}{\partial x^2},\ 0 < x < 5, t > 0 \\ \\ w(0,t) = 0,\ w(5,t) = 0,\ t > 0 \end{array} \right\}.$$

This explains why the solution of the original problem (11.10) should be in the form (11.11).

To solve (11.12), first note that the general solution of $0 = v''$ is given by $v = c_1 + c_2 x$, for some constants c_1, c_2. Substituting this into the boundary conditions in system (11.12) gives $300 = v(0) = c_1, 20 = v(2) = c_1 + c_2 5 = 300 + c_2 5$; hence, $-280 = c_2 5$. So, $v(x) = 300 - 56x$.

The only thing left is to arrange that the solution of the PDE-BVP, that is,

$$T(x,t) = 300 - 56x + \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{5}\right) e^{-\alpha\left(\frac{n\pi}{5}\right)^2 t},$$

satisfies the initial condition $T(x,0) = \left\{ \begin{array}{ll} 300, & 0 < x < \frac{5}{2} \\ 580 - 112x, & \frac{5}{2} < x < 5 \end{array} \right\}$, that is,

$$300 - 56x + \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{5}\right) = \left\{ \begin{array}{ll} 300, & 0 < x < \frac{5}{2} \\ 580 - 112x, & \frac{5}{2} < x < 5 \end{array} \right\},$$

that is,

$$\sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{5}\right) = \left\{ \begin{array}{ll} 56x, & 0 < x < \frac{5}{2} \\ 280 - 56x, & \frac{5}{2} < x < 5 \end{array} \right\}, \tag{11.13}$$

just as for Example 11.1. Using the solution of the latter problem, we find that the solution of Example 11.2 is

$$T(x,t) = 300 - 56x - \frac{1120}{\pi^2} \sum_{k=1}^{\infty} \frac{(-1)^k}{(2k-1)^2} \sin\left(\frac{(2k-1)\pi x}{5}\right) e^{-1.24\times 10^{-5}\left(\frac{(2k-1)\pi}{5}\right)^2 t},$$

using the thermal diffusivity for steel. ◯

Figure 11.2 depicts the time evolution of $T(x,t)$: Using the truncation of the infinite series to its first fifty terms, we graphed approximations of $T(x,t)$ versus x for $0 \le x \le 5$, for $t = 0, 10^3, 2.5 \times 10^4, 5 \times 10^4, 5 \times 10^5, 10^6$. The graphs clearly show the convergence of the temperature to its equilibrium distribution, that is, $T(x,t) \to v(x)$ as $t \to \infty$.

Example 11.3

For which values of the constant β does the PDE-BVP-IVP

$$\left\{ \begin{array}{c} \frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2} - x,\ 0 < x < L,\ t > 0, \\ \\ \frac{\partial T}{\partial x}(0,t) = 0,\ T(L,t) = \beta,\ t > 0, \\ \\ T(x,0) = 0,\ 0 < x < L \end{array} \right\} \tag{11.14}$$

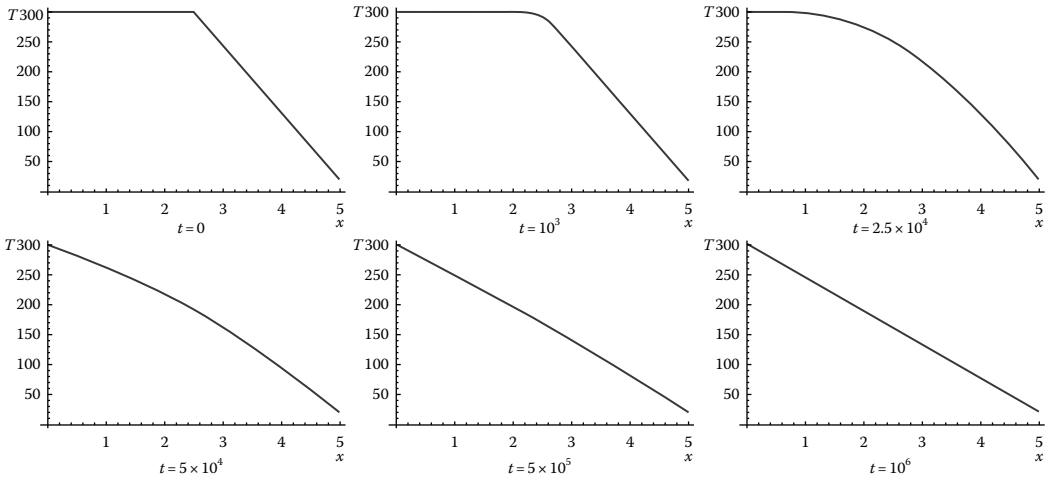


FIGURE 11.2
Temperature evolution for Example 11.1.2.

have a solution? What is the solution for such values of β? Your final conclusions should be in terms of β.

Method: If either of the BCs has an inhomogeneity or if the PDE has an inhomogeneity that is a function only of x, then we should use a two-step process, as in Example 11.2: First, find the equilibrium solution $v(x)$, a function only of x that satisfies the PDE and BCs, and then second, take the solution of the original problem to be in the form $T(x,t) = v(x) + w(x,t)$. We will see that $w(x,t)$ will satisfy a problem similar to that of (11.1) through (11.3).

The equilibrium solution, $v(x)$, satisfies $\dfrac{\partial v}{\partial t} \equiv 0$, and we want it to *also* satisfy the PDE in (11.14); hence, $0 = \alpha \dfrac{\partial^2 v}{\partial x^2} - x$. Because v is a function only of x, the partial derivatives are really ordinary derivatives, so $v(x)$ should satisfy ODE $0 = \alpha\, v'' - x$. The equilibrium solution of (11.14) should satisfy also the BCs $v'(0) = 0$, $v(L) = \beta$. So, altogether, $v(x)$ should satisfy

$$\left\{ \begin{array}{c} 0 = \alpha\, v''(x) - x,\ 0 < x < L, \\ \\ v'(0) = 0,\ v(L) = \beta \end{array} \right\}. \tag{11.15}$$

The ODE is solvable directly by integrating twice with respect to x:

$$v''(x) = \frac{x}{\alpha} \ \Rightarrow\ v'(x) = \int \frac{x}{\alpha} dx = \frac{1}{2\alpha}x^2 + c_1 \Rightarrow v(x) = \int \left(\frac{1}{2\alpha}x^2 + c_1\right) dx = \frac{1}{6\alpha}x^3 + c_1 x + c_2,$$

where c_1, c_2 are arbitrary constants. We use this to satisfy also the BCs:

$$v'(x) = \frac{1}{2\alpha}x^2 + c_1 \ \Rightarrow\ 0 = v'(0) = c_1\ ,\ \beta = v(L) = \frac{1}{6\alpha}L^3 + 0 \cdot x + c_2.$$

We take $c_2 = \beta - \frac{1}{6\alpha}L^3$, and the equilibrium solution is

$$v(x) = \frac{1}{6\alpha}x^3 + \beta - \frac{1}{6\alpha}L^3,$$

which exists for all β.

As in Example 11.2, the solution of the PDE and BVP is $T(x,t) = v(x) + w(x,t)$, that is, $w(x,t) = T(x,t) - v(x)$, so $w(x,t)$ should satisfy the PDE $\frac{\partial w}{\partial t} = \alpha \frac{\partial^2 w}{\partial x^2}$ because

$$\frac{\partial w}{\partial t} = \frac{\partial T}{\partial t} - \frac{\partial v}{\partial t} = \left(\alpha \frac{\partial^2 T}{\partial x^2} - x\right) - 0 = \left(\alpha \frac{\partial^2 (v+w)}{\partial x^2}\right) - x = \left(\alpha \frac{\partial^2 (v)}{\partial x^2} - x\right) + \alpha \frac{\partial^2 (w)}{\partial x^2}$$

$$= 0 + \alpha \frac{\partial^2 (w)}{\partial x^2}.$$

Also, $w(x,t)$ should satisfy BCs, for $t > 0$:

$$0 = \frac{\partial T}{\partial x}(0,t) = \frac{\partial (v+w)}{\partial x}(0,t) = v'(0) + \frac{\partial w}{\partial x}(0,t) = 0 + \frac{\partial w}{\partial x}(0,t),\ T(L,t) = \beta,\ t > 0;$$

hence,

$$\text{(i)} \quad \frac{\partial w}{\partial x}(0,t) = 0,$$

and

$$\beta = T(L,t) = v(L,t) + w(L,t) = \beta + w(L,t);$$

hence,

$$\text{(ii)} \quad w(L,t) = 0.$$

The PDE and BCs that $w(x,t)$ should satisfy are

$$\left\{ \begin{array}{l} \frac{\partial w}{\partial t} = \alpha \frac{\partial^2 w}{\partial x^2},\ 0 < x < L,\ t > 0, \\ \\ \frac{\partial w}{\partial x}(0,t) = 0,\ w(L,t) = 0,\ t > 0 \end{array} \right\}. \tag{11.16}$$

In Section 11.1.1 is Table 11.1, that is, Table 9.1, which we will use throughout this chapter.

Noting the two homogeneous BCs that $w(x,t)$ should satisfy and using the corresponding last line of Table 9.1,

$$w(x,t) = \sum_{n=1}^{\infty} a_n \cos\left(\frac{\left(n - \frac{1}{2}\right)\pi x}{L}\right) e^{-\alpha\left(\frac{\left(n-\frac{1}{2}\right)\pi}{L}\right)^2 t}.$$

So, the solution of the original problem should be in the form

$$T(x,t) = v(x) + w(x,t) = \frac{1}{6\alpha}x^3 + \beta - \frac{1}{6\alpha}L^3 + \sum_{n=1}^{\infty} A_n \cos\left(\frac{\left(n - \frac{1}{2}\right)\pi x}{L}\right) e^{-\alpha\left(\frac{\left(n-\frac{1}{2}\right)\pi}{L}\right)^2 t}.$$

We substitute this into the last remaining aspect of the problem, the initial condition (IC):

$$0 = T(x,0) = \frac{1}{6\alpha}x^3 + \beta - \frac{1}{6\alpha}L^3 + \sum_{n=1}^{\infty} a_n \cos\left(\frac{\left(n - \frac{1}{2}\right)\pi x}{L}\right),$$

TABLE 11.1

Eigenvalues and Eigenfunctions for $X'' + \lambda X = 0$

Boundary Conditions / Values of λ	**λ > 0**	**λ = 0**	**λ < 0**
$X(0) = X(L) = 0$	$\lambda_n = \left(\frac{n\pi}{L}\right)^2$, $X_n(x) = \sin\left(\frac{n\pi x}{L}\right)$, $n = 1, 2, \ldots$	None	None
$X'(0) = X'(L) = 0$	$\lambda_n = \left(\frac{n\pi}{L}\right)^2$, $X_n(x) = \cos\left(\frac{n\pi x}{L}\right)$, $n = 1, 2, \ldots$	$\lambda_0 = 0$, $X_0(x) \equiv 1$	None
$X(0) = X'(L) = 0$	$\lambda_n = \left(\frac{\left(n-\frac{1}{2}\right)\pi}{L}\right)^2$, $X_n(x) = \sin\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{L}\right)$, $n = 1, 2, \ldots$	None	None
$X'(0) = X(L) = 0$	$\lambda_n = \left(\frac{\left(n-\frac{1}{2}\right)\pi}{L}\right)^2$, $X_n(x) = \cos\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{L}\right)$, $n = 1, 2, \ldots$	None	None

that is,

$$\frac{1}{6\alpha}L^3 - \beta - \frac{1}{6\alpha}x^3 = \sum_{n=1}^{\infty} A_n \cos\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{L}\right).$$

This is a generalized Fourier series problem, and the coefficients are

$$A_n = \frac{2}{L}\int_0^L \left(\frac{1}{6\alpha}L^3 - \beta - \frac{1}{6\alpha}x^3\right)\cos\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{L}\right)$$

$$= \left(\frac{1}{6\alpha}L^3 - \beta\right)\frac{2}{\left(n-\frac{1}{2}\right)\pi}\left[\sin\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{L}\right)\right]_0^L - \frac{2}{L}\int_0^L \frac{1}{6\alpha}x^3\cos\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{L}\right).$$

Using integration by parts to do the latter integral and the fact that $\sin\left(\left(n-\frac{1}{2}\right)\pi\right) = -\cos n\pi = (-1)^{n+1}$, we see that

$$A_n = \frac{2}{\left(n-\frac{1}{2}\right)\pi}\left(\frac{1}{6\alpha}L^3 - \beta\right)(-1)^{n+1}$$

$$-\frac{L^3}{3\alpha}\left(\frac{6}{\left(n-\frac{1}{2}\right)^4\pi^4} - \frac{6}{\left(n-\frac{1}{2}\right)^3\pi^3}(-1)^{n+1} + \frac{1}{\left(n-\frac{1}{2}\right)\pi}(-1)^{n+1}\right)$$

$$= \frac{2\beta}{\left(n-\frac{1}{2}\right)\pi}(-1)^n - \frac{2L^3}{\alpha\left(n-\frac{1}{2}\right)^4\pi^4}\left(1 + \left(n-\frac{1}{2}\right)(-1)^n\right).$$

The solution of the original problem is

$$T(x,t) = \frac{1}{6\alpha}x^3 + \beta - \frac{1}{6\alpha}L^3 + \sum_{n=1}^{\infty} A_n \cos\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{L}\right) e^{-\alpha\left(\frac{\left(n-\frac{1}{2}\right)\pi}{L}\right)^2 t},$$

where A_n's given earlier. ◯

11.1.1 Easy Initial Conditions

Recall from our study of Fourier series that there are interesting special cases where a Fourier series, or Fourier sine series or Fourier cosine series, degenerates to just a finite sum of trigonometric functions. For example, if $f(x) = 2\sin x - \sin 4x$, then the general Fourier sine series representation, when $L = \pi$,

$$f(x) = \sum_{n=1}^{\infty} b_n \sin nx,$$

that is,

$$2\sin x - \sin 4x = \sum_{n=1}^{\infty} b_n \sin nx,$$

just says that $b_1 = 2$, $b_4 = -1$, and $b_n = 0$ for all n other than $n = 1, 4$.

Example 11.4

A two meter long steel rod, with thermal diffusivity α, has insulated lateral surface and its ends maintained at temperature 0°C. Determine the temperature if the initial temperature distribution is $T(x,0) = 2\sin(\pi x) - \sin\left(\frac{5\pi x}{2}\right),\ 0 < x < 2$.

Method: The physical problem translates into the mathematical problem

$$\left\{\begin{array}{c} \frac{\partial T}{\partial t} = \alpha\,\frac{\partial^2 T}{\partial x^2},\ 0 < x < 2,\ t > 0, \\ \\ T(0,t) = T(2,t) = 0,\ t > 0, \\ \\ T(x,0) = 2\sin\left(\frac{2\pi x}{2}\right) - \sin\left(\frac{5\pi x}{2}\right),\ 0 < x < 2 \end{array}\right\}. \tag{11.17}$$

The PDE and homogeneous BCs fit directly into the form of problem (11.1) through (11.3) with $L = 2$. The solution of the PDE and BVP is

$$T(x,t) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{2}\right) e^{-\alpha\left(\frac{n\pi}{2}\right)^2 t}.$$

The initial condition gives

$$2\sin\left(\frac{2\pi x}{2}\right) - \sin\left(\frac{5\pi x}{2}\right) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{2}\right),\ 0 < x < 2.$$

So, $b_2 = 2$, $b_5 = -1$, and $b_n = 0$ for all n other than $n = 2$ and 5. The solution of the PDE-BVP-IVP is simply

$$T(x,t) = 2\sin(\pi x)e^{-\alpha\pi^2 t} - \sin\left(\frac{5\pi x}{2}\right)e^{-\alpha\left(\frac{5\pi}{2}\right)^2 t}. \bigcirc$$

It would be a shame to cover up the simplicity of this solution by stating a final conclusion in terms of an infinite series. Also, we should not go through the unnecessary work of calculating the values of b_n for all n using the integrals

$$b_n = \frac{2}{2}\int_0^2 \left(2\sin\left(\frac{2\pi x}{2}\right) - \sin\left(\frac{5\pi x}{2}\right)\right)\sin\left(\frac{n\pi x}{2}\right)dx.$$

Table 11.1 gives for our convenience results about the eigenvalues and eigenfunctions for ODE $X''(x) + \lambda X(x) = 0$ for the four most common sets of separated homogeneous boundary conditions. Each of the four groups of entries corresponds to a different set of physical boundary conditions.

11.1.2 Composite Rod

Example 11.5

Solve

$$\left\{\begin{array}{l} \frac{\partial T}{\partial t} = \frac{1}{c\varrho}\frac{\partial}{\partial x}\left[\kappa(x)\frac{\partial T}{\partial x}\right],\ 0 < x < 1 \text{ and } 1 < x < 2,\ 0 < t < \infty, \\ \\ \frac{\partial T}{\partial x}(0,t) = \frac{\partial T}{\partial x}(2,t) = 0,\ 0 < t < \infty \\ \\ T(x,0) = f(x),\ 0 < x < 1 \text{ and } 1 < x < 2 \end{array}\right\},$$

where c, ϱ are constants and the thermal conductivity, $\kappa(x)$, is defined piecewise by

$$\kappa(x) = \left\{\begin{array}{ll} \frac{1}{2}, & 0 < x < 1 \\ \\ \frac{1}{3}, & 1 < x < 2 \end{array}\right\}.$$

As we saw in Section 10.2, the jump in the coefficient κ, as well as the interface conditions (9.75) in Section 9.6 at $x = 1$, corresponds physically to joining two types of materials, for example, two metals, at $x = 1$. This is why we refer to this as a problem for a "composite" rod.

Method: In Example 9.32 in Section 9.6, we found the eigenvalues, λ_n, $n = 0, 1, \ldots$, and corresponding eigenfunctions

$$X_n(x) = \left\{\begin{array}{ll} -\sqrt{2}\sin(\sqrt{3\lambda_n})\cos(\sqrt{2\lambda_n}\,x), & 0 < x < 1 \\ \\ \sqrt{3}\sin(\sqrt{2\lambda_n})\cos(\sqrt{3\lambda_n}\,(2-x)), & 1 < x < 2 \end{array}\right\},$$

$n = 0, 1, \ldots$, for the ODE-BVP with interface conditions (9.75) in Section 9.6. Note that $\lambda_0 = 0$ and $X_0(x) \equiv 1$. We look for product solutions $T = X_n(x)G_n(t)$, and, similar to previous work, we get $\dot{G}_n = (-\lambda_n/(c\varrho))\,G_n$; hence, $G_n(t) = \exp(-\lambda_n t/(c\varrho))$. The general

solution of the PDE plus the two BCs and the interface conditions is

$$T(x,t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n X_n(x) e^{-\lambda_n t/(c\varrho)}.$$

All that remains is to satisfy the initial condition

$$f(x) = T(x,0) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n X_n(x).$$

Orthogonality, with $L = 2$, gives us

$$a_0 = \int_0^2 f(x)\,dx \quad \text{and}$$

$$a_n = \frac{1}{N_n}\left(\int_0^1 X_n(x) f(x)\,dx + \int_1^2 X_n(x) f(x)\,dx\right)$$

$$= \frac{1}{N_n}\left(\int_0^1 -\sqrt{2}\sin(\sqrt{3\lambda_n})\cos(\sqrt{2\lambda_n}\,x) f(x)\,dx\right.$$

$$\left. + \int_1^2 \sqrt{3}\sin(\sqrt{2\lambda_n})\cos(\sqrt{3\lambda_n}\,(2-x)) f(x)\,dx\right),$$

where $N_n = \int_0^2 \big(X_n(x)\big)^2 dx = \cdots = \sin^2(\sqrt{3\lambda_n}) + \frac{3}{2}\sin^2(\sqrt{2\lambda_n})$. ○

11.1.3 Time-Dependent Boundary Conditions

For the heat equation, the boundary conditions could depend on time. For example, consider the following problem.

Example 11.6

A copper rod is 30.0 cm long and has diameter 0.400 cm. Assume that the approximate physical parameters of copper are mass density $\varrho \approx 8.96$ g/cm^3, specific heat $c \approx 0.3851$ J/(g · K), and thermal conductivity $\kappa \approx 3.98$ W/(cm · K). Its lateral surface is insulated, its left end is maintained at a temperature of $(20.0 - 2.00\sin(0.1t))$°C, and its right end is losing heat at a constant rate of 1 W. Throughout the rod, heat energy is added to the rod at a rate of 0.200 W/cm^3. Initially the temperature in the rod is a constant, 20.0°C.

(a) Translate the physical problem into a mathematical problem.

(b) Determine the temperature as a function of x and t.

Method:

(a) Measuring x in cm, the physical problem translates to a mathematical problem consisting of the PDE of a 1D heat equation for temperature $T(x,t)$, with source term $Q(x,t)/(c\varrho)$, a time-dependent boundary condition on $T(0,t)$, a boundary condition on $\frac{\partial T}{\partial x}(30,t)$, and the initial condition on $T(x,0)$.

First, the physical parameter of thermal diffusivity is

$$\alpha = \frac{\kappa}{c\varrho} = \frac{3.98\,\mathrm{J/s\cdot cm\cdot K}}{\left(0.3851\,\mathrm{J/(g\cdot{}^\circ K)}\right)\left(8.96\,\mathrm{g/cm^3}\right)} \approx 1.15\,\frac{\mathrm{cm}^2}{\mathrm{s}},$$

to the three significant digits of accuracy found in the given physical parameters. Next, the source term is

$$Q(x,t)/(c\varrho) = \frac{0.200\,\mathrm{J/(s\cdot cm^3)}}{\left(0.3851\,\mathrm{J/(g\cdot K)}\right)\left(8.96\,\mathrm{g/cm^3}\right)} \approx 0.0580\ \mathrm{K/s},$$

to the three significant digits of accuracy found in the given physical parameters. The left end boundary condition is $T(0,t) = (20.0 - 2.00\sin(0.1t))^\circ$C, and the initial condition is $T(x,0) = 20.0^\circ$C. As for the right end boundary condition, the $+1$ W heat flux out of the rod there is $-A\kappa\,\frac{\partial T}{\partial x}(30.0,t)$, where A is the area of cross section of the rod; hence,

$$\frac{\partial T}{\partial x}(30.0,t) = -\frac{1\,\mathrm{W}}{\pi(0.200\,\mathrm{cm})^2 3.98\,\mathrm{W/(cm\cdot{}^\circ K)}} \approx -2.00^\circ\mathrm{C/cm},$$

to the three significant digits of accuracy found in the physical condition at $x = 30$.

The mathematical problem is, with approximate physical parameters and suppressing trailing zeros that indicate the number of significant digits,

$$\left\{\begin{array}{c} \frac{\partial T}{\partial t} = 1.15\frac{\partial^2 T}{\partial x^2} + 0.058,\ 0 < x < 30,\ t > 0, \\ \\ T(0,t) = 20 - 2\sin(0.1t),\ \frac{\partial T}{\partial x}(30,t) = -2,\ t > 0, \\ \\ T(x,0) = 20,\ 0 < x < 30 \end{array}\right\}.$$

(b) Unfortunately, this problem does <u>not</u> directly fit into any problem we studied because of the time-dependent boundary condition. Nevertheless, we should proceed in a way that is close to the method of writing $T(x,t) = v(x) + w(x,t)$, where $v(x)$ is a steady state temperature distribution. But, in this problem, we don't expect that there is an equilibrium temperature distribution.

Instead, let's write $T(x,t) = v(x,t) + w(x,t)$, where $v(x,t)$ *only* satisfies the two inhomogeneous boundary conditions. Of course, an infinite number of functions $v(x,t)$ satisfy the two inhomogeneous boundary conditions. We might as well choose $v(x,t)$ to be as simple as possible, because of the basic principle to *keep things as simple as possible*. Also, we suspect that the simpler the function $v(x,t)$, the easier it will be to find $w(x,t)$.

As far as dependence on x is concerned, the simplest function v would have the form $v(x,t) = \beta(t) + \gamma(t)x$. The first boundary condition then yields $20 - 2\sin(0.1t) = v(0,t) = \beta(t)$, and the second boundary condition gives $-2 = \frac{\partial v}{\partial x}(30,t) = \gamma(t)$, so $v(x,t) = 20 - 2\sin(0.1t) - 2x$.

Now, $w(x,t) = T(x,t) - v(x,t) = T(x,t) - 20 + 2\sin(0.1t) + 2x$. First, we need to find the PDE to be satisfied by $w(x,t)$; the PDE itself will be affected by $v(x,t)$. We calculate

$$\frac{\partial w}{\partial t} = \frac{\partial T}{\partial t} - \frac{\partial v}{\partial t} = \left(1.15\frac{\partial^2 T}{\partial x^2} + 0.058\right) + 0.2\cos(0.1t)$$

$$= 1.15\frac{\partial^2 (w+v)}{\partial x^2} + 0.058 + 0.2\cos(0.1t),$$

that is,

$$\frac{\partial w}{\partial t} = 1.15\frac{\partial^2 w}{\partial x^2} + 0.058 + 0.2\cos(0.1t). \tag{11.18}$$

The boundary conditions that w should satisfy are $w(0,t) = T(0,t) - v(0,t) = 20 - 2\sin(0.1t) - (20 - 2\sin(0.1t)) = 0$ and $\frac{\partial w}{\partial x}(30,t) = \frac{\partial T}{\partial x}(30,t) - \frac{\partial v}{\partial x}(30,t) = -2 - (-2) = 0$, and the initial condition that w should satisfy is

$$w(x,0) = T(x,0) - v(x,0) = 20 - (20 - 2x) = 2x.$$

Because of the homogeneous boundary conditions

$$w(x,0) = 0, \quad \frac{\partial w}{\partial x}(30,t) = 0,$$

we should take $w(x,t)$ to be in terms of the functions $\sin\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{30}\right)$. But the PDE is not homogeneous; in fact, it involves a function of time, t. So, we should try to find w in the form

$$w(x,t) = \sum_{n=1}^{\infty} b_n(t) \sin\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{30}\right).$$

When we substitute this into PDE (11.18), we get

$$\sum_{n=1}^{\infty} \frac{db_n}{dt}(t) \sin\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{30}\right) = \frac{\partial w}{\partial t} = 1.15\,\frac{\partial^2 w}{\partial x^2} + 0.058 + 0.2\cos(0.1t) =$$

$$= 0.058 + 0.2\cos(0.1t) + 1.15\sum_{n=1}^{\infty} b_n(t)\left(-\left(\frac{\left(n-\frac{1}{2}\right)\pi}{30}\right)^2\right)\sin\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{30}\right). \tag{11.19}$$

Using a generalized Fourier series to express $(0.058 + 0.2\cos(0.1t))$ as

$$0.058 + 0.2\cos(0.1t)$$

$$\doteq \sum_{n=1}^{\infty}\left(\frac{2}{30}\int_0^{30}(0.058 + 0.2\cos(0.1t))\sin\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{30}\right)dx\right)\sin\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{30}\right)$$

$$\triangleq \sum_{n=1}^{\infty} R_n(t)\sin\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{30}\right),$$

and substituting that into (11.19), we get

$$\sum_{n=1}^{\infty} \frac{db_n}{dt}(t)\sin\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{30}\right)$$

$$= \sum_{n=1}^{\infty} R_n(t)\sin\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{30}\right) + 1.15\sum_{n=1}^{\infty} b_n(t)\left(-\left(\frac{\left(n-\frac{1}{2}\right)\pi}{30}\right)^2\right)\sin\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{30}\right).$$

Fourier analyzing, that is, breaking down the infinite sum into its components, gives

$$\frac{db_n}{dt}(t) = 1.15\left(-\left(\frac{\left(n-\frac{1}{2}\right)\pi}{30}\right)^2\right)b_n(t) + R_n(t), \tag{11.20}$$

for $n = 1, 2, \ldots$, where

$$R_n(t) = \frac{2}{30}\int_0^{30}(0.058 - 0.2\cos(0.1t))\sin\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{30}\right)dx$$

$$= \frac{2}{\cancel{30}}(0.058-0.2\cos(0.1t))\left[\frac{\cos\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{30}\right)}{-\frac{\left(n-\frac{1}{2}\right)\pi}{\cancel{30}}}\right]_0^{30} = \frac{-0.116+0.4\cos(0.1t))}{\left(n-\frac{1}{2}\right)\pi}.$$

To solve the first-order linear ODE (11.20), we could use an integrating factor or the method of undetermined coefficients. Let's use an integrating factor, after first putting the ODE into the standard form:

$$\frac{db_n}{dt}(t) + 1.15\left(\frac{\left(n-\frac{1}{2}\right)\pi}{30}\right)^2 b_n = R_n(t).$$

The integrating factor is $\mu(t) = e^{1.15\left(\frac{\left(n-\frac{1}{2}\right)\pi}{30}\right)^2 t}$, so the solution of (11.20) is

$$b_n(t) = b_n(0) + e^{-1.15\left(\frac{\left(n-\frac{1}{2}\right)\pi}{30}\right)^2 t}\cdot\frac{1}{\left(n-\frac{1}{2}\right)\pi}\int_0^t(0.116+0.4\cos(0.1s))e^{1.15\left(\frac{\left(n-\frac{1}{2}\right)\pi}{30}\right)^2 s}ds$$

$$= b_n(0) + \frac{1}{\left(n-\frac{1}{2}\right)\pi}e^{-1.15\left(\frac{\left(n-\frac{1}{2}\right)\pi}{30}\right)^2 t}$$

$$\cdot\left[e^{1.15\left(\frac{(n-\frac{1}{2})\pi}{30}\right)^2 s}\cdot\left(\frac{0.116}{1.15\left(\frac{(n-\frac{1}{2})\pi}{30}\right)^2} + \frac{.04\sin(0.1s)+0.46\left(\frac{(n-\frac{1}{2})\pi}{30}\right)^2\cos(0.1s)}{(.1)^2+\left(1.15\left(\frac{(n-\frac{1}{2})\pi}{30}\right)^2\right)^2}\right)\right]_0^t$$

$$= b_n(0) + \frac{1}{(n-\frac{1}{2})\pi}\left(\frac{0.116}{1.15\left(\frac{\left(n-\frac{1}{2}\right)}{30}\right)^2} + \frac{0.04\sin(0.1t)+0.46\left(\frac{(n-\frac{1}{2})\pi}{30}\right)^2\cos(0.1t)}{0.01+1.15^2\left(\frac{(n-\frac{1}{2})\pi}{30}\right)^4}\right)$$

$$- \frac{1}{(n-\frac{1}{2})\pi}e^{-1.15\left(\frac{(n-\frac{1}{2})\pi}{30}\right)^2 t}\left(\frac{1}{1.15\left(\frac{(n-\frac{1}{2})\pi}{30}\right)^2} + \frac{0.46\left(\frac{(n-\frac{1}{2})\pi}{30}\right)^2}{0.01+1.15^2\left(\frac{(n-\frac{1}{2})\pi}{30}\right)^4}\right)$$

As for $b_n(0)$, the initial condition of $w(x,0) = 2x$ yields

$$b_n(0) = \frac{2}{30}\int_0^{30}(2x)\sin\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{30}\right)dx = \cdots = \frac{120}{\left(n-\frac{1}{2}\right)^2\pi^2}(-1)^{n+1}.$$

The solution of the PDE-BVP-IVP, that is, the temperature distribution in the rod, is

$$\begin{aligned}T(x,t) = v + w &= 20 - 2\sin(0.1t) - 2x \\ &+ \sum_{n=1}^{\infty}\frac{1}{(n-\frac{1}{2})^2\pi^2}\left(120(-1)^{n+1} + \frac{104.4}{1.15(n-\frac{1}{2})\pi}\right)\sin\left(\frac{(n-\frac{1}{2})\pi x}{30}\right) \\ &+ \sum_{n=1}^{\infty}\frac{1}{(n-\frac{1}{2})\pi}\left(\frac{0.04\sin(0.1t) + 0.46\left(\frac{(n-\frac{1}{2})\pi}{30}\right)^2\cos(0.1t)}{0.01 + 1.15^2\left(\frac{(n-\frac{1}{2})\pi}{30}\right)^4}\right)\sin\left(\frac{(n-\frac{1}{2})\pi x}{30}\right) \\ &- \sum_{n=1}^{\infty}\frac{1}{(n-\frac{1}{2})\pi}e^{-1.15\left(\frac{(n-\frac{1}{2})\pi}{30}\right)^2 t}\left(\frac{1}{1.15\left(\frac{(n-\frac{1}{2})\pi}{30}\right)^2} + \frac{0.46\left(\frac{(n-\frac{1}{2})\pi}{30}\right)^2}{0.01 + 1.15^2\left(\frac{(n-\frac{1}{2})\pi}{30}\right)^4}\right)\cdot \\ &\cdot \sin\left(\frac{(n-\frac{1}{2})\pi x}{30}\right). \bigcirc\end{aligned}$$

Learn More About It

The approach of beginning with eigenfunction expansions is also found in *Separation of Variables for Partial Differential Equations: An Eigenfunction Approach*, by George Cain and Gunter H. Meyer, Chapman and Hall/CRC © 2005.

It is an unfortunate reality that we cannot expect to have the existence and uniqueness of a solution of a heat equation backward in time. See *Applied Partial Differential Equations: With Fourier Series and Boundary Value Problems*, 4*th* edn., by Richard Haberman, Prentice-Hall, Inc. © 2004, specifically Problem 2.5.2.14 and also p. 680.

11.1.4 Problems

1. Solve the PDE-BVP-IVP

$$\left\{\begin{array}{c} \frac{\partial T}{\partial t} = \alpha\,\frac{\partial^2 T}{\partial x^2},\ 0 < x < \pi,\ t > 0, \\ \\ T(0,t) = T(\pi,t) = 0,\ t > 0, \\ \\ T(x,0) = \sin x - \frac{1}{2}\sin 3x, 0 < x < \pi \end{array}\right\}.$$

2. Solve the PDE-BVP-IVP

$$\left\{\begin{array}{c} \frac{\partial T}{\partial t} = \alpha\, \frac{\partial^2 T}{\partial x^2},\ 0 < x < \pi,\ t > 0, \\ \\ T(0,t) = T(\pi,t) = 0,\ t > 0, \\ \\ T(x,0) = x,\ 0 < x < \pi \end{array}\right\}.$$

3. Solve the PDE-BVP-IVP

$$\left\{\begin{array}{c} \frac{\partial T}{\partial t} = \alpha\, \frac{\partial^2 T}{\partial x^2},\ 0 < x < \pi,\ t > 0, \\ \\ \frac{\partial T}{\partial x}(0,t) = \frac{\partial T}{\partial x}(\pi,t) = 1,\ t > 0, \\ \\ T(x,0) = 2x,\ 0 < x < \pi \end{array}\right\}.$$

4. Solve the PDE-BVP-IVP

$$\left\{\begin{array}{c} \frac{\partial T}{\partial t} = \alpha\, \frac{\partial^2 T}{\partial x^2},\ 0 < x < \pi,\ t > 0, \\ \\ \frac{\partial T}{\partial x}(0,t) = T(\pi,t) = 0,\ t > 0, \\ \\ T(x,0) = \sin x,\ 0 < x < \pi \end{array}\right\}.$$

   Hint: Use results from the last entries of Table 11.1. Also, you may assume the orthogonality relation

$$\int_0^L \cos\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{L}\right)\cos\left(\frac{\left(m-\frac{1}{2}\right)\pi x}{L}\right)dx = \left\{\begin{array}{ll} 0, & \text{if } n \neq m \\ \\ \frac{L}{2}, & \text{if } n = m \end{array}\right\}.$$

5. A five meter long copper rod with insulated surface has both its left and right ends also insulated. Determine the temperature as a function of x and t if the initial condition is given by $T(x,0) = x,\ 0 < x < 5$. Assume the thermal diffusivity of copper is about $1.15 \times 10^{-4}\ \mathrm{m}^2/\mathrm{s}$. Hint: See the second group of entries in Table 11.1. Also, you may assume the orthogonality relation

$$\int_0^L \cos\left(\frac{n\pi x}{L}\right)\cos\left(\frac{m\pi x}{L}\right)dx = \left\{\begin{array}{ll} 0, & \text{if } n \neq m \\ L, & \text{if } n = m = 0 \\ \frac{L}{2}, & n = m \geq 1 \end{array}\right\}.$$

6. Solve the PDE-BVP-IVP

$$\left\{\begin{array}{c} \frac{\partial T}{\partial t} = \alpha\, \frac{\partial^2 T}{\partial x^2},\ 0 < x < 5,\ t > 0, \\ \\ T(0,t) = 20,\ \frac{\partial T}{\partial x}(5,t) = 0,\ t > 0, \\ \\ T(x,0) = 0,\ 0 < x < 5 \end{array}\right\}.$$

Hints: First find the equilibrium solution. After that, continue using the method of Example 11.2, along with results from the third group of entries of Table 11.1. Also, you may assume the orthogonality relation

$$\int_0^L \sin\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{L}\right)\sin\left(\frac{\left(m-\frac{1}{2}\right)\pi x}{L}\right)dx = \begin{Bmatrix} 0, & \text{if } n \neq m \\ \frac{L}{2}, & \text{if } n = m \end{Bmatrix}.$$

7. Solve the PDE-BVP-IVP

$$\begin{Bmatrix} \frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2} - 1,\ 0 < x < 2,\ t > 0, \\ T(0,t) = \frac{\partial T}{\partial x}(2,t) = 0,\ t > 0, \\ T(x,0) = 0,\ 0 < x < 2 \end{Bmatrix}.$$

Hint: First find the equilibrium solution. After that, continue using the method of Example 11.2, along with results from the third group of entries of Table 11.1. Also, you may assume the orthogonality relation

$$\int_0^L \sin\left(\frac{\left(n-\frac{1}{2}\right)\pi x}{L}\right)\sin\left(\frac{\left(m-\frac{1}{2}\right)\pi x}{L}\right)dx = \begin{Bmatrix} 0, & \text{if } n \neq m \\ \frac{L}{2}, & \text{if } n = m \end{Bmatrix}.$$

8. Solve the PDE-BVP-IVP

$$\begin{Bmatrix} \frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2},\ 0 < x < \frac{1}{2},\ t > 0, \\ \frac{\partial T}{\partial x}(0,t) = 0,\ T\left(\frac{1}{2},t\right) = 10,\ t > 0, \\ T(x,0) = \begin{Bmatrix} 1, & 0 < x < \frac{1}{4} \\ 2-4x & \frac{1}{4} < x < \frac{1}{2} \end{Bmatrix} \end{Bmatrix}.$$

9. Solve the PDE-BVP-IVP

$$\begin{Bmatrix} \frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2},\ 0 < x < L,\ t > 0, \\ T(0,t) = 0,\ \frac{\partial T}{\partial x}(L,t) = 10,\ t > 0, \\ T(x,0) = f(x),\ 0 < x < L. \end{Bmatrix},$$

where the graph of the function f is given in Figure 11.3. Assume $f(x)$ is sinusoidal for $\frac{L}{2} < x < L$.

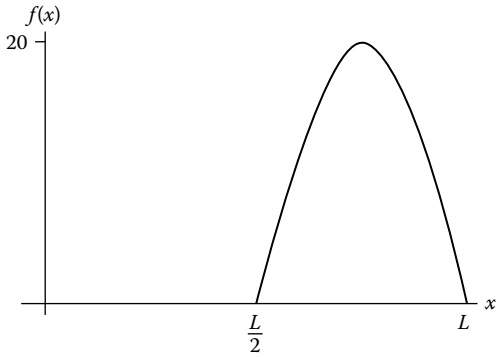


FIGURE 11.3
Initial temperature distribution for Problem 11.1.4.9.

10. Suppose the solution of the heat equation for a thin homogeneous rod is

$$T(x,t) = 80 + 20\cos\left(\frac{3\pi x}{L}\right) + \\ + \sum_{n=1}^{\infty} \frac{5}{(2n-1)^2} \sin\left(\frac{(2n-1)\pi x}{2L}\right) e^{-\alpha\left(\frac{(2n-1)\pi}{2L}\right)^2 t}:$$

    (a) Identify which part(s) of the solution is the equilibrium solution.
    (b) Identify which part(s) of the solution is the transient solution.
    (c) Find a formula for the initial temperature distribution.
    (d) Find boundary conditions satisfied by $T(0,t)$ and $\frac{\partial T}{\partial x}(L,t)$.

11. Solve the PDE-BVP-IVP

$$\left\{ \begin{array}{c} \frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2} + \cos(x),\ 0 < x < \pi,\ t > 0, \\ \\ \frac{\partial T}{\partial x}(0,t) = \frac{\partial T}{\partial x}(\pi,t) = 0,\ t > 0, \\ \\ T(x,0) = -x + \cos(x),\ 0 < x < \pi \end{array} \right\}.$$

12. Solve the PDE-BVP-IVP

$$\left\{ \begin{array}{c} \frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2},\ 0 < x < \pi,\ t > 0, \\ \\ T(0,t) = 20,\ T(\pi,t) = 70,\ t > 0, \\ \\ T(x,0) = 20 + 40x,\ 0 < x < \pi \end{array} \right\}.$$

13. Assume α is a positive constant. Solve the PDE-BVP-IVP

$$\left\{\begin{array}{c} \frac{\partial T}{\partial t} = \frac{\alpha}{t+1}\frac{\partial^2 T}{\partial x^2},\ 0 < x < \pi,\ t > 0, \\ T(0,t) = T(\pi,t) = 0,\ t > 0, \\ T(x,0) = f(x),\ 0 < x < \pi \end{array}\right\}.$$

14. Assume α is a positive constant. Solve the PDE-BVP-IVP

$$\left\{\begin{array}{c} \frac{\partial T}{\partial t} = \alpha\frac{\partial^2 T}{\partial x^2},\ 0 < x < \pi,\ t > 0, \\ T(0,t) = 20,\ \frac{\partial T}{\partial x}(\pi,t) = 0,\ t > 0, \\ T(x,0) = 0,\ 0 < x < \pi \end{array}\right\}.$$

15. Assuming $\frac{\pi}{L} > 1$, solve the PDE-BVP-IVP

$$\left\{\begin{array}{c} \frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2} + e^{-t},\ 0 < x < L, t > 0, \\ T(0,t) = T(L,t) = 0,\ t > 0, \\ T(x,0) = 0,\ 0 < x < L \end{array}\right\}.$$

16. Solve the PDE-BVP-IVP

$$\left\{\begin{array}{c} \frac{\partial T}{\partial t} = \alpha\frac{\partial^2 T}{\partial x^2} + g(t),\ 0 < x < L, t > 0, \\ \frac{\partial T}{\partial x}(0,t) = \frac{\partial T}{\partial x}(L,t) = 0,\ t > 0, \\ T(x,0) = f(x),\ 0 < x < L. \end{array}\right\}.$$

17. Solve the PDE-BVP-IVP

$$\left\{\begin{array}{c} \frac{\partial T}{\partial t} = \alpha\frac{\partial^2 T}{\partial x^2} + g(t),\ 0 < x < L, t > 0, \\ T(0,t) = T(L,t) = 0,\ t > 0, \\ T(x,0) = 0,\ 0 < x < L. \end{array}\right\}.$$

18. Solve the PDE-BVP-IVP

$$\left\{\begin{array}{c} \frac{\partial T}{\partial t} = \frac{\alpha}{t+1}\frac{\partial^2 T}{\partial x^2} - A,\ 0 < x < \pi,\ t > 0, \\ T(0,t) = 0,\ T(\pi,t) = B,\ t > 0, \\ T(x,0) = f(x),\ 0 < x < \pi \end{array}\right\}.$$

Assume the positive constant $\alpha > 1$. Your final conclusion should be stated in terms of the unspecified constants A, B and the unspecified function f.

19. (a) Find all eigenvalues and eigenfunctions of the ODE-BVP

$$\left\{ \begin{array}{c} \frac{d^2X}{dx^2} + \lambda X = 0,\ 0 < x < L \\ \\ X(-L) = X(L) = 0 \end{array} \right\}.$$

This is not the problem with periodic boundary conditions.

(b) Use the results of part (a) to solve the PDE-BVP-IVP

$$\left\{ \begin{array}{c} \frac{\partial T}{\partial t} = \alpha\, \frac{\partial^2 T}{\partial x^2} \\ \\ T(-L,t) = T(L,t) = 0,\ 0 < t < \infty \\ \\ T(x,0) = 100,\ 0 < x < L \end{array} \right\}.$$

20. Suppose we are *designing* heat flow in a thin rod, modeled by the PDE-BVP-IVP

$$\left\{ \begin{array}{c} \frac{\partial T}{\partial t} = \alpha\, \frac{\partial^2 T}{\partial x^2} \\ \\ \frac{\partial T}{\partial x}(0,t) = 10,\ T(L,t) = 0,\ 0 < t < \infty \\ \\ T(x,0) = 0,\ 0 < t < \infty \end{array} \right\}.$$

Find a formula for $g(t) \triangleq T\left(\frac{L}{2}, t\right)$, that is, the temperature measured by a sensor at the middle of the rod. Your formula should be in terms of L and α.

21. Assume β is a constant. Find a condition of the form $\beta <$ *constant* that guarantees that all solutions of the PDE-BVP-IVP

$$\left\{ \begin{array}{c} \frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2} + \beta T,\ 0 < x < L, t > 0 \\ \\ T(0,t) = T(L,t) = 0,\ t > 0 \\ \\ T(x,0) = f(x),\ 0 < x < L \end{array} \right\}$$

satisfy $\lim_{t\to\infty} T(x,t) = 0$, for all x in $[\,0, L\,]$. In that case, what is the time constant, in terms of β?

22. Assume β is a constant. Explain why all solutions of the PDE-BVP-IVP

$$\left\{ \begin{array}{c} \frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2} + \beta T,\ 0 < x < L, t > 0 \\ \\ \frac{\partial T}{\partial x}(0,t) = \frac{\partial T}{\partial x}(L,t) = 0,\ t > 0 \\ \\ T(x,0) = f(x),\ 0 < x < L \end{array} \right\}$$

satisfy $\lim_{t\to\infty} T(x,t) = 0$, for all x in $[\,0, L\,]$, if and only if $\beta < 0$. What is the time constant, in terms of β?

11.2 Wave Equation in One Space Dimension

In Section 10.4, we studied models for vibrating strings and membranes and for a vibrating air column. We saw how linear approximations led to the wave equation in one or more dimensions. We also saw how to get a linear model that predicted the speed of sound.

The PDE-BVP-IVP for the wave equation is

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2},\ 0 < x < L,\ t > 0, \tag{11.21}$$

$$u(0,t) = u(L,t) = 0,\ t > 0, \tag{11.22}$$

$$u(x,0) = f(x),\ 0 < x < L, \tag{11.23}$$

$$\frac{\partial u}{\partial t}(x,0) = g(x),\ 0 < x < L. \tag{11.24}$$

In this section, we will see how this problem (11.21) through (11.24) is solvable by the method of separation of variables. As in our study of the heat equation (11.1) in Section 11.1, because of the "homogeneous boundary conditions" $u(0,t) = u(L,t) = 0,\ t > 0$, we try "product solutions" of the PDE in the *assumed* form $u(x,t) = X(x)G(t)$, where X solves the ODE-BVP (11.4) in Section 11.1. Because $X(x)$ satisfies the boundary conditions $X(0) = X(L) = 0$, such a product solution $u(x,t)$ automatically satisfies the boundary conditions $u(0,t) = u(L,t) = 0$. Let $X = X_n(x) = \sin\left(\frac{n\pi x}{L}\right)$, and substitute $u(x,t) = \sin\left(\frac{n\pi x}{L}\right) G(t)$ into the PDE $\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$. This yields

$$\sin\left(\frac{n\pi x}{L}\right) \frac{d^2 G_n}{dt^2} = -c^2 \left(\frac{n\pi}{L}\right)^2 \sin\left(\frac{n\pi x}{L}\right) G_n(t),\ n = 1, 2, \ldots;$$

hence,

$$\ddot{G}_n + c^2 \left(\frac{n\pi}{L}\right)^2 G_n(t) = 0,\ n = 1, 2, \ldots\ .$$

This is the famous undamped harmonic oscillator ODE of Section 3.3. The solutions are

$$G_n(t) = a_n \cos\left(\frac{n\pi ct}{L}\right) + b_n \sin\left(\frac{n\pi ct}{L}\right),\ n = 1, 2, \ldots,$$

where a_n and b_n are arbitrary constants.

The product solutions are

$$u_n(x,t) = \left(a_n \cos\left(\frac{n\pi ct}{L}\right) + b_n \sin\left(\frac{n\pi ct}{L}\right)\right) \sin\left(\frac{n\pi x}{L}\right),\ n = 1, 2, \ldots .$$

What's left in solving the PDE-BVP-IVP is the two initial conditions $u(x,0) = f(x)$ and $\frac{\partial u}{\partial t}(x,0) = g(x),\ 0 < x < L$, that is, (11.23) and (11.24). The general method is to use *all* of the product solutions by using the infinite series:

$$u(x,t) = \sum_{n=1}^{\infty} u_n(x,t) = \sum_{n=1}^{\infty} \left(a_n \cos\left(\frac{n\pi ct}{L}\right) + b_n \sin\left(\frac{n\pi ct}{L}\right)\right) \sin\left(\frac{n\pi x}{L}\right). \tag{11.25}$$

Substitute (11.25) into the first initial condition (11.23), that is, $u(x,0) = f(x),\ 0 < x < L$, to get

$$f(x) = \sum_{n=1}^{\infty} a_n \sin\left(\frac{n\pi x}{L}\right).$$

Fourier analysis of this sine series yields

$$a_n = \frac{2}{L} \int_0^L f(x) \sin\left(\frac{n\pi x}{L}\right) dx.$$

Substitute (11.25) into the second initial condition (11.24), that is, $\frac{\partial u}{\partial t}(x,0) = g(x),\ 0 < x < L$. First, take the partial derivative of (11.24) with respect to t,

$$\frac{\partial u}{\partial t}(x,t) = \sum_{n=1}^{\infty} \left(\frac{n\pi c}{L}\right) \left(-a_n \sin\left(\frac{n\pi ct}{L}\right) + b_n \cos\left(\frac{n\pi ct}{L}\right)\right) \sin\left(\frac{n\pi x}{L}\right),$$

and, *after* that, substitute in $t = 0$, so the initial condition is satisfied by solving

$$g(x) = \frac{\partial u}{\partial t}(x,0) = \sum_{n=1}^{\infty} \left(\frac{n\pi c}{L} b_n\right) \sin\left(\frac{n\pi x}{L}\right),\ 0 < x < L.$$

This is just a Fourier sine series problem, so for all smooth enough functions g, there is exactly one solution for the coefficients, implicitly given by

$$\left(\frac{n\pi c}{L} \cdot b_n\right) = \frac{2}{L} \int_0^L g(x) \sin\left(\frac{n\pi x}{L}\right) dx.$$

The solution of the PDE-BVP-IVP is given by

$$u(x,t) = \sum_{n=1}^{\infty} \left(\frac{2}{L} \int_0^L f(x) \sin\left(\frac{n\pi x}{L}\right) dx \right) \cos\left(\frac{n\pi ct}{L}\right) \sin\left(\frac{n\pi x}{L}\right) + \sum_{n=1}^{\infty} \left(\frac{2}{n\pi c} \int_0^L g(x) \sin\left(\frac{n\pi x}{L}\right) dx \right) \sin\left(\frac{n\pi ct}{L}\right) \sin\left(\frac{n\pi x}{L}\right). \tag{11.26}$$

In the special case when the string starts from rest, that is, when $\frac{\partial u}{\partial t}(x,0) = 0,\ 0 < x < L$, we see that $b_n = 0$ for all n; hence, the solution of the PDE-BVP-IVP is

$$u(x,t) = \sum_{n=1}^{\infty} \left(\frac{2}{L} \int_0^L f(x) \sin\left(\frac{n\pi x}{L}\right) dx \right) \cos\left(\frac{n\pi ct}{L}\right) \sin\left(\frac{n\pi x}{L}\right). \tag{11.27}$$

Example 11.7

A guitar string has mass density of 4.466×10^{-3} kg/m, length 640.0 mm, and is under tension equal to the weight of a mass of 9.03 kg at sea level. The string is at rest but is given an initial velocity of $\sin\left(\frac{\pi x}{0.640}\right) - 0.2 \sin\left(\frac{2\pi x}{0.640}\right)$ from its equilibrium position. Find the displacement of the string.

Method: The tension is found by multiplying 9.03 kg by 9.80665 m/s^2, the acceleration due to gravity at sea level, so $T \approx 9.03 \cdot 9.80665\,\text{N} \approx 88.554\,\text{N}$. The wave speed is

$$c \approx \sqrt{T/\varrho} = \sqrt{\frac{88.554\,\text{N}}{(4.466 \times 10^{-3}\,\text{kg/m})}} \approx 140.81\,\text{m/s}.$$

The physical problem translates to the mathematical problem

$$\left\{ \begin{array}{c} \frac{\partial^2 u}{\partial t^2} = 140.81^2 \frac{\partial^2 u}{\partial x^2},\ 0 < x < 0.640,\ t > 0, \\ u(0,t) = u(0.640,t) = 0,\ t > 0, \\ u(x,0) = 0,\ 0 < x < 0.640, \\ \frac{\partial u}{\partial t}(x,0) = \sin\left(\frac{\pi x}{0.640}\right) - 0.2 \sin\left(\frac{2\pi x}{0.640}\right),\ 0 < x < 0.640 \end{array} \right\}.$$

This problem fits directly into the form of problem (11.21) through (11.24), in the special case of the string starting from the equilibrium position. Note especially that for a guitar string the BCs are $u(0,t) = u(0.640,t) = 0$, so $L = 0.640$ in this problem. Also, we are given that $c = 140.81$ m/s. We substitute $u(x,0) \equiv 0$ into the general solution (11.25) to get

$$u(x,t) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi ct}{0.640}\right) \sin\left(\frac{n\pi x}{0.640}\right).$$

The initial condition is satisfied by solving

$$\sin\left(\frac{\pi x}{0.640}\right) - 0.2\sin\left(\frac{2\pi x}{0.640}\right) = \frac{\partial u}{\partial t}(x,0) = \sum_{n=1}^{\infty}\left(\frac{n\pi c}{0.640}b_n\right)\sin\left(\frac{n\pi x}{0.640}\right),\ 0 < x < 0.640.$$

This is a special case of a Fourier sine series expansion because there are only two Fourier components in $\frac{\partial u}{\partial t}(x,0)$, specifically the $n = 1$ and $n = 2$ modes. So, $\frac{\pi c b_1}{0.640} = 1$, $\frac{2\pi c b_1}{0.640} = 0.2$, and all other b_n's are zero. The solution of the PDE-BVP-IVP gives the displacement of the string to be

$$\begin{aligned} u(x,t) \approx\ & 0.00145\sin\left(\frac{\pi x}{0.640}\right)\sin\left(\frac{140.81\pi t}{0.640}\right) \\ & - 0.000145\sin\left(\frac{2\pi x}{0.640}\right)\sin\left(\frac{281.62\pi t}{0.640}\right), \end{aligned}$$

to three significant digits. ○

Example 11.8

A taut string 4 m long with wave speed $c = 60$ m/s is initially in the position $u(x,0) = 50(1 - \frac{1}{5}\cos\left(\frac{\pi x}{2}\right))$ and is initially at rest. Determine the displacement of the string, as a function of position and time.

Method: The physical problem translates to a mathematical problem, specifically the PDE-BVP-IVP

$$\left\{\begin{array}{c} \frac{\partial^2 u}{\partial t^2} = 60^2\,\frac{\partial^2 u}{\partial x^2},\ 0 < x < 4,\ t > 0, \\ u(0,t) = u(4,t) = 0,\ t > 0, \\ u(x,0) = 50(1 - \frac{1}{5}\cos\left(\frac{\pi x}{2}\right)),\ 0 < x < 4, \\ \frac{\partial u}{\partial t}(x,0) = 0,\ 0 < x < 4 \end{array}\right\}.$$

This problem fits directly into the form of problem (11.21) through (11.24), in the special case of the string starting from rest. Note especially that the BCs are $u(0,t) = u(4,t) = 0$, so $L = 4$ in this problem. Also, we are given that $c = 60$ m/s, so the solution of the PDE-BVP, along with the initial condition on the velocity, is

$$u(x,t) = \sum_{n=1}^{\infty} a_n \sin\left(\frac{n\pi x}{4}\right)\cos\left(\frac{60n\pi t}{4}\right).$$

The initial condition is satisfied by solving

$$50\left(1 - \frac{1}{5}\cos\left(\frac{\pi x}{2}\right)\right) = u(x,0) = \sum_{n=1}^{\infty} a_n \sin\left(\frac{n\pi x}{4}\right),\ 0 < x < 4.$$

The coefficients in the Fourier sine series are given by

$$
\begin{aligned}
a_n &= \frac{2}{4}\int_0^4 \left(50 - 10\cos\left(\frac{\pi x}{2}\right)\right)\sin\left(\frac{n\pi x}{4}\right)dx \\
&= \frac{1}{2}\left(\int_0^4 50\sin\left(\frac{n\pi x}{4}\right)dx - \int_0^4 10\cos\left(\frac{\pi x}{2}\right)\sin\left(\frac{n\pi x}{4}\right)dx\right) \\
&= 25\left[\frac{4\cos\left(\frac{n\pi x}{4}\right)}{-n\pi}\right]_0^4 - 5\left\{\begin{array}{ll} \int_0^4 \cos\left(\frac{2\pi x}{4}\right)\sin\left(\frac{2\pi x}{4}\right)dx, & n=2 \\ \int_0^4 \cos\left(\frac{2\pi x}{4}\right)\sin\left(\frac{n\pi x}{4}\right)dx, & n\neq 2 \end{array}\right\} \\
&= 100\left(\frac{1-(-1)^n}{n\pi}\right) - 5\left\{\begin{array}{ll} \int_0^4 \frac{1}{2}\sin\pi x\,dx, & n=2 \\ \int_0^4 \frac{1}{2}\left(\sin\left(\frac{(n-2)\pi x}{4}\right)+\sin\left(\frac{(n+2)\pi x}{4}\right)\right)dx, & n\neq 2 \end{array}\right\} \\
&= \cdots = \frac{100}{n\pi}\left((-1)^n - 1\right) - \frac{40}{\pi}\left\{\begin{array}{ll} 0, & n=2 \\ \frac{n}{n^2-4}\left(1-(-1)^n\right), & n\neq 2 \end{array}\right\}.
\end{aligned}
$$

To simplify the conclusions, notice that *both* terms in a_2 are zero and further that $\left(1-(-1)^n\right)$ is zero for all *even* n. In addition, a little more algebraic simplification for $n = odd = 2k-1$ explains why the displacement as a function of position and time is given by

$$
u(x,t) = \sum_{k=1}^{\infty} \frac{160\left((2k-1)^2-5\right)}{\pi(2k-1)\left((2k-1)^2-4\right)}\sin\left(\frac{(2k-1)\pi x}{4}\right)\cos\left(15(2k-1)\pi t\right). \bigcirc
$$

Just as for the heat equation, the 1D wave equation could also have time-dependent boundary conditions, for example,

$$
u(0,t) = \delta(t),\ u(L,t) = \epsilon(t).
$$

Similarly to the method used in Example 11.5 in Section 11.1, we would assume $u(x,t) = v(x,t) + w(x,t)$, where v would be chosen to satisfy *only* the boundary conditions, and, in principle, we would choose v to be in a simple form, for example, $v(x,t) \triangleq \beta(t) + \gamma(t)x$.

11.2.1 Problems

1. Solve the "plucked" string model

$$
\left\{\begin{array}{c} \frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2},\ 0<x<3,\ t>0, \\ u(0,t) = u(3,t) = 0,\ t>0, \\ u(x,0) = \left\{\begin{array}{ll} x, & 0<x<1 \\ (3-x)/2, & 1<x<3 \end{array}\right\}, \\ \frac{\partial u}{\partial t}(x,0) = 0,\ 0<x<3 \end{array}\right\}.
$$

2. Solve the "plucked" string model

$$\left\{\begin{array}{c} \frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2},\ 0 < x < 4,\ t > 0, \\ \\ u(0,t) = u(4,t) = 0,\ t > 0, \\ \\ u(x,0) = \left\{\begin{array}{ll} x, & 0 < x < 1 \\ (4-x)/3, & 1 < x < 4 \end{array}\right\}, \\ \\ \frac{\partial u}{\partial t}(x,0) = 0,\ 0 < x < 4 \end{array}\right\}.$$

3. Solve, as easily and simply as possible,

$$\left\{\begin{array}{c} \frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2},\ 0 < x < L,\ t > 0, \\ \\ u(0,t) = \frac{\partial u}{\partial x}(L,t) = 0,\ t > 0, \\ \\ u(x,0) = \sin\left(\frac{\pi x}{2L}\right) - \frac{1}{3}\sin\left(\frac{5\pi x}{2L}\right), 0 < x < L, \\ \\ \frac{\partial u}{\partial t}(x,0) = 0,\ 0 < x < L \end{array}\right\}.$$

4. Suppose $g(x) = \sum_{n=3}^{10} B_n \sin nx$ for some unspecified constants B_n. Find, in as simple a form as possible, the solution of the problem

$$\left\{\begin{array}{c} \frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2},\ 0 < x < \pi,\ t > 0, \\ \\ u(0,t) = u(\pi,t) = 0,\ t > 0, \\ \\ u(x,0) = 0,\ 0 < x < \pi \\ \\ \frac{\partial u}{\partial t}(x,0) = g(x),\ 0 < x < \pi \end{array}\right\}.$$

   Your final conclusion should be in terms of the B_n's.

In problems 5–8, solve the given PDE-BVP-IVP:

5. $$\left\{\begin{array}{c} \frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2},\ 0 < x < L,\ t > 0, \\ \\ u(0,t) = u(L,t) = 0,\ t > 0, \\ \\ u(x,0) = 1 - \cos\left(\frac{\pi x}{L}\right),\ 0 < x < L, \\ \\ \frac{\partial u}{\partial t}(x,0) = 0,\ 0 < x < L \end{array}\right\}$$

6. $$\left\{\begin{array}{c} \frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2},\ 0 < x < 1,\ t > 0, \\ u(0,t) = u(1,t) = 0,\ t > 0, \\ u(x,0) = 4(x - x^2),\ 0 < x < 1, \\ \frac{\partial u}{\partial t}(x,0) = 0,\ 0 < x < 1 \end{array}\right\}$$

7. $$\left\{\begin{array}{c} \frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2},\ 0 < x < L,\ t > 0, \\ u(0,t) = u(L,t) = 0,\ t > 0, \\ u(x,0) = 0,\ 0 < x < L, \\ \frac{\partial u}{\partial t}(x,0) = \cos\left(\frac{2\pi x}{L}\right),\ 0 < x < L \end{array}\right\}$$

8. $$\left\{\begin{array}{c} \frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} - u,\ 0 < x < L,\ t > 0, \\ u(0,t) = u(L,t) = 0,\ t > 0, \\ u(x,0) = f(x),\ 0 < x < L, \\ \frac{\partial u}{\partial t}(x,0) = 0,\ 0 < x < L \end{array}\right\}$$

9. Solve

$$\left\{\begin{array}{c} \frac{\partial^2 u}{\partial t^2} = 4\frac{\partial^2 u}{\partial x^2},\ 0 < x < L,\ t > 0, \\ u(0,t) = \delta(t),\ u(L,t) = \epsilon(t),\ t > 0, \\ u(x,0) = f(x),\ 0 < x < L, \\ \frac{\partial u}{\partial t}(x,0) = g(x),\ 0 < x < L \end{array}\right\}$$

in terms of the unspecified functions $\delta(t)$, $\epsilon(t)$, $f(x)$, and $g(x)$. [Hint: It may be necessary to use the method of variation of parameters, for solutions of ODEs in t, or the result of Problem 5.4.1.12, at some point in your work.]

10. Assume h is a positive but unspecified constant. Solve the PDE-BVP-IVP of an inclined string:

$$\left\{\begin{array}{c} \frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2},\ 0 < x < L,\ t > 0, \\ u(0,t) = 1,\ u(L,t) = 1 + h,\ t > 0, \\ u(x,0) = \frac{\partial u}{\partial t}(x,0) = 0,\ 0 < x < L \end{array}\right\}.$$

[Hint: First, find the equilibrium solution satisfying the ODE-BVP $0 = v''$, $v(0) = 1$, $v(L) = 1 + h$.]

11. Find all solutions of the leaky transmission line modeled by the PDE+BCs

$$\left\{ \begin{array}{c} \frac{\partial^2 u}{\partial t^2} + 2\frac{\partial u}{\partial t} = c^2 \frac{\partial^2 u}{\partial x^2} - u,\ 0 < x < L,\ t > 0, \\ \\ u(0,t) = 0,\ u(L,t) = 0,\ t > 0 \end{array} \right\}.$$

12. A solution of PDE (11.21), BCs (11.22), and IC $\frac{\partial u}{\partial t}(x,0) \equiv 0$ is given by

$$u(x,t) = \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi ct}{L}\right) \sin\left(\frac{n\pi x}{L}\right).$$

Suppose that the string is further constrained to satisfy $u\left(\frac{L}{2}, t\right) \equiv 0$. What condition does this impose on the coefficients a_n? [Problems 11.2.1.12 and 11.2.1.13 are on page 150 of Pinsky (1998).]

13. A solution of PDE (11.21), BCs (11.22), and IC $\frac{\partial u}{\partial t}(x,0) \equiv 0$ is given by

$$u(x,t) = \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi ct}{L}\right) \sin\left(\frac{n\pi x}{L}\right).$$

Suppose that the string is further constrained to satisfy $u\left(\frac{L}{3}, t\right) \equiv 0$. What condition does this impose on the coefficients a_n?

14. Suppose we model the vibrations of a string forced by the wind by the PDE-BVP

$$\left\{ \begin{array}{c} \frac{\partial^2 y}{\partial t^2} = c^2 \frac{\partial^2 y}{\partial x^2} + \cos \omega t, \\ \\ y(0,t) = y(L,t) = 0,\ 0 < t < \infty \end{array} \right\}.$$

For what value(s) of the constant forcing frequency ω can a solution have $\max_{0<x<L} |y(x,t)| \to \infty$ as $t \to \infty$? What phenomenon for forced ODEs is involved in this situation?

15. Find all eigenvalues and eigenfunctions of the ODE-BVP

$$\left\{ \begin{array}{c} \frac{d^4 X}{dx^4} + \lambda X = 0,\ 0 < x < L, \\ \\ X(0) = X(L) = 0, \\ \\ X''(0) = X''(L) = 0 \end{array} \right\}$$

by following the steps as follows:

(a) First, explain why $\lambda = 0$ is not an eigenvalue. [Hint: Explain why $\lambda = 0$ implies that $X(x)$ is a third-degree polynomial, and then explain why it can't satisfy the four BCs.]

(b) Second, explain why $\lambda > 0$ cannot be an eigenvalue. [Hint: Use a method similar to that used in Examples 9.29 and 9.30 in Section 9.6. To be more explicit, multiply through by $X(x)$ the fourth-order ODE that $X(x)$ satisfies, use two integration by parts, and use the four BCs to get $0 = \int_0^L (X''(x))^2 dx + \lambda \int_0^L (X(x))^2 dx$.]

(c) Third, find all eigenvalues and eigenfunctions for the case $\lambda < 0$. [Hint: Define $\omega = (-\lambda)^{1/4}$ for convenience.]

16. Use the results of Problem 11.2.1.15 to solve the PDE-BVP-IVP for the vibrations of a beam whose ends are "pinned," modeled by

$$\left\{ \begin{array}{c} \frac{\partial^2 y}{\partial t^2} = -EI \frac{\partial^4 y}{\partial x^4} \\ y(0,t) = y(L,t) = 0,\ 0 < t < \infty \\ \frac{\partial^2 y}{\partial x^2}(0,t) = \frac{\partial^2 y}{\partial^2 x}(L,t) = 0,\ 0 < t < \infty, \\ y(x,0) = f(x),\ \frac{\partial y}{\partial t}(x,0) = g(x), 0 < x < L \end{array} \right\},$$

where E is the Young's modulus and I is the moment of inertia.

17. Suppose we model the vibrations of a string forced by the wind by the PDE-BVP

$$\left\{ \begin{array}{c} \frac{\partial^2 y}{\partial t^2} + 2\beta \frac{\partial y}{\partial t} = c^2 \frac{\partial^2 y}{\partial x^2} + \cos \omega t, \\ y(0,t) = y(L,t) = 0,\ 0 < t < \infty \end{array} \right\},$$

where β is a positive constant. What phenomenon analogous to that for forced ODEs is involved in this situation?

11.3 Laplace Equation in a Rectangle

In Section 10.3, we saw that the equilibrium temperature distribution in a 2- or 3D region satisfies the Laplace or Poisson equation. In addition, we saw that the electrostatic potential also satisfies the Laplace or Poisson equation.

Consider the equilibrium solution of the 2D heat equation $\frac{\partial T}{\partial t} = \alpha \nabla^2 T(x,y)$, that is, the solutions of $\nabla^2 T(x,y) = 0$, that is, the functions $T(x,y)$ that satisfy

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0.$$

Shown in Figure 11.4 is a picture of a slab $0 \le x \le 2, 0 \le y \le 1$, with boundary conditions indicated on the four sides (1) $0 < x < 2, y = 0$, (2) $0 < x < 2, y = 1$, (3) $0 < y < 1, x = 0$, (4) $0 < y < 1, x = 2$. Assume that the slab's top and bottom surfaces (in the z-direction) are insulated.

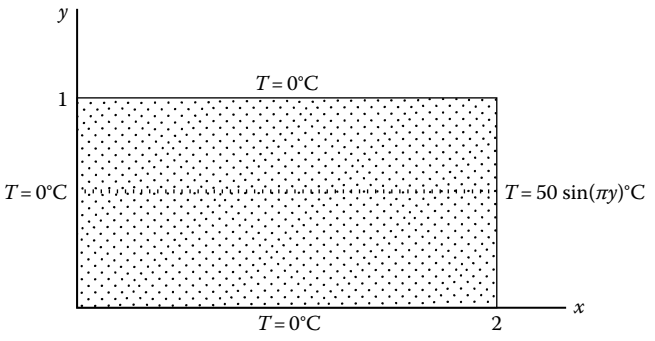


FIGURE 11.4
Example 11.9: equilibrium heat problem.

Example 11.9

Solve the mathematical problem to find the equilibrium solution of the heat equation in the slab that satisfies the boundary conditions given in the picture, that is,

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0,\ 0 < x < 2,\ 0 < y < 1, \tag{11.28}$$

$$T(x,0) = T(x,1) = 0,\ 0 < x < 2, \tag{11.29}$$

$$T(0,y) = 0,\ T(2,y) = 50\sin \pi y,\ 0 < y < 1. \tag{11.30}$$

We will see that we can reuse what we learned about separation of variables and ODE-BVP eigenvalue problems.

Method: The most important thing to notice first about the problem is that there are two homogeneous boundary conditions on *parallel* sides, specifically $T(x,0) = T(x,1) = 0$ on the horizontal sides with $y = 0$ and $y = 1$. This is like Example 11.1 in Section 11.1's two homogeneous boundary conditions on *parallel* sides, specifically $T(0,t) = T(L,t) = 0$ on the vertical sides with $x = 0$ and $x = L$, which led to an ODE-BVP whose eigenfunctions were $X_n(x) = \sin\left(\frac{n\pi x}{L}\right),\ n = 1, 2, \ldots$.

Here, the eigenfunctions are functions of y and $L = 1$ because the homogeneous boundary conditions are on the lines $y = 0$ and $y = 1$, so the eigenfunctions are $Y_n(y) = \sin(n\pi y),\ n = 1, 2, \ldots$, and correspondingly we look for solutions of the PDE in the form $T(x,y) = X(x)Y_n(y) = X(x)\sin(n\pi y),\ n = 1, 2, \ldots$. Such functions $T(x,y)$ automatically satisfy the first two (homogeneous) boundary conditions of the problem. Substituting $T(x,y) = X(x)\sin(n\pi y)$ into the PDE gives

$$0 = \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = \frac{d^2 X}{dx^2}(x)\sin(n\pi y) + X(x)(-(n\pi)^2 \sin(n\pi y)),\ n = 1, 2, \ldots,$$

which leads to ODE

$$\frac{d^2 X}{dx^2}(x) - (n\pi)^2 X(x) = 0, \tag{11.31}$$

whose general solution can be written in the form of either

$$X(x) = X_n(x) = c_1 e^{n\pi x} + c_2 e^{-n\pi x}$$

or

$$X(x) = X_n(x) = d_1 \cosh(n\pi x) + d_2 \sinh(n\pi x).$$

We will see that the latter is more convenient than the former.

We now have product solutions $T_n(x, y) = (d_1 \cosh(n\pi x) + d_2 \sinh(n\pi x)) \sin(n\pi y)$, where the d_1 and d_2 in fact can depend on n. Again we use the principle of linear superposition to "add up" all product solutions to get the general solution of the PDE + (two homogeneous BCs on parallel sides):

$$T(x, y) = \sum_{n=1}^{\infty} \sin(n\pi y)(a_n \cosh(n\pi x) + b_n \sinh(n\pi x)). \tag{11.32}$$

It is this general solution that we substitute into the remaining boundary conditions, which in this problem play the same role as two initial conditions did in the 1D wave equation. The third boundary condition is, using the facts $\cosh(0) = 1$ and $\sinh(0) = 0$,

$$0 = T(0, y) = \sum_{n=1}^{\infty} a_n \sin(n\pi y).$$

We can either recognize immediately that all a_n should be zero, or we can go to the extra work of noting that the Fourier sine series coefficients are given by $a_n = \frac{2}{1} \int_0^1 0 \sin(n\pi y) dy = 0,\ n = 1, 2, \ldots$. The fourth boundary condition is, using the facts that $\cosh(0) = 1$ and $\sinh(0) = 0$,

$$50 \sin \pi y = T(2, y) = \sum_{n=1}^{\infty} b_n \sinh(2n\pi) \sin(n\pi y).$$

Because the left-hand side is itself a Fourier sine series with $L = 1$, we don't need integrals to see that $b_1 = 50/\sinh(2\pi)$ and $b_n = 0,\ n = 2, 3, \ldots$. The final conclusion is that the equilibrium temperature distribution is given by

$$T(x, y) = 50 \frac{\sinh(\pi x)}{\sinh(2\pi)} \sin(\pi y). \bigcirc$$

What if, instead of using the general solution form (11.32), we had used the general solution form

$$T(x, y) = \sum_{n=1}^{\infty} \sin(n\pi y) \left(A_n e^{n\pi x} + B_n e^{-n\pi x}\right)?$$

We could still solve the problem, but it would take extra work: The third boundary condition is

$$0 = T(0, y) = \sum_{n=1}^{\infty} \sin(n\pi y)(A_n + B_n),$$

so Fourier sine series analysis implies that for all $n \geq 1$, $A_n + B_n = 0$, that is, $B_n = -A_n$. The fourth boundary condition is

$$50 \sin \pi y = T(2, y) = \sum_{n=1}^{\infty} \sin(n\pi y) \left(A_n e^{2n\pi} + B_n e^{-2n\pi} \right),$$

so, for $n = 1$, $50 = A_n e^{2\pi} + B_n e^{-2\pi}$, and, for all $n \geq 2$, $0 = A_n e^{2n\pi} + B_n e^{-2n\pi}$. Substituting $B_1 = -A_1$ gives $50 = A_1 \left(e^{2\pi} - e^{-2\pi} \right)$; hence, $A_1 = 50/\left(e^{2\pi} - e^{-2\pi} \right)$.

For $n \geq 2$, $B_n = -A_n$ gives $0 = A_n \left(e^{2n\pi} - e^{-2n\pi} \right)$; hence, $A_n = 0/\left(e^{2n\pi} - e^{-2n\pi} \right)$; hence, $A_n = B_n = 0$. Putting this all together, the series solution of the PDE-BVP reduces to a single term corresponding to $n = 1$:

$$T(x, y) = \sin \pi y \cdot \frac{50}{e^{2\pi} - e^{-2\pi}} \cdot \left(e^{\pi x} - e^{-\pi x} \right) = 50 \cdot \sin \pi y \cdot \frac{2}{e^{2\pi} - e^{-2\pi}} \cdot \frac{e^{\pi x} - e^{-\pi x}}{2}$$

$$= 50 \cdot \frac{\sinh(\pi x)}{\sinh(2\pi)} \cdot \sin(\pi y),$$

just as we found earlier using less work.

In a similar way, we can solve problems with other pairs of homogeneous boundary conditions on parallel sides by using other lines of Table 11.1.

Example 11.10

Find the electrostatic potential inside the rectangle $0 < x < 2, 0 < y < 1$, assuming the electric field normal to the horizontal sides is zero and the potential is specified on the vertical sides.

Method: It helps to translate the physical problem into the picture found in Figure 11.5.

The most important thing to notice first about the problem is that there are two homogeneous boundary conditions on *parallel* sides. Let $V = V(x, y)$ be the electrostatic potential. Because the electric field is given by $\mathbf{E} = \nabla V$, the boundary conditions are

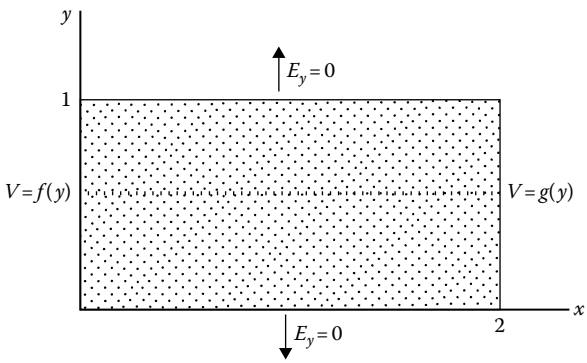


FIGURE 11.5
Example 11.10: electrostatic potential.

$\frac{\partial V}{\partial y}(x,0) = E_y(x,0) = 0$ and $\frac{\partial V}{\partial y}(x,1) = E_y(x,1) = 0$ on the horizontal sides with $y = 0$ and $y = 1$, similar to Example 11.9. So, $L = 1$ in this problem.

The physical problem translates into the mathematical problem

$$\left\{\begin{array}{l} \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = 0,\ 0 < x < 2,\ 0 < y < 1, \\ \\ \frac{\partial V}{\partial y}(x,0) = \frac{\partial V}{\partial y}(x,1) = 0,\ 0 < x < 2, \\ \\ V(0,y) = f(y),\ V(2,y) = g(y),\ 0 < y < 1 \end{array}\right\}.$$

We should express all of the constants found in the general solution in terms of integrals involving the functions $f(y)$ and $g(y)$.

This leads to an ODE-BVP whose eigenfunctions are $Y_n(y) = \cos(n\pi y),\ n = 0, 1, 2, \ldots$, as found in the second group of entries of Table 11.1. Note that $\lambda = 0$ is an eigenvalue, too, and the corresponding eigenfunction, $Y_0(y) = 1$, can be considered an "honorary" cosine function by taking $n = 0$ in $\cos\left(\frac{n\pi y}{L}\right)$. Correspondingly, we look for solutions of the PDE in the form $V(x,y) = X(x)Y_n(y) = X(x)\cos(n\pi y),\ n = 0, 1, 2, \ldots$. Such functions of (x,y) automatically satisfy the first two (homogeneous) boundary conditions of the problem. Substituting $V(x,y) = X(x)\cos(n\pi y)$ into the PDE gives

$$0 = \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = \frac{d^2 X}{dx^2}(x)\cos(n\pi y) + X(x)(-(n\pi)^2\cos(n\pi y)),\ n = 0, 1, 2, \ldots,$$

which leads to the familiar ODE (11.31), that is, $X''(x) - (n\pi)^2 X(x) = 0$. For $n = 1, 2, \ldots$, we get $X_n(x) = A_n\cosh(n\pi x) + B_n\sinh(n\pi x)$. For $n = 0$, the solution of $X''(x) - 0 = 0$ is $X_0(x) = A_0 + B_0 x$.

The general solution of the PDE and the first two (homogeneous) boundary conditions is

$$V(x,y) = \frac{a_0 + b_0 x}{2} + \sum_{n=1}^{\infty}\cos(n\pi y)\left(a_n\cosh(n\pi x) + b_n\sinh(n\pi x)\right),$$

where $A_0 = \frac{a_0}{2}$, $B(0) = \frac{b_0}{2}$, $A_n = a_n$, $B_n = b_n,\ n = 1, 2, \ldots$. It is this general solution that we substitute into the remaining boundary conditions. Using the facts $\cosh(0) = 1$ and $\sinh(0) = 0$, the third boundary condition is

$$f(y) = V(0,y) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n\cos(n\pi y).$$

This is a standard Fourier cosine series expansion problem with $L = 1$, so $a_0 = \frac{2}{1}\int_0^1 f(y)\,dy$ and, for $n = 1, 2, \ldots$, $a_n = \frac{2}{1}\int_0^1 f(y)\cos(n\pi y)\,dy$.

The fourth boundary condition is

$$g(y) = V(2,y) = \frac{a_0 + 2b_0}{2} + \sum_{n=1}^{\infty}(a_n\cosh(2n\pi) + b_n\sinh(2n\pi))\cos(n\pi y).$$

This is a standard Fourier cosine series with $L = 1$, so

$$a_0 + 2b_0 = \frac{2}{1}\int_0^1 g(y)\,dy,$$

and, for $n = 1, 2, \ldots$,

$$a_n \cosh(2n\pi) + b_n \sinh(2n\pi) = \frac{2}{1} \int_0^1 g(y) \cos(n\pi y)\, dy.$$

It is convenient to denote the *coefficients* of the Fourier cosine series of $g(y)$ by

$$g_0 = \frac{2}{1} \int_0^1 g(y)\, dy \text{ and } g_n = \frac{2}{1} \int_0^1 g(y) \cos(n\pi y)\, dy,\ n = 1, 2, \ldots, \tag{11.33}$$

and similarly for $f(y)$ by

$$f_0 = \frac{2}{1} \int_0^1 f(y)\, dy \text{ and } f_n = \frac{2}{1} \int_0^1 f(y) \cos(n\pi y)\, dy,\ n = 1, 2, \ldots. \tag{11.34}$$

This enables us to solve for the third and fourth boundary conditions in a shorter notation: $V(0, y) = f(y)$ is equivalent to

$$a_0 = f_0, \quad \text{and} \quad a_n = f_n,\ n \geq 1,$$

and $V(2, y) = g(y)$ is equivalent to

$$a_0 + 2b_0 = g_0, \quad \text{and} \quad a_n \cosh(2n\pi) + b_n \sinh(2n\pi) = g_n,\ n \geq 1.$$

Using the value $a_0 = f_0$, we find $b_0 = \frac{g_0 - f_0}{2}$. For $n = 1, 2, \ldots$, using the value $a_n = f_n$, it takes a little more work, but we can solve for b_n: $f_n \cosh(2n\pi) + b_n \sinh(2n\pi) = g_n$ yields $b_n \sinh(2n\pi) = g_n - f_n \cosh(2n\pi)$; hence,

$$b_n = \frac{g_n - f_n \cosh(2n\pi)}{\sinh(2n\pi)}.$$

Putting this all together, we know that the solution of the PDE-BVP is

$$V(x, y) = \frac{f_0}{2} + \frac{(g_0 - f_0)x}{4} + \sum_{n=1}^{\infty} \cos(n\pi y) \left(f_n \cosh(n\pi x) + \frac{g_n - f_n \cosh(2n\pi)}{\sinh(2n\pi)} \sinh(n\pi x) \right), \tag{11.35}$$

where the coefficients $f_n,\ g_n, n = 0, 1, 2, \ldots$, are given in (11.33) and (11.34) in terms of integrals involving the functions $f(y),\ g(y)$. ○

11.3.1 Using Clairvoyance to Choose Alternatives to cosh $\left(\frac{n\pi x}{L}\right)$ and sinh $\left(\frac{n\pi x}{L}\right)$

Let's combine all of the terms in (11.35) that have f_n as a factor:

$$f_n \cosh(n\pi x) - \frac{f_n \cosh(2n\pi)}{\sinh(2n\pi)} \sinh(n\pi x) = \frac{f_n \left(\cosh(n\pi x) \sinh(2n\pi) - \cosh(2n\pi) \sinh(n\pi x) \right)}{\sinh(2n\pi)}$$

$$= \frac{f_n}{\sinh(2n\pi)} \left(\sinh(2n\pi) \cosh(n\pi x) - \cosh(2n\pi) \sinh(n\pi x) \right) = \frac{\sinh(n\pi(2 - x))}{\sinh(2n\pi)} f_n,$$

the latter step by using a hyperbolic function identity,

$$\sinh(a - b) = \sinh(a) \cosh(b) - \cosh(a) \sinh(b).$$

So, we can rewrite the solution (11.35) in a nicer looking form:

$$V(x,y) = \frac{1}{4}\left(f_0 \cdot (2-x) + g_0 \cdot x\right) + \sum_{n=1}^{\infty} \cos(n\pi y)\left(\frac{\sinh(n\pi(2-x))}{\sinh(2n\pi)} f_n + \frac{\sinh(n\pi x)}{\sinh(2n\pi)} g_n\right), \tag{11.36}$$

where the coefficients $f_n,\ g_n, n = 0, 1, 2, \ldots,$ are given in (11.33) and (11.34) in terms of integrals involving the functions $f(y),\ g(y)$.

In (11.36), the final conclusion of the set of functions $\{\sinh(n\pi(2-x)),\ \sinh(n\pi x)\}$ appears instead of the original basic set of solutions $\{\cosh(n\pi x),\ \sinh(n\pi x)\}$ for the ODE $X''(x) - (n\pi)^2 X(x) = 0$. But, if we retrace our steps, we see that we got $\sinh(n\pi(2-x))$ from a linear combination of the functions $\cosh(n\pi x),\ \sinh(n\pi x)$. In Chapter 3, we saw that we can have many basic sets of solutions to the same ODE. In fact, if we choose *any two* of the four functions $\cosh(n\pi x),\ \sinh(n\pi x), \cosh(n\pi(2-x)),\ \sinh(n\pi(2-x))$, we will get a basic set of solutions of $X''(x) - (n\pi)^2 X(x) = 0$. So, in retrospect, it would have been better to use

$$V(x,y) = \frac{a_0 + b_0 x}{2} + \sum_{n=1}^{\infty} \cos(n\pi y)(c_n \sinh(n\pi(2-x)) + d_n \sinh(n\pi x)), \tag{11.37}$$

as the general solution of the PDE + (two homogeneous BCs on parallel sides), rather than

$$V(x,y) = \frac{a_0 + b_0 x}{2} + \sum_{n=1}^{\infty} \cos(n\pi y)(a_n \cosh(n\pi x) + b_n \sinh(n\pi x)).$$

But, why would we have known that in advance? The clue is found in the third and fourth boundary conditions in Example 11.10: $V(0,y) = f(y),\ V(2,y) = g(y),\ 0 < y < 1$. The function $h(x) = \sinh(n\pi x)$ is convenient because $h(0) = 0$, and the function $H(x) = \sinh(n\pi(2-x))$ is convenient because $H(2) = 0$.

For example, if the third and fourth boundary conditions were of the forms $\frac{\partial V}{\partial x}(0,y) = (\textit{function of } y)$ and $V(2,y) = (\textit{function of } y)$, then a convenient basic set of solutions of $X''(x) - (n\pi)^2 X(x) = 0$ would be $\{\cosh(n\pi x), \sinh(n\pi(2-x))\}$. Why? The function $h(x) = \cosh(n\pi x)$ is convenient because $\frac{dh}{dx}(0) = 0$, and the function $H(x) = \sinh(n\pi(2-x))$ is convenient because $H(2) = 0$.

Example 11.11

Solve the PDE-BVP

$$\left\{\begin{array}{c} \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0,\ 0 < x < 2,\ 0 < y < 1, \\ \\ \frac{\partial T}{\partial y}(x,0) = x,\ T(x,1) = 20,\ 0 < x < 2 \\ \\ T(0,y) = \frac{\partial T}{\partial x}(2,y) = 0,\ 0 < y < 1 \end{array}\right\}.$$

Do evaluate all integrals involved in order to reach a final conclusion.

Method: It helps to translate the mathematical symbols into the picture of the problem given in Figure 11.6.

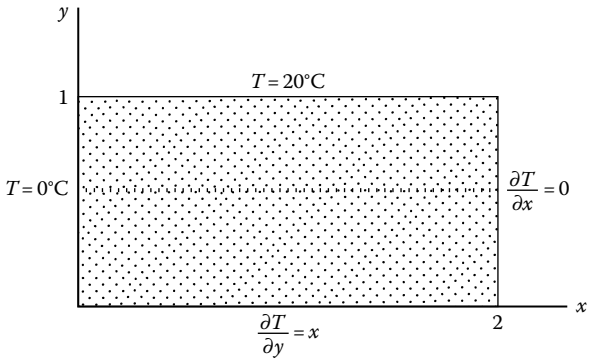


FIGURE 11.6
Example 11.11: equilibrium heat problem.

The most important thing to notice first about the problem is that there are two homogeneous boundary conditions on *parallel* sides, specifically $T(0,y) = \frac{\partial T}{\partial x}(2,y) = 0$ on the horizontal sides with $x = 0$ and $x = 2$. So, $L = 2$ in this problem. This leads to an ODE-BVP whose eigenfunctions are $X_n(x) = \sin\left(\frac{\left(n-\frac{1}{2}\right)\pi}{2}x\right)$, $n = 1, 2, \ldots$, as found in the third group of entries of Table 11.1. Correspondingly, we look for solutions of the PDE in the form $T(x,y) = X_n(x)Y(y) = \sin\left(\frac{\left(n-\frac{1}{2}\right)\pi}{2}x\right)Y(y)$, $n = 1, 2, \ldots$. Such functions $T(x,y)$ automatically satisfy the third and fourth (homogeneous) boundary conditions of the problem. Substituting $T(x,y) = \sin\left(\frac{\left(n-\frac{1}{2}\right)\pi}{2}x\right)Y(y)$ into the PDE gives, for $n = 1, 2, 3, \ldots$,

$$0 = \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = -\left(\frac{\left(n-\frac{1}{2}\right)\pi}{2}\right)^2 \sin\left(\frac{\left(n-\frac{1}{2}\right)\pi}{2}x\right)Y(y) + \sin\left(\frac{\left(n-\frac{1}{2}\right)\pi}{2}x\right)\frac{d^2Y}{dy^2},$$

which leads to ODE $\frac{d^2Y}{dy^2}(x) - \left(\frac{\left(n-\frac{1}{2}\right)\pi}{2}\right)^2 Y(y) = 0$. Because the other two boundary conditions are of the forms $\frac{\partial T}{\partial y}(x,0) = (\textit{function of } x)$, $T(x,1) = (\textit{function of } x)$, it is most convenient to use the basic set of solutions $\left\{\cosh\left(\frac{(2n-1)\pi}{4}y\right), \sinh\left(\frac{(2n-1)\pi}{4}(1-y)\right)\right\}$.

A convenient general solution of the PDE+(two homogeneous BCs on parallel sides) is

$$T(x,y) = \sum_{n=1}^{\infty} \sin\left(\frac{(2n-1)\pi x}{4}\right)\left(A_n \cosh\left(\frac{(2n-1)\pi}{4}y\right) + B_n \sinh\left(\frac{(2n-1)\pi}{4}(1-y)\right)\right).$$

Before applying the third boundary condition, we calculate that

$$\frac{\partial T}{\partial y}(x,y) = \sum_{n=1}^{\infty} \sin\left(\frac{(2n-1)\pi x}{4}\right)\cdot\frac{(2n-1)\pi}{4}$$

$$\cdot\left(A_n \sinh\left(\frac{(2n-1)\pi}{4}y\right) - B_n \cosh\left(\frac{(2n-1)\pi}{4}(1-y)\right)\right).$$

The third boundary condition is

$$x = \frac{\partial T}{\partial y}(x,0) = \sum_{n=1}^{\infty} \sin\left(\frac{(2n-1)\pi x}{4}\right) \cdot \frac{(2n-1)\pi}{4} \cdot (-B_n) \cosh\left(\frac{(2n-1)\pi}{4}\right),\ 0 < x < 2.$$

This is a *generalized Fourier series* with $L = 2$. Orthogonality gives

$$-\left(\frac{(2n-1)\pi}{4}\right) B_n \cosh\left(\frac{(2n-1)\pi}{4}\right) = \frac{2}{2}\int_0^2 x \sin\left(\frac{(2n-1)\pi x}{4}\right) dx$$

$$= \cdots = \frac{\sin\left(\frac{(2n-1)\pi}{4} \cdot 2\right)}{\left(\frac{(2n-1)\pi}{4}\right)^2} = \frac{(-1)^{n+1}}{\left(\frac{(2n-1)\pi}{4}\right)^2};$$

hence,

$$B_n = \frac{(-1)^n}{\left(\frac{(2n-1)\pi}{4}\right)^3 \cosh\left(\frac{(2n-1)\pi}{4}\right)}.$$

The fourth boundary condition gives

$$20 = T(x,1) = \sum_{n=1}^{\infty} \sin\left(\frac{(2n-1)\pi x}{4}\right)\left(A_n \cosh\left(\frac{(2n-1)\pi}{4}\right)\right).$$

For this *generalized Fourier series* with $L = 2$, orthogonality gives

$$A_n \cosh\left(\frac{(2n-1)\pi}{4}\right) = \frac{2}{2}\int_0^2 20 \sin\left(\frac{(2n-1)\pi x}{4}\right) dx = \cdots = \frac{20}{\left(\frac{(2n-1)\pi}{4}\right)};$$

hence,

$$A_n = \frac{20}{\left(\frac{(2n-1)\pi}{4}\right) \cosh\left(\frac{(2n-1)\pi}{4}\right)}.$$

So our final conclusion is that the solution of the PDE-BVP is

$$T(x,y) = \sum_{n=1}^{\infty} \sin\left(\frac{(2n-1)\pi x}{4}\right)$$
$$\times \frac{20\cosh\left(\frac{(2n-1)\pi}{4}\, y\right) + \frac{(-1)^n}{\left(\frac{(2n-1)\pi}{4}\right)^2} \sinh\left(\frac{(2n-1)\pi}{4}\,(1-y)\right)}{\left(\frac{(2n-1)\pi}{4}\right)\cosh\left(\frac{(2n-1)\pi}{4}\right)}. \bigcirc$$

11.3.2 Contour Plot and 3D Plot Using *Mathematica*

For Example 11.11, we entered a partial sum of the first 100 terms in the sum for $T(x,y)$. The *Mathematica* command

ContourPlot[T[x, y], {x, 0, 2}, {y, 0, 1}, Contours → {0, 2.5, 5, 7.5, 10, 12.5, 15, 17.5, 20}, ContourStyle → {Thick, Dashed}]

gave the contour plot shown in Figure 11.7. Because $T(x, y)$ is the temperature, the contours, that is, level curves of T, are called **isotherms**.

After that, the *Mathematica* command

```
Plot3D[T[x, y], {x, 0, 2}, {y, 0, 1}]
```

gave the graph shown in Figure 11.8. Using a mouse to rotate *Mathematica*'s 3D plot, we get Figure 11.9, which shows the same graph as for Figure 11.8 but from a different viewpoint.

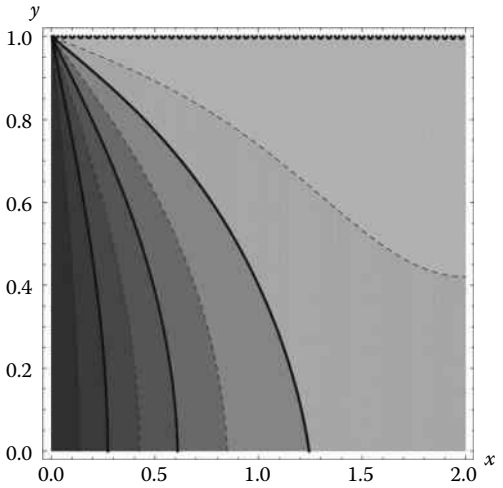


FIGURE 11.7
Contour plot for Example 11.11.

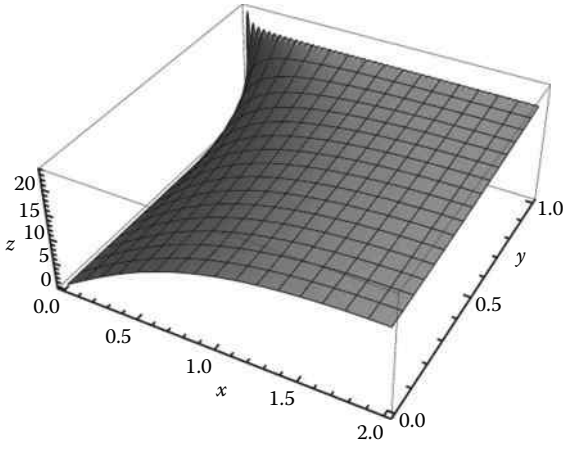


FIGURE 11.8
Graph of solution for Example 11.11.

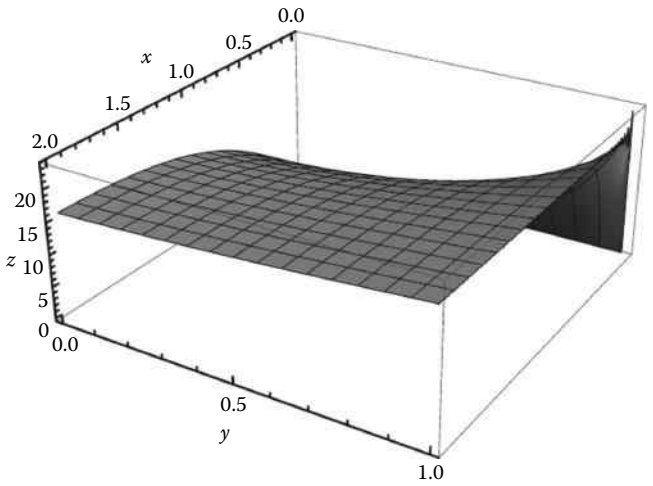


FIGURE 11.9
Rotated graph of solution for Example 11.11.

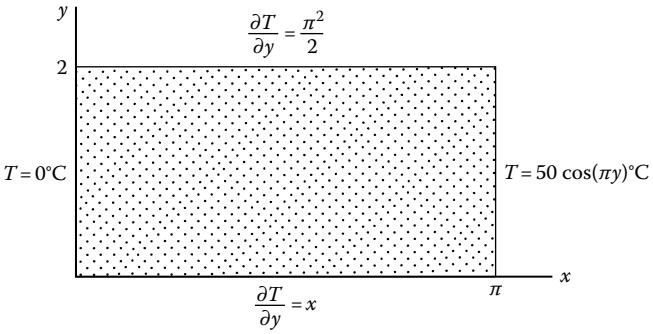


FIGURE 11.10
Example 11.12: equilibrium heat problem.

What about other problems, for example, what if our problem specifies that both the vertical and horizontal pairs of sides each have at least one nonzero boundary condition? We can use linear superposition, as illustrated in the two pictures in the following. Figure 11.10 shows a BVP in which there is no pair of homogeneous boundary conditions on parallel sides.

Figure 11.11 shows a "sum" of two problems: The solution $T(x, y)$ of the original problem can be expressed as $T(x, y) = T_1(x, y) + T_2(x, y)$ where $T_1(x, y)$ and $T_2(x, y)$ solve the two problems shown in Figure 11.11.

We must emphasize that this decomposition $T = T_1 + T_2$, which effectively decouples the contributions of the inhomogeneities on the horizontal sides from the inhomogeneity on the vertical sides, is possible because everything in the problem is linear.

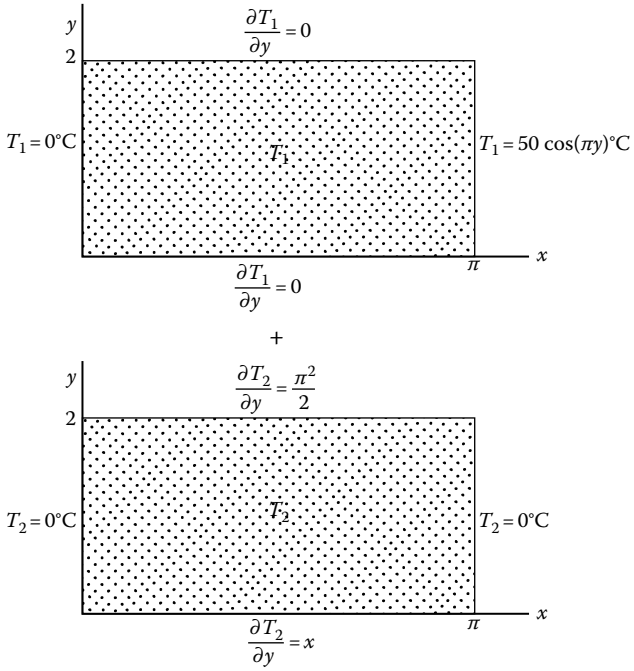


FIGURE 11.11
Example 11.12: sum of equilibrium heat problems.

Example 11.12

Solve the PDE-BVP

$$\left\{\begin{array}{c} \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0,\ 0 < x < \pi,\ 0 < y < 2, \\ \\ \frac{\partial T}{\partial y}(x,0) = x,\ \frac{\partial T}{\partial y}(x,2) = \frac{\pi^2}{2},\ 0 < x < \pi \\ \\ T(0,y) = 0,\ T(\pi,y) = 50\cos(\pi y),\ 0 < y < 2 \end{array}\right\}.$$

Do evaluate all integrals involved in order to reach a final conclusion.

Method: Figure 11.11 indicates that our solution is $T(x,y) = T_1(x,y) + T_2(x,y)$, where $T_1(x,y)$ solves

$$\left\{\begin{array}{c} \frac{\partial^2 T_1}{\partial x^2} + \frac{\partial^2 T_1}{\partial y^2} = 0,\ 0 < x < \pi,\ 0 < y < 2, \\ \\ \frac{\partial T_1}{\partial y}(x,0) = \frac{\partial T_1}{\partial y}(x,2) = 0,\ 0 < x < \pi, \\ \\ T_1(0,y) = 0,\ T_1(\pi,y) = 50\cos(\pi y),\ 0 < y < 2 \end{array}\right\}$$

and $T_2(x,y)$ solves

$$\left\{\begin{array}{c} \frac{\partial^2 T_2}{\partial x^2}+\frac{\partial^2 T_2}{\partial y^2}=0,\ 0<x<\pi,\ 0<y<2, \\ \\ \frac{\partial T_2}{\partial y}(x,0)=x,\ \frac{\partial T_2}{\partial y}(x,2)=\frac{\pi^2}{2},\ 0<x<\pi \\ \\ T_2(0,y)=T_2(\pi,y)=0,\ 0<y<2 \end{array}\right\}.$$

For the solution for $T_1(x,y)$, the two homogeneous boundary conditions on the lines $y=0$ and $y=2$ give us $L=2$. The general solution of the PDE and the two homogeneous boundary conditions is most conveniently written as

$$T_1(x,y)=\frac{a_0+b_0x}{2}+\sum_{n=1}^{\infty}\cos\left(\frac{n\pi y}{2}\right)\left(a_n\sinh\left(\frac{n\pi}{2}(\pi-x)\right)+b_n\sinh\left(\frac{n\pi x}{2}\right)\right).$$

We use the general solution also to satisfy the other two boundary conditions $T_1(0,y)=0, T_1(\pi,y)=50\cos(\pi y),\ 0<y<2$:

$$0=T_1(0,y)=\frac{a_0}{2}+\sum_{n=1}^{\infty}\cos\left(\frac{n\pi y}{2}\right)\sinh\left(\frac{n\pi}{2}\right)a_n;$$

hence, $a_n=0$ for $n\geq 0$, and

$$50\cos(\pi y)=T_1(\pi,y)=\frac{a_0+b_0\pi}{2}+\sum_{n=1}^{\infty}\cos\left(\frac{n\pi y}{2}\right)b_n\sinh\left(\frac{n\pi^2}{2}\right),$$

where T is measured in °C. Without having to calculate integrals, orthogonality gives

$$\frac{a_0+b_0\pi}{2}=0,\ \text{and, for}\ n\neq 2, b_n\sinh\left(\frac{n\pi^2}{2}\right)=0,$$

and, for $n=2$,

$$b_2\sinh(\pi^2)=50.$$

It follows that $b_n=0$ for $n\neq 2$ and

$$b_2=\frac{50}{\sinh(\pi^2)}.$$

So,

$$T_1(x,y)=50\cos(\pi y)\cdot\frac{\sinh(\pi x)}{\sinh(\pi^2)}.$$

When solving for $T_2(x,y)$, the two homogeneous boundary conditions on the lines $x=0$ and $x=\pi$ give us $L=\pi$. The general solution of the PDE and these two homogeneous boundary conditions is

$$T_2(x,y)=\sum_{n=1}^{\infty}\sin nx\left(a_n\cosh(ny)+b_n\cosh(n(2-y))\right);$$

the last term was chosen so as to more easily satisfy the fourth boundary condition. We use the general solution also to satisfy the third and fourth boundary conditions $\frac{\partial T_2}{\partial y}(x,0) = x$, $\frac{\partial T_2}{\partial y}(x,2) = \frac{\pi^2}{2}$, $0 < x < \pi$: first, we differentiate T_2 with respect to y to get

$$\frac{\partial T_2}{\partial y} = \sum_{n=1}^{\infty} n \sin nx \big(a_n \sinh(ny) - b_n \sinh(n(2-y))\big).$$

The third boundary condition is

$$x = \frac{\partial T_2}{\partial y}(x,0) = \sum_{n=1}^{\infty} \sin nx(-n \sinh(2n))b_n,\ 0 < x < \pi.$$

It follows that

$$(-n \sinh(2n))b_n = \frac{2}{\pi} \int_0^{\pi} x \sin nx\, dx = \cdots = \frac{2(-1)^{n+1}}{n};$$

hence,

$$b_n = \frac{2(-1)^n}{n \sinh(2n)}.$$

The fourth boundary condition is

$$\frac{\pi^2}{2} = \frac{\partial T_2}{\partial y}(x,2) = \sum_{n=1}^{\infty} \sin nx(n \sinh(2n))a_n,\ 0 < x < \pi;$$

hence,

$$n \sinh(2n)a_n = \frac{2}{\pi} \int_0^{\pi} \frac{\pi^2}{2} \sin nx\, dx = \cdots = \frac{\pi\big(1-(-1)^n\big)}{n},$$

which can be easily solved for b_n. So,

$$T_2(x,y) = \sum_{n=1}^{\infty} \frac{\pi}{n^2 \sinh(2n)} \sin nx\big(\big(1-(-1)^n\big)\cosh(ny) + 2(-1)^n \cosh(n(2-y))\big).$$

The solution to the original problem is

$$T(x,y) = 50\cos(\pi y) \cdot \frac{\sinh(\pi x)}{\sinh(\pi^2)}$$

$$+ \sum_{n=1}^{\infty} \frac{\pi \sin nx}{n^2 \sinh(2n)} \big(\big(1-(-1)^n\big)\cosh(ny) + 2(-1)^n \cosh(n(2-y))\big). \bigcirc$$

Learn More About It

Problem 11.3.3.13 leads to the concept of "Green's function." Much more about such functions can be found in *Green's Functions within Partial Differential Equations: with Complex Variables and Transform Methods*, by Hans Weinberger, Dover Publications Co. © 1995, and *Green's Functions with Applications*, by Dean G. Duffy, Chapman & Hall/CRC Press © 2001.

11.3.3 Problems

1. A thin slab measures 2π meters in the x-direction and π meters in the y-direction. Assume that the top and bottom surfaces (in the z-direction) are insulated and the temperature on the four edges satisfies $\frac{\partial T}{\partial x}(0,y)=0, T(x,0)=0, \frac{\partial T}{\partial x}(2\pi,y) = 10, T(x,\pi)=0$. Find the equilibrium temperature distribution inside the slab.
2. A thin slab measures 2π meters in the x-direction and π meters in the y-direction. Assume that the top and bottom surfaces (in the z-direction) are insulated and the temperature on the four edges satisfies $\frac{\partial T}{\partial x}(0,y)=0, T(x,0) = 100, \frac{\partial T}{\partial x}(2\pi,y)=0, T(x,\pi)=0$. Find the equilibrium temperature distribution inside the slab.
3. A thin slab measures 2π meters in the x-direction and π meters in the y-direction. Assume that the top and bottom surfaces (in the z-direction) are insulated and the temperature on the four edges satisfies $T(0,y) = 0, T(x,0) = 100, \frac{\partial T}{\partial y}(x,0) = -10, T(2\pi,y) = 0$. Find the equilibrium temperature distribution inside the slab. [No, there is no typographical error ... this is not one of the usual combination of boundary conditions on four sides.]
4. Find the equilibrium temperature at the center of a square, three of whose sides are kept at 0°C and whose fourth side is kept at 100°C.
5. Suppose $f(x) = \sum_{n=3}^{10} \alpha_n \sin\left((n-\frac{1}{2})x\right)$ for some unspecified constants α_n. Find, in as simple a form as possible, the solution of the problem

$$\left\{\begin{array}{c} \frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}=0, 0<x<\pi, 0<y<1, \\ \\ T(x,0)=20,\ T(x,1)=f(x), 0<x<\pi \\ \\ T(0,y)=\frac{\partial T}{\partial x}(\pi,y)=0, 0<y<1 \end{array}\right\}.$$

6. Solve

$$\left\{\begin{array}{c} \frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}=0, 0<x<a, 0<y<b, \\ \\ T(0,y)=0,\ T(a,y)=0, 0<y<b \\ T(x,0)=0, 0<x<a \\ \\ T(x,b)=\left\{\begin{array}{ll} \frac{2x}{a}, & 0<x<\frac{a}{2} \\ 2(1-\frac{x}{a}), & \frac{a}{2}<x<a \end{array}\right\} \end{array}\right\}.$$

7. Find the equilibrium temperature distribution in an a meters by b meters slab if the top and bottom surfaces (in the z-direction) are insulated, the edge $x=0$ is kept at temperature 0°C, and all the other edges are kept at temperature 20°C.

8. Solve

$$\left\{\begin{array}{c} \frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}=0, 0<x<2, 0<y<1, \\ \\ \frac{\partial T}{\partial x}(0,y)=2y, \frac{\partial T}{\partial x}(2,y)=2y, 0<y<1 \\ \\ \frac{\partial T}{\partial y}(x,0)=x, \frac{\partial T}{\partial y}(x,1)=x,\ 0<x<2 \end{array}\right\}.$$

   Check that the solvability condition that we discussed in Section 10.3 is verified in this problem: We have a Neumann problem for the Laplace equation, so $\oint_{\partial D} \frac{\partial u}{\partial n} ds = 0$ should hold.

9. Find the equilibrium temperature distribution $T = T(x, y)$ for the problem

$$\left\{\begin{array}{c} \frac{\partial T}{\partial t}=\left(\frac{\partial^2 T}{\partial x^2}+3\,\frac{\partial^2 T}{\partial y^2}\right), 0<x<a, 0<y<b \\ \\ \frac{\partial T}{\partial x}(0,y)=\frac{\partial T}{\partial x}(a,y)=0, 0<y<b, \\ \\ T(x,0)=0,\ T(x,b)=\frac{a}{2}-\left|x-\frac{a}{2}\right|,\ 0<x<a \end{array}\right\}.$$

   [Hint: Substitute product solutions $T(x,y) = \cos\left(\frac{n\pi x}{a}\right) Y(y)$ into the equilibrium version of the PDE earlier. Be careful about the resulting role played by the "3" found in the PDE earlier.]

10. Find the equilibrium temperature distribution $T = T(x, y)$ for the problem

$$\left\{\begin{array}{c} \frac{\partial T}{\partial t}=\left(2\,\frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}\right), 0<x<a, 0<y<b \\ \\ \frac{\partial T}{\partial x}(0,y)=\frac{\partial T}{\partial x}(a,y)=0, 0<y<b, \\ \\ T(x,0)=x,\ T(x,b)=0, 0<x<a \end{array}\right\}.$$

    [Hint: Substitute product solutions $T(x,y) = \cos\left(\frac{n\pi x}{a}\right) Y(y)$ into the equilibrium version of the PDE earlier. Be careful about the resulting role played by the "2" found in the PDE earlier.]

11. Find the equilibrium temperature distribution $T = T(x, y)$ for the problem

$$\left\{\begin{array}{c} \frac{\partial T}{\partial t}=\left(2\,\frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}\right), 0<x<\pi, 0<y<1 \\ \\ \frac{\partial T}{\partial x}(0,y)=\frac{\partial T}{\partial x}(\pi,y)=0, 0<y<1, \\ \\ \frac{\partial T}{\partial y}(x,0)=0, 0<x<\pi \\ T(x,1)=1-\frac{1}{3}\cos 2x+\frac{1}{5}\cos 5x, 0<x<\pi \end{array}\right\}.$$

12. Solve

$$\left\{\begin{array}{c} \frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}=0, 0<x<a, 0<y<b, \\ T(0,y)=0,\ T(a,y)=0, 0<y<b \\ T(x,0)=0, \frac{\partial T}{\partial y}(x,b)+hT(x,b)=1, 0<x<a \end{array}\right\}.$$

13. For the problem

$$\left\{\begin{array}{c} \frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}=0, 0<x<a, 0<y<b \\ T(0,y)=T(a,y)=0,\ 0<y<b \\ T(x,0)=0,\ T(x,b)=f(x), 0<x<a \end{array}\right\},$$

(a) Find the solution in the form of an infinite series in which there appears integrals involving the unspecified function $f(x)$.

(b) Do the following manipulations on your conclusion for part (a): Change the name of the variable of integration to z in those integrals mentioned in part (a), then move the definite integral symbol to the left of the summation notation, and then move around the order of multiplication of functions in order to rewrite your solution for $T(x,y)$ in the form

$$T(x,y)=\int_0^a G(x,y,z)f(z)\,dz.$$

As an aside, the function $G(x,y,z)$ is known as a **Green's function** for the PDE-BVP in part (a). The integration formula $T(x,y)=\int_0^a G(x,y,z)f(z)dz$ explains how the only nonzero boundary temperature values in $T(x,b)=f(x)$ influence the values of $T(x,y)$ inside the rectangle $0<x<a,\ 0<y<b$.

11.4 Eigenvalues of the Laplacian and Applications

An eigenvalue problem for the Laplacian has the form

$\Delta\phi+\lambda\phi=0$, in a domain $\mathcal{D}$, with homogeneous conditions on its boundary, $\partial\mathcal{D}$.

As an example, the PDE-BVP eigenvalue problem

$$\left\{\begin{array}{c} \dfrac{\partial^2\phi}{\partial x^2}+\dfrac{\partial^2\phi}{\partial y^2}+\lambda\phi(x,y)=0,\ 0<x<a,\ 0<y<b, \\ \\ \phi(x,0)=\phi(x,b)=0,\ 0<x<a \\ \\ \phi(0,y)=\phi(a,y)=0,\ 0<y<b \end{array}\right\} \tag{11.38}$$

has a nontrivial solution for ϕ exactly when λ is one of the infinite sets of eigenvalues

$$\left\{\lambda:\ \lambda=\left(\frac{m\pi}{a}\right)^2+\left(\frac{n\pi}{b}\right)^2,\ m=1,2\ldots,\ n=1,2,\ldots\right\}.$$

It is beyond the level of this course to explain why these are the *only* eigenvalues, but we can learn how people found those eigenvalues. If this were the PDE in Section 11.3 instead of the PDE we are considering now, we would see the two homogeneous boundary conditions for ϕ on the parallel sides $x=0$ and $x=a$ and decide to look for a solution of (11.38) in the form $\phi(x,y)=\sin\left(\frac{m\pi x}{a}\right)Y(y)$. Substituting that into the PDE $\lambda\phi(x,y)=-\Delta\phi(x,y)$ gives

$$\begin{aligned}\lambda\,\sin\left(\frac{m\pi x}{a}\right)Y(y) &= -\frac{\partial^2}{\partial x^2}\Big[\,\sin\left(\frac{m\pi x}{a}\right)Y(y)\Big]-\frac{\partial^2}{\partial y^2}\Big[\,\sin\left(\frac{m\pi x}{a}\right)Y(y)\Big] \\ &= \left(\frac{m\pi}{a}\right)^2\sin\left(\frac{m\pi x}{a}\right)Y(y)-\sin\left(\frac{m\pi x}{a}\right)\frac{d^2Y}{dy^2},\ m=1,2\ldots.\end{aligned}$$

After dividing out the factor $\sin\left(\frac{m\pi x}{a}\right)$, this gives the ODE-BVP, for $m=1,2\ldots$,

$$\left\{\begin{array}{c} Y''+\left(\lambda-\left(\frac{m\pi}{a}\right)^2\right)Y(y)=0, \\ \\ Y(0)=Y(b)=0 \end{array}\right\},$$

which is similar to ODE-BVP (11.4) in Section 11.1 except with $\left(\lambda-\left(\frac{m\pi}{a}\right)^2\right)$ replacing λ, $Y(y)$ replacing $X(x)$, and b replacing L. Using the results of Table 11.1, there is a nontrivial solution for $Y(y)$, namely, $Y_n(y)=\sin\left(\frac{n\pi y}{b}\right)$, exactly when

$$\lambda-\left(\frac{m\pi}{a}\right)^2=\left(\frac{n\pi}{b}\right)^2,\ n=1,2\ldots.$$

Solving for λ gives

$$\lambda=\lambda_{m,n}\triangleq\left(\frac{m\pi}{a}\right)^2+\left(\frac{n\pi}{b}\right)^2,\ m=1,2\ldots,\ n=1,2\ldots,$$

as we promised.

For each $m = 1, 2 \ldots,\ n = 1, 2, \ldots$, corresponding to each eigenvalue

$$\lambda_{m,n} = \left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2,$$

we have an eigenfunction

$$\phi_{m,n}(x, y) = \sin\left(\frac{m\pi x}{a}\right) \sin\left(\frac{n\pi y}{b}\right).$$

In retrospect, we see that the boundary conditions $\phi(x, 0) = \phi(x, b) = 0,\ 0 < x < a$, could have told us to try the y-dependence of $\phi(x, y)$ as $\sin\left(\frac{n\pi y}{b}\right)$, and the boundary conditions $\phi(0, y) = \phi(a, y) = 0,\ 0 < y < b$, could have told us to try the x-dependence of $\phi(x, y)$ as $\sin\left(\frac{m\pi y}{a}\right)$. So, from the four homogeneous boundary conditions, that is, two pairs of two homogeneous boundary conditions, we could have jumped to the assumption to substitute $\phi(x, y) = \sin\left(\frac{m\pi x}{a}\right) \sin\left(\frac{n\pi y}{b}\right)$ into the "eigenvalues of the Laplacian" PDE $\frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} + \lambda\phi(x, y) = 0$. If we had done that, we would have also arrived at the formula for the eigenvalues: $\lambda = \lambda_{m,n} \triangleq \left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2,\ m = 1, 2 \ldots,\ n = 1, 2 \ldots$.

Of course, we can find eigenvalues and eigenfunctions for the same PDE with different pairs of homogeneous boundary conditions. For example, the PDE-BVP

$$\left\{\begin{array}{l} \dfrac{\partial^2\phi}{\partial x^2} + \dfrac{\partial^2\phi}{\partial y^2} + \lambda\phi(x, y) = 0,\ 0 < x < a,\ 0 < y < b, \\ \\ \dfrac{\partial\phi}{\partial y}(x, 0) = \phi(x, b) = 0,\ 0 < x < a, \\ \\ \dfrac{\partial\phi}{\partial x}(0, y) = \dfrac{\partial\phi}{\partial x}(a, y) = 0,\ 0 < y < b \end{array}\right\}$$

has eigenvalues

$$\lambda_{m,n} = \left(\frac{m\pi}{a}\right)^2 + \left(\frac{\left(n - \frac{1}{2}\right)\pi}{b}\right)^2,\ m = 0, 1, 2 \ldots,\ n = 1, 2, \ldots,$$

and corresponding eigenfunctions

$$\phi_{m,n}(x, y) = \cos\left(\frac{m\pi x}{a}\right) \cos\left(\frac{\left(n - \frac{1}{2}\right)\pi y}{b}\right),\ m = 0, 1, 2 \ldots,\ n = 1, 2.$$

To get these results, use Table 11.1.

We can use these eigenvalues and eigenfunctions of the Laplacian in many ways.

11.4.1 Application to Time-Dependent Heat Flow in a Slab

Example 11.13

Solve the PDE-BVP-IVP

$$\left\{\begin{array}{c} \dfrac{\partial T}{\partial t} = \alpha\left(\dfrac{\partial^2 T}{\partial x^2} + \dfrac{\partial^2 T}{\partial y^2}\right),\ 0 < x < a,\ 0 < y < b,\ t > 0, \\ T(0,y,t) = T(a,y,t) = 0,\ 0 < y < b,\ t > 0, \\ T(x,0,t) = T(x,b,t) = 0,\ 0 < x < a,\ t > 0, \\ T(x,y,0) = f(x,y),\ 0 < x < a,\ 0 < y < b \end{array}\right\}. \tag{11.39}$$

Method: We look at the four homogeneous boundary conditions on the four sides of the rectangle and think that we should try to find product solutions of the PDE in the form

$$T(x,y,t) = G(t)\phi_{m,n}(x,y) = G(t)\sin\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right),\ m = 1,2\ldots,\ n = 1,2,\ldots.$$

Noting that $\Delta\phi_{m,n}(x,y) = -\lambda_{m,n}\phi_{m,n}(x,y)$ and substituting $T(x,y,t)$ into the PDE gives

$$\frac{dG}{dt}\phi_{m,n}(x,y) = \frac{\partial T}{\partial t} = \alpha\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}\right) = \alpha\, G(t)\ \left(-\lambda_{m,n}\phi_{m,n}(x,y)\right),$$

so we want $G(t)$ to satisfy

$$\frac{dG}{dt} = -\alpha\,\lambda_{m,n}G(t).$$

So, similarly to work in Section 11.1, the general solution of the PDE-BVP is

$$T(x,y,t) = \sum_{\substack{n=1\\m=1}}^{\infty} T_{m,n}(x,y,t) = \sum_{\substack{n=1\\m=1}}^{\infty} b_{m,n}\sin\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right)e^{-\alpha\left(\left(\frac{m\pi}{a}\right)^2+\left(\frac{n\pi}{b}\right)^2\right)t}.$$

Applying the initial condition gives

$$f(x,y) = T(x,y,0) = \sum_{\substack{n=1\\m=1}}^{\infty} b_{m,n}\sin\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right),\ 0 < x < a,\ 0 < y < b.$$

This is an example of a **double Fourier series**. Very much like the orthogonality used to find the coefficients in a Fourier sine series, we get the orthogonality relation

$$\int_0^b\int_0^a \sin\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right)\sin\left(\frac{m'\pi x}{a}\right)\sin\left(\frac{n'\pi y}{b}\right)dxdy = \left\{\begin{array}{ll} 0, & \text{if } m \neq m' \text{ or } n \neq n' \\ \frac{ab}{4}, & \text{if } m = m' \text{ and } n = n' \end{array}\right\}.$$

It follows, after a little more thought, that

$$b_{m,n} = \frac{4}{ab}\int_0^b\int_0^a f(x,y)\sin\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right)dxdy. \tag{11.40}$$

The solution of the PDE-BVP-IVP is

$$T(x,y,t) = \sum_{\substack{n=1\\m=1}}^{\infty} \left(\frac{4}{ab} \int_0^b \int_0^a f(x,y) \sin\left(\frac{m\pi x}{a}\right) \sin\left(\frac{n\pi y}{b}\right) dxdy \right)$$

$$\times \sin\left(\frac{m\pi x}{a}\right) \sin\left(\frac{n\pi y}{b}\right) e^{-\alpha\left(\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2\right)t}. \bigcirc$$

11.4.2 Special Case

As an aside, if $f(x,y) = F(x)G(y)$, that is, the initial temperature distribution function is itself a product of a function of x times a function of y, then the double integrals (11.40) are a lot easier to calculate. In Theorem 7.12 in Section 7.3, we noted that the double integral on a rectangle of a product function can be calculated as a product of integrals, each with respect to a single variable. In this situation, that is, *if* $f(x,y) = F(x)G(y)$, then

$$f(x,y) \sin\left(\frac{m\pi x}{a}\right) \sin\left(\frac{n\pi y}{b}\right) = \left(F(x) \sin\left(\frac{m\pi x}{a}\right)\right) \left(G(y) \sin\left(\frac{n\pi y}{b}\right)\right);$$

hence,

$$b_{m,n} = \frac{4}{ab} \left(\int_0^a F(x) \sin\left(\frac{m\pi x}{a}\right) dx \right) \left(\int_0^b G(y) \sin\left(\frac{n\pi y}{b}\right) dy \right).$$

11.4.3 Application to Transverse Vibrations of a Rectangular Membrane

Example 11.14

Solve the PDE-BVP-IVP

$$\left\{ \begin{array}{c} \frac{\partial^2 u}{\partial t^2} = c^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right),\ 0 < x < a,\ 0 < y < b,\ t > 0, \\ u(0,y,t) = u(a,y,t) = 0,\ 0 < y < b,\ t > 0, \\ u(x,0,t) = u(x,b,t) = 0,\ 0 < x < a,\ t > 0, \\ u(x,y,0) = f(x,y),\ 0 < x < a,\ 0 < y < b \\ \frac{\partial u}{\partial t}(x,y,0) = g(x,y),\ 0 < x < a,\ 0 < y < b \end{array} \right\}. \tag{11.41}$$

Method: This is like Example 11.13 combined with Example 11.7 in Section 11.2, that is, the 1D wave equation, that is, (11.21) through (11.24). Use eigenfunctions $\phi_{mn,n}(x,y) \triangleq \sin\left(\frac{m\pi x}{a}\right) \sin\left(\frac{n\pi y}{b}\right)$ and substitute $u(x,y,t) = G_{m,n}(t)\phi_{mn,n}(x,y)$ into the PDE to get

$$\ddot{G}_{m,n}(t)\phi_{m,n}(x,y) = c^2 G_{m,n}(t) \Delta\phi_{mn,n}(x,y) = -c^2 \lambda_{m,n} G_{m,n}(t) \phi_{m,n}(x,y).$$

So, the PDE-BVP has the general solution

$$\begin{aligned} u(x,y,t) &= \sum_{\substack{n=1\\m=1}} u_{m,n}(x,y,t) \\ &= \sum_{\substack{n=1\\m=1}}^{\infty} \left(a_{m,n} \cos\left(\sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2}\, ct \right) + b_{m,n} \sin\left(\sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2}\, ct \right) \right) \\ &\quad \times \sin\left(\frac{m\pi x}{a}\right) \sin\left(\frac{n\pi y}{b}\right). \end{aligned}$$

The two initial conditions give

$$a_{m,n} = \frac{4}{ab} \int_0^b \int_0^a f(x,y) \sin\left(\frac{m\pi x}{a}\right) \sin\left(\frac{n\pi y}{b}\right) dxdy$$

and

$$b_{m,n} = \frac{4}{cab\sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2}} \int_0^b \int_0^a g(x,y) \sin\left(\frac{m\pi x}{a}\right) \sin\left(\frac{n\pi y}{b}\right) dxdy. \bigcirc$$

Of course, the method can be applied to the vibrations of a rectangular plate with other combinations of homogeneous boundary conditions on the four sides.

11.4.4 Application to Steady-State Temperature in a Slab with a Steady Source or Sink

Example 11.15

Solve the PDE-BVP

$$\left\{ \begin{array}{c} 0 = \kappa \left(\dfrac{\partial^2 T}{\partial x^2} + \dfrac{\partial^2 T}{\partial y^2} \right) + Q(x,y),\ 0 < x < a,\ 0 < y < b \\ \\ T(0,y) = T(a,y) = 0,\ 0 < y < b, \\ \\ T(x,0) = T(x,b) = 0,\ 0 < x < a \end{array} \right\}. \tag{11.42}$$

Method: If we assume $T(x,y)$ has the form of the double Fourier series

$$T(x,y) = \sum_{\substack{n=1\\m=1}}^{\infty} b_{m,n} \sin\left(\frac{m\pi x}{a}\right) \sin\left(\frac{n\pi y}{b}\right),$$

then it is guaranteed to satisfy the four homogeneous boundary conditions. Substituting $T(x,y)$ into the PDE gives

$$0 = Q(x,y) + \kappa \sum_{\substack{n=1\\m=1}}^{\infty} b_{m,n} \left(-\lambda_{m,n}\right) \sin\left(\frac{m\pi x}{a}\right) \sin\left(\frac{n\pi y}{b}\right),$$

that is,

$$-\frac{Q(x,y)}{\kappa} = \sum_{\substack{n=1\\m=1}}^{\infty} b_{m,n}\left(-\lambda_{m,n}\right)\sin\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right),$$

which is just a double Fourier series problem. So, we find that

$$b_{m,n} = \frac{4}{\kappa ab\left(\left(\frac{m\pi}{a}\right)^2+\left(\frac{n\pi}{b}\right)^2\right)}\int_0^b\int_0^a Q(x,y)\sin\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right)dxdy.$$

The solution for the equilibrium temperature is

$$T(x,y) = \sum_{\substack{n=1\\m=1}}^{\infty}\left(\frac{4}{\kappa ab\left(\left(\frac{m\pi}{a}\right)^2+\left(\frac{n\pi}{b}\right)^2\right)}\int_0^b\int_0^a Q(x,y)\sin\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right)dxdy\right)$$
$$\times\sin\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right). \bigcirc$$

By the way, a PDE of the form

$$\frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2} = p(x,y)$$

is called a 2D *Poisson equation*. It came up in the study of problems with source terms, for example, the previously mentioned problems of finding an equilibrium temperature distribution in Section 10.2, but also in electro- or magnetostatics.

11.4.5 Application to Surface Waves

Consider fluid in a 3D rectangular basin whose base is $0 \le x \le a,\ 0 \le y \le b,\ z = -d$. The surface of the fluid is given by $z = \eta(x,y,t)$. If we assume irrotational potential flow, then the potential function $\Phi = \Phi(x,y,z,t)$ satisfies the PDE $\frac{\partial^2\Phi}{\partial x^2}+\frac{\partial^2\Phi}{\partial y^2}+\frac{\partial^2\Phi}{\partial z^2} = 0$ in the spatial region $0 < x < a,\ 0 < y < b, -d < z < \eta(x,y,t)$, along with the boundary conditions that the derivative of Φ in the direction normal to the spatial region is zero on the bottom and on the sides of the basin, that is, on the planes $x = 0$, $x = a$, $y = 0$, $y = b$, and $z = -d$. These boundary conditions are, for $t > 0$,

$$\left\{\begin{array}{l}\frac{\partial\Phi}{\partial x}(0,y,z,t) = \frac{\partial\Phi}{\partial x}(a,y,z,t) = 0,\ 0 < y < b, -d < z < \eta(x,y,t),\\ \\ \frac{\partial\Phi}{\partial y}(x,0,z,t) = \frac{\partial\Phi}{\partial y}(x,b,z,t) = 0,\ 0 < x < a, -d < z < \eta(x,y,t), \text{and}\\ \\ \frac{\partial\Phi}{\partial z}(x,y,-d,t) = 0,\ 0 < x < a,\ 0 < y < b.\end{array}\right\}.$$

In addition, the potential is specified on the free surface $z = \eta(x,y,t)$, that is,

$$\Phi(x,y,\eta(x,y,t)) = p(x,y,t),\ 0 < x < a,\ 0 < y < b.$$

Because of the four homogeneous boundary conditions on the sides of the basin and the homogeneous boundary condition on the bottom of the basin, the solution for Φ should be in the "clairvoyant" form:

$$\Phi(x,y,z,t) = \sum_{\substack{n=0\\ m=0\\ m+n>0}}^{\infty} b_{m,n}(t)\cos\left(\frac{m\pi x}{a}\right)\cos\left(\frac{n\pi y}{b}\right)\cosh(\sqrt{\lambda_{m,n}}\,(z+d)), \tag{11.43}$$

where the eigenvalues of the Laplacian with the four homogeneous boundary conditions are given by

$$\lambda_{mn} = \left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2, m = 0,1,2,\ldots,\ n = 0,1,2,\ldots,\ \text{with } m+n>0.$$

We arbitrarily assume that the average value of the potential Φ is zero, that is, that $b_{0,0} = 0$; the value of $b_{0,0}$ does not affect the fluid velocity given by $u\,\hat{\mathbf{i}} + v\,\hat{\mathbf{j}} + w\,\hat{\mathbf{k}} = \nabla\Phi$. So, we only need the eigenvalues for $m+n>0$.

Now, we will not pretend to solve the free surface boundary value problem, which involves other, nonlinear PDEs called the "kinematic" condition and the "Bernoulli equation." This is because both the potential, $u(x,y,t)$, on the free surface and also the shape of the free surface itself, $z = \eta(x,y,t)$, are unknowns.

Nevertheless, we can make some contribution to that work by solving the earlier problem for $\Phi(x,y,z,t)$ in terms of $u(x,y,t)$ for the "nominal" case when the free surface is $z=0$, that is, when $\eta(x,y,t) \equiv 0$. While this solution, denoted $\Phi_0(x,y,z,t)$, cannot give an actual solution for the general case where η is not identically zero, the so-called "perturbation" methods would use our solution, $\Phi_0(x,y,z,t)$, as one ingredient in the "secret sauce" of a method for the real situation. So, let's find the solution for Φ_0 in the form of (11.42) that satisfies the one last boundary condition:

$$u(x,y,t) = \Phi_0(x,y,0,t) = \sum_{\substack{n=0\\ m=0\\ m+n>0}}^{\infty} b_{m,n}(t)\cos\left(\frac{m\pi x}{a}\right)\cos\left(\frac{n\pi y}{b}\right)\cosh(\sqrt{\lambda_{m,n}}\,d),$$

with $b_{0,0} = 0$ assumed, is a double Fourier series. Using the orthogonality relation

$$\int_0^b\int_0^a \cos\left(\frac{m\pi x}{a}\right)\cos\left(\frac{n\pi y}{b}\right)\cos\left(\frac{m'\pi x}{a}\right)\cos\left(\frac{n'\pi y}{b}\right)dxdy$$

$$= \begin{cases} 0, & \text{if } m \neq m' \text{ or } n \neq n' \\ ab, & \text{if } m = m = 0' \text{ and } n = n = 0' \\ \frac{ab}{2}, & \text{if } m = m' \text{ and } n = n', \text{ and, } m = 0 \text{ or } n = 0 \\ \frac{ab}{4}, & \text{if } m = m' \geq 1 \text{ and } n = n \geq 1' \end{cases},$$

we see that for $m = 0,\ n \geq 1$ and for $m \geq 1,\ n = 0$,

$$\cosh(\sqrt{\lambda_{m,n}}\, d) b_{m,n}(t) = \frac{2}{ab} \int_0^b \int_0^a \cos\left(\frac{m\pi x}{a}\right) \cos\left(\frac{n\pi y}{b}\right) u(x,y,t) dxdy,$$

and for $m \geq 1$ and $n \geq 1$,

$$\cosh(\sqrt{\lambda_{m,n}}\, d) b_{m,n}(t) = \frac{4}{ab} \int_0^b \int_0^a \cos\left(\frac{m\pi x}{a}\right) \cos\left(\frac{n\pi y}{b}\right) u(x,y,t)\, dxdy.$$

The solution is, for $0 < x < a,\ 0 < y < b,\ -d < z < 0, t > 0$,

$$\begin{aligned}
\Phi_0(x,y,z,t) = &\sum_{n=1}^{\infty} \left(\frac{2}{ab} \int_0^b \int_0^a \cos\left(\frac{n\pi y}{b}\right) u(x,y,t) dxdy \right) \cos\left(\frac{n\pi y}{b}\right) \frac{\cosh\left(\frac{n\pi}{b}(z+d)\right)}{\cosh\left(\frac{n\pi d}{b}\right)} \\
&+ \sum_{m=1}^{\infty} \left(\frac{2}{ab} \int_0^b \int_0^a \cos\left(\frac{m\pi x}{a}\right) u(x,y,t) dxdy \right) \cos\left(\frac{m\pi x}{a}\right) \frac{\cosh\left(\frac{m\pi}{a}(z+d)\right)}{\cosh\left(\frac{m\pi d}{a}\right)} \\
&+ \sum_{\substack{n=1 \\ m=1}}^{\infty} \left(\frac{4}{ab} \int_0^b \int_0^a \cos\left(\frac{m\pi x}{a}\right) \cos\left(\frac{n\pi y}{b}\right) u(x,y,t) dxdy \right) \\
&\times \cos\left(\frac{m\pi x}{a}\right) \cos\left(\frac{n\pi y}{b}\right) \frac{\cosh(\sqrt{\lambda_{m,n}}\,(z+d))}{\cosh(\sqrt{\lambda_{m,n}}\, d)}. \bigcirc
\end{aligned}$$

11.4.6 Problems

In problems 1 and 2, find the eigenvalues and eigenfunctions of the Laplacian with the given boundary conditions:

1. $$\left\{ \begin{array}{c} \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = \lambda \phi(x,y),\ 0 < x < a,\ 0 < y < b, \\ \phi(x,0) = \phi(x,b) = 0,\ 0 < x < a \\ \frac{\partial \phi}{\partial x}(0,y) = \frac{\partial \phi}{\partial x}(a,y) = 0,\ 0 < y < b \end{array} \right\}.$$

2. $$\left\{ \begin{array}{c} \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = \lambda \phi(x,y),\ 0 < x < a,\ 0 < y < b, \\ \frac{\partial \phi}{\partial y}(x,0) = \frac{\partial \phi}{\partial y}(x,b) = 0,\ 0 < x < a \\ \frac{\partial \phi}{\partial x}(0,y) = \frac{\partial \phi}{\partial x}(a,y) = 0,\ 0 < y < b \end{array} \right\}.$$

3. Solve the PDE-BVP-IVP

$$\left\{\begin{array}{c} \frac{\partial T}{\partial t} = \alpha\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}\right),\ 0 < x < \pi,\ 0 < y < 2,\ t > 0, \\ \\ T(0,y,t) = \frac{\partial T}{\partial x}(\pi,y,t) = 0,\ 0 < y < 2,\ t > 0, \\ \\ T(x,0,t) = T(x,2,t) = 0,\ 0 < x < \pi,\ t > 0, \\ \\ T(x,y,0) = \sin\left(\frac{x}{2}\right)(1 - \cos \pi y),\ 0 < x < \pi,\ 0 < y < 2 \end{array}\right\}.$$

4. Solve the PDE-BVP-IVP

$$\left\{\begin{array}{c} \frac{\partial^2 u}{\partial t^2} = c^2\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right),\ 0 < x < 2,\ 0 < y < 1,\ t > 0, \\ \\ u(0,y,t) = u(2,y,t) = 0,\ 0 < y < 1,\ t > 0, \\ u(x,0,t) = u(x,1,t) = 0,\ 0 < x < 2,\ t > 0, \\ \\ u(x,y,0) = 0,\ 0 < x < 2,\ 0 < y < 1 \\ \\ \frac{\partial u}{\partial t}(x,y,0) = (1 - |x-1|)y,\ 0 < x < 2,\ 0 < y < 1 \end{array}\right\}.$$

5. Solve

$$\left\{\begin{array}{c} 0 = \kappa\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}\right) - 1,\ 0 < x < a,\ 0 < y < b \\ \\ T(0,y) = T(a,y) = 0,\ 0 < y < b, \\ T(x,0) = T(x,b) = 0,\ 0 < x < a \end{array}\right\}.$$

6. Solve

$$\left\{\begin{array}{c} 0 = \kappa\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}\right) - 1,\ 0 < x < a,\ 0 < y < b \\ \\ T(0,y) = y,\ T(a,y) = 0,\ 0 < y < b, \\ T(x,0) = T(x,b) = 0,\ 0 < x < a \end{array}\right\}.$$

[Hint: Let $T(x,y) = T_1(x,y) + T_2(x,y)$, where T_1 solves Problem 11.4.6.5 and T_2 solves another problem.]

7. Solve an example of Laplace equation in three dimensions:

$$\left\{\begin{array}{c} 0 = \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2},\ 0 < x < a,\ 0 < y < b,\ 0 < z < c \\ \\ T(0,y,z) = 0, T(a,y,z) = f(y,z),\ 0 < y < b,\ 0 < z < c \\ T(x,0,z) = T(x,b,z) = 0,\ 0 < x < a,\ 0 < z < c \\ \\ \frac{\partial T}{\partial z}(x,y,0) = \frac{\partial T}{\partial z}(x,y,c) = 0,\ 0 < x < a,\ 0 < y < b \end{array}\right\}.$$

8. Suppose $T = T(x, y, t)$ satisfies the heat equation

$$\frac{\partial T}{\partial t} = \alpha \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} \right), \ 0 < x < \pi, \ 0 < y < \pi, \ t > 0,$$

the four sides of the square $\{(x, y) : 0 < x < \pi, \ 0 < y < \pi\}$ are kept at temperature zero, and T satisfies the initial condition $T(x, y, 0) = f(x, y)$ in the square:
(a) Find the solution $T(x, y, t)$.
(b) Suppose f is symmetric in x and y, that is, $f(x, y) \equiv f(y, x)$. Explain why T is also symmetric in x and y.

9. Solve $\left\{ \begin{array}{c} -1 + 6\cos 4y = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \ 0 < x < \pi, \\ 0 < y < \frac{\pi}{2} \\ \\ u(0, y) = u(\pi, y) = 0, \ 0 < y < \frac{\pi}{2} \\ u(x, 0) = u(x, \frac{\pi}{2}) = 2, \ 0 < x < \pi \end{array} \right\}$.

Note that while we could use eigenvalues of the Laplacian, an easier way seems to be to write the solution in the form $u(x, y) = v(y) + w(x, y)$, where first we find $v(y)$, the solution of the problem $-1 + 6\cos 4y = v''(y)$, $0 < y < \frac{\pi}{2}$, and then let $w(x, y) = u(x, y) - v(y)$.

11.5 PDEs in Polar Coordinates

Now we will explain how the method of separation of variables can be used in polar coordinates. First, we recall from Section 6.7 the Laplacian differential operator in polar coordinates:

$$\Delta u \triangleq \frac{1}{r} \frac{\partial}{\partial r} \left[r \frac{\partial u}{\partial r} \right] + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} . \tag{11.44}$$

We notice that the second term mixes the r and θ variables. So, when using the method of separation of variables, it may be useful to note that

$$r^2 \Delta u = r \frac{\partial}{\partial r} \left[r \frac{\partial u}{\partial r} \right] + \frac{\partial^2 u}{\partial \theta^2}, \tag{11.45}$$

because that would help isolate the θ-dependence of u.

11.5.1 Laplace Equation in Polar Coordinates

Example 11.16

Solve $\Delta u = 0$ for the equilibrium temperature, $u(r, \theta)$, in the half disk $\mathcal{D}$ shown in Figure 11.12. Assume that the temperature satisfies the boundary conditions as shown in the picture of $\mathcal{D}$.

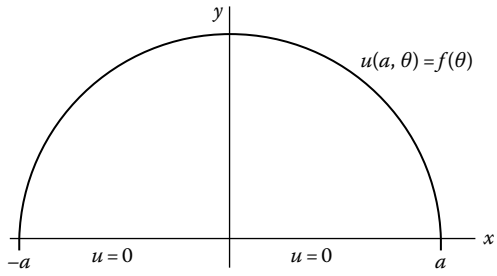


FIGURE 11.12
Example 11.16: equilibrium heat problem in a half disk.

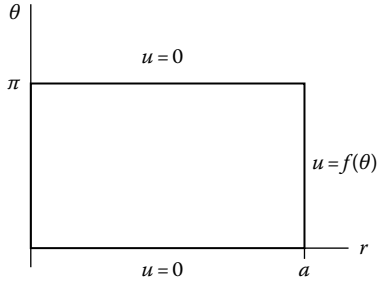


FIGURE 11.13
Example 11.16: equilibrium heat problem in polar coordinates.

Method: Solving in $\mathcal{D}$, the PDE

$$0 = r^2 \Delta u, \text{ that is, } 0 = r\,\frac{\partial}{\partial r}\left[r\,\frac{\partial u}{\partial r}\right] + \frac{\partial^2 u}{\partial \theta^2} \tag{11.46}$$

seems to have no relationship to the method of separation of variables as we used it so far in this chapter. But if we rewrite $\mathcal{D}$ in polar coordinates, as we did for double integrals in Section 11.5, the picture becomes clear! Our domain

$$\mathcal{D} = \{(r, \theta) : 0 \le r < a,\ 0 < \theta < \pi\}$$

is actually a rectangle in the $r\theta$-plane, and u satisfies boundary conditions on sides of the rectangle, as shown in Figure 11.13.

So, just as in Section 11.2, we will use the two homogeneous boundary conditions on the parallel sides $\theta = 0$ and $\theta = \pi$ to give the form of the product solutions

$$u(r, \theta) = R(r)\Theta(\theta),$$

where $\Theta(\theta)$ satisfies the ODE-BVP eigenvalue problem

$$\left\{\begin{array}{l} \Theta''(\theta) + \lambda\Theta(\theta) = 0, \\ \\ \Theta(0) = \Theta(\pi) = 0 \end{array}\right\}.$$

The first entry of Table 11.1, with $L = \pi$, gives the eigenfunctions and eigenvalues

$$\Theta_n(\theta) = \sin n\theta,\ \lambda_n = n^2, \text{for } n = 1, 2, \ldots.$$

So we substitute

$$u(r,\theta) = R_n(r)\sin n\theta$$

into the PDE (11.46) to get

$$\begin{aligned} 0 &= r\frac{\partial}{\partial r}\left[r\frac{\partial}{\partial r}R_n(r)\right]\sin n\theta + \frac{\partial^2}{\partial\theta^2}\left[R_n(r)\sin n\theta\right] \\ &= \sin n\theta\left(r\frac{d}{dr}\left[r\frac{dR_n}{dr}\right]\right) + R_n(r)\frac{d^2}{d\theta^2}\sin n\theta \\ &= \sin n\theta\left(r\frac{d}{dr}\left[r\frac{dR_n}{dr}\right] - n^2R_n(r)\right). \end{aligned}$$

So, to have a nontrivial product solution, $R_n(r)$ should satisfy the ODE

$$0 = r\frac{d}{dr}\left[r\frac{dR_n}{dr}\right] - n^2R_n(r),$$

that is,

$$0 = r^2\frac{d^2R_n}{dr^2} + r\frac{dR_n}{dr} - n^2R_n, \tag{11.47}$$

for $n = 1, 2, \ldots$, which we recognize as Cauchy–Euler ODEs we studied in Section 3.5. We substitute $R(r) = r^m$ into (11.47) and get the characteristic equation

$$0 = m(m-1) + m - n^2 = m^2 - n^2,$$

whose roots are $m = \pm n$ and give corresponding solution of (11.47): $R_n(r) = c_1r^{-n} + c_2r^n$. The product solutions are of the form

$$u_n(r,\theta) = (c_1r^{-n} + c_2r^n)\sin n\theta.$$

Using linearity, we get the general solution of the PDE plus the two homogeneous BCs: The general solution of the problem

$$\left\{\begin{array}{c} \Delta u(r,\theta) = 0, \text{ in } \mathcal{D} = \{(r,\theta) : 0 \le r < a,\ 0 < \theta < \pi\} \\ \\ u(r,0) = u(r,\pi) = 0, 0 < r < a \end{array}\right\}$$

is

$$u(r,\theta) = \sum_{n=1}^{\infty}(a_nr^{-n} + b_nr^n)\sin n\theta. \tag{11.48}$$

But now, (11.48) stares us in the face with the danger of what happens at $r = 0$, or, more properly, what happens as $r \to 0^+$. Since clearly we don't want our solution $u(r,\theta)$ to blow up as $r \to 0^+$, we should require $a_n = 0$, for $n = 1, 2, \ldots$. That leaves us with the general solution

$$u(r,\theta) = \sum_{n=1}^{\infty} b_nr^n\sin n\theta. \tag{11.49}$$

We substitute (11.49) into the remaining BC, $u(a,\theta) = f(\theta)$, for $0 < \theta < \pi$:

$$f(\theta) = u(a,\theta) = \sum_{n=1}^{\infty} b_n a^n \sin n\theta = \sum_{n=1}^{\infty} B_n \sin n\theta \tag{11.50}$$

where $B_n \triangleq b_n a^n$. But, (11.50) is just a standard Fourier sine series expansion problem. As we learned in Section 9.2,

$$b_n a^n = B_n = \frac{2}{\pi} \int_0^{\pi} f(\theta) \sin n\theta \, d\theta;$$

hence,

$$b_n = a^{-n} \left(\frac{2}{\pi} \int_0^{\pi} f(\theta) \sin n\theta \, d\theta \right).$$

So the solution to our physical problem is the temperature distribution

$$u(r,\theta) = \frac{2}{\pi} \sum_{n=1}^{\infty} a^{-n} \left(\int_0^{\pi} f(\theta) \sin n\theta \, d\theta \right) r^n \sin n\theta,$$

that is,

$$u(r,\theta) = \frac{2}{\pi} \sum_{n=1}^{\infty} \left(\int_0^{\pi} f(\theta) \sin n\theta \, d\theta \right) \left(\frac{r}{a} \right)^n \sin n\theta. \bigcirc \tag{11.51}$$

Along the way, we had to deal with the mathematical singularity as $r \to 0^+$ that was *not* in the original problem. The mathematical singularity was an artifact of using polar coordinates, but we have both mathematical and physical grounds for paying attention to it. Except for this annoying issue, we have seen once again, just as when doing double integrals in Chapter 7, how useful polar coordinates can be if appropriate for the geometry of the domain.

The need to address the issue of the mathematical singularity as $r \to 0^+$ can be restated physically: We expect that the temperature, $u(r,\theta)$, as well as the flux vector

$$\mathbf{q} = -\kappa \left(\frac{\partial u}{\partial r} \hat{\mathbf{e}}_r + \frac{1}{r} \frac{\partial u}{\partial \theta} \hat{\mathbf{e}}_\theta \right),$$

should not blow up as $r \to 0^+$, that is,

$$| u(0^+,\theta) | < \infty, \quad \left| \frac{\partial u}{\partial r}(0^+,\theta) \right| < \infty, \text{ and } \left| \lim_{r \to 0^+} \frac{1}{r} \frac{\partial u}{\partial \theta}(r,\theta) \right| < \infty.$$

Example 11.17

Solve $\Delta u = 0$ for the equilibrium temperature, $u(r,\theta)$, in the disk $\mathcal{D}_a$: $0 \le r < a$, depicted in Figure 11.14.

Method: We depict this problem in the (r,θ) plane in Figure 11.14. Mathematically, this consists of

$$(PDE) \quad 0 = r \frac{\partial}{\partial r} \left[r \frac{\partial u}{\partial r} \right] + \frac{\partial^2 u}{\partial \theta^2}, 0 \le r < a, -\pi < \theta \le \pi \tag{11.52}$$

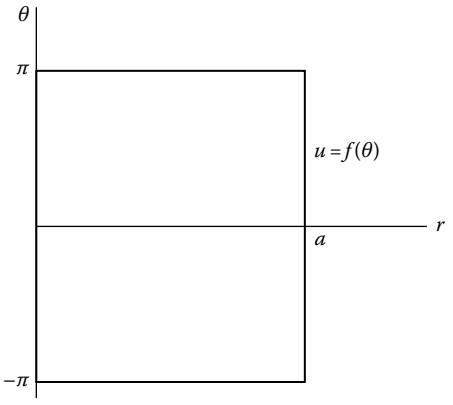


FIGURE 11.14
Equilibrium heat problem in a half disk with boundary conditions.

and

$$(BC)\ \ u(a,\theta) = f(\theta),\ \ -\pi < \theta < \pi. \tag{11.53}$$

In addition, the change of variables to (r,θ) introduces *two* mathematical issues: The first, as in Example 11.16, is the boundedness of the temperature and the flux vector as $r \to 0^+$. The second is that, physically, the ray $\theta = -\pi$ is the same location as the ray $\theta = \pi$! It makes sense that the temperature and the flux should be continuous across the ray $\theta = \pm\pi$, that is,

$$\lim_{\theta \to -\pi^+} u(r,\theta) = \lim_{\theta \to \pi^-} u(r,\theta),$$

and similarly for the flux vector $-\kappa\left(\frac{\partial u}{\partial r}\hat{\mathbf{e}}_r + \frac{1}{r}\frac{\partial u}{\partial \theta}\hat{\mathbf{e}}_\theta\right)$. This can be expressed mathematically as the additional "continuity" boundary conditions:

$$u(r,-\pi^+) = u(r,\pi^-) \text{ and } \frac{\partial u}{\partial \theta}(r,-\pi^+) = \frac{\partial u}{\partial \theta}(r,\pi^-).^* \tag{11.54}$$

From now on, we will write $u(r,-\pi)$ instead of $u(r,-\pi^+)$, etc., for convenience.

Let's solve (11.52) through (11.54) using separation of variables. First, the disk $\mathcal{D}_a$ is a rectangle in (r,θ) coordinates. If we substitute a product solution $u(r,\theta) = R(r)\Theta(\theta)$ into the continuity BCs (11.54), we get

$$R(r)\Theta(-\pi) = R(r)\Theta(\pi) \text{ and } R(r)\Theta'(-\pi) = R(r)\Theta'(\pi), 0 < r < a,$$

that is,

$$R(r)(\Theta(\pi) - \Theta(-\pi)) = R(r)(\Theta'(\pi) - \Theta'(-\pi)) = 0, 0 < r < a.$$

Since having $R(r) \equiv 0$ would be useless for getting product solutions, we get the **periodic boundary conditions**

$$\Theta(\pi) - \Theta(-\pi) = 0 \quad \text{and} \quad \Theta'(\pi) - \Theta'(-\pi) = 0$$

* Mathematically, if $u(r,\theta)$ is continuously differentiable and $u(r,-\pi^+) \equiv u(r,\pi^-)$, then $\frac{\partial u}{\partial r}(r,-\pi^+) \equiv \frac{\partial u}{\partial r}(r,\pi^-)$ follows, because, for example, it is okay to interchange the operations of taking limit as $\theta \to -\pi^+$ and the limit involved in taking the derivative with respect to r.

that we studied in Section 9.3. Indeed, we saw there that the ODE-BVP eigenvalue problem

$$\left\{\begin{array}{c} \Theta''(\theta)+\lambda\Theta(\theta)=0, \\ \\ \Theta(\pi)=\Theta(-\pi), \quad \Theta'(\pi)=\Theta'(-\pi) \end{array}\right\}$$

has eigenvalues and corresponding eigenfunctions

$$\lambda_0 = 0: \ \Theta_0(\theta) = 1$$

$$\lambda_n = n^2: \ \Theta_n(\theta): \ \{\cos n\theta, \sin n\theta\} \quad \text{for} \quad n = 1, 2, 3, \ldots.$$

Recall from Section 9.3 that the latter means that both $\cos n\theta$ and $\sin n\theta$ are eigenfunctions corresponding to the same eigenvalue λ_n.

Corresponding to eigenvalues λ_0 and λ_n, we get product solutions

$$u_0(r,\theta) = R_0(r), \ u_{n,c}(r,\theta) = R_{n,c}(r)\cos n\theta, \text{ and } u_{n,s}(r,\theta) = R_{n,s}(r)\sin n\theta.$$

Substituting $u_0(r,\theta)$ into the PDE (11.52) gives, for $n = 1, 2, \ldots$,

$$0 = r\frac{d}{dr}\left[r\frac{dR_0}{dr}\right] + 0 = r^2\frac{d^2R_0}{dr^2} + r\frac{dR_0}{dr},$$

a Cauchy–Euler ODE. Substituting in $R_0(r) = r^m$ gives characteristic equation $0 = m(m-1) + m = m^2$, so the roots are $m = 0, 0$. The double root $m = 0$ gives

$$R_0(r) = c_1 r^0 + c_2 r^0 \ln r = c_1 + c_2 \ln r.$$

Substituting $u_{n,c}(r,\theta)$ into the PDE (11.52) gives

$$0 = \cos n\theta\left(r\frac{d}{dr}\left[r\frac{dR_{n,c}}{dr}\right] - n^2R_{n,c}\right) = \cos n\theta\left(r\frac{d^2R_{n,c}}{dr^2} + r\frac{dR_{n,c}}{dr} - n^2R_{n,c}\right);$$

hence, we want

$$0 = r^2\frac{d^2R_{n,c}}{dr^2} + r\frac{dR_{n,c}}{dr} - n^2R_{n,c},$$

a Cauchy–Euler ODE. Substituting in $R_{n,c}(r) = r^m$ gives characteristic equation $0 = m(m-1) + m - n^2 = m^2 - n^2$, so the roots are $m = \pm n$. This gives

$$R_{n,c}(r) = \widetilde{c}_1 r^{-n} + \widetilde{c}_2 r^n.$$

The work for $u_{n,s}$ is similar.

Using linearity to include all product solutions, the general solution of the PDE + BCs (11.52) + (11.54) is

$$u(r,\theta) = \frac{a_0}{2} + \frac{c_0\ln r}{2} + \sum_{n=1}^{\infty}(a_n\cos n\theta + b_n\sin n\theta)r^n$$

$$+ \sum_{n=1}^{\infty}(c_n\cos n\theta + d_n\sin n\theta)r^{-n}. \tag{11.55}$$

At this point, the boundedness of $u(0^+,\theta)$ yields $c_0 = 0, c_n = 0, d_n = 0$. The general solution of (11.52) + (11.54) is

$$u(r,\theta) = \frac{a_0}{2} + \sum_{n=1}^{\infty} r^n (a_n \cos n\theta + b_n \sin n\theta).$$

We substitute the general solution into the final BC, (11.53), which gives the temperature on the physical boundary of the disk:

$$f(\theta) = u(a,\theta) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a^n a_n \cos n\theta + a^n b_n \sin n\theta), \quad -\pi < \theta < \pi. \tag{11.56}$$

But (11.56) is just a standard Fourier series expansion problem. As we learned in Section 9.1,

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} f(\theta)d\theta,\ a^n a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(\theta)\cos n\theta\, d\theta, a^n b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(\theta)\sin n\theta\, d\theta.$$

So the solution of the problem is

$$u(r,\theta) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\theta)d\theta + \frac{1}{\pi}\sum_{n=1}^{\infty}\left(\frac{r}{a}\right)^n \left(\left(\int_{-\pi}^{\pi} f(\theta)\cos n\theta\, d\theta\right)\cos n\theta + \left(\int_{-\pi}^{\pi} f(\theta)\sin n\theta\, d\theta\right)\sin n\theta\right). \bigcirc$$

Example 11.18

Solve $\Delta u = 0$ for the equilibrium temperature, $u(r,\theta)$, in the annulus $A_{a,b}$: $a < r < b$. Assume the temperatures on the circles $r = a$ and $r = b$ are given.

Method: Mathematically, this consists of

$$(PDE) \quad 0 = r\frac{\partial}{\partial r}\left[r\frac{\partial u}{\partial r}\right] + \frac{\partial^2 u}{\partial \theta^2},\ a < r < b, -\pi < \theta \le \pi, \tag{11.57}$$

$$(BCs)\ u(a,\theta) = f(\theta),\ u(b,\theta) = g(\theta), -\pi < \theta < \pi. \tag{11.58}$$

As in Example 11.17, continuity of temperature and flux across the ray $\theta = \pm\pi$ implies the additional, periodic boundary conditions (11.54), that is,

$$u(r,-\pi^+) = u(r,\pi^-) \quad \text{and} \quad \frac{\partial u}{\partial \theta}(r,-\pi^+) = \frac{\partial u}{\partial \theta}(r,\pi^-).$$

At first, the solution proceeds exactly the same as that for Example 11.17, which we summarize now: The periodic boundary conditions (11.54) imply that the product solutions are given:

$$u_0(r,\theta) = R_0(r),\ u_{n,c}(r,\theta) = R_{n,c}(r)\cos n\theta, \text{ and } u_{n,s}(r,\theta) = R_{n,s}(r)\sin n\theta.$$

Substituting them into the PDE of Laplace equation leads to Cauchy–Euler ODEs whose solutions are the functions $R_0(r)$, $R_{n,c}(r)$, $R_{n,s}(r)$, and this gives product solutions that we add up to see that the general solution of (11.52) and (11.54) is (11.55), that is,

$$u(r,\theta)=\frac{a_0+c_0\ln r}{2}+\sum_{n=1}^{\infty}(a_n\cos n\theta+b_n\sin n\theta)r^n+\sum_{n=1}^{\infty}(c_n\cos n\theta+d_n\sin n\theta)r^{-n}.$$

Because $r=0$ is *not* in the domain $A_{a,b}$, we do *not* need to worry about boundedness of $|\,u(0^+,\theta)\,|$, so we do *not* require $c_0=0, c_n=0, d_n=0$.

At this point, we could substitute (11.55) into the remaining boundary conditions (11.58). But, it will make our algebraic work a lot easier if we first use some *clairvoyance*, as we did in Section 11.3. Instead of writing $R_{n,c}(r)$ as a linear combination of the functions r^{-n} and r^n, we could write $R_{n,c}(r)$ as a linear combination of the functions

$$H_n(r)=\left(\frac{r}{a}\right)^n-\left(\frac{r}{a}\right)^{-n},\ h_n(r)=\left(\frac{r}{b}\right)^n-\left(\frac{r}{b}\right)^{-n},\quad \text{for}\quad n=1,2,\ldots.$$

Why choose these functions? The functions $H_n(r)$ satisfy $H_n(a)=0$, and the functions $h_n(r)$ satisfy $h_n(b)=0$. And, both $H_n(r)$ and $h_n(r)$ are linear combinations of the functions r^n and r^{-n}. Similarly, we can use clairvoyance to rewrite the terms $\frac{a_0+c_0\ln r}{2}$ as

$$\frac{C_0\,\ln\left(\frac{b}{r}\right)+A_0\,\ln\left(\frac{r}{a}\right)}{2}.$$

Why? If we define $H_0(r)=\ln\left(\frac{r}{a}\right)$ and $h_0(r)=\ln\left(\frac{b}{r}\right)$, then these will be useful in satisfying BCs at $r=a$ and $r=b$ because $H_0(a)=0$ and $h_0(b)=0$. We also see that $H_0(r)$ and $h_0(r)$ are linear combinations of 1 and $\ln r$ because $H_0(r)=\ln(r)-\ln(a)=-\ln(a)\cdot 1+1\cdot\ln r$ and similarly $h_0(r)=\ln(b)-\ln(r)=\ln(b)\cdot 1-1\cdot\ln r$.

So, we take the general solution of (11.57) + (11.54) in the form

$$u(r,\theta)=\frac{C_0\,\ln\left(\frac{b}{r}\right)+A_0\,\ln\left(\frac{r}{a}\right)}{2}$$
$$+\sum_{n=1}^{\infty}(A_nH_n(r)+C_nh_n(r))\cos n\theta+\sum_{n=1}^{\infty}(B_nH_n(r)+E_nh_n(r))\sin n\theta,$$

that is,

$$u(r,\theta)=\frac{C_0\,\ln\left(\frac{b}{r}\right)+A_0\,\ln\left(\frac{r}{a}\right)}{2}$$
$$+\sum_{n=1}^{\infty}H_n(r)\,(A_n\cos n\theta+B_n\sin n\theta)+\sum_{n=1}^{\infty}h_n(r)\,(C_n\cos n\theta+E_n\sin n\theta)\,,$$

(11.59)

and substitute that into the remaining boundary conditions (11.58):

$$f(\theta) = u(a,\theta) = \frac{C_0 \ln\left(\frac{b}{a}\right) + A_0 \ln\left(\frac{a}{a}\right)}{2}$$

$$+ \sum_{n=1}^{\infty} H_n(a)\left(A_n \cos n\theta + B_n \sin n\theta\right) + \sum_{n=1}^{\infty} h_n(a)\left(C_n \cos n\theta + E_n \sin n\theta\right)$$

$$= \frac{C_0 \ln\left(\frac{b}{a}\right)}{2} + \sum_{n=1}^{\infty}\left(C_n h_n(a)\cos n\theta + E_n h_n(a)\sin n\theta\right),$$

because $H_n(a) = 0$ and $\ln 1 = 0$. This is a standard Fourier series expansion problem from Section 9.1, so the coefficients satisfy

$$C_0 \ln\left(\frac{b}{a}\right) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(\theta)d\theta,\ C_n h_n(a) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(\theta)\cos n\theta\, d\theta,\ E_n h_n(a) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(\theta)\sin n\theta\, d\theta\,.$$

When we substitute (11.59) into the last boundary condition in (11.58), we get

$$g(\theta) = u(b,\theta) = \frac{C_0 \ln\left(\frac{b}{b}\right) + A_0 \ln\left(\frac{b}{a}\right)}{2}$$

$$+ \sum_{n=1}^{\infty} H_n(b)\left(A_n \cos n\theta + B_n \sin n\theta\right) + \sum_{n=1}^{\infty} h_n(b)\left(C_n \cos n\theta + E_n \sin n\theta\right)$$

$$= \frac{A_0 \ln(\frac{b}{a})}{2} + \sum_{n=1}^{\infty}\left(A_n H_n(b)\cos n\theta + B_n H_n(b)\sin n\theta\right).$$

This is a standard Fourier series expansion problem from Section 9.1, so the coefficients satisfy

$$A_0 \ln\left(\frac{b}{a}\right) = \frac{1}{\pi}\int_{-\pi}^{\pi} g(\theta)d\theta,\ A_n H_n(b) = \frac{1}{\pi}\int_{-\pi}^{\pi} g(\theta)\cos n\theta\, d\theta,$$

$$B_n H_n(b) = \frac{1}{\pi}\int_{-\pi}^{\pi} g(\theta)\sin n\theta\, d\theta.$$

The solution of the whole problem is

$$u(r,\theta) = \frac{\ln\left(\frac{b}{r}\right)\left(\int_{-\pi}^{\pi} f(\theta)d\theta\right) + \ln\left(\frac{r}{a}\right)\left(\int_{-\pi}^{\pi} g(\theta)d\theta\right)}{2\pi \ln\left(\frac{b}{a}\right)}$$

$$+ \frac{1}{\pi}\sum_{n=1}^{\infty}\frac{H_n(r)}{H_n(b)}\left(\left(\int_{-\pi}^{\pi} g(\theta)\cos n\theta d\theta\right)\cos n\theta + \left(\int_{-\pi}^{\pi} g(\theta)\sin n\theta d\theta\right)\sin n\theta\right)$$

$$+ \frac{1}{\pi}\sum_{n=1}^{\infty}\frac{h_n(r)}{h_n(a)}\left(\left(\int_{-\pi}^{\pi} f(\theta)\cos n\theta d\theta\right)\cos n\theta + \left(\int_{-\pi}^{\pi} f(\theta)\sin n\theta d\theta\right)\sin n\theta\right). \bigcirc$$

11.5.2 Heat Equation in Polar Coordinates

Example 11.19

Suppose the temperature, T, in a disk $\mathcal{D}_a : 0 \le r < a$ satisfies the heat equation $\frac{\partial T}{\partial t} = \Delta T$, the temperature is 0 on the circle that bounds the disk, and the initial temperature distribution, $T|_{t=0} = f(r)$, is circularly symmetric. Find the temperature distribution as a function of space and time.

Method: Mathematically, this is the PDE-BVP-IVP

$$\frac{\partial T}{\partial t}(r,\theta,t) = \alpha \left(\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial T}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 T}{\partial \theta^2}\right), 0 \le r < a, -\pi < \theta \le \pi, \tag{11.60}$$

$$T(a,\theta,t) = 0, -\pi < \theta < \pi,\ 0 < t < \infty, \tag{11.61}$$

$$T(r,\theta,0) = f(r), 0 < r < a, -\pi < \theta \le \pi. \tag{11.62}$$

Because of the circular symmetry of *both* the boundary condition and the initial condition, it makes sense to hope that the solution does not depend on θ, that is, $T = T(r,t)$, where (r,t) lies in the infinite slab, $0 < r < a,\ 0 < t < \infty$. The PDE becomes just

$$\frac{\partial T}{\partial t}(r,t) = \alpha \frac{1}{r}\frac{\partial}{\partial r}\left[r\frac{\partial T}{\partial r}\right]. \tag{11.63}$$

The closest things we have to two homogeneous boundary conditions on parallel sides are the physical BC $T(a,t) = 0$ and the mathematical/physical BC $|\,T(0^+,t)\,| < \infty$. So, we will try product solutions $T(r,t) = R(r)G(t)$, where $R(r)$ solves the ODE-BVP

$$\left\{ \begin{array}{l} \frac{1}{r}\frac{d}{dr}\left[r\frac{dR}{dr}\right] + \lambda R(r) = 0, \\ \\ |\,R(0^+)\,| < \infty,\ R(a) = 0 \end{array} \right\}. \tag{11.64}$$

The ODE can be rewritten as

$$r\frac{d}{dr}\left[r\frac{dR}{dr}\right] + r^2\lambda R(r) = 0, 0 < r < a,$$

that is,

$$r^2R''(r) + rR'(r) + \lambda r^2 R(r) = 0, 0 < r < a, \tag{11.65}$$

which is a form of Bessel's equation of order 0, as we saw in Section 9.6. For $\lambda > 0$, the solutions of (11.65) are given by

$$R(r) = c_1 J_0(\sqrt{\lambda}\, r) + c_2 Y_0(\sqrt{\lambda}\, r), 0 < r < a.$$

But $\lim_{r\to 0^+} |Y_0(\sqrt{\lambda}\, r)| = \infty$. So, the mathematical/physical BC requires that $c_2 = 0$; hence, $R(r) = c_1 J_0(\sqrt{\lambda}\, r)$, $0 < r < a$. The second, physical BC gives a characteristic equation $J_0(\sqrt{\lambda}\, a) = 0$. This is entirely analogous to the characteristic equation $\sin(\sqrt{\lambda}\, L) = 0$ that we got for the ODE-BVP (11.4) in Section 11.1, that is,

$$X''(x) + \lambda X(x) = 0, \quad X(0) = X(L) = 0.$$

When we graph $J_0(\sqrt{\lambda}\, a)$ versus λ in Figure 11.15, we see that there are infinitely many eigenvalues $\lambda_n = \left(\frac{\gamma_{n,0}}{a}\right)^2$, where $\gamma_{n,0}$ satisfy $J_0(\gamma_{n,0}) = 0$, for (11.64). (This is analogous

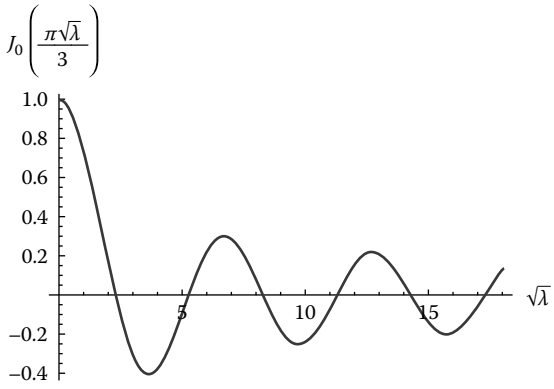


FIGURE 11.15
Finding eigenvalues.

to the eigenvalues $\lambda_n = \left(\frac{n\pi}{L}\right)^2$ for the ODE-BVP (11.4) in Section 11.1, found by solving characteristic equation $\sin(\sqrt{\lambda}\, L) = 0$.) There are no eigenvalues $\lambda \leq 0$ for (11.64), as we explained in Example 9.30 in Section 9.6.1.

The eigenfunctions for the ODE-BVP (11.64) are $R_n(r) = J_0(\gamma_{n,0}\, r/a)$, corresponding to eigenvalues $\lambda_n = \left(\frac{\gamma_{n,0}}{a}\right)^2$. We substitute our hoped for product solutions $u(r,t) = R_n(r)G_n(t)$ into the PDE to get

$$R_n(r)\frac{dG_n}{dt} = \alpha\, G_n(t) \cdot \frac{1}{r}\frac{d}{dr}\left[\, r\,\frac{dR_n}{dr}\,\right] = \alpha\, G_n(t) \cdot (-\lambda_n R_n(r)),$$

so we need $G_n(t)$ to satisfy the ODE

$$\frac{dG_n}{dt} = -\alpha\, \lambda_n G_n(t).$$

Solving this gives product solutions

$$T_n(r,t) = J_0\left(\frac{\gamma_{n,0}\, r}{a}\right)\exp\left(-\alpha\left(\frac{\gamma_{n,0}}{a}\right)^2 t\right),$$

and linearity allows us to add them up to get the general solution of (11.60) and (11.61) in the form

$$T(r,t) = \sum_{n=1}^{\infty} a_n J_0\left(\frac{\gamma_{n,0}\, r}{a}\right)\exp\left(-\alpha\left(\frac{\gamma_{n,0}}{a}\right)^2 t\right).$$

We substitute this into the IC, (11.62), to get

$$f(r) = T(r,0) = \sum_{n=1}^{\infty} a_n J_0\left(\frac{\gamma_{n,0}\, r}{a}\right). \tag{11.66}$$

To analyze (11.66), we use the orthogonality relation, as mentioned in (9.71):

$$\int_0^a J_0\left(\frac{\gamma_{m,0}\, r}{a}\right) J_0\left(\frac{\gamma_{n,0}\, r}{a}\right) r\, dr = \left\{\begin{matrix} 0, & \text{if } n \neq m \\ N_m, & \text{if } n = m \end{matrix}\right\},$$

where

$$N_m \triangleq \int_0^a \left(J_0\left(\frac{\gamma_{m,0}\, r}{a}\right)\right)^2 rdr = \frac{1}{2}\left(J_1(\gamma_{m,0})\right)^2 \tag{11.67}$$

does depend on m.

To solve (11.66) for the coefficients a_n, multiply through by $rJ_0(\gamma_{m,0}\, r/a)$, and then integrate with respect to r on the interval $0 < r < a$. This gives

$$\int_0^a f(r)J_0\left(\frac{\gamma_{m,0}\, r}{a}\right) rdr = \int_0^a \left(\sum_{n=1}^{\infty} a_n J_0\left(\frac{\gamma_{n,0}\, r}{a}\right)\right) J_0\left(\frac{\gamma_{m,0}\, r}{a}\right) r\, dr.$$

We ignore the mathematical issue of justifying the interchange of integration and summation to get, using (11.67),

$$\int_0^a f(r)J_0\left(\frac{\gamma_{m,0}\, r}{a}\right) r\, dr = \sum_{n=1}^{\infty} a_n \int_0^a J_0\left(\frac{\gamma_{m,0}\, r}{a}\right) J_0\left(\frac{\gamma_{n,0}\, r}{a}\right) r\, dr = \sum_{n=1}^{\infty} a_n \begin{Bmatrix} 0, & n \neq m \\ N_m, & n = m \end{Bmatrix}$$

$$= a_m \cdot \frac{1}{2}\left(J_1(\gamma_{m,0})\right)^2.$$

It follows that $a_m = \dfrac{1}{N_m}\int_0^a f(r)J_0\left(\dfrac{\gamma_{m,0}\, r}{a}\right) rdr$. So, the solution of the original problem is

$$T(r,t) = \sum_{n=1}^{\infty} \frac{\int_0^a f(r)J_0\left(\dfrac{\gamma_{n,0}\, r}{a}\right) r\, dr}{\frac{1}{2}\left(J_1(\gamma_{n,0})\right)^2} J_0\left(\frac{\gamma_{n,0}\, r}{a}\right) \exp\left(-\alpha\left(\frac{\gamma_{n,0}}{a}\right)^2 t\right). \bigcirc$$

Note that $T(r,t) \to 0$ as $t \to \infty$. This makes sense physically because the equilibrium temperature distribution is zero.

11.5.3 Problems

1. Solve the problem

$$\left\{ \begin{array}{c} \dfrac{1}{r}\dfrac{\partial}{\partial r}\left[r\dfrac{\partial u}{\partial r}\right] + \dfrac{1}{r^2}\dfrac{\partial^2 u}{\partial \theta^2} = 0, 0 \leq r < a, -\pi < \theta \leq \pi \\ \\ u(a,\theta) = |\theta|, \ -\pi < \theta < \pi \end{array} \right\},$$

along with any other boundary condition(s) based on mathematical and/or physical grounds.

2. Solve Laplace equation for $a < r < \infty$, that is, in the **exterior** of the disk $0 \leq r < a$. Assume that the solution $u(r,\theta)$ is bounded as $r \to \infty$ and also satisfies the boundary condition $u(a,\theta) = 1 - 2\sin(3\theta)$. Write your final conclusion in as simple a form as possible.
3. Find the electrostatic potential, $V(r,\theta)$, that satisfies Laplace equation in the annulus $a < r < b$. Assume that the potential is constant on the circles $r = a$ and $r = b$, specifically $V(a,\theta) \equiv V_a$ and $V(b,\theta) \equiv V_b$, for $-\pi < \theta < \pi$.

4. Solve Laplace equation in the half disk $0 < r < a, 0 < \theta < \pi$, shown in Figure 11.12, assuming instead that the bottom, flat side is insulated and the temperature is prescribed on the top, half circle side.
5. For the problem

$$\left\{ \begin{array}{c} 0 = \Delta u, 0 < r < a, -\frac{\pi}{2} < \theta < \frac{\pi}{2} \\ \\ u(r, -\frac{\pi}{2}) = u(r, \frac{\pi}{2}) = 0, 0 < r < a \end{array} \right\},$$

   (a) explain why the general solution for the PDE + (BCs on the rays $\theta = \pm\frac{\pi}{2}$) is given by

$$u(r,\theta) = \sum_{k=1}^{\infty} \left(a_k r^{2k} \sin(2k\theta) + b_k r^{2k-1} \cos(2k-1)\theta \right),$$

   and (b) find the solution of the problem where the last boundary condition is $u(a, \theta) = |\theta|, -\frac{\pi}{2} < \theta < \frac{\pi}{2}$.
6. Solve

$$\left\{ \begin{array}{c} \frac{\partial T}{\partial t} = \alpha \frac{1}{r} \frac{\partial}{\partial r} \left[r \frac{\partial T}{\partial r} \right], 0 \le r < a, 0 < t < \infty \\ \\ T(a,t) = 0, 0 < t < \infty, \\ \\ T(r,0) = J_0(\gamma_{3,0}\, r/a) - \frac{1}{2} J_0(\gamma_{5,0}\, r/a), 0 < r < a. \end{array} \right\}.$$

7. Find the equilibrium temperature distribution, $T = T(r, z)$, in a circularly symmetric rod that is insulated everywhere except on the two ends, modeled by

$$\left\{ \begin{array}{c} \frac{\partial T}{\partial t} = \alpha \left(\frac{1}{r} \frac{\partial}{\partial r} \left[r \frac{\partial T}{\partial r} \right] + \frac{\partial^2 T}{\partial z^2} \right), \\ \\ 0 \le r < a,\ 0 < z < L, \\ \\ \frac{\partial T}{\partial r}(a,z) = 0,\ 0 < z < L, \\ \\ T(r,0) = f(r), T(r,L) = g(r), 0 < r < a, \end{array} \right\},$$

   along with any other boundary condition(s) based on mathematical and/or physical grounds. [Hint: You may assume that $\frac{d}{dx}\left[J_0(x) \right] \equiv J_1(x)$.]

8. Solve

$$\left\{\begin{array}{c} 0=\dfrac{1}{r}\dfrac{\partial}{\partial r}\left[r\dfrac{\partial T}{\partial r}\right]+\dfrac{\partial^2 T}{\partial z^2},\ 0\le r<2, 0<z<3 \\ \\ T(2,z)=0,\ 0<z<3, \\ \\ T(r,0)=J_0(\gamma_{1,0}\,r/2)-\frac{1}{4}J_0(\gamma_{4,0}\,r/2), 0<r<2 \\ \\ T(r,3)=-\frac{1}{2}J_0(\gamma_{2,0}\,r/2), 0<r<2 \end{array}\right\},$$

along with any other boundary condition(s) based on mathematical and/or physical grounds.

9. In Section 6.7, we explained why in polar coordinates the vector differential operators of *grad* and *div* are given by, respectively,

$$\boldsymbol{\nabla} u=\frac{\partial u}{\partial r}\,\hat{\mathbf{e}}_r+\frac{1}{r}\frac{\partial u}{\partial \theta}\,\hat{\mathbf{e}}_\theta \quad \text{and} \quad \boldsymbol{\nabla}\bullet(F_r\,\hat{\mathbf{e}}_r+F_\theta\,\hat{\mathbf{e}}_\theta)=\frac{1}{r}\frac{\partial}{\partial r}\big[rF_r\big]+\frac{1}{r}\frac{\partial}{\partial \theta}\big[F_\theta\big].$$

In Section 10.2, we learned that the equations

$$0=\boldsymbol{\nabla}\bullet\mathbf{q} \text{ and } \mathbf{q}=-\kappa\boldsymbol{\nabla}T$$

model the equilibrium temperature, T, where $\mathbf{q}$ is the heat flux vector:

(a) In the special case that the physical domain is the disk $\mathcal{D}_a : 0 \le r < a$ and the thermal conductivity is given by $\kappa = c\varrho r$ and c and ϱ are constants, write down the single second-order PDE that T should satisfy.

(b) Find the equilibrium temperature that satisfies the PDE you wrote in part (a), along with the mathematical boundary conditions appropriate for a disk and the physical boundary condition prescribing the temperature on the boundary, that is, $T(a,\theta)=f(\theta), -\pi<\theta<\pi$.

10. Find the equilibrium temperature distribution in the quarter sector of an annulus given by $a \le r \le b,\ 0 \le \theta \le \frac{\pi}{2}$. Assume that the sides $\theta = 0$ and $\theta = \frac{\pi}{2}$ are kept insulated, that is, have zero flux in the normal direction, and that the temperature is prescribed on the inner and outer quarter circles $r = a$ and $r = b$, for $0 \le \theta \le \frac{\pi}{2}$. Assume that there are no sources or sinks inside the quarter sector of the annulus.

11. Solve

$$\left\{\begin{array}{c} 0=\dfrac{1}{r}\dfrac{\partial}{\partial r}\left[r\dfrac{\partial T}{\partial r}\right]+\dfrac{\partial^2 T}{\partial z^2},\ 0\le r<2, 0<z<4, \\ \\ T(2,z)=0,\ 0<z<4, \\ \\ T(r,0)=50, T(r,4)=0, 0<r<2, \end{array}\right\},$$

along with any other boundary condition(s) based on mathematical and/or physical grounds. [Hint: $\int J_0(x)\,x\,dx = x J_1(x)$.]

12. Let the domain $\mathcal{D}$ be the three-quarters of a disk given by

$$\left\{(r,\theta):\ 0<r<1, -\frac{3\pi}{4} < \theta < \frac{3\pi}{4}\right\}.$$

You may assume that the solution of the problem

$$\left\{\begin{array}{c} \Delta u(r,\theta) = 0,\ (r,\theta) \text{ in } \mathcal{D} \\ u(r,-\tfrac{3\pi}{4}) = u(r,\tfrac{3\pi}{4}) = 0,\ 0<r<1 \\ u(a,\theta) = f(\theta), -\tfrac{3\pi}{4} < \theta < \tfrac{3\pi}{4} \end{array}\right\}$$

is given by

$$u(r,\theta) = \sum_{n=1}^{\infty} a_n r^{2n/3} \sin\left(n\left(\frac{2\theta}{3}+\frac{\pi}{2}\right)\right).$$

(a) Use orthogonality to find a formula for the generalized Fourier coefficients a_n.

(b) Rewrite the solution as the sum of the first term and the sum of the terms for $n \geq 2$, and then find $\frac{\partial u}{\partial r}$ in the same type of form.

(c) Find a condition on $f(\theta)$ that must be satisfied if we are to have the flux vector $||\nabla u(0^+,\theta)|| < \infty$.

13. Solve the problem

$$\left\{\begin{array}{c} 0 = \dfrac{1}{r}\dfrac{\partial u}{\partial r}\left[r\dfrac{\partial u}{\partial r}\right] + \dfrac{1}{r^2}\dfrac{\partial^2 u}{\partial \theta^2}, a<r<b, -\pi < \theta \leq \pi \\ u(a,\theta) = f(\theta),\ -\pi < \theta < \pi \\ b\dfrac{\partial u}{\partial r}(b,\theta) = a\dfrac{\partial u}{\partial r}(a,\theta),\ -\pi < \theta < \pi \end{array}\right\},$$

along with any other boundary condition(s) based on mathematical and/or physical grounds.

14. Let the domain $\mathcal{D}$ be the half annulus given by $\{(r,\theta):\ a<r<b,\ 0<\theta<\pi\}$. For the problem

$$\left\{ \begin{array}{c} \Delta u(r,\theta) = 0,\ (r,\theta) \text{ in } \mathcal{D} \\ \left\{ \begin{array}{l} \dfrac{1}{r}\dfrac{\partial u}{\partial \theta}(r,0) = f(r) \\ \dfrac{1}{r}\dfrac{\partial u}{\partial \theta}(r,\pi) = g(r) \end{array} \right\}, a < r < b \\ \dfrac{\partial u}{\partial r}(a,\theta) = \dfrac{\partial u}{\partial r}(b,\theta) = 0,\ 0 < \theta < \pi \end{array} \right\},$$

(a) What does the solvability condition for this Neumann problem require of the functions f and g?

(b) Solve the problem. You may find useful the results of Problem 9.6.4.3.

11.6 PDEs in Cylindrical and Spherical Coordinates

First, in two different examples, we will explain how the method of separation of variables can be used in cylindrical coordinates. We recall from Section 6.7 the Laplacian differential operator in cylindrical coordinates:

$$\Delta u \triangleq \frac{1}{r}\frac{\partial}{\partial r}\left[r\frac{\partial u}{\partial r}\right] + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} + \frac{\partial^2 u}{\partial z^2}. \tag{11.68}$$

Suppose the domain is the inside of a cylinder,

$$\mathcal{D} = \{\,(r,\theta,z) : -\pi < \theta \le \pi,\ 0 \le r < a,\ 0 < z < H\,\}, \tag{11.69}$$

or the region between two co-axial cylinders,

$$\{\,(r,\theta,z) : -\pi < \theta \le \pi,\ a < r < b,\ 0 < z < H\,\},$$

or the region(s) exterior to either of the aforementioned. When we separate variables, we will separate the θ-dependence from the r and/or z-dependence at some point. This is because the mathematical boundary conditions of periodicity with respect to θ imply two separated, homogeneous boundary conditions (see (11.73) in the following) analogous to those of (11.54) in Section 11.5.

Example 11.20

Find the equilibrium temperature distribution in the cylinder (11.69), assuming the lateral surface is insulated and the temperature is given on the two circular ends.

Method: Let T be the temperature. Because it is in equilibrium, $\dfrac{\partial T}{\partial t} \equiv 0$, so

$$0 = \frac{1}{r}\frac{\partial}{\partial r}\left[r\frac{\partial T}{\partial r}\right] + \frac{1}{r^2}\frac{\partial^2 T}{\partial \theta^2} + \frac{\partial^2 T}{\partial z^2},\ -\pi < \theta \le \pi,\ 0 \le r < a,\ 0 < z < H. \tag{11.70}$$

Insulation of the lateral surface implies flux in the normal direction is zero, that is,

$$\frac{\partial T}{\partial r}(a,\theta,z) \equiv 0, \ -\pi < \theta \leq \pi, \ 0 < z < H. \tag{11.71}$$

On physical grounds,

$$|\, T(0^+,\theta,z)\,| < \infty, \ -\pi < \theta \leq \pi, \ 0 < z < H. \tag{11.72}$$

Because of periodicity with respect to θ,

$$T(r,-\pi^+,z) = T(r,\pi^-,z) \quad \text{and}$$

$$\frac{\partial T}{\partial \theta}(r,-\pi^+,z) = \frac{\partial T}{\partial \theta}(r,\pi^-,z), \ 0 \leq r < a, \ 0 < z < H. \tag{11.73}$$

Finally, we specify the temperatures on the circular ends:

$$\left\{ \begin{array}{l} T(r,\theta,0) = f(r,\theta) \\ \\ T(r,\theta,H) = g(r,\theta) \end{array} \right\}, \ -\pi < \theta \leq \pi, \ 0 \leq r < a. \tag{11.74}$$

The physical cylinder is a box (also known as parallelepiped) in (r,θ,z) coordinates, so (11.73), combined with (11.71) and (11.72), expresses pairs of "homogeneous" boundary conditions on parallel sides of that box. This suggests separating both the r- and θ-dependence away from the z-dependence. As in the application to surface waves in Section 11.4, we can do this by finding eigenvalues of the (r,θ) *part* of the Laplacian

$$\frac{1}{r}\frac{\partial}{\partial r}\Big[\, r\,\frac{\partial \phi}{\partial r}\,\Big] + \frac{1}{r^2}\frac{\partial^2 \phi}{\partial \theta^2} + \lambda\phi = 0, \tag{11.75}$$

along with the boundary conditions

$$\left\{ \begin{array}{c} \phi(r,-\pi^+) = \phi(r,\pi^-) \quad \text{and} \quad \dfrac{\partial \phi}{\partial \theta}(r,-\pi^+) = \dfrac{\partial \phi}{\partial \theta}(r,\pi^-), 0 \leq r < a, \\ \\ |\, \phi(0^+,\theta)\,| < \infty \quad \text{and} \quad \dfrac{\partial \phi}{\partial r}(a,\theta) \equiv 0, -\pi < \theta < \pi. \end{array} \right\}. \tag{11.76}$$

The periodicity in θ suggests looking for solutions in the forms $\phi(r,\theta) = 1 \cdot R_0(r)$ and $\phi(r,\theta) = \{\cos n\theta, \sin n\theta\} R(r)$, where we recall from Sections 9.3 and 11.5 that the notation $\{\cos n\theta, \sin n\theta\}$ means both $\cos n\theta$ and $\sin n\theta$, for $n = 1, 2, 3, \ldots$.

First, if we substitute $\phi(r,\theta) = 1 \cdot R_0(r)$ into the PDE (11.75), we get ODE

$$\frac{1}{r}\frac{d}{dr}\left[\, r\,\frac{dR_0}{dr}\,\right] + \lambda R_0 = 0, \quad \text{that is,} \quad r^2 R_0'' + rR' + \lambda r^2 R_0 = 0,$$

plus the boundary conditions $|\, R_0(0^+)\,| < \infty$, $\frac{dR_0}{dr}(a) = 0$. For $\lambda > 0$, the ODE is a form of Bessel's equation of order zero. Similar to Example 11.19 in Section 11.5, the solutions of the ODE are $R(r) = c_1 J_0(\sqrt{\lambda}\, r) + c_2 Y_0(\sqrt{\lambda}\, r), 0 < r < a$. When we apply the boundary condition $|\, R_0(0^+)\,| < \infty$, we conclude that $c_2 = 0$, because $|\, Y_0(0^+)\,| = \infty$. When we apply the last boundary condition by graphing $J_0'(\sqrt{\lambda}\, a)$ versus $\lambda > 0$, we find there are infinitely many eigenvalues $\lambda_{m,0}, m = 1, 2, \ldots$. In fact, $J_0'(x) = -J_1(x)$, so

$$\lambda_{m,0} = \left(\frac{\gamma_{m,1}}{a}\right)^2,$$

where $\gamma_{m,1}$, $m = 1, 2, \ldots$, are the zeros of the function $J_1(x)$.

So, we get solutions

$$\phi_{m,0}(r,\theta) = J_0(\sqrt{\lambda_{m,0}}\, r),\ m = 1, 2, \ldots.$$

Second, if we substitute $\phi(r,\theta) = \cos n\theta R_n(r)$ into the PDE (11.75), we get

$$\cos n\theta \,\frac{1}{r}\,\frac{d}{dr}\left[r\,\frac{dR_n}{dr}\right] + R_n(r)\frac{1}{r^2}\cdot(-n^2)\cos n\theta + \lambda\cos n\theta R_n(r) = 0;$$

hence, $R_n(r)$ should satisfy the ODE

$$\frac{1}{r}\,\frac{d}{dr}\left[r\,\frac{dR_n}{dr}\right] - n^2\cdot\frac{1}{r^2}\,R_n + \lambda R_n = 0,$$

that is, the ODE

$$r^2R_n'' + rR_n' + (-n^2 + \lambda r^2)R_n = 0,$$

plus the boundary conditions $|\,R_n(0^+)\,| < \infty$ and $\frac{dR_n}{dr}(a) = 0$. For $\lambda > 0$, the ODE is a form of Bessel's equation of order n. Similar to Example 11.19, the solutions are $R_n(r) = c_1J_n(\sqrt{\lambda}\, r) + c_2Y_n(\sqrt{\lambda}\, r), 0 < r < a$. When we apply the boundary condition $|\,R_n(0^+)\,| < \infty$, we conclude that $c_2 = 0$, because $|\,Y_n(0^+)\,| = \infty$. When we apply the last boundary condition by graphing $J_n'(\sqrt{\lambda}\, a)$ versus $\lambda > 0$, we find there are infinitely many eigenvalues $\lambda_{m,n}, m = 1, 2, \ldots$.

The work for $\phi(r,\theta) = \sin n\theta R_n(r)$ is the same as for $\phi(r,\theta) = \cos n\theta R_n(r)$, except for using the sine function instead of the cosine function.

So, we get solutions

$$\begin{aligned}\phi_{m,n,c}(r,\theta) &= J_n(\sqrt{\lambda_{m,n}}\, r)\cos n\theta,\ m = 1, 2, \ldots,\ n = 1, 2, \ldots\\ \phi_{m,n,s}(r,\theta) &= J_n(\sqrt{\lambda_{m,n}}\, r)\sin n\theta,\ m = 1, 2, \ldots,\ n = 1, 2, \ldots.\end{aligned}$$

When we substitute product solutions $\phi_{m,0}(r,\theta)Z(z)$ into the original PDE, we get

$$\begin{aligned}0 &= \frac{1}{r}\,\frac{\partial}{\partial r}\left[r\,\frac{\partial}{\partial r}\phi_{m,0}(r,\theta)Z(z)\right] + \frac{1}{r^2}\,\frac{\partial^2}{\partial\theta^2}\big[\phi_{m,0}(r,\theta)Z(z)\big] + \frac{\partial^2}{\partial z^2}\big[\phi_{m,0}(r,\theta)Z(z)\big]\\ &= \cdots = -\lambda_{m,0}\,\phi_{m,0}(r,\theta)Z(z) + \phi_{m,0}(r,\theta)Z''(z);\end{aligned}$$

hence, $Z(z)$ should satisfy the ODE

$$Z''(z) - \lambda_{m,0}Z(z) = 0.$$

Using *clairvoyance*, as in Section 11.3, we get product solutions

$$J_0(\sqrt{\lambda_{m,0}}\, r)\sinh(\sqrt{\lambda_{m,0}}(H - z)) \text{ and } J_0(\sqrt{\lambda_{m,0}}\, r)\sinh(\sqrt{\lambda_{m,0}}\, z).$$

When we substitute product solutions $\phi_{m,n,c}(r,\theta)Z(z)$ into the original PDE, we get

$$0 = \frac{1}{r}\,\frac{\partial}{\partial r}\left[r\,\frac{\partial}{\partial r}\phi_{m,n,c}(r,\theta)Z(z)\right] + \frac{1}{r^2}\,\frac{\partial^2}{\partial\theta^2}\big[\phi_{m,n,c}(r,\theta)Z(z)\big] + \frac{\partial^2}{\partial z^2}\big[\phi_{m,n,c}(r,\theta)Z(z)\big];$$

hence,

$$0 = -\lambda_{m,n}\,\phi_{m,n,c}(r,\theta)Z(z) + \phi_{m,n,c}(r,\theta)Z''(z); \tag{11.77}$$

hence, $Z(z)$ should satisfy the ODE

$$Z''(z) - \lambda_{m,n}Z(z) = 0.$$

Using clairvoyance, for $n = 1, 2, 3, \ldots$, we get product solutions

$$J_n(\sqrt{\lambda_{m,n}}\,r)\sinh(\sqrt{\lambda_{m,n}}(H-z))\{\cos n\theta, \sin n\theta\}$$

and

$$J_n(\sqrt{\lambda_{m,n}}\,r)\sinh(\sqrt{\lambda_{m,n}}\,z)\{\cos n\theta, \sin n\theta\}.$$

We can put all product solutions together to get the general solution of (11.70) through (11.73) in the form

$$T(r,\theta,z) = \sum_{m=1}^{\infty} J_0(\sqrt{\lambda_{m,0}}\,r)\left(\frac{a_{m,0}}{2}\,\sinh(\sqrt{\lambda_{m,0}}(H-z)) + \frac{c_{m,0}}{2}\,\sinh(\sqrt{\lambda_{m,0}}\,z)\right)$$

$$+ \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} J_n(\sqrt{\lambda_{m,n}}\,r)\sinh(\sqrt{\lambda_{m,n}}(H-z))\left(a_{m,n}\cos n\theta + b_{m,n}\sin n\theta\right)$$

$$+ \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} J_n(\sqrt{\lambda_{m,n}}\,r)\sinh(\sqrt{\lambda_{m,n}}\,z)\left(c_{m,n}\cos n\theta + d_{m,n}\sin n\theta\right).$$

It is this general solution that we will substitute into the remaining two, nonhomogeneous boundary conditions (11.74):

$$f(r,\theta) = T(r,\theta,0) = \sum_{m=1}^{\infty} J_0(\sqrt{\lambda_{m,0}}\,r)\,\frac{a_{m,0}}{2}\sinh(\sqrt{\lambda_{m,0}}\,H)$$

$$+ \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} J_n(\sqrt{\lambda_{m,n}}\,r)\sinh(\sqrt{\lambda_{m,n}}\,H)\left(a_{m,n}\cos n\theta + b_{m,n}\sin n\theta\right).$$

Using

$$\int_{-\pi}^{\pi}\int_{0}^{a} J_n\left(\sqrt{\lambda_{m,n}}\,r\right)\cos n\theta\; J_{n'}\left(\sqrt{\lambda_{n',m'}}\,r\right)\cos n'\theta\; r\,dr\,d\theta =$$

$$= \left(\int_{-\pi}^{\pi}\cos n\theta\;\cos n'\theta\,d\theta\right)\cdot\left(\int_{0}^{a} J_n\left(\sqrt{\lambda_{m,n}}\,r\right)J_{n'}\left(\sqrt{\lambda_{n',m'}}\,r\right)r\,dr\right)$$

and similar results, we get the appropriate orthogonality relation

$$\int_{-\pi}^{\pi}\int_{0}^{a} J_n\left(\sqrt{\lambda_{m,n}}\,r\right)\cos n\theta\, J_{n'}\left(\sqrt{\lambda_{n',m'}}\,r\right)\cos(n'\theta)r\,dr\,d\theta$$

$$= \left\{\begin{array}{ll} 0, & \text{if } n\neq n' \text{ or } m\neq m' \\ 2\pi\displaystyle\int_0^a \left(J_0\left(\sqrt{\lambda_{m,0}}\,r\right)\right)^2 r\,dr, & \text{if } n=n'=0 \text{ and } m=m' \\ \pi\displaystyle\int_0^a \left(J_n\left(\sqrt{\lambda_{m,n}}\,r\right)\right)^2 r\,dr, & \text{if } n=n'\geq 1 \text{ and } m=m' \end{array}\right\},$$

$$\int_{-\pi}^{\pi}\int_{0}^{a} J_n\left(\sqrt{\lambda_{m,n}}\,r\right)\sin n\theta\, J_{n'}\left(\sqrt{\lambda_{n',m'}}\,r\right)\sin n'\theta\, r\,dr\,d\theta$$

$$= \left\{\begin{array}{ll} 0, & n\neq n' \text{ or } m\neq m' \\ 2\pi\displaystyle\int_0^a \left(J_0\left(\sqrt{\lambda_{m,0}}\,r\right)\right)^2 r\,dr, & n=n'=0 \text{ and } m=m' \\ \pi\displaystyle\int_0^a \left(J_n\left(\sqrt{\lambda_{m,n}}\,r\right)\right)^2 r\,dr, & n=n'\geq 1 \text{ and } m=m' \end{array}\right\},$$

and

$$\int_{-\pi}^{\pi}\int_{0}^{a} J_n\left(\sqrt{\lambda_{m,n}}\,r\right)\cos n\theta\, J_{n'}\left(\sqrt{\lambda_{n',m'}}\,r\right)\sin n'\theta\, r dr\,d\theta = 0, \text{ all } n\geq 0,\ n',m,m'\geq 1.$$

Using the orthogonality relation, the Fourier–Bessel coefficients satisfy

$$a_{m,0} = \frac{1}{\sinh(\sqrt{\lambda_{m,0}}\,H)}\,\frac{\int_{-\pi}^{\pi}\int_0^a f(r,\theta)J_0(\sqrt{\lambda_{m,0}}\,r)r dr\,d\theta}{2\pi\int_0^a (J_0(\sqrt{\lambda_{m,0}}\,r))^2\,r\,dr}$$

$$a_{m,n} = \frac{1}{\sinh(\sqrt{\lambda_{m,n}}\,H)}\,\frac{\int_{-\pi}^{\pi}\int_0^a f(r,\theta)\cos n\theta J_n(\sqrt{\lambda_{m,n}}\,r)r\,dr\,d\theta}{\pi\int_0^a (J_n(\sqrt{\lambda_{m,n}}\,r))^2\,r\,dr}$$

$$b_{m,n} = \frac{1}{\sinh(\sqrt{\lambda_{m,n}}\,H)}\,\frac{\int_{-\pi}^{\pi}\int_0^a f(r,\theta)\sin n\theta J_n(\sqrt{\lambda_{m,n}}\,r)r\,dr\,d\theta}{\pi\int_0^a (J_n(\sqrt{\lambda_{m,n}}\,r))^2\,r\,dr}.$$

Finally, substitute the general solution into the last boundary condition

$$g(r,\theta) = T(r,\theta,H) = \sum_{m=1}^{\infty} J_0(\sqrt{\lambda_{m,0}}\, r)\frac{c_{m,0}}{2}\sinh(\sqrt{\lambda_{m,0}}\, H)$$

$$+ \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} J_n(\sqrt{\lambda_{m,n}}\, r)\sinh(\sqrt{\lambda_{m,n}} H)\left(c_{m,n}\cos n\theta + d_{m,n}\sin n\theta\right)$$

and use orthogonality to get

$$c_{m,0} = \frac{1}{\sinh(\sqrt{\lambda_{m,0}}\, H)}\frac{\int_{-\pi}^{\pi}\int_0^a g(r,\theta)J_0(\sqrt{\lambda_{m,0}}\, r)r dr\, d\theta}{2\pi\int_0^a (J_0(\sqrt{\lambda_{m,0}}\, r))^2\, r\, dr}$$

$$c_{m,n} = \frac{1}{\sinh(\sqrt{\lambda_{m,n}}\, H)}\frac{\int_{-\pi}^{\pi}\int_0^a g(r,\theta)\cos n\theta J_n(\sqrt{\lambda_{m,n}}\, r)r\, dr\, d\theta}{\pi\int_0^a (J_n(\sqrt{\lambda_{m,n}}\, r))^2\, r\, dr}$$

$$d_{m,n} = \frac{1}{\sinh(\sqrt{\lambda_{m,n}}\, H)}\frac{\int_{-\pi}^{\pi}\int_0^a g(r,\theta)\sin n\theta J_n(\sqrt{\lambda_{m,n}}\, r)r\, dr\, d\theta}{\pi\int_0^a (J_n(\sqrt{\lambda_{m,n}}\, r))^2\, r\, dr}.$$

In the interest of chopping down fewer trees to make paper, this time we won't bother to substitute the Fourier–Bessel coefficients in to write the final conclusion explicitly. ◯

Example 11.21

Suppose the electrostatic potential is given on the lateral side of the cylinder (11.69), that is, $u(a,\theta,z) = f(\theta,z)$, $-\pi < \theta \le \pi, 0 < z < H$, and the electric field is zero in the direction normal to the circular ends, that is, $\frac{\partial u}{\partial z}(r,\theta,0) \equiv 0$, $\frac{\partial u}{\partial z}(r,\theta,H) \equiv 0, 0 \le r < a, -\pi < \theta \le \pi$. Find the electric field inside of the cylinder.

Method: The potential does not depend on time in electro*statics*, so $\dfrac{\partial u}{\partial t} \equiv 0$. We recall that the electric field $\mathbf{E} = \mathbf{E}(r,\theta,z) = \nabla u$, where u is the electric potential. Mathematically, our problem translates again to PDE (11.70), that is,

$$0 = \frac{1}{r}\frac{\partial}{\partial r}\left[r\frac{\partial u}{\partial r}\right] + \frac{1}{r^2}\frac{\partial^2 u}{\partial\theta^2} + \frac{\partial^2 u}{\partial z^2},\ -\pi < \theta \le \pi,\ 0 \le r < a,\ 0 < z < H. \tag{11.78}$$

Periodicity with respect to θ requires (11.73), that is,

$$u(r,-\pi^+,z) = u(r,\pi^-,z) \quad \text{and}$$

$$\frac{\partial u}{\partial\theta}(r,-\pi^+,z) = \frac{\partial u}{\partial\theta}(r,\pi^-,z),\ 0 \le r < a,\ 0 < z < H. \tag{11.79}$$

The zero component of the electric field normal to the circular ends requires

$$\frac{\partial u}{\partial z}(r,\theta,0) = \frac{\partial u}{\partial z}(r,\theta,H) = 0,\ 0 \le r < a, -\pi < \theta \le \pi. \tag{11.80}$$

As usual, on physical grounds, we require that

$$|u(0+,\theta,z)| < \infty, \ -\pi < \theta \le \pi, \ 0 < z < H. \tag{11.81}$$

Finally, we specify the potential specified on the lateral surface of the cylinder:

$$u(a,\theta,z) = f(\theta,z), \ -\pi < \theta \le \pi, \ 0 < z < H. \tag{11.82}$$

The physical cylinder is a box (also known as parallelepiped) in (r,θ,z) coordinates, and so the periodicity in θ, as well as the zero normal electric field on the circular ends, expresses a pair of homogeneous boundary conditions on parallel sides of that box. This suggests separating both the θ- and z-dependence away from the r-dependence. Unfortunately, the r- and θ-dependence are tied together in the Laplacian operator. While we *could* just separate the θ-dependence from the (r,z)-dependence, instead we will try separating *all* of the variables at once (Arscott and Darai 1990). So, let's try our product solutions in the form

$$u(r,\theta,z) = R(r)\Theta(\theta)Z(z).$$

The boundary conditions on u naturally imply boundary conditions on the factors of the product solutions:

$$\text{periodicity in } \theta \Longrightarrow \Theta(-\pi^+) = \Theta(\pi^-), \Theta'(-\pi^+) = \Theta'(\pi^-)$$

$$\text{zero normal electric field on circular ends} \Longrightarrow Z'(0) = Z'(H) = 0.$$

So, let's use eigenfunctions for the ODE-BVPs

$$\left\{ \begin{array}{c} \Theta''(\theta) + \lambda\Theta(\theta) = 0, \\ \\ \Theta(\pi) - \Theta(-\pi) = 0 \quad \text{and} \quad \Theta'(\pi) - \Theta'(-\pi) = 0 \end{array} \right\}$$

and

$$Z'' + \mu Z = 0, Z'(0) = Z'(H) = 0.$$

This gives product solutions of the form

$$u_{00}(r,\theta,z) = R_{00}(r) \cdot 1 \cdot 1,$$

$$u_{nm}(r,\theta,z) = R_{nm}(r)\{\cos n\theta, \sin n\theta\} \cos\left(\tfrac{m\pi z}{H}\right), \ m = 0,1,2,\ldots, \ n = 0,1,2,\ldots; \ n+m > 0.$$

When we substitute the first of these into the PDE (11.78), we get

$$0 = \frac{1}{r}\frac{d}{dr}\left[r\frac{dR_{00}}{dr}\right] + R_{00}(r)\frac{1}{r^2} \cdot 0 + R_{00}(r), \quad \text{that is,} \quad r^2R_{00}'' + rR_{00}' = 0;$$

hence, $R_{00}(r) = c_1 + c_2 \ln r$.

When we substitute $u_{nm}(r,\theta,z) = R_{nm}(r)\cos n\theta \cos\left(\tfrac{m\pi z}{H}\right)$ into the PDE, we get

$$0 = \cos n\theta \ \cos\left(\frac{m\pi z}{H}\right)\frac{1}{r}\frac{d}{dr}\left[r\frac{dR_{nm}}{dr}\right] - \frac{n^2}{r^2}R_{nm}\cos n\theta \cos\left(\frac{m\pi z}{H}\right)$$

$$- \left(\frac{m\pi}{H}\right)^2 \cos\left(\frac{m\pi z}{H}\right)\cos n\theta R_{nm}.$$

So, for $m = 0, 1, 2, \ldots$, $n = 0, 1, 2, \ldots$, $n + m > 0$, R_{nm} should satisfy

$$\frac{1}{r}\frac{d}{dr}\left[r\,\frac{dR_{nm}}{dr} \right] - \frac{n^2}{r^2}R_{nm} - \left(\frac{m\pi}{H}\right)^2 R_{nm} = 0,$$

that is,

$$r^2 R''_{nm} + rR'_{nm} + \left(-n^2 - \left(\frac{m\pi}{H}\right)^2 r^2 \right) R_{nm} = 0,$$

an ODE which is, for $m = 0$, a Cauchy–Euler equation whose solutions are $R_{n0} = c_1 r^n + c_2 r^{-n}$. For $m \geq 1$, the ODE is a form of the *modified* Bessel's equation of order n, whose solutions are

$$R_{nm} = c_1 I_n\left(\frac{m\pi r}{H}\right) + c_2 K_n\left(\frac{m\pi r}{H}\right).$$

Because of the boundary condition $|\, u(0^+, \theta, z)\,| < \infty$, $-\pi < \theta \leq \pi$, $0 < z < H$, we must leave out the terms involving r^{-n}, $\ln r$ and $K_n\left(\frac{m\pi r}{H}\right)$ because they blow up as $r \to 0^+$.

We can put all product solutions together to get the general solution of (11.78) through (11.81) in the form

$$u(r,\theta,z) = \frac{a_{0,0}}{4} + \sum_{n=1}^{\infty} r^n \left(\frac{a_{0,n}}{2}\cos n\theta + \frac{b_{0,n}}{2}\sin n\theta \right) + \sum_{m=1}^{\infty} \frac{a_{m,0}}{2} I_0\left(\frac{m\pi r}{H}\right)\cos\left(\frac{m\pi z}{H}\right)$$

$$+ \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} I_n\left(\frac{m\pi r}{H}\right)\cos\left(\frac{m\pi z}{H}\right)\left(a_{m,n}\cos n\theta + b_{m,n}\sin n\theta\right).$$

We substitute this general solution into the remaining boundary condition, (11.82):

$$f(\theta, z) = u(a,\theta,z) = \frac{a_{0,0}}{4} + \sum_{n=1}^{\infty} a^n \left(\frac{a_{0,n}}{2}\cos n\theta + \frac{b_{0,n}}{2}\sin n\theta \right)$$

$$+ \sum_{m=1}^{\infty} \frac{a_{m,0}}{2} I_0\left(\frac{m\pi a}{H}\right)\cos\left(\frac{m\pi z}{H}\right)$$

$$+ \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} I_n\left(\frac{m\pi a}{H}\right)\cos\left(\frac{m\pi z}{H}\right)\left(a_{m,n}\cos n\theta + b_{m,n}\sin n\theta\right).$$

The appropriate orthogonality relation involving double integrals is

$$\int_{-\pi}^{\pi}\int_{0}^{H} \cos\left(\frac{m\pi z}{H}\right)\cos\left(\frac{m'\pi z}{H}\right)\cos n\theta\cos n'\theta\, dz\, d\theta$$

$$= \left(\int_{-\pi}^{\pi} \cos n\theta\cos n'\theta\, d\theta\right)\cdot\left(\int_{0}^{H}\cos\left(\frac{m\pi z}{H}\right)\cos\left(\frac{m'\pi z}{H}\right)dz\right)$$

$$= \begin{cases} 0, & \text{if } n \neq n' \text{ or } m \neq m' \\ 2\pi H, & \text{if } n = n' = 0 \text{ and } m = m' = 0 \\ \pi H, & \text{if } n = n' = 0 \text{ and } m = m' \geq 1, \text{ or} \\ & \text{if } m = m' = 0 \quad \text{and} \quad n = n' \geq 1 \\ \frac{\pi H}{2}, & \text{if } n = n' \geq 1 \quad \text{and} \quad m = m' \geq 1 \end{cases}$$

as well as

$$\int_{-\pi}^{\pi}\int_{0}^{H}\cos\left(\frac{m\pi z}{H}\right)\cos\left(\frac{m'\pi z}{H}\right)\sin n\theta\sin n'\theta dz\,d\theta$$

$$=\begin{Bmatrix}0, & n\neq n' \text{ or } m\neq m'\\ \pi H, & m=m'=0 \text{ and } n=n'\geq 1\\ \frac{\pi H}{2}, & m=m'\geq 1 \text{ and } n=n'\geq 1\end{Bmatrix},$$

and for all $m, m', n\geq 0,\ n'\geq 1$,

$$\int_{-\pi}^{\pi}\int_{0}^{H}\cos\left(\frac{m\pi z}{H}\right)\cos\left(\frac{m'\pi z}{H}\right)\cos n\theta\sin n'\theta dz\,d\theta=0.$$

It follows that

$$a_{0,0}=\frac{1}{\pi H}\int_{-\pi}^{\pi}\int_{0}^{H}f(\theta,z)dz\,d\theta,$$

$$a_{0,n}=\frac{1}{\pi Ha^n}\int_{-\pi}^{\pi}\int_{0}^{H}\cos n\theta\, f(\theta,z)dz\,d\theta,$$

$$b_{0,n}=\frac{1}{\pi Ha^n}\int_{-\pi}^{\pi}\int_{0}^{H}\sin n\theta\, f(\theta,z)dz\,d\theta,$$

$$a_{m,0}=\frac{1}{\pi H I_0\left(\frac{m\pi a}{H}\right)}\int_{-\pi}^{\pi}\int_{0}^{H}\cos\left(\frac{m\pi z}{H}\right)f(\theta,z)dz\,d\theta,$$

and for $m=1,2,\ldots;\ n=1,2,\ldots,$

$$a_{m,n}=\frac{1}{\pi H I_n\left(\frac{m\pi a}{H}\right)}\int_{-\pi}^{\pi}\int_{0}^{H}\cos\left(\frac{m\pi z}{H}\right)\cos n\theta\, f(\theta,z)dz\,d\theta,$$

$$b_{m,n}=\frac{1}{\pi H I_n\left(\frac{m\pi a}{H}\right)}\int_{-\pi}^{\pi}\int_{0}^{H}\cos\left(\frac{m\pi z}{H}\right)\sin n\theta\, f(\theta,z)dz\,d\theta.\ \bigcirc$$

As an aside, it is known that the modified Bessel functions $I_n(x)$ are positive for all non-negative integers n and $x>0$, so there is no difficulty in dividing by $I_n\left(\frac{m\pi a}{H}\right)$.

11.6.1 Spherical Coordinates

Recall from Section 6.7 the formulas for differential operators in spherical coordinates.

Example 11.22

Suppose the electric field in the normal direction is given on the surface of a sphere,

$$x^2+y^2+z^2=a^2,\ \text{that is,}\ \rho=a, \tag{11.83}$$

and has no dependence on the angle theta, that is, the potential u satisfies

$$\frac{\partial u}{\partial\rho}(a,\phi,\theta)=f(\phi),\ 0<\phi<\pi,-\pi<\theta\leq\pi. \tag{11.84}$$

Find the electric field inside and outside of the sphere.

Method: We recall that the electric field $\mathbf{E} = \mathbf{E}(r,\theta,z) = \nabla u$, where u is the electric potential. Given the lack of θ-dependence in boundary condition (11.84), when we look for a solution $u(\rho,\phi)$ of Laplace equation, u should satisfy

$$\frac{1}{\rho^2}\frac{\partial}{\partial\rho}\left[\rho^2\frac{\partial u}{\partial\rho}\right]+\frac{1}{\rho^2\sin\phi}\frac{\partial}{\partial\phi}\left[\sin\phi\,\frac{\partial u}{\partial\phi}\right]\equiv 0,\ 0<\phi<\pi,\ \text{either } 0\le\rho<a \text{ or } a<\rho<\infty, \tag{11.85}$$

depending on whether we are solving inside or outside the sphere.

We will see later that the change of variables from rectangular coordinates to spherical coordinates brings artificial mathematical boundary conditions that

$$|\,u(\rho,0^+)\,|\triangleq\lim_{\phi\to 0^+}|u(\rho,\phi)|<\infty \quad\text{and}\quad |u(\rho,\pi^-)|<\infty,$$
$$0\le\rho<a \text{ or } a<\rho<\infty. \tag{11.86}$$

On physical grounds, this makes sense because where there are no sources, the potential and electric field should be finite. In addition, when we solve the problem inside the sphere, we will require

$$|\,u(0^+,\phi)\,|\triangleq\lim_{\rho\to 0^+}|u(\rho,\phi)|<\infty,\ 0<\phi<\pi$$

on physical grounds; when we solve the problem on the outside of the sphere, we will require, instead,

$$\lim_{\rho\to\infty}|u(\rho,\phi)|=0,\ 0<\phi<\pi.$$

Our product solutions will be easier to find if we separate the ρ- and ϕ- dependence as much as possible in (11.85) by multiplying through by ρ^2 to get

$$\frac{\partial}{\partial\rho}\left[\rho^2\frac{\partial u}{\partial\rho}\right]+\frac{1}{\sin\phi}\frac{\partial}{\partial\phi}\left[\sin\phi\,\frac{\partial u}{\partial\phi}\right]\equiv 0. \tag{11.87}$$

Our physical domain in the (ρ,ϕ)-plane is either the rectangle $0\le\rho<a,\ 0<\phi<\pi$ or the half-infinite strip $a<\rho<\infty,\ 0<\phi<\pi$. In either case, the closest thing we have to two separated homogeneous boundary conditions on parallel sides is (11.86). So, our eigenvalue problem is the ODE-BVP

$$\left\{\begin{array}{l}\dfrac{1}{\sin\phi}\dfrac{d}{d\phi}\left[\sin\phi\,\dfrac{d\Phi}{d\phi}\right]+\lambda\Phi(\phi)=0,\\[2ex] |\,\Phi(0^+)\,|<\infty,\ |\Phi(\pi^-)|<\infty,\end{array}\right\}. \tag{11.88}$$

First, we note that

$$\frac{1}{\sin\phi}\frac{d}{d\phi}\left[\sin\phi\,\frac{d\Phi}{d\phi}\right]=\frac{1}{\sin\phi}\left[\cos\phi\,\frac{d\Phi}{d\phi}+\sin\phi\,\frac{d^2\Phi}{d\phi^2}\right]=\frac{\cos\phi}{\sin\phi}\frac{d\Phi}{d\phi}+\frac{d^2\Phi}{d\phi^2}.$$

We will see that our ODE is a form of Legendre's equation after making the change of variables $x=\cos\phi$ and defining a new function Ψ by $\Phi(\phi)=\Psi(\cos\phi)$. We have

$$\frac{d\Phi}{d\phi}=\frac{d}{d\phi}\big[\,\Psi(\cos\phi)\,\big]=-\sin\phi\,\frac{d\Psi}{dx}(\cos\phi),$$

and

$$\frac{d^2\Phi}{d\phi^2}=-\cos\phi\,\frac{d\Psi}{dx}(\cos\phi)+(\sin\phi)^2\,\frac{d^2\Psi}{dx^2}(\cos\phi),$$

so our ODE is

$$0 = \lambda\Psi(\cos\phi) + \frac{\cos\phi}{\sin\phi}\left(-\sin\phi\,\frac{d\Psi}{dx}(\cos\phi)\right) - \cos\phi\,\frac{d\Psi}{dx}(\cos\phi) + (\sin\phi)^2\,\frac{d^2\Psi}{dx^2}(\cos\phi)$$

$$= \lambda\Psi(\cos\phi) - 2\cos\phi\frac{d\Psi}{dx}(\cos\phi) + \sin^2(\phi)\frac{d^2\Psi}{dx^2}(\cos\phi),$$

that is,

$$(1-x^2)\Psi''(x) - 2x\Psi'(x) + \lambda\Psi(x) = 0,$$

which is Legendre's equation, was discussed in Section 9.6, and will be discussed in Appendix C.

In addition to satisfying the ODE, we have boundary conditions for the function $\Psi = \Psi(x)$ corresponding to the BCs for $\Phi = \Phi(\phi)$: As $\phi \to 0^+, x = \cos\phi \to 1^-$, and as $\phi \to \pi^-, x = \cos\phi \to -1^+$. The ODE-BVP eigenvalue problem is

$$\left\{\begin{array}{c}(1-x^2)\Psi''(x) - 2x\Psi'(x) + \lambda\Psi(x) = 0,\\ \\ |\Psi(-1^+)| < \infty,\ |\Psi(1^-)| < \infty.\end{array}\right\}. \tag{11.89}$$

We explained in Appendix C why this ODE-BVP eigenvalue problem has a nontrivial solution for Ψ for eigenvalues $\lambda = n(n+1)$, $n = 0, 1, 2, \ldots$. For nonnegative integers n, the corresponding eigenfunctions, $P_n(x)$, are in fact polynomials in x, called the Legendre polynomials. During the nineteenth century, there was extensive mathematical research on special ODEs that have polynomial solutions. During the twentieth and twenty-first centuries, those special polynomial functions were used in numerical methods for approximating the solutions of very complicated problems.

The Legendre polynomials automatically satisfy the boundary condition requirements that they have all of the finite limits as in (11.89) because they are polynomials. Now, the second-order ODE has two basic solutions, not just one, so one would seem to have, for $\lambda = n(n+1)$, solutions

$$\Psi(x) = c_1 P_n(x) + c_2 Q_n(x).$$

But the second solution, $Q_n(x)$, is not bounded on the interval $(-1, 1)$, so the BCs (11.89) force $c_2 = 0$.

Corresponding to eigenvalues $\lambda = n(n+1)$, we get eigenfunctions $\Phi_n(\phi) = \Psi_n(x) = P_n(x) = P_n(\cos\phi)$. For future reference, note that P_n is a polynomial of degree n; in particular, $P_0(x) \equiv 1$. In addition, for $n = 1, 3, 5, \ldots$, P_n has only odd powers of x, and for $n = 0, 2, 4, \ldots$, P_n has only even powers of x. You'll use these last facts in a problem in this section.

By the way, the first four Legendre polynomials are $P_0(x) \equiv 1$, $P_1(x) = x$, $P_2(x) = \frac{1}{2}(3x^2 - 1)$, and $P_3(x) = \frac{1}{2}(5x^3 - 3x)$.

When we substitute a product solution $u(\rho, \phi) = R_n(\rho)\Phi_n(\phi)$ into the original PDE, we get

$$0 = \frac{1}{\rho^2}\frac{\partial}{\partial\rho}\left[\rho^2\frac{\partial}{\partial\rho}R_n(\rho)\Phi_n(\phi)\right] + \frac{1}{\rho^2\sin\phi}\frac{\partial}{\partial\phi}\left[\sin\phi\,\frac{\partial}{\partial\phi}R_n(\rho)\Phi_n(\phi)\right]$$

$$= \Phi_n(\phi)\left(\frac{1}{\rho^2}\frac{d}{d\rho}\left[\rho^2\frac{d}{d\rho}R_n\right] - n(n+1)R_n\right).$$

Omitting the $\Phi_n(\phi)$ factor and multiplying through by ρ^2 explain why R_n satisfies the Cauchy–Euler ODE $\rho^2 R_n'' + 2\rho R_n' - n(n+1)R_n = 0$. When $n = 0$, the solutions are

$$R_0(\rho) = c_1 + c_2\rho^{-1}.$$

For $n \geq 1$, the solutions are

$$R_n(\rho) = c_1\rho^n + c_2\rho^{-(n+1)}.$$

For a solution inside the sphere, that is, for $0 < \rho < a$, boundedness as $\rho \to 0^+$ implies we must leave out the terms involving $\rho^{-1}, \rho^{-(n+1)}$; for solutions outside the sphere, decay to zero as $\rho \to \infty$ implies we must leave out the terms involving ρ^n.

Put together all the product solutions $R_n(\rho)\Phi_n(\phi)$ and recall that $P_0(x) \equiv 1$, so the general solution for the potential inside the sphere is

$$u(\rho, \phi) = a_0 + \sum_{n=1}^{\infty} a_n \rho^n \Phi_n(\phi)$$

and the general solution for the potential outside the sphere is

$$u(\rho, \phi) = A_0 + \sum_{n=1}^{\infty} A_n \rho^{-(n+1)} \Phi_n(\phi).$$

In order for the solution *inside* the sphere to match the boundary condition (11.84) that the electric field in the normal direction, that is, $\dfrac{\partial u}{\partial \rho}$, is given on the surface of the sphere, take the partial derivative of the solution with respect to ρ. The constants $a_n, n = 0, 1, \ldots$, should be chosen to satisfy

$$f(\phi) = \frac{\partial u}{\partial \rho}(a, \phi) = \sum_{n=1}^{\infty} n a_n a^{n-1} \Phi_n(\phi),\ 0 < \phi < \pi.$$

The appropriate orthogonality relation is

$$\int_0^{\pi} \Phi_n(\phi)\Phi_{n'}(\phi) \sin\phi\, d\phi = \left\{ \begin{array}{ll} 0, & \text{if } n \neq n' \\ 2, & \text{if } n = n' = 0 \\ \displaystyle\int_0^{\pi} (\Phi_n(\phi))^2 \sin\phi\, d\phi, & \text{if } n = n' \geq 1 \end{array} \right\}.$$

So, we calculate that for $n = 1, 2, \ldots$,

$$a_n = \frac{\displaystyle\int_0^{\pi} \Phi_n(\phi) f(\phi) \sin\phi\, d\phi}{n a^{n-1} \displaystyle\int_0^{\pi} (\Phi_n(\phi))^2 \sin\phi\, d\phi},$$

along with the solvability condition that

$$0 = \int_0^{\pi} \Phi_0(\phi) f(\phi) \sin\phi\, d\phi = \int_0^{\pi} f(\phi) \sin\phi\, d\phi.$$

In order for the solution *outside* the sphere to match the boundary condition (11.84) that the electric field in the normal direction, that is, $\dfrac{\partial u}{\partial \rho}$, is given on the surface of the

sphere, take the partial derivative of the solution with respect to ρ. The constants $A_n, n = 0, 1, \ldots$, should be chosen to satisfy

$$f(\phi) = \frac{\partial u}{\partial \rho}(a, \phi) = \sum_{n=1}^{\infty} A_n(-(n+1)a^{-n-2})\Phi_n(\phi),\ 0 < \phi < \pi.$$

So, we calculate that for $n = 1, 2, \ldots$,

$$A_n = -\frac{\int_0^\pi \Phi_n(\phi) f(\phi) \sin\phi\, d\phi}{(n+1)a^{-n-2}\int_0^\pi (\Phi_n(\phi))^2 \sin\phi\, d\phi},$$

again, along with the solvability condition that

$$0 = \int_0^\pi \Phi_0(\phi) f(\phi) \sin\phi\, d\phi = \int_0^\pi f(\phi) \sin\phi\, d\phi.$$

The solution for the electric field in the normal direction inside, on, and outside the sphere is

$$E_\rho(\rho, \phi)$$

$$= \begin{cases} \displaystyle\sum_{n=1}^{\infty} \left(\frac{\int_0^\pi \Phi_n(\phi) f(\phi) \sin\phi\, d\phi}{n a^{n-1} \int_0^\pi (\Phi_n(\phi))^2 \sin\phi\, d\phi} \right) \rho^n \Phi_n(\phi), & 0 < \rho < a,\ 0 < \phi < \pi \\ f(\phi), & \rho = a,\ 0 < \phi < \pi \\ \displaystyle\sum_{n=1}^{\infty} \left(\frac{-\int_0^\pi \Phi_n(\phi) f(\phi) \sin\phi\, d\phi}{(n+1) a^{-n-2} \int_0^\pi (\Phi_n(\phi))^2 \sin\phi\, d\phi} \right) \rho^{-(n+1)} \Phi_n(\phi), & a < \rho < \infty,\ 0 < \phi < \pi \end{cases}. \bigcirc$$

The solvability condition that we discussed in Section 10.3 applies to Example 11.22 because we have a Neumann problem for the Laplace equation: it is specifically in the form $\Delta u = 0$ *in* $\mathcal{D}$, $\dfrac{\partial u}{\partial n} = 0$ *on* $\partial\mathcal{D}$.

11.6.2 Polar Coordinates Again

Now that we saw how to use Bessel functions to solve problems in cylindrical coordinates that involve r, θ, and a third independent variable, we can also solve many problems in polar coordinates that involve r, θ, and t. We will explore this later in some of the problems.

11.6.3 Problems

1. Find the equilibrium temperature distribution in the half cylinder $0 \le \theta \le \pi,\ 0 \le r \le a,\ 0 < z < H$ modeled by

$$\left\{ \begin{array}{c} 0 = \dfrac{1}{r}\dfrac{\partial}{\partial r}\left[r\dfrac{\partial T}{\partial r}\right] + \dfrac{1}{r^2}\dfrac{\partial^2 T}{\partial \theta^2} + \dfrac{\partial^2 T}{\partial z^2},\ 0 < \theta < \pi,\ 0 \le r < a,\ 0 < z < H \\ \\ T(a,\theta,z) = 0,\ 0 < \theta < \pi,\ 0 < z < H \\ \\ T(r,0,z) = T(r,\pi,z) = 0,\ 0 \le r < a,\ 0 < z < H \\ \\ T(r,\theta,0) = f(r,\theta),\ T(r,\theta,H) = 0,\ 0 < \theta < \pi,\ 0 \le r < a \end{array} \right\}.$$

   Assume also any other boundary condition(s) needed in a cylindrical coordinate system based on mathematical and/or physical grounds.

2. Find the equilibrium temperature distribution in the cylinder (11.69) assuming the lateral surface is insulated and the temperature is given on the two circular ends. Assume also that there is heat generation within the cylinder proportional to the temperature, modeled by the PDE

$$0 = \frac{1}{r}\frac{\partial}{\partial r}\left[r\frac{\partial T}{\partial r}\right] + \frac{1}{r^2}\frac{\partial^2 T}{\partial \theta^2} + \frac{\partial^2 T}{\partial z^2} + \eta T,\ -\pi < \theta \le \pi,\ 0 \le r < a,\ 0 < z < H,$$

   where η is an unspecified positive constant of proportionality satisfying $\eta < \frac{\gamma_{1,1}}{a}$, where $\gamma_{m,1}$ are the zeros of the Bessel function $J_1(x)$. Assume also any other boundary condition(s) needed in a cylindrical coordinate system based on mathematical and/or physical grounds. [This problem might be a model for the temperature in a nuclear fuel rod if it is laterally insulated.]

   You may assume that all of the zeros, $j'_{m,n}$, of the functions $J'_n(x)$ satisfy* $j'_{m,n} > \gamma_{1,1}$, for $n = 1, 2, \ldots$, $m = 1, 2, \ldots$.

3. Find the temperature distribution, $T = T(r,\theta,t)$, in an annulus whose boundary circles are kept at 0°C. temperature, modeled by

$$\left\{ \begin{array}{c} \dfrac{\partial T}{\partial t} = \alpha\left(\dfrac{1}{r}\dfrac{\partial}{\partial r}\left[r\dfrac{\partial T}{\partial r}\right] + \dfrac{1}{r^2}\dfrac{\partial^2 T}{\partial \theta^2}\right),\ a < r < b,\ -\pi < \theta \le \pi,\ 0 < t < \infty \\ \\ T(a,\theta,t) = T(b,\theta,t) = 0,\ -\pi < \theta < \pi,\ 0 < t < \infty \\ \\ T(r,\theta,0) = f(r,\theta),\ a < r < b,\ -\pi < \theta < \pi \end{array} \right\}.$$

   Assume also any other boundary condition(s) needed in a polar coordinate system based on mathematical and/or physical grounds. [Hint: Theorem 9.17 in Section 9.6 applies to the regular Sturm–Liouville problem involving the ODE $r(rR')' + \left(\mu r^2 - n^2\right)R(r) = 0$ on the interval $a < r < b$.]

---

* By Rayleigh's argument that $j'_{m,n}$ is increasing in n because of the behavior of frequencies of vibration of circular membranes, as mentioned in Section 15.61 of Watson (1992).

4. Suppose the temperature is given on the lateral side of the cylinder (11.69), that is, $T(a,\theta,z) = f(\theta,z)$, $-\pi < \theta \le \pi, 0 < z < H$, and the circular ends are insulated; hence, $\dfrac{\partial T}{\partial z}(r,\theta,0) = \dfrac{\partial T}{\partial z}(r,\theta,H) = 0$, $0 \le r < a, -\pi < \theta \le \pi$. Assume also that the thermal conductivity in the angular direction is different from that in the other directions, modeled by the PDE

$$0 = \frac{1}{r}\frac{\partial}{\partial r}\left[r\frac{\partial T}{\partial r} \right] + \eta\frac{1}{r^2}\frac{\partial^2 T}{\partial \theta^2} + \frac{\partial^2 T}{\partial z^2}, \quad -\pi < \theta \le \pi,\ 0 \le r < a,\ 0 < z < H,$$

where η is a positive but unspecified parameter. Find the temperature distribution inside of the cylinder.

5. Suppose the steady state temperature $T = T(r,\theta,z)$ satisfies Laplace equation in the half cylinder $0 < \theta < \pi$, $0 < z < H$, $0 \le r < a$. Assume the temperature is 0°C. on the side sheet $r = a$ and the rest of the boundary is insulated except for the top half disk at $z = H$, where the temperature is specified to be $f(r,\theta)$. (a) Write down the PDE and BCs to model this situation and (b) solve for T in terms of f.

6. Find the solution of the problem of vibrations of an annular membrane held fixed at zero displacement on the boundary, modeled by

$$\left\{ \begin{array}{c} \dfrac{\partial^2 u}{\partial t^2} = c^2\left(\dfrac{1}{r}\dfrac{\partial}{\partial r}\left[r\dfrac{\partial u}{\partial r}\right] + \dfrac{1}{r^2}\dfrac{\partial^2 u}{\partial \theta^2}\right),\ a < r < b,\ -\pi < \theta \le \pi,\ 0 < t < \infty \\ \\ u(a,\theta,t) = u(b,\theta,t) = 0,\ -\pi < \theta < \pi,\ 0 < t < \infty \\ \\ \left\{ \begin{array}{c} u(r,\theta,0) = f(r,\theta) \\ \\ \dfrac{\partial u}{\partial t}(r,\theta,0) = g(r,\theta) \end{array} \right\},\ a < r < b,\ -\pi < \theta \le \pi \end{array} \right\}.$$

7. Find the solution of the problem of vibrations of a circular membrane held fixed at zero displacement on the boundary, modeled by

$$\left\{ \begin{array}{c} \dfrac{\partial^2 u}{\partial t^2} = c^2\left(\dfrac{1}{r}\dfrac{\partial}{\partial r}\left[r\dfrac{\partial u}{\partial r}\right] + \dfrac{1}{r^2}\dfrac{\partial^2 u}{\partial \theta^2}\right), 0 \le r < a,\ -\pi < \theta \le \pi,\ 0 < t < \infty \\ \\ u(a,\theta,t) = 0,\ -\pi < \theta < \pi,\ 0 < t < \infty \\ \\ \left\{ \begin{array}{c} u(r,\theta,0) = 0 \\ \\ \dfrac{\partial u}{\partial t}(r,\theta,0) = \left(1 - \frac{r}{a}\right)\cos 2\theta \end{array} \right\},\ 0 \le r < a,\ -\pi < \theta \le \pi \end{array} \right\}.$$

8. Find the equilibrium temperature distribution in the solid hemisphere $0 < \phi < \frac{\pi}{2}$ assuming the flat bottom is insulated and the temperature is prescribed on the curved hemispherical surface and is independent of θ there. Here's how you can do it. Extend the prescribed temperature on the curved hemispherical surface to be given on the *whole sphere* by using an *even extension*: If a function $f(\phi)$ is defined for $0 < \phi < \frac{\pi}{2}$ and satisfies $f'\left(\frac{\pi}{2}\right) = 0$, extend f to be even about $\phi = \frac{\pi}{2}$. In terms of the

change of variable $x = \cos\phi$ that we used to arrive at the Legendre polynomials, this corresponds to an even extension of a function $F(x)$ defined for $0 < x < 1$ that satisfies $F'(0) = 0$ to give an even function $\breve{F}(x)$ defined on $-1 < x < 1$. Earlier in Section 11.6, we noted a fact that implies that the Legendre polynomials $P_n(x)$ are even for $n = 0, 2, 4, \ldots$ and odd for $n = 1, 3, 5, \ldots$, so half of the Fourier–Legendre coefficients should be zero. You may use the result of Problem 11.6.3.9, and some of the calculations in doing Problem 11.6.3.9 may be useful models for calculations in doing Problem 11.6.3.8.

9. A function $f(\phi)$ is *even* about $\phi = \frac{\pi}{2}$ if

$$f\left(\frac{\pi}{2} + \varphi\right) \equiv f\left(\frac{\pi}{2} - \varphi\right), \text{ for } 0 < \varphi < \frac{\pi}{2}.$$

A function $f(\phi)$ is *odd* about $\phi = \frac{\pi}{2}$ if

$$f\left(\frac{\pi}{2} + \varphi\right) \equiv -f\left(\frac{\pi}{2} - \varphi\right), \text{ for } 0 < \varphi < \frac{\pi}{2}.$$

(a) Explain why $\sin(\phi)$ is even about $\phi = \frac{\pi}{2}$.

(b) Suppose $f(\phi)$ is even about $\phi = \frac{\pi}{2}$. Explain why, for all odd n, $a_n = 0$ in the expansion of $f(\phi)$ in Legendre functions $\Phi_n(\phi)$ given by

$$f(\phi) = a_0 + \sum_{n=1}^{\infty} a_n \Phi_n(\phi).$$

10. Solve the inhomogeneous heat problem

$$\left\{ \begin{array}{c} 0 = \nabla^2 u + Q = \dfrac{1}{r}\dfrac{\partial}{\partial r}\left[r\dfrac{\partial u}{\partial r} \right] + \dfrac{1}{r^2}\dfrac{\partial^2 u}{\partial \theta^2} + Q(r,\theta),\ 0 \le r < a,\ -\pi < \theta \le \pi \\ \\ u(a,\theta) = 0,\ -\pi < \theta < \pi \end{array} \right\}.$$

Assume also any other boundary condition(s) needed in a polar coordinate system based on mathematical and/or physical grounds.

11. Solve

$$\left\{ \begin{array}{c} 0 = \dfrac{1}{\rho^2}\dfrac{\partial}{\partial \rho}\left[\rho^2 \dfrac{\partial V}{\partial \rho} \right] + \dfrac{1}{\rho^2 \sin\phi}\dfrac{\partial}{\partial \phi}\left[\sin\phi\, \dfrac{\partial V}{\partial \phi} \right],\ 0 \le \rho < 3,\ 0 < \phi < \pi \\ \\ V(3,\phi) = 5\left(1 - \cos^2\phi\right),\ 0 < \phi < \pi \end{array} \right\},$$

along with any other boundary condition(s) needed in a polar coordinate system based on mathematical and/or physical grounds.

12. Find the solution of the problem

$$\left\{\begin{array}{c} \dfrac{\partial T}{\partial t} = \alpha \dfrac{1}{r}\dfrac{\partial}{\partial r}\left[r\dfrac{\partial T}{\partial r}\right] - \dfrac{4}{r^2}T, 0 \le r < a, 0 < t < \infty \\ \\ \dfrac{\partial T}{\partial r}(a,t) = 5,\ 0 < t < \infty, \\ \\ T(r,0) = f(r),\ 0 < r < a \end{array}\right\},$$

along with any other boundary condition(s) based on mathematical and/or physical grounds.

Key Terms

composite rod: Example 11.5 in Section 11.1
double Fourier series: Example 11.13 in Section 11.4
exterior of disk: Problem 11.5.3.2
Green's function: Problem 11.3.3.13
isotherms: after Example 11.11 in Section 11.3
method of separation of variables: after (11.4) in Section 11.1
periodic boundary conditions: after (11.54) in Section 11.5
rate of decay: before Example 11.2 in Section 11.1
surface waves: after Example 11.15 in Section 11.4
temperature in a slab with a steady source or sink: Example 11.15 in Section 11.4
time constant: before Example 11.2 in Section 11.1
time dependent boundary conditions: Example 11.6 in Section 11.1
transverse vibrations of a rectangular membrane: Example 11.14 in Section 11.4
spherical coordinates: Example. 11.22 in Section 11.6

Mathematica Commands

ContourPlot[T[x, y], {x, 0, 2}, {y, 0, 1}, Contours → {0, 2.5, 5, 7.5, 10, 12.5, 15, 17.5, 20}, ContourStyle → {Thick, Dashed}] : after Example 11.11 in Section 11.3.1
Plot3D[T[x, y], {x, 0, 2}, {y, 0, 1}]: after Example 11.11

References

F.M. Arscott, and A. Darai, A three-dimensional parabolic punch problem in linear elasticity. *Journal of Engineering Mathematics*, 24, 369–383, 1990.

Pinsky, M.A. *Partial Differential Equations and Boundary Value Problems, with Applications*, 3rd edn. McGraw-Hill, New York, 1998.

Watson, G.N. *The Theory of Bessel Functions*. Cambridge University Press, Cambridge, U.K., 1992.

12

Numerical Methods II

12.1 Finite Difference Methods for Heat Equations

For numerical approximation of solutions of ordinary differential equations (ODEs) our methods have included finite differences in Section 8.8 and spline functions in Section 8.9. In Chapter 12, we will see analogous methods for approximating solutions of partial differential equations (PDEs).

For a heat equation

$$\frac{\partial T}{\partial t}(x,t) = \alpha \frac{\partial^2 T}{\partial x^2}(x,t) + g(x,t), \quad 0 < x < L, \quad t > 0, \tag{12.1}$$

we can use finite differences for both the space variable, x, and the time variable, t. Partition the interval $[\,0, L\,]$ using

$$0 = x_0 < x_1 < \cdots < x_N = L.$$

Denoting $h = x_j - x_{j-1}$, we have

$$x_j = jh, \quad j = 0, 1, \ldots, N.$$

We may also write $h = \triangle x$.

Similarly, we can partition the time interval $[0, \infty)$ or $[0, T]$ using

$$t_m = m \triangle t, \quad m = 0, 1, 2, \ldots .$$

The first and simplest example of a finite difference method for (12.1) approximates the time partial derivative using the forward difference, that is,

$$\frac{\partial T}{\partial t}(x_j, t_m) \approx \frac{T(x_j, t_{m+1}) - T(x_j, t_m)}{\triangle t},$$

and approximates the second space derivative using the central difference, that is,

$$\frac{\partial^2 T}{\partial x^2}(x_j, t_m) \approx \frac{T(x_{j+1}, t_m) - 2T(x_j, t_m) + T(x_{j-1}, t_m)}{h^2}.$$

Using the forward difference to approximate the time partial derivative makes this method analogous to Euler's method for ODEs.

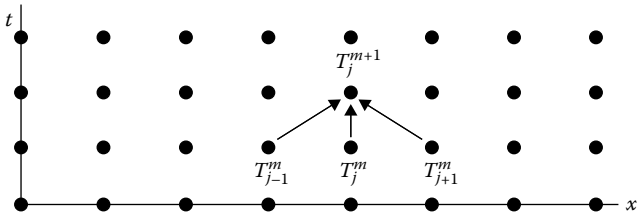


FIGURE 12.1
Finite difference stencil for heat equation.

Denote

$$T_j^m = T(x_j, t_m) \quad \text{and} \quad g_j^m = g(x_j, t_m).$$

It is customary to use the superscript m rather than (m). Keep in mind that T_j^m is not the mth power of T_j; in fact, T_j without a superscript should never appear in our work.

The **stencil** or **replacement equations** for (12.1) are

$$\frac{T_j^{m+1} - T_j^m}{\triangle t} = \alpha \frac{T_{j+1}^m - 2T_j^m + T_{j-1}^m}{h^2} + g_j^m, \quad m = 0, 1, 2, \ldots, \quad j = 0, 1, \ldots, N. \tag{12.2}$$

Equivalently,

$$T_j^{m+1} = T_j^m + \epsilon \cdot \left(T_{j+1}^m - 2T_j^m + T_{j-1}^m\right) + \triangle t \cdot g_j^m, \quad m = 0, 1, 2, \ldots, \quad j = 0, 1, \ldots, N, \tag{12.3}$$

where

$$\epsilon \triangleq \frac{\alpha \triangle t}{h^2} = \frac{\alpha \triangle t}{(\triangle x)^2}.$$

Figure 12.1 illustrates that the approximation of the heat equation in (12.2) uses the temperature and source values at time $t = t_m$ to *explicitly* find the temperature values at time $t = t_{m+1}$.

Example 12.1

For the PDE–boundary value problem–initial value problem (PDE-BVP-IVP)

$$\left\{ \begin{array}{l} \frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2}, \quad 0 < x < \pi, \quad t > 0, \\ T(0, t) = T(\pi, t) = 0, \quad t > 0, \\ T(x, 0) = 4(\pi x - x^2), \quad 0 < x < \pi \end{array} \right\},$$

use the preceding finite difference method to find an approximate solution.

Method: The boundary conditions (BCs) on $T(x,t)$ directly imply BCs on $T_j^m \triangleq T(jh, m\triangle t)$:

$$T_0^m = T_N^m = 0, \quad m = 1,2,\dots . \tag{12.4}$$

So, we need only to approximate the temperature values at $x = x_j = \frac{j\pi}{N}$, for $j = 1,\dots,N-1$. Define

$$\mathbf{T}^m = \begin{bmatrix} T_1^m \\ \vdots \\ T_{N-1}^m \end{bmatrix}.$$

The initial condition for the temperature directly implies

$$T_j^0 = 4\left(\pi x_j - x_j^2\right) = 4\left(\pi\,\frac{j\pi}{N} - \left(\frac{j\pi}{N}\right)^2\right) = \frac{4\pi^2 j(N-j)}{N^2}, \quad j = 0,1,\dots,N.$$

For this PDE with $g(x,t) \equiv 0$ and BCs $T(0,t) = T(\pi,t) = 0$, the replacement equations (12.3) are

$$T_j^{m+1} = \epsilon T_{j+1}^m + (1-2\epsilon)T_j^m + \epsilon T_{j-1}^m, \quad j = 1,\dots,N-1, \quad m = 0,1,2,\dots . \tag{12.5}$$

As earlier, $\epsilon = \frac{\alpha\triangle t}{(\triangle x)^2}$. In terms of the vector $\mathbf{T}^m$, we see that the replacement equations are the system of linear, homogeneous partial difference equations

$$\mathbf{T}^{m+1} = A\mathbf{T}^m, \quad m = 0,1,\dots , \tag{12.6}$$

where the constant, $(N-1)\times(N-1)$, **discretization matrix** A is given by

$$A(\epsilon) = \begin{bmatrix} 1-2\epsilon & \epsilon & 0 & 0 & & & . & . & . & 0 \\ \epsilon & 1-2\epsilon & \epsilon & 0 & & & & & & 0 \\ 0 & \epsilon & 1-2\epsilon & \epsilon & & & & & & . \\ . & & & . & . & . & & & & . \\ . & & & & . & . & . & & & . \\ . & & & & & . & . & . & & . \\ & & & & & & . & 1-2\epsilon & \epsilon & 0 \\ & & & & & & & \epsilon & 1-2\epsilon & \epsilon \\ 0 & & & & . & . & . & 0 & \epsilon & 1-2\epsilon \end{bmatrix}.$$

Because the matrix $A = A(\epsilon)$ depends on the parameter $\epsilon \triangleq \alpha\frac{\triangle t}{(\triangle x)^2}$, in principle the approximation we get could depend on ϵ. As we will see, "... could depend..." turns out to be "...significantly depends...!"

In effect, the BCs $T_0^m = T_N^m = 0$, $m = 1,2,\dots$, are "built into" the relationship between the partial difference equations (12.6) and the matrix $A(\epsilon)$:

T_j^m satisfies both the system of partial difference equations (12.6) and the BCs (12.4) $\Longleftrightarrow$ T_j^m satisfies both (12.4) and (12.5).

The solution of (12.6) is given by

$$\mathbf{T}^m = A^m\,\mathbf{T}^0, \quad m = 0,1,\dots , \tag{12.7}$$

where A^m *does* denote the mth power of the matrix A. ◯

Example 12.2

For the PDE-BVP-IVP of Example 12.1 with $\alpha = 1$, find the numerical, that is, approximate, solution for grid size $h = \triangle x = \frac{\pi}{6}$ and two values of ϵ of your choice.

Method: Here $N = 6$, so the solution vector is

$$\mathbf{T}^m = \begin{bmatrix} T_1^m \\ \vdots \\ T_5^m \end{bmatrix}.$$

(a) For $\triangle t = \frac{1}{8}$, that is, for $\epsilon = \frac{1/8}{(\pi/6)^2} = \frac{9}{2\pi^2}$, the 5×5 matrix $A(\epsilon)$ is

$$A\left(\frac{9}{2\pi^2}\right) = \begin{bmatrix} 1-\frac{9}{\pi^2} & \frac{9}{2\pi^2} & 0 & 0 & 0 \\ \frac{9}{2\pi^2} & 1-\frac{9}{\pi^2} & \frac{9}{2\pi^2} & 0 & 0 \\ 0 & \frac{9}{2\pi^2} & 1-\frac{9}{\pi^2} & \frac{9}{2\pi^2} & 0 \\ 0 & 0 & \frac{9}{2\pi^2} & 1-\frac{9}{\pi^2} & \frac{9}{2\pi^2} \\ 0 & 0 & 0 & \frac{9}{2\pi^2} & 1-\frac{9}{\pi^2} \end{bmatrix}.$$

Mathematica™ says that the eigenvalues of A are $\lambda \approx 0.87783, -0.701611, 0.544055, -0.367836$, and 0.0881093, all of which have magnitude less than 1. It follows that, for all initial temperature distributions and thus any initial vector $\mathbf{T}^0$, the approximate solution has

$$\mathbf{T}^m \to \mathbf{0}, \quad \text{as } m \to \infty.$$

This agrees with the exact solution, $T(x,t)$, using separation of variables, which has

$$T(x,t) \to 0, \quad \text{as } t \to \infty.$$

In fact, the decay rate for the exact solution is $r = 1$ because the product solutions are of the form $\sin nx \cdot e^{-n^2 t}$. The maximum value $T(x,1)$ behaves like e^{-1} times $T(x,0)$. For the finite difference solution with $\triangle t = \frac{1}{8}$, after eight time steps, we have

$$\begin{bmatrix} T\left(\frac{\pi}{6},1\right) \\ \vdots \\ T\left(\frac{5\pi}{6},1\right) \end{bmatrix} \approx \begin{bmatrix} T_1^8 \\ \vdots \\ T_5^8 \end{bmatrix} = \mathbf{T}^8 = A^8 \mathbf{T}^0.$$

So, the maximum magnitude eigenvalue λ_1 of the matrix $A\left(\frac{9}{2\pi^2}\right)$ should have

$$|\lambda_1|^8 \approx e^{-1}.$$

In fact,

$$|\lambda_1|^8 = 0.87783^8 \approx 0.3526008467 \quad \text{versus} \quad e^{-1} \approx 0.3678794412.$$

So, even this coarse finite difference method, with $\triangle x = \frac{\pi}{6}$ and $\triangle t = \frac{1}{8}$, gives a good approximation to the rate of decay of solutions.

Table 12.1 has the numerical results for the approximate solution, which actually was prepared using an Excel spreadsheet using (12.3).

TABLE 12.1

Example 12.2(a): Finite Differences for a Heat Equation with $\triangle t = \frac{1}{8}$

i	t_i	T_1^m	T_2^m	T_3^m	T_4^m	T_5^m
0	0	5.483113556	8.77298169	9.869604401	8.77298169	5.483113556
1	0.125	4.483113556	7.77298169	8.869604401	7.77298169	4.483113556
2	0.25	3.939058883	6.77298169	7.869604401	6.77298169	3.939058883
3	0.375	3.435177254	5.980867831	6.869604401	5.980867831	3.435177254
4	0.5	3.029619961	5.225387395	6.05917383	5.225387395	3.029619961
5	0.625	2.649428799	4.604398524	5.298851774	4.604398524	2.649428799
6	0.75	2.33279343	4.029671928	4.665586346	4.029671928	2.33279343
7	0.875	2.042850989	3.545930314	4.085701932	3.545930314	2.042850989
8	1	1.796744622	3.106714667	3.593489239	3.106714667	1.796744622
...	...	...	...	...	...	...
16	2	0.633079741	1.096212821	1.266159481	1.096212821	0.633079741
...	...	...	...	...	...	...
24	3	0.22319749	0.386570987	0.44639498	0.386570987	0.22319749
...	...	...	...	...	...	...
32	4	0.078697932	0.136307736	0.157395864	0.136307736	0.078697932

(b) For $\triangle t = \frac{1}{4}$, that is, for $\epsilon = \frac{1/4}{(\pi/6)^2} = \frac{9}{\pi^2}$, the 5×5 matrix $A(\epsilon)$ is

$$A\left(\frac{9}{\pi^2}\right) = \begin{bmatrix} 1-\frac{18}{\pi^2} & \frac{9}{\pi^2} & 0 & 0 & 0 \\ \frac{9}{\pi^2} & 1-\frac{18}{\pi^2} & \frac{9}{\pi^2} & 0 & 0 \\ 0 & \frac{9}{\pi^2} & 1-\frac{18}{\pi^2} & \frac{9}{\pi^2} & 0 \\ 0 & 0 & \frac{9}{\pi^2} & 1-\frac{18}{\pi^2} & \frac{9}{\pi^2} \\ 0 & 0 & 0 & \frac{9}{\pi^2} & 1-\frac{18}{\pi^2} \end{bmatrix}.$$

Mathematica says that the eigenvalues of A are $\lambda \approx -2.40322, -1.73567, -0.823781$, 0.75566, and 0.0881093, at least one of which has magnitude greater than one. It follows that, at least for some initial temperature distributions, the approximate solution vector $\mathbf{T}^m$ has

$$||\mathbf{T}^m|| \to \infty, \quad \text{as } m \to \infty.$$

This is very bad! In fact, because the largest magnitude eigenvalue is less than -1, for some initial conditions, the solution produced by finite differences oscillates between larger and larger positive and negative values in at least one of the entries T_j^m. This is true even though the exact solution has $T(x,t) \to 0$, as $t \to \infty$.

Table 12.2 has the numerical results for the approximate solution, which again was prepared using an Excel spreadsheet using (12.3). ◯

In Problem 12.1.5.5, you will address the natural question raised by the aforementioned "... for some initial conditions ...."

TABLE 12.2

Example 12.2(b): Finite Differences for a Heat Equation with $\triangle t = \frac{1}{4}$

i	t_i	T_1^m	T_2^m	T_3^m	T_4^m	T_5^m
0	0	5.483113556	8.77298169	9.869604401	8.77298169	5.483113556
1	0.25	3.483113556	6.77298169	7.869604401	6.77298169	3.483113556
2	0.5	3.306894862	4.77298169	5.869604401	4.77298169	3.306894862
3	0.75	1.628279222	4.436070815	3.869604401	4.436070815	1.628279222
4	1	2.703865528	1.359116479	4.902715257	1.359116479	2.703865528
5	1.25	−0.988028262	5.81675517	−1.56003395	5.81675517	−0.988028262
6	1.5	6.118163881	−7.115288281	11.89361614	−7.115288281	6.118163881
7	1.75	−11.5283939	22.28621531	−22.77446838	22.28621531	−11.5283939
8	2	29.81946681	−49.63942703	59.40636416	−49.63942703	29.81946681
...	...	...	...	...	...	...
16	4	32489.63605	−56273.47252	64979.22277	−56273.47252	32489.63605
...	...	...	...	...	...	...
24	6	36148921.36	−62611768.41	72297842.71	−62611768.41	36148921.36
...	...	...	...	...	...	...
32	8	40220441029	−69663847365	80440882058	−69663847365	40220441029

12.1.1 Incompatibility of Initial Condition with Boundary Conditions

In Examples 12.1 and 12.2, the initial temperature distribution $T(x,0) = 4(\pi x - x^2)$, $0 < x < \pi$, just happened to also satisfy the BCs $T(0,t) = T(\pi,t) = 0$, at $t = 0$, in the limits as $x \to 0^+$ and as $x \to \pi^-$. We do not demand that this always be true. If the initial temperature distribution $T(x,0)$ happens to not satisfy one or more of the BCs at $t = 0$, then the replacement equations would have to be broken into two cases: $m = 0$ versus $m \geq 1$.

Example 12.3

With $N = 3$, for the PDE-BVP-IVP

$$\left\{ \begin{array}{l} \frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2},\ 0 < x < \pi,\ t > 0, \\ \\ T(0,t) = T(\pi,t) = 0,\ t > 0, \\ \\ T(x,0) = f(x),\ 0 < x < \pi \end{array} \right\},$$

set up, but do not solve, the replacement equations.

Method: With $N = 3$, hence $h = \frac{\pi}{3}$, and $\epsilon = \frac{\alpha \triangle t}{(\pi/3)^2}$, the replacement equations for the PDE-BVP-IVP are (12.5), that is,

$$T_j^{m+1} = \epsilon T_{j+1}^m + (1 - 2\epsilon) T_j^m + \epsilon T_{j-1}^m, \quad j = 1, \ldots, N-1,$$

for $m \geq 1$ but are

$$T_j^1 = \epsilon f(x_{j+1}) + (1 - 2\epsilon) f(x_j) + \epsilon f(x_{j-1}), \quad j = 1, \ldots, N-1,$$

for $m = 0$. (If, for example, $f(x)$ is undefined at $x = 0$, then for $j = 1$, replace $f(x_0)$ by $\lim_{x \to 0^+} f(x)$.) ○

12.1.2 Time-Dependent Boundary Conditions

Instead of the homogeneous Dirichlet BCs $T(0,t) = T(L,t) = 0$, the temperature might satisfy a nonhomogeneous Dirichlet BC $T(0,t) = \phi(t)$ and/or $T(L,t) = \psi(t)$. This is an extension of our previous work by replacing

$$T_0^m = \phi(t_m) \quad \text{and/or} \quad T_N^m = \psi(t_m), \text{ respectively.}$$

The matrix $A(\epsilon)$ is not affected, but (12.6) is replaced by the nonhomogeneous system of difference equations

$$\begin{bmatrix} T_1^{m+1} \\ T_2^{m+1} \\ \vdots \\ T_{N-2}^{m+1} \\ T_{N-1}^{m+1} \end{bmatrix} = A(\epsilon) \begin{bmatrix} T_1^m \\ T_2^m \\ \vdots \\ T_{N-2}^{m+1} \\ T_{N-1}^m \end{bmatrix} + \epsilon \begin{bmatrix} \phi(t_m) \\ 0 \\ \vdots \\ 0 \\ \psi(t_m) \end{bmatrix}.$$

12.1.3 Other Boundary Conditions

What is less straightforward is what to do with Neumann or Robin BCs. In Problem 8.6.1.8 in Section 8.6, we found that the central difference approximation

$$\frac{\partial T}{\partial x}(x_j, t_m) \approx \frac{T_{j+1}^m - T_{j-1}^m}{2h} \tag{12.8}$$

has accuracy $\mathcal{O}(h^2)$. So, for example, we can replace the Neumann BC

$$\frac{\partial T}{\partial x}(0, t_m)\phi(t_m) \triangleq \phi^m$$

by

$$\frac{T_1^m - T_{-1}^m}{2h} = \phi^m. \tag{12.9}$$

But then it appears that we need to know $T_{-1}^m \approx T(-h, t_m)$. To deal with this, we also need to use the replacement equation for the PDE at $x = 0$, that is,

$$T_0^{m+1} = T_0^m + \epsilon\left(T_1^m - 2T_0^m + T_{-1}^m\right) + \triangle t \cdot g_0^m. \tag{12.10}$$

Note that when we had Dirichlet BCs we never used the replacement equation for the PDE at $x = 0$ or $x = L$, even though (12.2) did allow for $j = 0$ and $j = N$, respectively.

We can combine (12.9) and (12.10) by solving (12.9) for T_{-1}^m to get

$$T_{-1}^m = T_1^m - 2h\,\phi^m.$$

Substitute this into (12.10) to get

$$T_0^{m+1} = T_0^m + \epsilon\left(T_1^m - 2T_0^m + T_1^m - 2h\,\phi^m\right) + \triangle\, t \cdot g_0^m,$$

that is,

$$T_0^{m+1} = (1-2\epsilon)T_0^m + 2\epsilon T_1^m - 2h\,\epsilon\,\phi^m + \triangle\, t \cdot g_0^m. \tag{12.11}$$

We include (12.11) in the system of difference equations.

Example 12.4

For the PDE-BVP-IVP with $N = 3$

$$\left\{\begin{array}{c} \frac{\partial T}{\partial t} = \alpha\frac{\partial^2 T}{\partial x^2}, \quad 0 < x < \pi, \quad t > 0, \\ \\ \frac{\partial T}{\partial x}(0,t) = \phi(t), \quad T(\pi,t) = 0, \quad t > 0, \\ \\ T(x,0) = f(x), \quad 0 < x < \pi \end{array}\right\},$$

set up, but do not solve, the replacement equations in vector form.

Method: With $N = 3$, hence $h = \frac{\pi}{3}$, and $\epsilon = \frac{\alpha\triangle t}{(\pi/3)^2}$, the PDE-BVP-IVP are discretized by

$$\mathbf{T}^{m+1} = \begin{bmatrix} 1-2\epsilon & 2\epsilon & 0 \\ \epsilon & 1-2\epsilon & \epsilon \\ 0 & \epsilon & 1-2\epsilon \end{bmatrix}\mathbf{T}^m - 2h\epsilon\begin{bmatrix} \phi^m \\ 0 \\ 0 \end{bmatrix},$$

where $\mathbf{T}^m \triangleq \begin{bmatrix} T_0^m \\ T_1^m \\ T_2^m \end{bmatrix}$ and the initial condition is $\mathbf{T}^0 \triangleq \begin{bmatrix} f(x_0) \\ f(x_1) \\ f(x_2) \end{bmatrix}$. ◯

12.1.4 Nonlinearity

As for ODEs, numerical methods for PDEs also handle nonlinear PDEs. For example, for a PDE of the form

$$\frac{\partial T}{\partial t} = \alpha\frac{\partial^2 T}{\partial x^2} + q\left(x, t, T, \frac{\partial T}{\partial x}\right),$$

we can define $g(x,t) \triangleq q\big(x, t, T(x,t), \frac{\partial T}{\partial x}(x,t)\big)$, which could have replacement equations

$$T_j^{m+1} = \epsilon T_{j+1}^m + (1-2\epsilon)T_j^m + \epsilon T_{j-1}^m + q\left(x_j, t_m, T_j^m, \frac{T_{j+1}^m - T_{j-1}^m}{2h}\right), \quad j = 1, \ldots, N-1,$$

$$m = 0, 1, 2, \ldots.$$

Handling a nonlinear BC is a trickier matter. In principle it would require solving a nonlinear equation at each time step. For example, a BC of the form

$$T(0,t) - \eta\left(\frac{\partial T}{\partial x}(x_j,t)\right)^3 = \phi(t)$$

would require solving

$$T_0^m - \eta\left(\frac{T_1^m - T_{-1}^m}{2h}\right)^3 - \phi^m = 0$$

for T_{-1}^m in terms of T_0^m, T_1^m, and $\phi^m \triangleq \phi(t_m)$, in order to substitute for T_{-1}^m in the replacement equation at $j = 0, m-1, 2, \ldots$.

Example 12.5

For the nonlinear PDE-BVP-IVP that follows, find the numerical, that is, approximate, solution for grid size $h = \triangle x = \frac{\pi}{6}$ and $\triangle t = \frac{1}{8}$:

$$\left\{\begin{array}{c} \frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2} + 1.5T(1-T), \quad 0 < x < \pi, \quad t > 0, \\ \\ T(0,t) = T(\pi,t) = 0, \quad t > 0, \\ \\ T(x,0) = \pi x - x^2, \quad 0 < x < \pi \end{array}\right\}.$$

Method: Here $\epsilon = \frac{1/8}{(\pi/6)^2} = \frac{9}{2\pi^2}$ and $N = 6$, so the solution vector is

$$\mathbf{T}^m = \begin{bmatrix} T_1^m \\ \vdots \\ T_5^m \end{bmatrix}.$$

The replacement equations are

$$T_j^{m+1} = T_j^m + \epsilon \cdot \left(T_{j+1}^m - 2T_j^m + T_{j-1}^m\right) + \triangle t \cdot T_j^m(1 - T_j^m), \quad m = 0, 1, 2, \ldots,$$
$$j = 1, 2, \ldots, N-1.$$

Table 12.3 has the numerical results for the approximate solution, which again was prepared using an Excel spreadsheet.

It appears that the approximate solution agrees with $T(x,t) \to v(x)$ as $t \to \infty$, where the equilibrium temperature distribution $v(x)$ satisfies the nonlinear ODE-BVP

$$\left\{\begin{array}{c} 0 = v'' + 1.5v(1-v), \quad 0 < x < \pi, \\ \\ v(0) = v(\pi) = 0 \end{array}\right\}. \bigcirc$$

TABLE 12.3

Example 12.5: Finite Differences for a Nonlinear Heat Equation with $\triangle t = \frac{1}{8}$

i	t_i	T_1^m	T_2^m	T_3^m	T_4^m	T_5^m
0	0	1.370778389	2.193245422	2.4674011	2.193245422	1.370778389
1	0.125	1.025480576	1.452542911	1.538526021	1.452542911	1.025480576
2	0.25	0.747735232	1.173778558	1.304768271	1.173778558	0.747735232
3	0.375	0.636428921	1.001004452	1.110760232	1.001004452	0.636428921
4	0.5	0.555863729	0.884632053	0.987607201	0.884632053	0.555863729
...	...	...	...	...	...	...
8	1	0.395780439	0.645035453	0.726666641	0.645035453	0.395780439
...	...	...	...	...	...	...
16	2	0.289638291	0.479833303	0.543923601	0.479833303	0.289638291
...	...	...	...	...	...	...
128	16	0.212938681	0.356952076	0.406565109	0.356952076	0.212938681
...	...	...	...	...	...	...
296	37	0.212908658	0.356903395	0.406510433	0.356903395	0.212908658
...	...	...	...	...	...	...
320	40	0.212908658	0.356903395	0.406510432	0.356903395	0.212908658

12.1.5 Problems

In problems 1–4, set up the replacement equations for the PDE-BVP-IVP using $N = 3$ and $\epsilon = \frac{9}{2\pi^2}$:

(a) Find the eigenvalues of the matrix $A(\epsilon)$.

(b) In the special case $\phi(t) \equiv 0$ and $\psi(t) \equiv 0$, discuss whether the decay rate of the numerical solutions should agree well with the decay rate for the exact solution.

1. $$\left\{ \begin{array}{c} \frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2}, \quad 0 < x < \pi, \quad t > 0, \\ T(0,t) = 0, \quad \frac{\partial T}{\partial x}(\pi, t) = \psi(t), \quad t > 0, \\ T(x,0) = f(x), \quad 0 < x < \pi \end{array} \right\}$$

2. $$\left\{ \begin{array}{c} \frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2}, \quad 0 < x < \pi, \quad t > 0, \\ T(0,t) = \phi(t), \quad \frac{\partial T}{\partial x}(\pi, t) = \psi(t), \quad t > 0, \\ T(x,0) = f(x), \quad 0 < x < \pi \end{array} \right\}$$

3. $$\left\{ \begin{array}{c} \frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2}, \quad 0 < x < \pi, \quad t > 0, \\ \frac{\partial T}{\partial x}(0,t) = \phi(t), \quad \frac{\partial T}{\partial x}(\pi, t) = \psi(t), \quad t > 0, \\ T(x,0) = f(x), \quad 0 < x < \pi \end{array} \right\}$$

4. $$\left\{\begin{array}{l} \frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2}, \quad 0 < x < \pi, \quad t > 0, \\ \\ 3T(0,t) - \frac{\partial T}{\partial x}(0,t) = \phi(t), \\ \\ \frac{\partial T}{\partial x}(\pi,t) = \psi(t), \quad t > 0, \\ \\ T(x,0) = f(x),\ 0 < x < \pi \end{array}\right\}$$

5. In Example 12.2(b), we saw that for some initial temperature distributions, the approximate solution has $||\mathbf{T}^m|| \to \infty$, as $m \to \infty$. For the PDE and BCs of Example 12.2(b), are there some nonzero initial temperature distributions, $\mathbf{T}^0$, for which $||\mathbf{T}^m|| \to 0$, as $m \to \infty$? If so, find one such nonzero $\mathbf{T}^0$.
6. For the following nonlinear PDE-BVP-IVP, find the numerical, that is, approximate, solution for grid size $h = \triangle x = \frac{\pi}{6}$ and $\triangle t = \frac{1}{8}$:

$$\left\{\begin{array}{l} \frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2} + 2T(1-T), \quad 0 < x < \pi, \quad t > 0, \\ \\ T(0,t) = T(\pi,t) = 0, \quad t > 0, \\ \\ T(x,0) = \pi x - x^2, \quad 0 < x < \pi \end{array}\right\}.$$

12.2 Numerical Stability

As we remarked in Chapter 8, astute users of numerical methods noticed that if the time step size, $\triangle t$, is too large, then a numerical solution might behave qualitatively in a significantly incorrect way(s).

For example, we saw in Example 12.2 in Section 12.1 that while the exact solution of a heat equation should have $T(x,t) \to 0$, as $t \to \infty$, an approximate solution using finite differences can oscillate between larger and larger positive and negative values as m, the number of time steps, increases.

There are three general ways to discuss the accuracy of an approximation solution *method* that depends on a time step size, $\triangle t$, and/or a space grid size, h. A *method* is **consistent** if a limiting process results in the terms in the difference equation converging to the appropriate terms in the original PDE. It is usually easy to design a numerical method to be consistent. For example, in Section 8.6, we saw how to get finite differences that were consistent approximations of derivatives, and then in Section 8.7, we replaced terms in ODEs by those finite differences to get the "replacement equations," that is, a numerical method.

A numerical method is **convergent** if all numerical solutions converge to exact solutions by a limiting process. For example, Theorem 8.6(c) in Section 8.3 implies that the numerical integration method of Simpson's rule is convergent because $S_n \to I$ as $n \to \infty$, that is, as $h = \frac{b-a}{n} \to 0$.

A convergent *method* is successful in the sense that theoretically it converges to the exact solution for a broad set of conditions. For example, Simpson's rule is a convergent method for *all* functions $f(x)$ that are four times continuously differentiable on $[a, b]$.

We italicize the word *method* to emphasize that its success is *not* "hit or miss," that is, its success is predictably related to the conditions under which it is asked to work.

In Example 12.2(b) in Section 12.1, we saw that a numerical method could be consistent yet fail to be convergent.

There is a third way to discuss the accuracy of an approximate solution *method*: a method is **stable** if it does not *catastrophically* magnify errors as the number of time steps increases.

Consider the heat equation in Section 12.1. We used the approximation

$$T(x_j, t_m) \approx T_j^m.$$

A stable method is allowed to have a certain amount of magnification of errors.

The solution of the heat problem

$$\left\{ \begin{array}{l} \frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2}, \quad 0 < x < \pi, \quad t > 0, \\ \\ T(0,t) = T(\pi,t) = 0, \quad t > 0, \\ \\ T(x,0) = f(x), \quad 0 < x < \pi \end{array} \right\} \tag{12.12}$$

depends linearly on the initial temperature distribution $f(x)$: Suppose $T_1(x,t)$ solves (12.12) when the initial condition is $T(x,t) = f_1(x)$ and $T_2(x,t)$ solves (12.12) when the initial condition is $T(x,t) = f_2(x)$. Then, for any constants β_1, β_2, when the initial condition is $T(x,t) = \beta_1 f_2(x) + \beta_2 f_2(x)$, then the solution of (12.12) is

$$T(x,t) = \beta_1 T_1(x,t) + \beta_2 T_2(x,t).$$

Likewise, the approximate solution given by the finite difference method, that is,

$$\mathbf{T}^m = A^m \mathbf{T}^0, \tag{12.13}$$

depends linearly on the initial vector $\mathbf{T}^0$.

Suppose there is an error in the initial condition, that is, we use an initial vector

$$\mathbf{T}^0 + (\textbf{initial error})$$

rather than $\mathbf{T}^0$. The approximate solution depends linearly on the initial vector, so the initial error leads to an error

$$A^m(\textbf{initial error})$$

in the approximate solution given by (12.13).

We see that the method of approximating the solution of (12.12) by (12.13) is numerically unstable exactly when the discretization matrix $A = A(\epsilon)$ has an eigenvalue whose magnitude is greater than 1 or a deficient eigenvalue whose magnitude is 1. This happens exactly when the linear system of difference equations (12.6) in Section 12.1 is unstable.

So, let us study the eigenvalues of the matrix $A(\epsilon)$. We could find a formula for the characteristic polynomial $0 = |A(\epsilon) - \lambda I|$ using the technique of Example 4.38 in Section 4.6.

Instead, we will work in a way similar to what we did in Chapter 11: substitute a product solution,

$$T_j^m = Q^m \sin \frac{nj\pi}{N}, \quad j = 1, \ldots, N-1, \tag{12.14}$$

into the partial difference equation (12.5) in Section 12.1 and then divide by Q^m to get

$$Q \sin \frac{nj\pi}{N} = \sin \frac{nj\pi}{N} + \epsilon \left(\sin \frac{n(j+1)\pi}{N} - 2 \sin \frac{nj\pi}{N} + \sin \frac{n(j-1)\pi}{N} \right), \quad j = 1, \ldots, N-1. \tag{12.15}$$

We use a sum of sines formula to rewrite

$$\begin{aligned} &\sin \frac{n(j+1)\pi}{N} + \sin \frac{n(j-1)\pi}{N} \\ &\quad = 2 \sin \left(\frac{1}{2} \left(\frac{n(j+1)\pi}{N} + \frac{n(j-1)\pi}{N} \right) \right) \cos \left(\frac{1}{2} \left(\frac{n(j+1)\pi}{N} - \frac{n(j-1)\pi}{N} \right) \right) \\ &\quad = 2 \sin \frac{nj\pi}{N} \cos \frac{n\pi}{N}. \end{aligned}$$

After dividing through by $\sin \frac{nj\pi}{N}$, we have the **amplification factor**

$$Q_n = 1 + 2\epsilon \left(-1 + \cos \frac{n\pi}{N} \right), \tag{12.16}$$

where n can be any integer other than an integer multiple of N.

Because n is an integer, $T_j^m = Q^m \sin \frac{nj\pi}{N}$, $j = 1, \ldots, N-1$, satisfies the BCs (12.4) in Section 12.1, that is, $T_0^m = T_N^m = 0$, $m = 1, 2, \ldots$. This guarantees that the vector $[\sin \frac{nj\pi}{N}]_{j=1}^{N-1}$ is an eigenvector of the matrix $A(\epsilon)$ if, and only if, T_j^m satisfies the partial difference equation (12.5) in Section 12.1.

Note that the logical equivalence between Q satisfying (12.15) and Q satisfying (12.16) does not allow n to be an integer multiple of N.

Theorem 12.1

(von Neumann–Goldstine criterion)

(a) If $\frac{1}{2} \geq \epsilon \triangleq \alpha \frac{\Delta t}{(\Delta x)^2} > 0$, then the numerical method of linear partial difference equations (12.5) in Section 12.1 and BCs (12.4) in Section 12.1 is stable.

(b) If $\frac{1}{2} < \epsilon$, then the numerical method of linear partial difference equations (12.5) in Section 12.1 and BCs (12.4) in Section 12.1 is unstable if N is sufficiently large.

(c) If $\epsilon < 0$, that is, we are going backward in time, then the numerical method of linear partial difference equations (12.5) in Section 12.1 and BCs (12.4) in Section 12.1 is unstable.

Why? First, for each of $n = 1$, $n = 2, \ldots$, $n = N - 1$, the vector $\mathbf{T}^{(n)} \triangleq [\sin \frac{nj\pi}{N}]_{j=1}^{N-1}$ is an eigenvector of the matrix $A(\epsilon)$. This gives a complete set of n basic solutions

$$Q_n^m \mathbf{T}^{(n)}, \quad n = 1, \ldots, N-1,$$

for the system of difference equations $\mathbf{T}^{m+1} = A\mathbf{T}^m$ and thus assures the stability of the numerical method if, and only if,

$$|Q_n| = \left|1 + 2\epsilon\left(-1 + \cos\frac{n\pi}{N}\right)\right| \leq 1, \quad \text{for all } n = 1, \ldots, N-1. \tag{12.17}$$

Note that $-1 + \cos\frac{n\pi}{N} < 0$ for $1 \leq n \leq N-1$, so for $\epsilon > 0$,

$$Q_n = 1 + 2\epsilon\left(-1 + \cos\frac{n\pi}{N}\right) < 1.$$

So, in cases (a) and (b), stability is decided by whether $Q_n \geq -1$ for *all* $n = 1, \ldots, N-1$.

(a) Suppose $0 < \epsilon \leq \frac{1}{2}$. Then for $1 \leq n \leq N-1$, we have $\cos\frac{n\pi}{N} > -1$, so

$$Q_n = 1 + 2\epsilon\left(-1 + \cos\frac{n\pi}{N}\right) \geq 1 + 2 \cdot \frac{1}{2}\big(-1 - 1\big) = -1;$$

thus, the numerical method is stable.

(b) Suppose $\epsilon > \frac{1}{2}$. Then for $n = N-1$ and sufficiently large N, $\frac{n\pi}{N} = \frac{\pi(N-1)}{N}$ will be sufficiently close to π that we have

$$\cos\frac{\pi(N-1)}{N}$$

sufficiently close to the smallest value it can be, -1. Then

$$Q_{N-1} = 1 + 2\epsilon\left(-1 + \cos\frac{n\pi}{N}\right) < 1 + 2\epsilon\big(-1 + (\approx -1)\big) < -1,$$

because $2\epsilon > 1$. Thus, the numerical method is unstable for sufficiently large N.

(c) Suppose $\epsilon < 0$, that is, $\triangle t < 0$, that is, we are going backward in time. Then because $-1 + \cos\frac{\pi}{N} < 0$, for $n = 1$, we have

$$Q_1 = 1 + 2\epsilon\left(-1 + \cos\frac{\pi}{N}\right) > 1,$$

which implies that the numerical method is unstable. □

In case (b), that is, when the time step size is too large, the "mechanism" of instability is that A has an eigenvalue $Q_{N-1} < -1$. Because Q_{N-1}^m as a function of m alternates

in sign, this causes the numerical solution to oscillate between positive and negative values of larger and larger magnitude, exactly what we saw in Table 12.2. It is great when the numerical results show exactly what we expect from a theoretical analysis.

To check numerical results often one takes a finer and finer grid. Theorem 12.1 tells us that to avoid numerical instability, we must take the time step size to be very small if the grid size is small. For example, if $\alpha = 1$ and $h = \triangle x = 10^{-7}$, then we need to use $\triangle t = 5.0 \times 10^{-15}$. If time is measured in seconds, then numerically we must work in femtoseconds, and we need 2 *petasteps* to predict what happens during 1 s on time. If we insist on a very accurate description in space, then our numerical work may take as much computer time as the age of the universe. Another danger is that if we take a huge number of time steps, then the *accumulated* round off errors would make the numerical results meaningless.

12.2.1 Crank–Nicholson Method

We saw in Section 8.4 that solving a system of linear algebraic equations can be accomplished using an implicit, iterative method such as Gauss–Seidel's. Here we will see that an implicit method can help us overcome the numerical instability issue.

Consider the heat equation, $\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2}$. Instead of approximating

$$\frac{\partial^2 T}{\partial x^2}(x_j, t_m) \approx \frac{T(x_{j+1}, t_m) - 2T(x_j, t_m) + T(x_{j-1}, t_m)}{h^2},$$

use the **Crank–Nicholson method** of approximating the PDE:

$$\frac{\partial^2 T}{\partial x^2}(x_j, t_m) \approx \frac{1}{2}\,\frac{T_{j+1}^m - 2T_j^m + T_{j-1}^m}{h^2} + \frac{1}{2}\,\frac{T_{j+1}^{m+1} - 2T_j^{m+1} + T_{j-1}^{m+1}}{h^2}. \tag{12.18}$$

Substitute this, along with the usual forward difference approximation of $\frac{\partial T}{\partial t}(x_j, t_m)$, into the heat equation. This gives the **implicit** partial difference equation

$$-\frac{\epsilon}{2}\,T_{j-1}^{m+1} + (1+\epsilon)T_j^{m+1} - \frac{\epsilon}{2}\,T_{j+1}^{m+1} = \frac{\epsilon}{2}\,T_{j+1}^m + (1-\epsilon)T_j^m + \frac{\epsilon}{2}\,T_{j-1}^m, \quad j = 1, \ldots, N-1, \tag{12.19}$$

where, as usual, $\epsilon \triangleq \frac{\triangle t}{(\pi/N)^2}$. We call (12.19) "implicit" because it does not explicitly give a formula for T_j^{m+1}, the temperature at time t_{m+1} and space point x_j, in terms of the temperature values at the earlier time, t_m.

We can express the vector $\mathbf{T}^{m+1}$ in terms of $\mathbf{T}^m$:

$$\begin{bmatrix} 1+\epsilon & -\frac{1}{2}\epsilon & 0 & . & . & . & 0 \\ -\frac{1}{2}\epsilon & 1+\epsilon & -\frac{1}{2}\epsilon & & & & . \\ 0 & -\frac{1}{2}\epsilon & 1-\epsilon & . & & & . \\ . & & . & . & . & & . \\ . & & & . & . & . & 0 \\ . & & & & -\frac{1}{2}\epsilon & 1+\epsilon & -\frac{1}{2}\epsilon \\ 0 & . & . & & 0 & -\frac{1}{2}\epsilon & 1+\epsilon \end{bmatrix} \mathbf{T}^{m+1}$$

$$= \begin{bmatrix} 1-\epsilon & \frac{1}{2}\epsilon & 0 & . & . & . & 0 \\ \frac{1}{2}\epsilon & 1-\epsilon & \frac{1}{2}\epsilon & & & & . \\ 0 & \frac{1}{2}\epsilon & 1-\epsilon & . & & & . \\ . & & . & . & . & & . \\ . & & & . & . & . & 0 \\ . & & & & \frac{1}{2}\epsilon & 1-\epsilon & \frac{1}{2}\epsilon \\ 0 & . & . & & 0 & \frac{1}{2}\epsilon & 1-\epsilon \end{bmatrix} \mathbf{T}^m. \tag{12.20}$$

To study numerical stability of the Crank–Nicholson method, again we substitute into the partial difference equation (12.19) the product solution (12.14), that is,

$$T_j^m = Q^m \sin\frac{nj\pi}{N}, \quad j = 1, \ldots, N-1,$$

for the partial difference equation (12.5) in Section 12.1. Substitute that into (12.20) to get

$$Q \cdot \left(\sin\frac{nj\pi}{N} - \frac{\epsilon}{2}\left(\sin\frac{n(j+1)\pi}{N} - 2\sin\frac{nj\pi}{N} + \sin\frac{n(j-1)\pi}{N} \right) \right)$$
$$= \sin\frac{nj\pi}{N} + \frac{\epsilon}{2}\left(\sin\frac{n(j+1)\pi}{N} - 2\sin\frac{nj\pi}{N} + \sin\frac{n(j-1)\pi}{N} \right), \quad j = 1, \ldots, N-1. \tag{12.21}$$

In Problem 12.2.2.1, you will derive that (12.21) can be rewritten as

$$Q_n = \frac{1 + \epsilon\left(-1 + \cos\frac{n\pi}{N}\right)}{1 - \epsilon\left(-1 + \cos\frac{n\pi}{N}\right)}, \quad \text{for } n = 1, \ldots, N-1. \tag{12.22}$$

Theorem 12.2

(a) If $\epsilon > 0$, then the Crank–Nicholson method of linear partial difference equations (12.20) and BCs (12.4) in Section 12.1 is stable.

(b) If $\epsilon < 0$, that is, we are going backward in time, then the numerical method of linear partial difference equations (12.20) and BCs (12.4) in Section 12.1 is unstable.

Why? You will explain (a) in Problem 12.2.2.2 and (b) in Problem 12.2.2.3. □

12.2.2 Problems

1. Explain why $\epsilon > 0$ implies Q_n given in (12.22), that is, for the Crank–Nicholson method, satisfies $-1 < Q_n < 1$ for $n = 1, \ldots, N-1$. Also, explain why Theorem 12.2(a) is true.
2. Derive that (12.21) can be rewritten as (12.22).
3. Use the result of Problem 12.2.2.2 and define $\beta = -\epsilon > 0$ to help explain why Theorem 12.2(b) is true.

12.3 Finite Difference Methods for Potential Equations

On a spatial rectangle $0 \le x \le L, 0 \le y \le H$, we can discretize the potential equation

$$F(x,y) = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \tag{12.23}$$

by defining

$$x_n \triangleq n \triangle x, \quad y_m \triangleq m \triangle y, \quad u_{n,m} \triangleq u(x_n, y_m), \quad \text{and} \quad F_{n,m} \triangleq F(x_n, y_m).$$

We can use the central difference approximations

$$\frac{\partial^2 u}{\partial x^2}(x_n, y_m) \approx (\triangle x)^{-2} \left(u_{n-1,m} - 2u_{n,m} + u_{n+1,m}\right)$$

and

$$\frac{\partial^2 u}{\partial y^2}(x_n, y_m) \approx (\triangle y)^{-2} \left(u_{n,m-1} - 2u_{n,m} + u_{n,m+1}\right).$$

Let N be an integer with $L = N \triangle x$ and M be an integer with $H = M \triangle y$. The replacement equations for (12.23) are, for $n = 1, \ldots, N-1$, and $m = 1, \ldots, M-1$,

$$F_{n,m} = (\triangle x)^{-2} \left(u_{n-1,m} - 2u_{n,m} + u_{n+1,m}\right) + (\triangle y)^{-2} \left(u_{n,m-1} - 2u_{n,m} + u_{n,m+1}\right). \tag{12.24}$$

Example 12.6

Set up and solve a finite differences approximate solution of

$$-1 + 6\cos 4y = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, 0 < x < \pi, \quad 0 < y < \frac{\pi}{2}, \tag{12.25}$$

along with BCs

$$\left.\begin{cases} u(0,y) = u(\pi,y) = 0, & 0 < y < \frac{\pi}{2} \\ u(x,0) = u(x,\frac{\pi}{2}) = 2, & 0 < x < \pi \end{cases}\right\}. \tag{12.26}$$

Method: The instructions did not say how fine or coarse we should make the grids in the x and y directions. $N = 8$ and $M = 4$ look large enough to make the work informative and also seems to be a reasonable compromise between accuracy and tediousness in the calculations.

So, $x_n = n \cdot \frac{\pi}{8}$, $y_m = m \cdot \frac{\pi}{8}$, and $u_{n,m} = u(x_n, y_m)$, $n = 1,\ldots,7$, $m = 1,2,3$. For convenience, define $h = \triangle x = \triangle y = \frac{\pi}{8}$. The central difference approximation of the Laplacian is, in this example,

$$\nabla^2 u(x_n, y_m) \approx h^{-2}\left(u_{n,m-1} + u_{n,m+1} + u_{n-1,m} + u_{n+1,m} - 4u_{n,m}\right), \tag{12.27}$$

because $\triangle x = h = \triangle y$. This **five-point Laplacian** stencil * is illustrated in Figure 12.2.

The replacement equations are

$$\begin{aligned} -1 + 6\cos(4y_m) &= F_{n,m} \approx \nabla^2 u(x_n, y_m) \\ &\approx h^{-2}\left(u_{n,m-1} + u_{n,m+1} + u_{n-1,m} + u_{n+1,m} - 4u_{n,m}\right), \end{aligned}$$

along with the BCs

$$u_{0,m} = u_{8,m} = 0, \quad m = 1,2,3, \quad \text{and} \quad u_{n,0} = u_{n,4} = 2, \quad n = 1,\ldots,7.$$

The approximate solution's values and the boundary values can be displayed in a rectangular array:

$$\begin{array}{ccccccccc} & 2 & 2 & 2 & 2 & 2 & 2 & 2 & \\ 0 & u_{13} & u_{23} & u_{33} & u_{43} & u_{53} & u_{63} & u_{73} & 0 \\ 0 & u_{12} & u_{22} & u_{32} & u_{42} & u_{52} & u_{62} & u_{72} & 0 \\ 0 & u_{11} & u_{21} & u_{31} & u_{41} & u_{51} & u_{61} & u_{71} & 0 \\ & 2 & 2 & 2 & 2 & 2 & 2 & 2 & \end{array}.$$

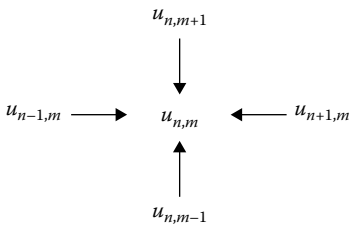


FIGURE 12.2
Five-point Laplacian.

* Discretizations of the Laplacian and other partial differential operators are found in Section 25.3 of Abramowitz and Stegun (1964).

We could set up a system of 21 equations in the 21 unknowns $u_{n,m}$, $n = 1, \ldots, 7$, $m = 1, 2, 3$. For that purpose, it would help to define the great vector

$$\mathbf{u} = [u_{11} \quad \ldots \quad u_{71} \;\vdots\; u_{12} \quad \ldots \quad u_{72} \;\vdots\; u_{13} \quad \ldots \quad u_{73}]^T \quad \text{in} \quad \mathbb{R}^{21}.$$

But we can use the apparent symmetry satisfied by *both* the BCs (12.26) and the inhomogeneity in the PDE (12.25): We suspect that the solution $u(x, y)$ satisfies

$$u(x, y) = u(\pi - x, y) \quad \text{and} \quad u(x, y) = u\left(x, \frac{\pi}{2} - y\right), \quad 0 < x < \pi, \quad 0 < y < \frac{\pi}{2},$$

which would manifest in the approximate solution satisfying

$$u_{n,m} = u_{N-n,m} \quad \text{and} \quad u_{n,m} = u_{n,M-m}, \quad n = 1, \ldots, 7, \quad m = 1, 2, 3.$$

Using symmetry, the rectangular array would display the solution as

$$\begin{array}{ccccccccc}
 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & \\
0 & u_{11} & u_{21} & u_{31} & u_{41} & u_{31} & u_{21} & u_{11} & 0 \\
0 & u_{12} & u_{22} & u_{32} & u_{42} & u_{32} & u_{22} & u_{12} & 0 \\
0 & u_{11} & u_{21} & u_{31} & u_{41} & u_{31} & u_{21} & u_{11} & 0 \\
 & 2 & 2 & 2 & 2 & 2 & 2 & 2 &
\end{array}.$$

For example, at $(n, m) = (1, 2)$, using $u_{1,3} = u_{1,1}$, the BC $u_{0,2} = 0$, and the inhomogeneity value $F_{1,2}$ and multiplying through the replacement equations by h^2, the replacement equation becomes

$$\begin{aligned}
\frac{\pi^2}{64} \cdot (-7) &= \frac{\pi^2}{64}\left(-1 + 6\cos\left(4 \cdot \frac{\pi}{4}\right)\right) = \frac{\pi^2}{64}\left(-1 + 6\cos(4y_2)\right) = h^2 F_{1,2} \\
&\approx h^2 \nabla^2 u(x_1, y_2) \approx u_{1,1} + u_{1,3} + u_{0,2} + U_{2,2} - 4u_{1,2} = 2u_{1,1} + 0 + u_{2,2} - 4u_{1,2}.
\end{aligned}$$

Define the vector

$$\mathbf{U} = [u_{11} \quad u_{21} \quad u_{31} \quad u_{41} \quad u_{12} \quad u_{22} \quad u_{32} \quad u_{42}]^T \text{ in } \mathbb{R}^8.$$

Incorporating $h = \frac{\pi}{8}$, the symmetry assumption, the BCs, and the inhomogeneity values $F_{n,m} = -1 + 6\cos(4y_m)$ and multiplying through the replacement equations by h^2 gives us the system

$$\begin{bmatrix}
-4 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & -4 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & -4 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 2 & -4 & 0 & 0 & 0 & 1 \\
2 & 0 & 0 & 0 & -4 & 1 & 0 & 0 \\
0 & 2 & 0 & 0 & 1 & -4 & 1 & 0 \\
0 & 0 & 2 & 0 & 0 & 1 & -4 & 1 \\
0 & 0 & 0 & 2 & 0 & 0 & 2 & -4
\end{bmatrix}
\begin{bmatrix} u_{11} \\ u_{21} \\ u_{31} \\ u_{41} \\ u_{12} \\ u_{22} \\ u_{32} \\ u_{42} \end{bmatrix}
= \begin{bmatrix} -2 \\ -2 \\ -2 \\ -2 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}
+ \frac{\pi^2}{64}
\begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \\ -7 \\ -7 \\ -7 \\ -7 \end{bmatrix}. \qquad (12.28)$$

The numerical solution of (12.28) was done by *Mathematica*, although MATLAB® would have done just as well. The numerical results are in the rectangular array

$$\begin{array}{cccccccccc} & 2 & 2 & 2 & 2 & 2 & 2 & 2 & & \\ 0 & 1.46580 & 2.10184 & 2.36989 & 2.44390 & 2.36989 & 2.10184 & 1.46580 & 0 & \\ 0 & 1.60714 & 2.41748 & 2.77959 & 2.88162 & 2.77959 & 2.41748 & 1.60714 & 0 & . \bigcirc \\ 0 & 1.46580 & 2.10184 & 2.36989 & 2.44390 & 2.36989 & 2.10184 & 1.46580 & 0 & \\ & 2 & 2 & 2 & 2 & 2 & 2 & 2 & & \end{array}$$

By the way, the exact solution can be found using separation of variables in the form $u(x,y) = v(y) + w_1(x,y) + w_2(x,y)$, where first we find $v(y) = -\frac{1}{2}y^2 - \frac{3}{8}\cos(4y)$, the solution of the problem where the BCs have been replaced by zero on all four sides of the rectangle. The exact solution values on our rectangular array are, after truncating the infinite series for w_1 and w_2 at 80 terms each,

$$\begin{array}{cccccccccc} & 1.9625 & 1.9797 & 1.98446 & 1.98564 & 1.98446 & 1.9797 & 1.9625 & & \\ -0.100339 & 1.41052 & 2.06223 & 2.32217 & 2.39234 & 2.32217 & 2.06223 & 1.41052 & -0.100339 & \\ -0.007893 & 1.5117 & 2.30692 & 2.65895 & 2.75649 & 2.65895 & 2.30692 & 1.5117 & -0.007893 & . \\ 0.078020 & 1.44556 & 2.06953 & 2.32375 & 2.39296 & 2.32375 & 2.06953 & 1.44556 & 0.078020 & \\ & 1.97532 & 1.98664 & 1.98977 & 1.99055 & 1.98977 & 1.98664 & 1.97532 & & \end{array}$$

Increasing the number of terms in each series from 80 to 200 improves the agreement with the exact BCs but has little effect on the values inside the rectangle. So, we see that the finite difference values are correct to within an error of about 5%. The chief problem seems to be that using $\triangle y = \frac{\pi}{8}$ is too coarse to capture the oscillation of the term $\cos(4y)$ in the inhomogeneity in the PDE.

12.3.1 Other Boundary Conditions

If a BC on an edge of the rectangle involves the normal derivative, for example,

$$\frac{\partial u}{\partial y}\left(x, \frac{\pi}{2}\right), \quad 0 < x < \pi,$$

on the horizontal boundary $y = \frac{\pi}{2}$, we can approximate that derivative in many different ways. Pick any n with $0 < n < N$. At a boundary point (x_n, y_M), we can use $\dfrac{\partial u}{\partial y}(x_n, y_M) \approx$

(a) $\dfrac{u_{n,M} - u_{n,M-1}}{h}$, using forward difference

(b) $\dfrac{u_{n,M+1} - u_{n,M}}{h}$, using forward difference

(c) $\dfrac{u_{n,M+1} - u_{n,M-1}}{2h}$, using central difference

The simplest choice is to use (a) because it uses only approximate solution values at points that will also be used in the five-point Laplacian at (x_n, y_{M-1}), that is,

$$\nabla^2 u(x_n, y_{M-1}) \approx h^{-2}\left(u_{n,M-2} + u_{n,M} + u_{n-1,M-1} + u_{n+1,M-1} - 4u_{n,M-1}\right).$$

Both (b) and (c) involve the fictitious approximate solution value $u_{n,M+1} \approx u(x_n, y_{M+1})$ at a point that is not in the rectangle. But the central difference approximation, (c), is appealing because it has higher accuracy than a forward difference approximation. If we use either (b) or (c), we must supplement the equations by using an approximation of the PDE $\nabla^2 u = F(x, y)$ at the boundary point (x_n, y_M), for example,

$$F_{n,M} \approx F(x_n, y_M) = \nabla^2 u(x_n, y_M) \approx h^{-2}\left(u_{n,M-1} + u_{n,M+1} + u_{n-1,M} + u_{n+1,M} - 4u_{n,M}\right).$$

Remarks

We note that the methods we used in two space dimensions extend to finite difference methods for regions in three space dimensions.

We also note that we can use iterative methods mentioned in Section 8.4, for example, SOR or SSOR, to solve the system of linear algebraic equations produced by finite difference methods for potential equations. Indeed, such systems of PDEs were the original context for the SOR method of David M. Young, Jr., whose thesis, "Iterative methods for solving PDEs of elliptical type," can be downloaded from the Internet.

12.3.2 Problems

In problems 1–7, find a numerical approximate solution of the problem using the specified $\triangle x$ and $\triangle y$. *If* possible, use symmetry to reduce the tedium of the work.

1. $$\left\{\begin{array}{c} 0 = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \quad 0 < x < 1, \quad 0 < y < 1 \\ \\ u(0, y) = 0, \quad u(1, y) = 1, \quad 0 < y < 1 \\ u(x, 0) = u(x, 1) = 0, \quad 0 < x < 1 \end{array}\right\},$$
   with $\triangle x = \triangle y = \frac{1}{3}$.
2. $0 = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$, (x, y) in $\mathcal{D}$, with $\triangle x = \triangle y = \frac{1}{2}$, where Figure 12.3 shows the shaded region $\mathcal{D}$, which is a rectangle from which the lower right has been cut out. Consecutive dots in Figure 12.3 are $\frac{1}{2}$ unit apart. The numerical values of $u(x, y)$ on the boundary of $\mathcal{D}$ are shown in Figure 12.3.

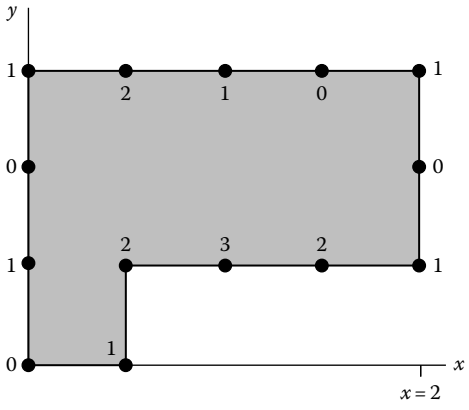


FIGURE 12.3
Problem 12.3.2.2.

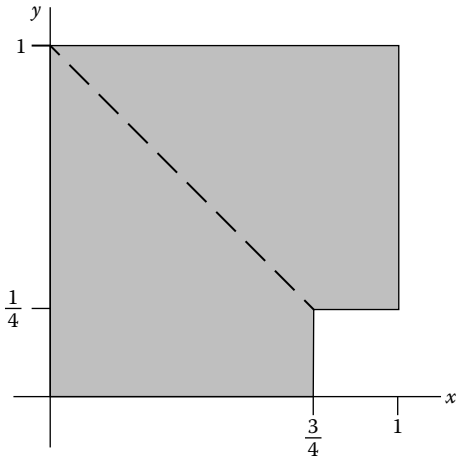


FIGURE 12.4
Problem 12.3.2.3.

3. $\left\{\begin{array}{l} -1 = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \quad (x, y) \text{ in } \mathcal{D} \\ \\ u = 0 \text{ on the boundary of } \mathcal{D} \end{array}\right\}$,
with $\triangle x = \triangle y = \frac{1}{4}$, where Figure 12.4 shows the shaded region $\mathcal{D}$, which is a 1×1 square from which the lower right $\frac{1}{4} \times \frac{1}{4}$ corner square has been cut out. The dashed diagonal line shows symmetry in the problem.

4. $\left\{\begin{array}{l} -1 = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \quad (x, y) \text{ in } \mathcal{D} \\ \\ u = 0 \text{ on the outer boundary of } \mathcal{D} \\ u = 1 \text{ on the inner boundary of } \mathcal{D} \end{array}\right\}$,

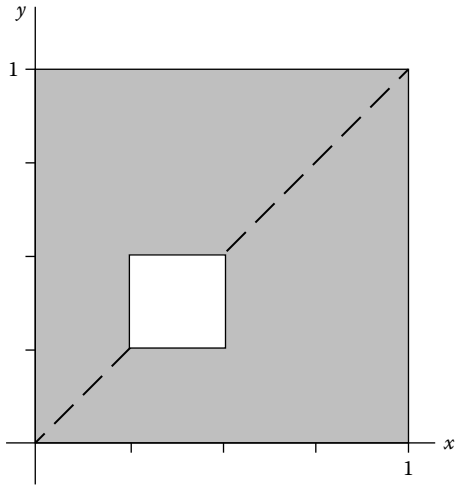


FIGURE 12.5
Problem 12.3.2.4.

with $\triangle x = \triangle y = \frac{1}{4}$, where Figure 12.5 shows the shaded region $\mathcal{D}$, which is a 1×1 square from which a $\frac{1}{4} \times \frac{1}{4}$ square has been cut out. The dashed diagonal line shows symmetry in the problem.

5. $$\left\{ \begin{array}{c} -1 = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \quad 0 < x < \pi, \quad 0 < y < \pi \\ \\ u(0,y) = u(\pi,y) = 4 \\ \\ u(x,0) = 0, \quad \frac{\partial u}{\partial y}(x,\pi) = 1 \end{array} \right\},$$
with $\triangle x = \triangle y = \frac{\pi}{3}$.

6. $$\left\{ \begin{array}{c} -1 = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \, 0 < x < \pi, \, 0 < y < \pi \\ \\ u(0,y) = u(\pi,y) = 2 \\ \\ \frac{\partial u}{\partial y}(x,0) = \frac{\partial u}{\partial y}(x,\pi) = 1 \end{array} \right\},$$
with $\triangle x = \triangle y = \frac{\pi}{3}$.

7. $$\left\{ \begin{array}{c} 0 = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \quad 0 < x < \pi, \quad 0 < y < \pi \\ \\ u(0,y) = u(\pi,y) = 0 \\ \\ \frac{\partial u}{\partial y}(x,0) = -5, \\ \\ 3u(x,\pi) + \frac{\partial u}{\partial y}(x,\pi) = 6 \end{array} \right\},$$
with $\triangle x = \triangle y = \frac{\pi}{3}$.

12.4 Finite Difference Methods for the Wave Equation

The wave equation is

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2},$$

and we know from Section 10.5 that it has D'Alembert wave solutions

$$u(x,t) = f(x+ct) + g(x-ct),$$

whose two **wave speeds** are $\mp c$, respectively. We have seen also that the linearized equations of gas dynamics, the system of two first-order PDEs in (10.66) in Section 10.4, that is,

$$\left\{ \begin{array}{l} \frac{\partial \varrho_1}{\partial t} = -\varrho_0 \frac{\partial u_1}{\partial x} \\ \\ \varrho_0 \frac{\partial u_1}{\partial t} = -P'(\varrho_0) \cdot \frac{\partial \varrho_1}{\partial x} \end{array} \right\}, \tag{12.29}$$

can be rewritten as the wave equation, where $c = \sqrt{P'(\varrho_0)}$.

Now we will rewrite (12.29) in another way that is useful for our discussion of numerical methods: Define $w_1(x,t) = \varrho_0 u_1(x,t)$ and $w_2(x,t) = c\,\varrho_1(x,t)$. In Problem 12.4.3.1, you will explain why (12.29) can be rewritten as

$$\frac{\partial}{\partial t}\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = -c\,B \frac{\partial}{\partial x}\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \tag{12.30}$$

where $B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.

12.4.1 Scalar Hyperbolic Problem

We will begin by studying the **scalar hyperbolic problem**

$$\frac{\partial w}{\partial t} = -c\,\frac{\partial w}{\partial x},$$

that is,

$$\frac{\partial w}{\partial t} + c\,\frac{\partial w}{\partial x} = 0. \tag{12.31}$$

Here, the wave speed c may be either positive or negative.

As we used for the heat equation, discretize space and time by

$$x_j = jh, \quad j = 0,1,2,\ldots, \quad \text{and} \quad t_m = m\,\triangle t, \quad m = 0,1,2,\ldots,$$

and define

$$w_j^m = w(x_j, t_m).$$

We may also write $h = \triangle x$.

It is customary to use the superscript m rather than (m); keep in mind that w_j^m is not the mth power of w_j.

The natural way to discretize (12.31) would be to use the **forward Euler scheme**: use the forward time difference approximation

$$\frac{\partial w}{\partial t}(x_j, t_m) \approx \frac{w_j^{m+1} - w_j^m}{\triangle t}. \tag{12.32}$$

It is also natural to use the central space difference approximation

$$\frac{\partial w}{\partial x}(x_j, t_m) \approx \frac{w_{j+1}^m - w_{j-1}^m}{2h}, \tag{12.33}$$

which we know from Problem 8.6.1.8 has accuracy $\mathcal{O}(h^2)$.

The replacement equations,

$$\frac{w_j^{m+1} - w_j^m}{\triangle t} + c\,\frac{w_{j+1}^m - w_{j-1}^m}{2h} = 0, \tag{12.34}$$

are consistent with the PDE (12.31) and simple to implement; unfortunately, if $c \neq 0$, they are unstable and thus useless! This is an example of the saying that "For every problem, there is a simple answer that is wrong."

There are no BCs, so for the stability analysis, we can substitute into the PDE (12.34) a product solution in the general form

$$w_j^m = Q^m e^{i\alpha j}, \quad \text{for } j = 0, 1, 2, \ldots, \quad m = 0, 1, 2, \ldots, \tag{12.35}$$

where $i = \sqrt{-1}$ and α is real. From (12.34), this gives

$$Q^{m+1} e^{i\alpha j} - Q^m e^{i\alpha j} + \frac{\epsilon}{2}\left(Q^m e^{i\alpha(j+1)} - Q^m e^{i\alpha(j-1)}\right) = 0,$$

where

$$\epsilon \triangleq \frac{c\,\triangle t}{\triangle x}.$$

Dividing through by $Q^m e^{i\alpha j}$ and solving for Q gives the expression for the amplification factor, as in Section 12.2, also known as the **dispersion equation**,

$$Q = 1 - \frac{\epsilon}{2}\left(e^{i\alpha} - e^{-i\alpha}\right) = 1 - i\epsilon \sin\alpha; \tag{12.36}$$

hence

$$|Q|^2 = 1 + \epsilon^2 \sin^2 \alpha.$$

Because we are free to choose $\alpha \neq 0$, $|Q|^2 > 1$; hence, $|Q| > 1$, which shows that the numerical scheme is unstable.

In Problem 12.4.3.2 you will explain why using a forward Euler scheme consisting of the forward time difference approximation (12.32) and the forward space difference approximation,

$$\frac{\partial w}{\partial x}(x_j, t_m) \approx \frac{w_{j+1}^m - w_j^m}{h}, \tag{12.37}$$

is also unstable if the wave speed $c > 0$.

If the wave speed $c < 0$, then the forward Euler scheme consisting of the forward time difference approximation (12.32) and the forward space difference approximation (12.37) is stable.

The latter suggests the scheme of the forward time difference approximation (12.32) and the backward space difference approximation,

$$\frac{\partial w}{\partial x}(x_j, t_m) \approx \frac{w_j^m - w_{j-1}^m}{h}. \tag{12.38}$$

In fact, this is stable, as long as the wave speed $c > 0$, as you will explain in Problem 12.4.3.3.

12.4.2 Lax Scheme

Peter Lax proposed using an averaged approximation of w_j^m in approximating the time derivative:

$$\frac{\partial w}{\partial t}(x_j, t_m) \approx (\triangle t)^{-1} \cdot \left(w_j^{m+1} - \frac{w_{j+1}^m + w_{j-1}^m}{2} \right). \tag{12.39}$$

Along with using the central space difference approximation for the space derivative, this gives the **Lax scheme** for (12.31):

$$(\triangle t)^{-1} \cdot \left(w_j^{m+1} - \frac{w_{j+1}^m + w_{j-1}^m}{2} \right) + c\, \frac{w_{j+1}^m - w_{j-1}^m}{2h} = 0. \tag{12.40}$$

The replacement equations are

$$(\textit{Lax scheme}) \qquad w_j^{m+1} = \frac{1 - \epsilon}{2}\, w_{j+1}^m + \frac{1 + \epsilon}{2}\, w_{j-1}^m, \tag{12.41}$$

where, as earlier,

$$\epsilon \triangleq \frac{c \triangle t}{\triangle x}.$$

In Problem 12.4.3.4, you will explain why the Lax scheme is numerically stable as long as $|\epsilon| \leq 1$, which is the Courant–Friedrichs–Lewy (CFL) criterion.

Learn More About It

Computational Fluid Mechanics and Heat Transfer, 2*nd* edn., by John C. Tannehill, Dale A. Anderson, and Richard H. Pletcher, Taylor & Francis, © 1997, and *Numerical Recipes: The Art of Scientific Computing*, 3*rd* edn., by William H. Press et al., Cambridge University Press © 2007, are among many good resources for learning about finite difference methods for wave equations and, more generally, conservation laws such as the Euler PDE (7.41) in Section 7.6, that is, (10.2) in Section 10.1. As you may see, our study barely scratch the surface of this topic's practical and theoretical aspects, which include the CFL criterion.

12.4.3 Problems

1. Explain why (12.29) can be rewritten as (12.30).
2. Use the following steps to help explain why using a forward Euler scheme consisting of the forward time difference approximation (12.32) and the forward space difference approximation (12.37) is unstable if the wave speed $c > 0$.
   (a) Substitute into the replacement a product solution of the form (12.35), perform algebraic manipulations to arrive at the dispersion equation

   $$Q = 1 + \epsilon - \epsilon\, e^{i\alpha},$$

   and then find the real and imaginary parts of Q.
   (b) Perform algebraic manipulations to get an expression for the square of the modulus of the amplification factor, namely,

   $$|Q|^2 = 1 + 2\epsilon(1+\epsilon)(1-\cos\alpha),$$

   and use that to show that the numerical scheme is unstable if the wave speed $c > 0$.
   (c) Show that the numerical scheme is stable if $c < 0$.
3. Explain why using the scheme for (12.31) of using the forward time difference approximation (12.32) and the backward space difference approximation (12.38), is stable, as long as the wave speed $c > 0$.
4. Explain why the Lax scheme (12.41), that is, replacement equations

   $$w_{j+1}^m = \frac{1-\epsilon}{2}\, w_{j+1}^m + \frac{1+\epsilon}{2}\, w_{j-1}^m,$$

   is numerically stable as long as $|\epsilon| \leq 1$.

5. Study the stability of the scheme for (12.31) of using the backward time difference approximation

$$\frac{\partial w}{\partial t}(x_j, t_m) \approx \frac{w_j^m - w_j^{m-1}}{\triangle\, t}$$

and the forward space difference approximation (12.38).

12.5 Short Take: Galerkin Method

In this section, we will use "Galerkin methods" to approximate the solution of an ODE-BVP or a PDE-BVP using a finite dimensional subspace of functions.

Suppose $L < \infty$ and functions $f(x)$ and $g(x)$ are square integrable on the interval $[\,0, L\,]$, for example,

$$\int_0^L |f(x)|^2\, dx < \infty.$$

In particular, if $f(x)$ is continuous on $[\,0, L\,]$, then $f(x)$ is square integrable on $[\,0, L\,]$.

Define an inner product by

$$\langle f, g\rangle \triangleq \int_0^L f(x)\, g(x)\, dx.$$

Suppose $u(x)$ satisfies an ODE-BVP of the form

$$\left\{ \begin{array}{c} -\big(p(x)u'\big)' + q(x)u = f(x) \\ \\ u(0) = u(L) = 0 \end{array} \right\}. \tag{12.42}$$

For convenience, we define a differential operator A on a function u by

$$(Au)(x) \triangleq -\big(p(x)u'(x)\big)' + q(x)u(x). \tag{12.43}$$

Define a space of functions $\mathcal{V} \triangleq C_0^2[0, L]$ by

$$u \text{ is in } \mathcal{V} \text{ if } u(x) \text{ is twice continuously differentiable on [0, L] and } u(0) = u(L) = 0. \tag{12.44}$$

In the notation C_0^2, the subscript 0 refers to u satisfying the zero Dirichlet BCs $u(0) = u(L) = 0$ and the superscript 2 refers to u being twice continuously differentiable.

ODE-BVP (12.42) can be stated abstractly as

$$Au = f \quad \text{and} \quad u \text{ is in } \mathcal{V}. \tag{12.45}$$

The **Galerkin method** is to pick an N-dimensional vector subspace $\mathcal{V}_N$ and a set of functions $\{\phi_1(x), \ldots, \phi_N(x)\}$ that is a basis for $\mathcal{V}_N \subseteq \mathcal{V}$. We approximate

$$u(x) \approx u_N(x) \triangleq \sum_{n=1}^{N} c_n \phi_n(x)$$

and substitute that into the abstract form of the ODE-BVP to get

$$\sum_{n=1}^{N} c_n A\phi_n(x) = f(x), \tag{12.46}$$

using the linearity of A. Take the inner product of both sides of (12.46) with $\phi_m(x)$ to get

$$\sum_{n=1}^{N} \langle \phi_m,\, A\phi_n \rangle c_n = \langle \phi_m,\, f \rangle, \quad m = 1, \ldots, N. \tag{12.47}$$

This is a system of N linear equations in N unknowns $c_1, \ldots, c_N$.

But (12.47) can be rewritten as

$$\mathrm{A}\,\mathbf{c} = \mathbf{f}, \tag{12.48}$$

where $\mathbf{c} = [\, c_1 \quad \ldots \quad c_N \,]^T$ and

$$\mathbf{f} = [\, \langle \phi_1, f \rangle \quad \ldots \quad \langle \phi_N, f \rangle \,]^T$$

are vectors in $\mathbb{R}^N$ and A is the $N \times N$ matrix defined by

$$\mathrm{A} \triangleq [\, \langle \phi_m,\, A\phi_n \rangle \,]_{\substack{1 \le m \le N \\ 1 \le n \le N}}. \tag{12.49}$$

Definition 12.1

Operator A is **coercive** if there exists a positive constant α such that for all v in $\mathcal{V}$,

$$\langle Av, v \rangle \ge \alpha \langle v, v \rangle. \tag{12.50}$$

Note that if the vector space is $\mathbb{R}^n$ and A is a real $n \times n$ matrix, then A being coercive is the same as A being positive definite, as in Definition 2.19 in Section 2.6.

Theorem 12.3

If A coercive and $\{\phi_1(x), \ldots, \phi_N(x)\}$ is a linearly independent set of functions in $\mathcal{V}$, then the $N \times N$ matrix A given by (12.49) is invertible; hence, we can solve (12.48) for $\mathbf{c}$.

Why? If A is not invertible, then there must exist a vector $\mathbf{c} \neq \mathbf{0}$ such that $A\mathbf{c} = \mathbf{0}$. Defining a function $v(x) \triangleq \sum_{n=1}^{N} c_n \phi_n(x)$, we have

$$0 = \sum_{n=1}^{N} \langle \phi_m,\ A\phi_n \rangle c_n = \Big\langle \phi_m, A\Big(\sum_{n=1}^{N} c_n \phi_n \Big) \Big\rangle = \langle \phi_m, Av \rangle, \quad \text{for } m = 1, \ldots, N; \tag{12.51}$$

the latter follows from using linearity of A. From (12.51) we conclude that

$$0 + \cdots + 0 = \sum_{m=1}^{N} c_m \langle \phi_m, Av \rangle = \sum_{m=1}^{N} \langle c_m \phi_m, Av \rangle = \Big\langle \sum_{m=1}^{N} c_m \phi_m, Av \Big\rangle$$

$$= \langle v, Av \rangle \geq \alpha \langle v, v \rangle = \alpha ||v||^2 \geq 0;$$

hence, $||v|| = 0$; hence, $0 = v \triangleq \sum_{n=1}^{N} c_n \phi_n$. But this would contradict the linear independence of $\{\phi_1(x), \ldots, \phi_N(x)\}$ because $\mathbf{c} \neq \mathbf{0}$. So, A must be invertible. □

Example 12.7

Use the Galerkin method to approximate the solution of the ODE-BVP

$$\left\{ \begin{array}{ll} -y''(x) + xy = 1 + x, & 0 < x < 1 \\ y(0) = y(1) = 0 & \end{array} \right\}. \tag{12.52}$$

Method: It makes sense for the functions $\phi_n(x)$ to satisfy the BCs. Let's use

$$\phi_1(x) = \sin \pi x, \quad \phi_2(x) = \sin 2\pi x, \quad \phi_3(x) = \sin 3\pi x,$$

that is, the approximate solution is to be of the form

$$y(x) = c_1 \sin \pi x + c_2 \sin 2\pi x + c_3 \sin 3\pi x.$$

The differential operator A is defined by

$$Ay = -y''(x) + xy$$

and the inner product is given by $\langle u, v \rangle \triangleq \int_0^1 u(x)v(x)dx$. Define $f(x) = 1 + x$. Using *Mathematica*, for example, to calculate

$$a_{12} \triangleq \langle \phi_1, A\phi_2 \rangle = \int_0^1 \sin \pi x(-(\sin 2\pi x)'' + x \sin 2\pi x)dx$$

$$= \int_0^1 (-4 + x) \sin \pi x \sin 2\pi x\, dx = -\frac{8}{9\pi^2},$$

we calculated the matrix

$$\mathrm{A} \triangleq [\langle \phi_m,\ A\phi_n \rangle]_{\substack{1 \leq m \leq 3 \\ 1 \leq n \leq 3}} = \begin{bmatrix} \frac{1}{4} + \frac{\pi^2}{2} & -\frac{8}{9\pi^2} & 0 \\ -\frac{8}{9\pi^2} & \frac{1}{4} + 2\pi^2 & -\frac{24}{25\pi^2} \\ 0 & -\frac{24}{25\pi^2} & \frac{1}{4} + 9\pi^2 \end{bmatrix}$$

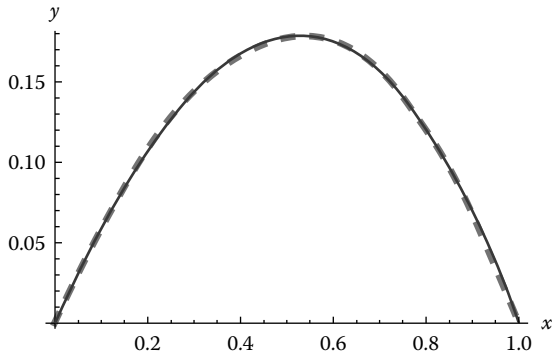


FIGURE 12.6
Example 12.7.

and the vector

$$\mathbf{f} = [\,\langle\phi_1, f\rangle \ldots \langle\phi_N, f\rangle\,]^T = \left[\,\frac{3}{\pi} \quad -\frac{1}{2\pi} \quad \frac{1}{\pi}\,\right]^T.$$

So,

$$\mathbf{c} = [\,c_1 \quad c_2 \quad c_3\,]^T = \mathsf{A}^{-1}\mathbf{f} \approx [\,0.184055 \quad -0.00709816 \quad 0.00711143\,]^T.$$

In Figure 12.6, we graph the approximate solution

$$y(x) \approx 0.184055\sin(\pi x) - 0.00709816\sin(2\pi x) + 0.00711143\sin(3\pi x)$$

as a dashed curve; *Mathematica* produced a numerical approximate solution of the ODE-BVP using NDSolve, and then we graphed it as the solid curve. ○

So, even the Galerkin method using only a three dimensional subspace $\mathcal{V}$ of the infinite dimensional space $C_0^2[0, L]$ can produce a very good approximate solution of an ODE-BVP.

12.5.1 A Generalization of the Galerkin Method

Consider again (12.45), that is, $Au = f$, u in $\mathcal{V}$, where A is linear. As earlier, suppose we approximate

$$u(x) \approx u_N(x) \triangleq \sum_{n=1}^{N} c_n\phi_n(x),$$

and substitute that into $Au = f$ to get (12.46), that is,

$$\sum_{n=1}^{N} c_n A\phi_n(x) = f(x).$$

Let $\{\psi_m\}_{m=1}^N$ be a basis for an N dimensional subspace of $\mathcal{V}$. Take the inner product of both sides of (12.46) with $\psi_m(x)$ to get

$$\sum_{n=1}^{N} \langle \psi_m,\, A\phi_n \rangle c_n = \langle \psi_m,\, f \rangle, \quad m = 1, \ldots, N. \tag{12.53}$$

This is a system of N linear equations in N unknowns $c_1, \ldots, c_N$.

A particularly useful case is when we choose the functions $\{\psi_m\}_{m=1}^N$ to be a **dual basis** for the subspace spanned by the set of functions $\{A\phi_n\}_{n=1}^N$, meaning that

$$\langle \psi_m,\, A\phi_n \rangle = \delta_{mn} = \begin{Bmatrix} 1, & \text{if } m = n \\ 0, & \text{if } m \neq n \end{Bmatrix}.$$

In this very special case, the solution of (12.53) is given simply by

$$c_n = \langle \psi_n,\, f \rangle;$$

hence, the approximate solution is

$$u(x) \approx u_N(x) \triangleq \sum_{n=1}^{N} \langle \psi_n,\, f \rangle \phi_n(x). \tag{12.54}$$

As an aside, suppose that, instead of a function space, we have $\mathcal{V} = \mathbb{R}^m$ with the usual, Euclidean inner product, so an element of $\mathcal{V}$ would be a vector $\mathbf{u}$ and the right-hand side of $Au = f$ would be a vector $\mathbf{f}$. In that situation, (12.54) would be

$$\mathbf{u} = \left(\sum_{n=1}^{N} \boldsymbol{\phi}_n \boldsymbol{\psi}_n^T \right) \mathbf{f}.$$

This formula is reminiscent of formula (2.61) in the context of the SVD (singular value decomposition).

12.5.2 The Galerkin Method for PDEs

Of course, we can approximate the solution of a PDE, too.

Example 12.8

Use the Galerkin method to approximate the solution of the PDE-BVP

$$\begin{Bmatrix} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = -(1 + x + y),\ 0 < x < 1,\ \ 0 < y < 1 \\ \\ u(0, y) = u(1, y) = 0 \\ u(x, 0) = u(x, 1) = 0 \end{Bmatrix}. \tag{12.55}$$

Method: It makes sense for the functions $\phi_n(x)$ to satisfy the BCs. Let's use

$$\phi_1(x,y) = xy(1-x)(1-y), \quad \phi_2(x,y) = x^2y(1-x)(1-y),$$
$$\phi_3(x,y) = xy^2(1-x)(1-y),$$

and so the approximate solution is to be of the form

$$y(x) = c_1xy(1-x)(1-y) + c_2x^2y(1-x)(1-y) + c_3xy^2(1-x)(1-y).$$

The Laplacian differential operator $A = \Delta$ is defined by

$$Ay = \frac{\partial^2 y}{\partial x^2} + \frac{\partial^2 y}{\partial y^2}$$

for functions that satisfy the zero Dirichlet BCs in (12.55). The inner product is given by $\langle u, v\rangle \triangleq \int_0^1 \int_0^1 u(x,y)v(x,y)dx\,dy$. Using *Mathematica* we calculated the matrix

$$\mathrm{A} \triangleq [\langle \phi_m,\, A\phi_n\rangle]_{\substack{1 \le m \le 3 \\ 1 \le n \le 3}} = \begin{bmatrix} -\frac{1}{45} & -\frac{1}{90} & -\frac{1}{90} \\ -\frac{1}{90} & -\frac{4}{525} & -\frac{1}{180} \\ -\frac{1}{90} & -\frac{1}{180} & -\frac{4}{525} \end{bmatrix}$$

and the vector

$$\mathbf{f} = [\,\langle\phi_1,f\rangle \quad \langle\phi_2,f\rangle \quad \langle\phi_3,f\rangle\,]^T = \begin{bmatrix} -\frac{1}{18} & -\frac{7}{240} & -\frac{7}{240} \end{bmatrix}^T,$$

where $f(x,y) = -(1+x+y)$. So,

$$\mathbf{c} = [\,c_1 \quad c_2 \quad c_3\,]^T = \mathrm{A}^{-1}\mathbf{f} = \begin{bmatrix} \frac{95}{52} & \frac{35}{52} & \frac{35}{52} \end{bmatrix}^T.$$

So, the approximate solution is

$$u(x,y) = \frac{5}{52}\, xy(1-x)(1-y)\left(19+7x+7y\right). \bigcirc$$

Figure 12.7a shows the graph of the approximate solution. Figure 12.7b shows the graph of the absolute value of the error, $u(x,y) - v(x,y)$, where $v(x,y)$ is a truncation of the exact solution obtained by the method of Section 11.4. The exact solution would be

$$v(x,y) = \sum_{m=1}^{\infty}\sum_{m=1}^{\infty} \frac{4(1-2(-1)^m - 2(-1)^n + 3(-1)^{m+n})}{mn\left(m^2+n^2\right)\pi^4} \sin(m\pi x)\sin(n\pi y)\,.$$

It appears that the error is not very big, although the error is relatively great near the sides $x = 1$ and $y = 1$.

12.5.3 Nonlinear Problems

Certainly Galerkin method is useful for problems such as Example 12.7 where the non-constant coefficients ODE does not have simple solutions. Even more useful is Galerkin

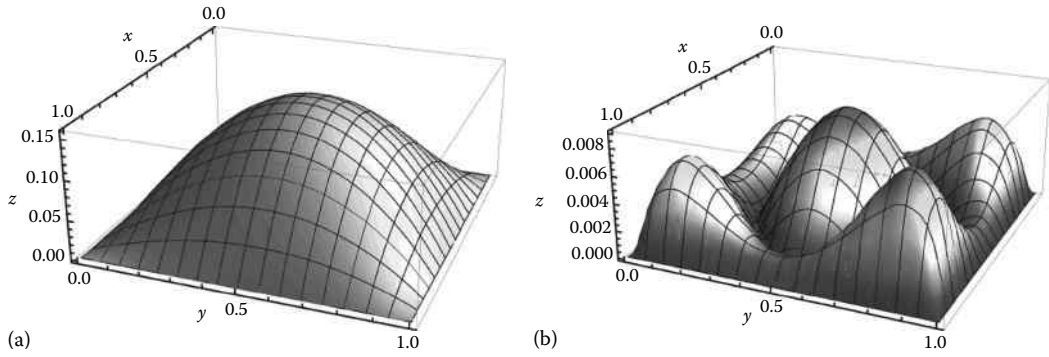


FIGURE 12.7
Example 12.8.

method for nonlinear problems. While we will not include such an example, we note that having a *nonlinear* differential operator A means that

$$A\left(\sum_{n=1}^{N} c_n\phi_n(x)\right) = f(x)$$

cannot be reduced to (12.46). Here, the Galerkin method is to solve the system of nonlinear equations

$$\left\langle \phi_m,\ A\Big(\sum_{n=1}^{N} c_n\phi_n\Big)\right\rangle = \langle \phi_m,\ f\rangle, \quad m = 1,\ldots,N, \tag{12.56}$$

for $c_1,\ldots,c_N$. Newton's method or its variants, such as the secant method, can be used to solve (12.56). Or you can use *Mathematica*'s NSolve command to solve a nonlinear system of equations.

12.5.4 Problems

In problems 1–5, use Galerkin method to approximate the solution of the ODE- or PDE-BVP.

1. $\left\{\begin{array}{ll} -y''(x) + x^2y = 1 + x, & 0 < x < 1 \\ \\ y(0) = y(1) = 0 \end{array}\right\}$
2. $\left\{\begin{array}{ll} -y''(x) + xy = 1 + x, & 0 < x < 1 \\ \\ y(0) = y'(1) = 0 \end{array}\right\}$

   [Hint: See Table 9.1, which implicitly gives functions that satisfy the BCs in Problem 12.5.4.2.]

3. $\left\{\begin{array}{l} -y''(x) + xy = 1 + x,\ 0 < x < 1 \\ \\ y'(0) = y(1) = 0 \end{array}\right\}$

   [Hint: See Table 9.1, which implicitly gives functions that satisfy the BCs in Problem 12.5.4.3.]

4. $\left\{\begin{array}{l} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = -(1 + xy), 0 < x, y < 1 \\ \\ \frac{\partial u}{\partial x}(0, y) = u(1, y) = 0 \\ u(x, 0) = u(x, 1) = 0 \end{array}\right\}$

   [Hint: $\phi_3(x, y) = x^2 y(1 - x)(1 - y)$ is one example of a function that satisfies all of the BCs. See Table 9.1, which implicitly gives product functions that satisfy the BCs in Problem 12.5.4.4.]

5. $\left\{\begin{array}{l} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = -(1 + xy), 0 < x, y < 1 \\ \\ \frac{\partial u}{\partial x}(0, y) = u(1, y) = 0 \\ \\ \frac{\partial u}{\partial y} u(x, 0) = u(x, 1) = 0 \end{array}\right\}$

   [Hint: $\phi_1(x, y) = \cos\left(\frac{\pi x}{2}\right) \cos\left(\frac{\pi y}{2}\right)$ and $\phi_3(x, y) = x^2 y^2 (1-x)(1-y)$ are two examples of a function that satisfies all of the BCs.]

Key Terms

amplification factor: (12.16) in Section 12.2
coercive: Definition 12.1 in Section 12.5
consistent numerical method: beginning of Section 12.2
convergent numerical method: beginning of Section 12.2
Crank–Nicholson method: (12.18) in Section 12.2
discretization matrix: after (12.6) in Section 12.1
dispersion equation: (12.36) in Section 12.4
dual basis: after (12.53) in Section 12.5
five point Laplacian: (12.27) in Section 12.3
forward Euler scheme: (12.32) in Section 12.4
Galerkin method: before and in (12.47) in Section 12.5
generalization of the Galerkin method: (12.53) in Section 12.5
implicit partial difference equation: (12.19) in Section 12.2
Lax scheme: (12.40) in Section 12.4
replacement equations: (12.2) in Section 12.1
scalar hyperbolic problem: (12.31) in Section 12.4
stable numerical method: beginning of Section 12.2
stencil: (12.2) in Section 12.1

Reference

Abramowitz, M. and Stegun I. (eds.). *Handbook of Mathematical Functions*. Dover Publications, New York, 1964.

13

Optimization

13.1 Functions of a Single Variable

In this chapter, we will study methods for some of the more practical problems engineers face to optimize use of resources, performance characteristics, or financial gain. In addition, some of the techniques and ideas are relevant to the subject of calculus of variations in Chapter 14.

Companies spend millions of dollars on research to save 1% or 2% of the weight of an automobile or get a 1% or 2% improvement of the fuel efficiency of an airplane engine. Often there is an interplay between numerical methods for modeling and solution of equations and the optimization of a performance characteristic(s).

Definition 13.1

Suppose x is a single real variable and $f(x)$ is a function whose domain is an interval $\mathcal{I}$.

(a) $x^\star$ is a **global minimizer** of $f(x)$ on $\mathcal{I}$ if

$$f(x^\star) \leq f(x) \text{ for all } x \text{ in } \mathcal{I}.$$

In this case, we say $f(x^\star)$ is the **global minimum (value)** of $f(x)$ on $\mathcal{I}$.

(b) $x^\star$ is a **local (or relative) minimizer** of $f(x)$ on $\mathcal{I}$ if there is a $\delta > 0$ such that

$$f(x^\star) \leq f(x) \text{ for all } x \text{ in } \mathcal{I} \text{ with } |x - x^\star| < \delta.$$

In this case, we say $f(x^\star)$ is a **local minimum (value)** of $f(x)$ on $\mathcal{I}$.

The definitions of "global maximizer" and "local (or relative) maximizer" are similar, for example, $x^\star$ is a **global maximizer** of $f(x)$ on $\mathcal{I}$ if $f(x^\star) \geq f(x)$ for all x in $\mathcal{I}$.

Theorem 13.1

Suppose $f(x)$ is a continuous function on a closed, finite interval $\mathcal{I} = [a, b]$. Then f has a global minimizer (global maximizer) in $\mathcal{I}$.

Definition 13.2

$f(x)$ has a **critical number** x_0 if $f'(x_0)$ is either zero or does not exist.

Theorem 13.2

(Fermat's theorem) Suppose $f(x)$ is a function on an interval $\mathcal{I}$ and $x^\star$ is strictly inside $\mathcal{I}$, that is, there is a $\delta > 0$ such that the interval $(x^\star - \delta, x^\star + \delta) \subseteq \mathcal{I}$. If f has a local minimum or maximum at $x^\star$, then $x^\star$ is a critical number of f.

Corollary 13.1

If $f(x)$ is a continuous function defined on a closed, finite interval $[a, b]$ then each of its global minimizer(s) and maximizer(s) is either a critical number or an endpoint a or b.

Corollary 13.1 gives us a theoretical method for finding the global minimizer and maximizer of a continuous function on a closed, finite interval.

Example 13.1

A cylinder of radius r_i and length L is carrying matter whose temperature is T_i. The cylinder has insulation from radius r_i to radius r_o, where $r_o > r_i$. As depicted in Figure 13.1, the temperature outside the cylinder is $T_o < T_i$. Here, i stands for the inside of the insulation, and o stands for its outside.

According to the classical theory, the rate of convective and conductive heat loss from the insulated cylinder is (See Aziz, 1997)

$$q = \frac{\beta}{h \ln (r_o/r_i) + (k/r_o)},$$

where

h is the convective heat transfer coefficient for the insulation material
k is the conductive heat transfer coefficient for the insulation material
$\beta = 2\pi khL(T_i - T_\infty)$
T_∞ is the (constant) temperature of the medium

Discuss the dependence of the convective heat loss on the outer radius, r_o.

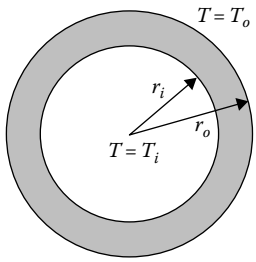


FIGURE 13.1
Example 13.1.

Method: Let $r = r_o$ for notational convenience and define the function

$$f(r) \triangleq \frac{\beta}{h \ln (r/r_i) + (k/r)}.$$

Let h, k, L, r_i, T_i, and T_o be constant parameters. On its domain $r_i \le r < \infty$, we calculate

$$f'(r) = \cdots = -\frac{\beta}{\left(h \ln (r/r_i) + (k/r)\right)^2} \cdot \left(\frac{h}{r} - \frac{k}{r^2}\right),$$

so the only possible critical point is at $r^\star = \frac{k}{h}$, which is called the **critical radius of insulation**. It exists only if $\frac{k}{h} > r_i$.

Because

$$f'(r) = \cdots = -\frac{\beta(hr - k)}{r^2 \left(h \ln (r/r_i) + (k/r)\right)^2},$$

we see that $f'(r) > 0$ for $r < r^\star$ and $f'(r) < 0$ for $r \ge r^\star$. By Theorem 13.3 mentioned later, the global maximum of the convective heat transfer is at $r_o = r^\star$, assuming the critical radius of insulation exists.

The paradoxical result is that if $r_i < r^\star = \frac{k}{h}$, then an insulated cylinder can lose more heat by convection than an uninsulated cylinder. This can be useful if the cylinder is an electrical wire that needs to dissipate heat.

On the other hand, if $r_i > r^\star = \frac{k}{h}$, then adding insulation has the usually desired effect of decreasing heat dissipation from a pipe of radius r_i carrying, for example, hot water or steam. ◯

Aziz goes on to discuss a more recent theory of forced convection and shows that the classical theory greatly underestimates the maximum heat transfer rate.

13.1.1 Global Optimization Result

In Section 4.2, we discussed the maximum amplitude of steady-state oscillations for a sinusoidally forced second-order linear ordinary differential equation. At that time, we stated essentially the same result as the next theorem.

Theorem 13.3

Suppose a function $f(x)$ is defined and differentiable for x in the interval (a, ∞), for some a. If $f'(x)$ is positive for $a < x < x^\star$ and $f'(x)$ is negative for $x^\star < x < \infty$, then the global maximum value of f on (a, ∞) is $f(x^\star)$.

13.1.2 Convex Functions and Optimization

Definition 13.3

A function $f = f(x)$ defined on an interval $\mathcal{I}$ is

(a) **Convex** if

$$f\big((1-t)x + ty\big) \le (1-t)f(x) + tf(y), \text{ for all } 0 \le t \le 1,\ x, y \text{ in } \mathcal{I} \tag{13.1}$$

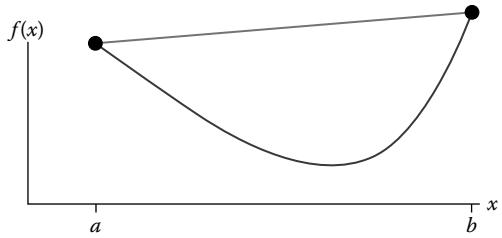


FIGURE 13.2
Secant line above graph of convex function.

(b) **Strongly (or strictly) convex** if

$$f\big((1-t)x+ty\big) < (1-t)f(x)+tf(y), \text{ for all } 0<t<1,\ x\neq y \text{ in } \mathcal{I} \tag{13.2}$$

Figure 13.2 illustrates that the graph of a convex function on an interval $[a,b]$ lies below the secant line connecting the points $(a,f(a))$, $(b,f(b))$.

In Problem 13.1.3.9, you will explain why a twice continuously differentiable function of a single variable defined on an interval is convex if, and only if, its second derivative is non-negative on that interval. So, "convex" is what is called "concave up" in Calculus I.

Theorem 13.4 *

Suppose $f(x)$ is a convex function on an interval $\mathcal{I}$ and $x^\star$ is in $\mathcal{I}$. If $x^\star$ is a local minimizer for f, then it is a global minimizer for f on $\mathcal{I}$.

Why? Suppose, to the contrary, that there is an $x^\star$ that is a local minimizer for f but is *not* a global minimizer for f on $\mathcal{I}$. $x^\star$ being a local minimizer implies there is a $\delta>0$ such that $f(x^\star)\le f(x)$ for all x in $\mathcal{I}$ with $|x-x^\star|<\delta$. $x^\star$ not being a global minimizer implies there is a $\widetilde{x}$ such that $f(\widetilde{x})<f(x^\star)$, which implies that $|x^\star-\widetilde{x}|\ge\delta$. Define $t=0.9\delta/|x^\star-\widetilde{x}|$. We see that $0<t<1$. Note for future reference that $|x^\star-\widetilde{x}|=0.9\delta/t$.

Define $x\triangleq t\widetilde{x}+(1-t)x^\star$. By the convexity of f,

$$f(x)=f\big(t\widetilde{x}+(1-t)x^\star\big)\le tf\big(\widetilde{x}\big)+(1-t)f\big(x^\star\big)<tf\big(x^\star\big)+(1-t)f\big(x^\star\big)=f\big(x^\star\big). \tag{13.3}$$

But, $x=t\widetilde{x}+(1-t)x^\star$ satisfies

$$|x^\star-x|=\big|x^\star-t\widetilde{x}-(1-t)x^\star\big|=\big|t\big(x^\star-\widetilde{x}\big)\big|=t\cdot\big|x^\star-\widetilde{x}\big|=t\cdot(0.9\delta/t)=0.9\delta;$$

by the local minimizer assumption, this implies $f(x)\ge f(x^\star)$, contradicting (13.3). So, the local minimizer must be the global minimizer on the interval $\mathcal{I}$. □

13.1.3 Problems

1. A page of a book is to contain 40 in.^{2} of print. If the margins of both the top and bottom sides must be 2 in. and the margins of both the left and right sides must be 1 in., find the dimensions of the paper that minimize its area.

* This theorem is from Cheney (1982).

2. Inside a circular pipe whose inner diameter is 25 cm, a machine part is rotating anomalously because its motion is not perfectly circular. Suppose a point on it describes the parametric motion

$$\mathbf{r}(t) = (10\cos t + \sin^3 t)\hat{\boldsymbol{\jmath}} + (10\sin t + \cos^3 t)\hat{\boldsymbol{\jmath}}.$$

   Find the closest that the point gets to the inner wall of the pipe. At what times does that happen?
3. A rectangular beam is to be cut from a log with a circular cross section of diameter 1 m. If the strength of the beam is proportional to the product of its cross section's width and the square of its cross section's length, find the cross-section dimensions of the strongest beam.
4. A closed rectangular box has a square base and a volume of 20 ft^3. Because of the weight of its contents, the box is constructed from two different types of material. The material for the sides costs \$2/ft^2, and the stronger material for the top and bottom costs \$3/ft^2. Design the dimensions of the box whose total material cost is the least.
5. Suppose that in a small town, the number of people who realize they have contracted a severe influenza is $N(t) \triangleq \frac{2000}{1+999e^{-2t}}$, where t is measured in days. Then $\dot{N}(t)$ measures the rate at which people become aware that they have the illness. When will the town's small hospital have to be ready for the greatest influx of patients?
6. In the paper by Aziz mentioned in Example 13.1, he also studied heat dissipation from an insulated sphere. He stated that the rate of convective and conductive heat loss is

$$q = \beta / \left(\frac{h(r_o - r_i)}{r_i r_o} + \frac{k}{r_o^2} \right),$$

   where
   $\beta = 4\pi k h (T_i - T_\infty)$,
   h is the convective heat transfer coefficient for the insulation material
   k is the conductive heat transfer coefficient for the insulation material

   Discuss the dependence of the convective heat loss on the outer radius, r_o.
7. In the tropical island nation of Quovideo, every person is entitled to a plot of land in the shape of a right triangle and having 1000 m^2 area. Find the dimensions of the plot that uses the least amount of perimeter fencing.
8. Suppose that a rectangle has its base on the x-axis and two of its vertices on the curve $y = \frac{1}{x^2+1}$, and suppose that the rectangle is symmetric about the y-axis. Find the largest possible area of such a rectangle.
9. Explain why a twice continuously differentiable function of a single variable defined on an interval is convex if, and only if, its second derivative is nonnegative on that interval.
10. (a) Use the result of Problem 13.1.3.9 to explain why the function $f(x) \triangleq -\ln x$ is convex on the interval $(0, \infty)$.

(b) Use the result of part (a) to explain why $\ln\big((1-t)x+ty\big) \geq (1-t)\ln(x) + t\ln(y)$, for all $0 \leq t \leq 1$, positive x, y.

(c) Use the result of part (b) to explain why $x^{1-t}y^t \geq (1-t)x + ty$, for all $0 \leq t \leq 1$, positive x, y.

11. Suppose $f(x)$ is a convex function of x in an interval $\mathcal{I}$ and M is any real number. Explain why $S \triangleq \{x \text{ in } \mathcal{I} : f(x) \leq M\}$ is either an interval or empty.

12. (a) Suppose both $f(x)$ and $g(x)$ are convex functions defined on a common interval $\mathcal{I}$. Explain why $f(x) + g(x)$ is also convex on $\mathcal{I}$.

    (b) Give an example where $f(x)$ is convex on an interval $\mathcal{J}$, $g(x)$ is convex on an interval $\mathcal{I}$, and $g(\mathcal{I}) \subseteq \mathcal{J}$, yet $f(g(x))$ is not convex on $\mathcal{I}$.

13. Explain why a twice continuously differentiable function of a single variable defined on an interval is strictly convex if its second derivative is greater than or equal to a positive constant γ on that interval.

14. Suppose $f(x)$ is strictly convex on an interval $\mathcal{I}$. Explain why f cannot have two distinct global minimizers in $\mathcal{I}$.

13.2 Functions of Several Variables

Suppose $f(\mathbf{x})$ is a function which has domain $\mathcal{D}$ in $\mathbb{R}^n$. The **minimization** problem is

$$\left\{\begin{array}{ll}\text{Minimize} & f(\mathbf{x}) \\ \\ \text{Subject to} & \mathbf{x} \text{ is in } \mathcal{D}\end{array}\right\}. \tag{13.4}$$

We call $f = f(\mathbf{x})$ an **objective function**.

Definition 13.4

Suppose $f(\mathbf{x})$ is a function which has domain $\mathcal{D}$ in $\mathbb{R}^n$.

(a) $\mathbf{x}^\star$ is a **global minimizer** of $f(\mathbf{x})$ on $\mathcal{D}$ if

$$f(\mathbf{x}^\star) \leq f(\mathbf{x}) \text{ for all } \mathbf{x} \text{ in } \mathcal{D}.$$

In this case we say $f(\mathbf{x}^\star)$ is the **global minimum (value)** of $f(\mathbf{x})$ on $\mathcal{D}$.

(b) $\mathbf{x}^\star$ is a **local (or relative) minimizer** of $f(\mathbf{x})$ on $\mathcal{D}$ if there is a $\delta > 0$ such that

$$f(\mathbf{x}^\star) \leq f(\mathbf{x}) \text{ for all } \mathbf{x} \text{ in } \mathcal{D} \text{ with } ||\mathbf{x} - \mathbf{x}^\star|| < \delta.$$

In this case, we say $f(\mathbf{x}^\star)$ is a **local minimum (value)** of $f(\mathbf{x})$ on $\mathcal{D}$.

(c) A unit vector $\widehat{\mathbf{d}}$ is a **feasible direction** at $\mathbf{x}_0$ in $\mathcal{D}$ if there is an $\widetilde{\alpha} > 0$ such that the line segment $\{\mathbf{x}_0 + \alpha\widehat{\mathbf{d}} : 0 \leq \alpha \leq \widetilde{\alpha}\}$ is in $\mathcal{D}$.

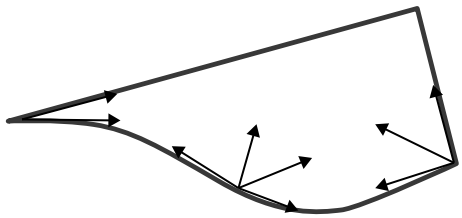


FIGURE 13.3
Feasible directions.

Definition 13.4(c) is illustrated in Figure 13.3. The concept of "feasible direction" is based only on the nature of $\mathcal{D}$, that is, is not influenced by the choice of objective function f.

Theorem 13.5

If $\mathbf{x}^\star$ is a local minimizer for (13.4) and $\mathbf{d}$ is a feasible direction at $\mathbf{x}^\star$ in $\mathcal{D}$, then either

$$\mathbf{d} \bullet \nabla f(\mathbf{x}^\star) \geq 0$$

or $\nabla f(\mathbf{x}^\star)$ does not exist.

Why? If $\nabla f(\mathbf{x}^\star)$ exists, define a function of a single variable by $g(\alpha) = f(\mathbf{x}^\star + \alpha \mathbf{d})$. The chain rule gives

$$g'(\alpha) = \mathbf{d} \bullet \nabla f(\mathbf{x}^\star + \alpha \mathbf{d}),$$

which is the directional derivative of f in the direction of $\mathbf{d}$ at the point $\mathbf{x}^\star + \alpha \mathbf{d}$. Because $\mathbf{x}^\star$ is a local minimizer for (13.4),

$$0 \geq \lim_{\alpha \to 0^+} \frac{g(\alpha) - g(0)}{\alpha} = g'(0^+) = \mathbf{d} \bullet \nabla f(\mathbf{x}^\star). \square$$

Theorem 13.5 says that there is an interaction between the domain and the objective function at a minimizer.

Definition 13.5

Suppose $\mathcal{D}$ is a set in $\mathbb{R}^n$ and $\mathbf{x}_0$ is in $\mathcal{D}$. If there is some solid, open ball $B_r(\mathbf{x}_0) \triangleq \{\mathbf{x} : ||\mathbf{x} - \mathbf{x}_0|| < r\}$ that is contained in $\mathcal{D}$ for some $r > 0$, then we say $\mathbf{x}_0$ is in the **interior** of $\mathcal{D}$.

If $\mathbf{x}_0$ is in the interior of $\mathcal{D}$, then *every* direction is feasible at $\mathbf{x}^\star$ in $\mathcal{D}$, by using $\widetilde{\alpha} = \frac{r}{2}$.

Definition 13.6

(a) $\mathbf{x}_0$ is a **critical point** of $f = f(\mathbf{x})$ if $\nabla f(\mathbf{x}_0)$ is either $\mathbf{0}$ or does not exist.
(b) $\mathbf{x}_0$ is a **stationary point** of $f = f(\mathbf{x})$ if $\nabla f(\mathbf{x}_0) = \mathbf{0}$.

Example 13.2

Find all of the critical points of $f(x,y) \triangleq (2-x^2)(3y-y^2)$.

Method: $\nabla f(\mathbf{x})$ exists at all $\mathbf{x}$, so we need only to find all (x,y) for which both $\frac{\partial f}{\partial x} = 0$ and $\frac{\partial f}{\partial y} = 0$, that is,

$$\left\{ \begin{array}{ll} -2x(3y-y^2) & = 0 \\ (2-x^2)(3-2y) & = 0 \end{array} \right\}. \tag{13.5}$$

The first equation is true when (1) $x = 0$ or (2) $y = 0$ or (3) $y = 3$. The second equation is true when (4) $x = -\sqrt{2}$ or (5) $x = \sqrt{2}$ or (6) $y = \frac{3}{2}$. For both equations in (13.5) to be simultaneously true, we must have both one of (1), (2), or (3) *and* one of (4), (5), or (6). This is like a strangely dictatorial restaurant that insists that you buy exactly one item from "column A" *and* exactly one item from "column B."

In principle, there are nine possibilities:

$$\begin{array}{llllll}
(1) & \text{and} & (4): & x = 0 & \text{and} & x = -\sqrt{2}: \textit{ impossible} \\
(1) & \text{and} & (5): & x = 0 & \text{and} & x = \sqrt{2}: \textit{ impossible} \\
(1) & \text{and} & (6): & x = 0 & \text{and} & y = \dfrac{3}{2}: \ (x,y) = (0, \dfrac{3}{2}) \\
(2) & \text{and} & (4): & y = 0 & \text{and} & x = -\sqrt{2}: \ (x,y) = (-\sqrt{2}, 0) \\
(2) & \text{and} & (5): & y = 0 & \text{and} & x = \sqrt{2}: \ (x,y) = (\sqrt{2}, 0) \\
(2) & \text{and} & (6): & y = 0 & \text{and} & y = \dfrac{3}{2}: \textit{ impossible} \\
(3) & \text{and} & (4): & y = 3 & \text{and} & x = -\sqrt{2}: \ (x,y) = (-\sqrt{2}, 3) \\
(3) & \text{and} & (5): & y = 3 & \text{and} & x = \sqrt{2}: \ (x,y) = (\sqrt{2}, 3) \\
(3) & \text{and} & (6): & y = 3 & \text{and} & y = \dfrac{3}{2}: \textit{ impossible.}
\end{array}$$

So, there are exactly five critical points: $(x,y) = (\pm\sqrt{2}, 0)$, $(\pm\sqrt{2}, 3)$, and $(0, \frac{3}{2})$. ○

Theorem 13.6

(Fermat's theorem) If $\mathbf{x}^\star$ is in the interior of $\mathcal{D}$ and is a local minimizer for (13.4), then $\mathbf{x}^\star$ is a critical point of f.

Why? Because $\mathbf{x}^\star$ is in the interior of $\mathcal{D}$, for every $j = 1, \ldots, n$, both the standard unit basis vector $\mathbf{e}^{(j)}$ and its opposite, $-\mathbf{e}^{(j)}$, are feasible direction. Theorem 13.5 implies

$$\mathbf{e}^{(j)} \bullet \nabla f(\mathbf{x}^\star) \geq 0 \quad \text{and} \quad \left(-\mathbf{e}^{(j)}\right) \bullet \nabla f(\mathbf{x}^\star) \geq 0,$$

or $\nabla f(\mathbf{x}^\star)$ does not exist.

The latter of the two inequalities is equivalent to $\mathbf{e}^{(j)} \bullet \nabla f(\mathbf{x}^\star) \leq 0$, so we get

$$0 \leq \mathbf{e}^{(j)} \bullet \nabla f(\mathbf{x}^\star) \leq 0;$$

hence,

$$\frac{\partial f}{\partial x_j}(\mathbf{x}^\star) = \mathbf{e}^{(j)} \bullet \nabla f(\mathbf{x}^\star) = 0.$$

Because this is true for $j = 1, \ldots, n$, $\nabla f(\mathbf{x}^\star)$ is either $\mathbf{0}$ or does not exist. □

A local maximizer for $f(\mathbf{x})$ that lies in the interior of $\mathcal{D}$ must also be a critical point.

Definition 13.7

Suppose $\mathbf{x}_0$ is a critical point of $f = f(\mathbf{x})$ and there are two feasible directions $\mathbf{d}_1, \mathbf{d}_2$ such that for some $\widetilde{\alpha}$,

$$f(\mathbf{x}_0 + \alpha \mathbf{d}_1) < f(\mathbf{x}_0) < f(\mathbf{x}_0 + \alpha \mathbf{d}_2)$$

for $0 < \alpha \leq \widetilde{\alpha}$. Then we say f has a **saddle point** at $\mathbf{x}_0$.

A critical point may be a local minimizer, a local maximizer, a saddle point, or "none of the above." To find out the nature of a critical point, it helps to use the multivariable generalization of Taylor's theorem. We will state it for $\mathbf{x} = [\, x \quad y \,]^T$ in $\mathbb{R}^2$ and then indicate how it generalizes to $\mathbb{R}^n$. Denote $\triangle \mathbf{r} = \begin{bmatrix} x - a \\ y - b \end{bmatrix}$ and recall the $\mathcal{O}$ notation from Definition 8.5 in Section 8.6.

Definition 13.8

If $f = f(x, y)$ is at least twice differentiable at (a, b), the **Hessian matrix** is

$$[D^2 f(a,b)] \triangleq \begin{bmatrix} \dfrac{\partial^2 f}{\partial x^2}(a,b) & \dfrac{\partial^2 f}{\partial x \partial y}(a,b) \\ \\ \dfrac{\partial^2 f}{\partial x \partial y}(a,b) & \dfrac{\partial^2 f}{\partial y^2}(a,b) \end{bmatrix}.$$

The 2×2 real, symmetric Hessian matrix $[D^2 f(a,b)]$ is a generalization of the concept of second derivative to a function of two variables.

Note that

$$\begin{aligned} &\frac{1}{2!} \triangle \mathbf{r}^T [D^2 f(a,b)\,] \triangle \mathbf{r} \\ &\quad = \frac{1}{2} \left(\frac{\partial^2 f}{\partial x^2}(a,b)(x-a)^2 + 2 \frac{\partial^2 f}{\partial x \partial y}(a,b)(x-a)(y-b) + \frac{\partial^2 f}{\partial x \partial y}(a,b)(y-b)^2 \right). \end{aligned}$$

Theorem 13.7

(Taylor's theorem in $\mathbb{R}^2$) Suppose $f = f(x, y)$ is $(N+1)$ times continuously differentiable at (a, b) for some $N \geq 2$. Then for (x, y) sufficiently near (a, b),

$$f(x,y) = f(a,b) + \frac{1}{1!}\,\triangle\mathbf{r} \bullet \nabla f(a,b) + \frac{1}{2!}\,\triangle\mathbf{r}^T[\,D^2 f(a,b)\,]\triangle\mathbf{r} + \cdots + \frac{1}{N!}\sum_{i=0}^{N}\binom{N}{i}\frac{\partial^N f}{\partial x^i \partial y^{N-i}}(a,b)(x-a)^i(y-b)^{N-i} + R_{N+1}(a,b;x,y), \qquad (13.6)$$

where the remainder term $R_{N+1}(a, b; x, y)$ satisfies

$$|R_{N+1}(a,b;x,y)| = \mathcal{O}\left(\left|\left|\begin{bmatrix} x-a \\ y-b \end{bmatrix}\right|\right|^{N+1}\right)$$

and depends upon the $(N+1)$-th order derivatives $\dfrac{\partial^{N+1} f}{\partial x^i \partial y^{N+1-i}}(x, y)$ for (x, y) near (a, b).

Theorem 13.8

(Second derivative test) Suppose $f = f(x, y)$ has a stationary point at (a, b) and is at least three * times continuously differentiable at (a, b).

(a) If the Hessian matrix $[D^2 f(a, b)]$ is positive definite,† that is, has only positive eigenvalues, then (a, b) is a local minimizer.
(b) If the Hessian matrix $[D^2 f(a, b)]$ is negative definite, that is, has only negative eigenvalues, then (a, b) is a local maximizer.
(c) If the Hessian matrix $[D^2 f(a, b)]$ has one positive eigenvalue and one negative eigenvalue then (a, b) is a saddle point.
(d) If the Hessian matrix has a zero eigenvalue, then further information is needed in order to determine the nature of the critical point at (a, b).

Why? If (a, b) is a stationary point of f, then (13.6) with $N = 2$ gives

$$f(x,y) = f(a,b) + \frac{1}{2!}\;\triangle\mathbf{r}^T[\,D^2 f(a,b)\,]\,\triangle\mathbf{r} + R_3(a,b;x,y).$$

When discussing *local* behavior of f, as long as $[D^2 f(a, b)]$ has no zero eigenvalue, then we may assume that (x, y) is so close to (a, b) that the remainder term, $R_3(a, b; x, y)$, is insignificant compared to the quadratic terms. Then parts (a), (b), and (c) make sense. For example, if $[D^2 f(a, b)]$ has only positive eigenvalues, then positive definiteness of the 2×2 matrix

* It is enough to assume that f is two, rather than three, times continuously differentiable, but "three times" is appropriate when using the $\mathcal{O}$ notation, as in Definition 8.5 in Section 8.6.
† See Section 2.6 for the definition of positive definite.

$[D^2f(a,b)] = [\,a_{ij}\,]$ implies that $a_{11}(x-a)^2 + 2a_{12}(x-a)(y-b) + a_{22}(y-b)^2 > 0$ for $(x,y) \neq (a,b)$; hence, $f(x,y) > f(a,b)$, that is, (a,b) is a local minimizer for f. □

This theorem generalizes to $\mathbb{R}^n$ in place of $\mathbb{R}^2$ in a clear way, except in (c) we change the hypothesis to "have at least one positive eigenvalue and at least one negative eigenvalue."

By Theorem 2.31 in Section 2.6 and discussion before it, a 2×2 symmetric matrix $[\,a_{ij}\,]$ is positive definite if, and only if, $a_{11} > 0$ and $0 < \det\left([\,a_{ij}\,]\right) = a_{11}a_{22} - a_{12}^2$.

On the other hand, *if* a symmetric 2×2 matrix has negative determinant you will explain in Problem 13.2.3.9 why the matrix has one negative and one positive eigenvalue.

Example 13.3

Use Theorem 13.8 to determine, if possible, the nature of the critical points of $f(x,y) \triangleq (2-x^2)(3y-y^2)$.

Method: In Example 13.2, we found that the critical points are $(x,y) = (\pm\sqrt{2}, 0)$, $(\pm\sqrt{2}, 3)$, and $(0, \frac{3}{2})$. The Hessian matrix is

$$D^2f(x,y) = \begin{bmatrix} -2(3y-y^2) & -2x(3-2y) \\ -2x(3-2y) & -2(2-x^2) \end{bmatrix},$$

so

$$\det D^2f(\pm\sqrt{2}, 0) = \begin{vmatrix} 0 & \mp 6\sqrt{2} \\ \mp 6\sqrt{2} & 0 \end{vmatrix} = -72 < 0,$$

so both $(\sqrt{2}, 0)$ and $(-\sqrt{2}, 0)$ are saddle points.

Also,

$$\det D^2f(\pm\sqrt{2}, 3) = \begin{vmatrix} 0 & \pm 6\sqrt{2} \\ \pm 6\sqrt{2} & 0 \end{vmatrix} = -72 < 0,$$

so both $(\sqrt{2}, 3)$ and $(-\sqrt{2}, 3)$ are saddle points.

Finally,

$$\det A = \det D^2f\left(0, \frac{3}{2}\right) = \begin{vmatrix} -\frac{9}{2} & 0 \\ 0 & -4 \end{vmatrix} = 18 > 0,$$

and $a_{11} < 0$. It follows that the matrix $-A$ is positive definite; hence, A is negative definite. So, $\left(0, \frac{3}{2}\right)$ is a local maximum point. ○

Figure 13.4 illustrates the saddle point at $(-\sqrt{2}, 0)$ and the local maximum point at $(0, \frac{3}{2})$ on the graph of $z = f(x,y) = (2-x^2)(3y-y^2)$.

For functions of more than one variable, all sorts of remarkable things can happen. For example, in Problem 13.2.3.8, you will derive that the function $f(x,y) \triangleq e^{-x}\sin(x+y)$ has no critical points yet takes on all values between $-\infty$ and ∞. In Problem 13.2.3.14, you will derive that the function $k(x,y) \triangleq 5xe^y - x^5 - e^{5y}$ has a local maximum at $(x,y) = (1,0)$, but the function does not have a global maximum there, despite the fact that the function has no other critical point.

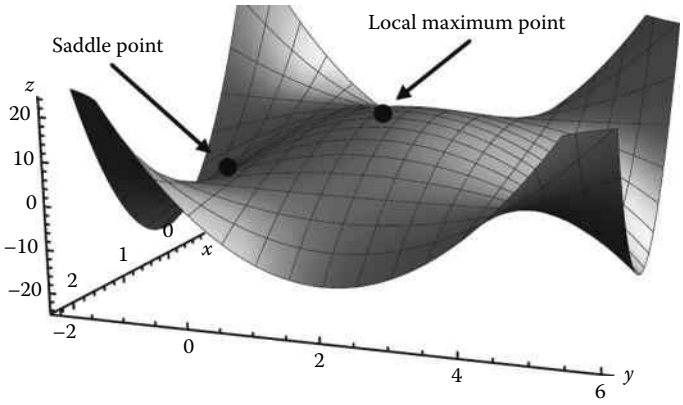


FIGURE 13.4
Example 13.3.

13.2.1 Global Optimization and Lagrange Multipliers

Suppose we want to find the minimum (or maximum) value that $f = f(\mathbf{x})$ has on its domain $\mathcal{D}$ and we want to find a global minimizer (or global maximizer). As in the study of functions of one variable, we have a basic existence result.

Theorem 13.9

Suppose $f = f(\mathbf{x})$ is continuous on a closed, bounded domain $\mathcal{D}$ in $\mathbb{R}^n$. Then there exist both a global minimizer $\mathbf{x}_0^\star$ and a global maximizer $\mathbf{x}_1^\star$ in $\mathcal{D}$, that is, for all $\mathbf{x}$ in $\mathcal{D}$

$$f(\mathbf{x}_0^\star) \leq f(\mathbf{x}) \leq f(\mathbf{x}_1^\star).$$

This seems to be a great result, but unfortunately it does not give a method for finding a global minimizer or maximizer.

One of the most powerful theoretical methods uses geometry: For example, suppose $\mathcal{D}$ is a bounded closed domain in $\mathbb{R}^3$ and suppose that a surface $g(x, y, z) = 0$ is the boundary of $\mathcal{D}$. The global maximizer of an objective function $f = f(x, y, z)$ is either in the interior of $\mathcal{D}$ or is on its boundary. If the global maximizer is in the interior, then (Fermat's) Theorem 13.6 implies it must be a critical point. So, one method of finding the global maximizer is, in principle, to evaluate $f(x, y, z)$ at all critical points and also minimize $f(x, y, z)$ at all (x, y, z) satisfying $g(x, y, z) = 0$. The latter gives us a **constrained optimization problem**:

$$\left\{\begin{array}{ll}\text{Minimize} & f(x, y, z) \\ \text{Subject to} & g(x, y, z) = 0\end{array}\right\}. \tag{13.7}$$

Theorem 13.10

(Lagrange multiplier) Suppose constrained minimization problem (13.7) has solution $\mathbf{x}^\star = (x^\star, y^\star, z^\star)$. Then there exists a **Lagrange multiplier** λ such that

$$\nabla f(\mathbf{x}^\star) = \lambda \nabla g(\mathbf{x}^\star), \tag{13.8}$$

as long as

$$\nabla g(\mathbf{x}^\star) \neq \mathbf{0}. \tag{13.9}$$

Why? Because $\nabla g(\mathbf{x}^\star) \neq \mathbf{0}$, at the point $\mathbf{x}^\star$ on the surface $g(x, y, z) = 0$, there exists the tangent plane whose normal vector is $\nabla g(\mathbf{x}^\star)$.

On the other hand, let $\mathcal{C} : \mathbf{x} = \mathbf{x}(t)$ be *any* continuously differentiable curve on the surface $g(x, y, z) = 0$ that passes through $\mathbf{x}^\star$ at $t = 0$, that is, $\mathbf{x}(0) = \mathbf{x}^\star$. Because $f(x, y, z)$ is minimized at $\mathbf{x}^\star$, the function of a single variable given by

$$\phi(t) \triangleq f(\mathbf{x}(t))$$

must have a local minimum at $t = 0$; hence, by the chain rule

$$0 = \dot{\phi}(0) = \nabla f(\mathbf{x}^\star) \bullet \dot{\mathbf{x}}(0). \tag{13.10}$$

Denote by $\mathbf{T} = \dot{\mathbf{x}}(0)$, the tangent vector to the curve $\mathcal{C}$ at the point $\mathbf{x}^\star$. Because of (13.10), $\nabla f(\mathbf{x}^\star)$ is orthogonal to every tangent vector to the surface $g(x, y, z) = 0$ at $\mathbf{x}(0)$; hence, $\nabla f(\mathbf{x}^\star)$ is either (a) the zero vector or (b) normal to the surface $g(x, y, z) = 0$ at $\mathbf{x}(0)$. In case (a), $\nabla f(\mathbf{x}^\star) = 0 \cdot \nabla g(\mathbf{x}^\star)$, and in case (b), $\nabla f(\mathbf{x}^\star)$ is parallel to $\nabla g(\mathbf{x}^\star)$. In either case, (a) or (b), there exists a scalar λ such that $\nabla f(\mathbf{x}^\star) = \lambda \nabla g(\mathbf{x}^\star)$, that is, (13.8) is true. □

We may refer to (13.9) as a technical requirement. Note that, in principle, it can be verified only after the candidate for the global minimizer has been found.

The method of Lagrange multipliers generalizes to $\mathbb{R}^n$:

Corollary 13.2

Suppose $\mathbf{x}$ is in $\mathbb{R}^n$, $n \geq 2$, and the constrained optimization problem

$$\left\{ \begin{array}{ll} \text{Minimize} & f(\mathbf{x}) \\ \text{Subject to} & g(\mathbf{x}) = 0 \end{array} \right\}$$

has a solution $\mathbf{x}^\star$. Then there exists a Lagrange multiplier λ such that

$$\nabla f(\mathbf{x}^\star) = \lambda \nabla g(\mathbf{x}^\star),$$

as long as

$$\nabla g(\mathbf{x}^\star) \neq \mathbf{0}.$$

Lemma 13.1

Suppose $f(\mathbf{x})$ is continuous on closed, bounded sets $\mathcal{D}$ and $\widetilde{\mathcal{D}}$ in $\mathbb{R}^n$. If $\mathcal{D} \subseteq \widetilde{\mathcal{D}}$, that is, $\mathcal{D}$ is contained in $\widetilde{\mathcal{D}}$, then

$$\max_{\mathbf{x} \text{ in } \mathcal{D}} f(\mathbf{x}) \leq \max_{\mathbf{x} \text{ in } \widetilde{\mathcal{D}}} f(\mathbf{x}).$$

Let's see how Theorem 13.10, Corollary 13.2, and Lemma 13.1 can help us solve a "real-world" problem.

Example 13.4

The U.S. Postal Services "Global Express Guaranteed" international shipping option will deliver a rectangular boxed package as long as its height is at least 5.5 in. and no more than 46 in., its width is no more than 35 in., its length is at least 9.5 in. and no more than 46 in., its weight is no more than 70 lb, and the sum of its girth, that is, the perimeter of the longitudinal side, and its length is no more than 108 in. Find the dimensions of the most voluminous box that can be shipped.

Method: Let $h, w,$ and ℓ be the height, width, and length of the box, in inches. A direct translation of the problem is to

$$\left\{\begin{array}{ll} \text{Maximize} & hw\ell \\ \\ \text{Subject to} & 5.5 \leq h \leq 46 \\ & 0 \leq w \leq 35 \\ & 9.5 \leq \ell \leq 46 \\ & 0 \leq 2w + 2h + \ell \leq 108 \end{array}\right\}. \tag{13.11}$$

Here, our closed, bounded domain is $\mathcal{D} = \{(h, w, \ell) : 5.5 \leq h \leq 46, 0 \leq w \leq 35, 9.5 \leq \ell \leq 46, 0 \leq 2h + 2w + \ell \leq 108\}$. Physically, it is nonsense to allow the width to be zero, but, mathematically, allowing w to be zero, puts our problem in the category discussed in Theorem 13.9.

We used the *Mathematica*^{TM} command

```
RegionPlot3D[2h + 2w + ℓ ≥ 0&&108 ≥ 2w+ 2h + ℓ, {h, 5.5, 46}, {w, 0, 35}, {ℓ, 9.5, 46},
  AxesLabel → {h, w, ℓ}, PlotStyle → Thick,LabelStyle → (FontSize → 14),
  ColorFunction → "LightTemperatureMap,"AxesStyle → Thickness[0.00315],
    PlotPoints → 100]
```

to produce the picture of $\mathcal{D}$ in Figure 13.5. As we see, the boundary of $\mathcal{D}$ consists of many pieces, so maximizing

$$f(h, w, \ell) \triangleq hw\ell$$

over $\mathcal{D}$ would at least involve, in principle, many constrained maximization problems, one for each 2D piece of the boundary of $\mathcal{D}$, as well as evaluating $f(h, w, \ell)$ at all critical points in the interior of $\mathcal{D}$. And, at least one of those 2D pieces of the boundary of $\mathcal{D}$ itself has many 1D line segments bounding it.

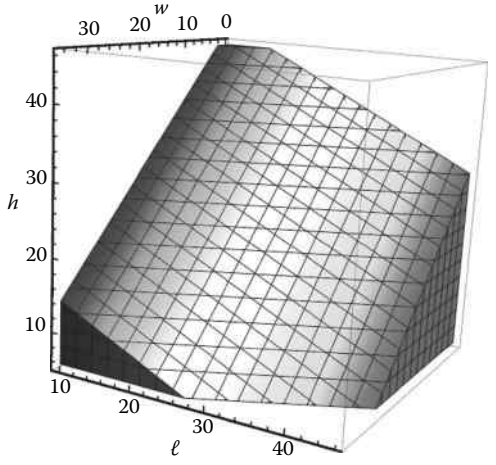


FIGURE 13.5
Example 13.4.

Instead, we will maximize $f(h,w,\ell)$ over the tetrahedron

$$\widetilde{\mathcal{D}} \triangleq \{(h,w,\ell):\ 0 \le h \le 54, 0 \le w \le 54, 0 \le \ell \le 108, 0 \le 2h+2w+\ell \le 108\}$$

by evaluating f at all of its critical points in the interior of $\widetilde{\mathcal{D}}$ and maximizing f on the boundary of $\widetilde{\mathcal{D}}$. It will turn out that the maximum of $f(h,w,\ell)$ over $\widetilde{\mathcal{D}}$ occurs at $(h^\star,w^\star,\ell^\star) = (18,18,36)$. Because the maximizer of f on $\widetilde{\mathcal{D}}$ actually is a point in $\mathcal{D}$ and Lemma 13.1 says that the maximum of f on $\mathcal{D}$ is less than or equal to the maximum of f on $\widetilde{\mathcal{D}}$, we conclude that the maximum of f on $\mathcal{D}$ equals the maximum of f on $\widetilde{\mathcal{D}}$ and occurs at $(h^\star,w^\star,\ell^\star) = (18,18,36)$, too.

The boundary of the tetrahedron consists of three different triangles in the planes $h = 0$, $w = 0$, and $\ell = 0$ *and* part of the plane $2h^\star + 2w^\star + \ell^\star = 108$. But, we know the maximizer can't be on any of the planes $h = 0$, $w = 0$, and $\ell = 0$ because the volume of the box is zero there.

We claim that a maximizer $\mathbf{x}^\star = (h^\star,w^\star,\ell^\star)$ must satisfy $2h^\star + 2w^\star + \ell^\star = 108$: because, if not, that is, $2h^\star + 2w^\star + \ell^\star < 108$, then it cannot be true that $h^\star = 46$, $w^\star = 35$, *and* $\ell^\star = 46$. So, we could increase the height to be $h > h^\star$ or the width to be $w > w^\star$ or the length $\ell > \ell^\star$ to get a box with a higher volume and still satisfy the constraint $2w+2h+\ell \le 108$. So, we may assume that $2h^\star + 2w^\star + \ell^\star = 108$. This gives us a problem of the form (13.7):

$$\left\{\begin{array}{ll}\text{Maximize} & hw\ell \\ \\ \text{Subject to} & 2w+2h+\ell = 108\end{array}\right\}, \tag{13.12}$$

along with the additional requirements that $0 \le h \le 54$, $0 \le w \le 54$, and $0 \le \ell \le 108$.

With $f(h,w,\ell) = hw\ell$ and $g(h,w,\ell) = 2h + 2w + \ell - 108$, the Lagrange multiplier theorem, 13.10, tells us to solve

$$\left\{\begin{array}{ll} w\ell\,\hat{\mathbf{w}} + h\ell\,\hat{\mathbf{h}} + hw\,\hat{\boldsymbol{\ell}} = \nabla f = \lambda \nabla g & = \lambda\left(2\,\hat{\mathbf{w}} + 2\,\hat{\mathbf{h}} + \hat{\boldsymbol{\ell}}\right) \\ \\ 2h + 2w + \ell & = 108 \end{array}\right\};$$

hence, $w\ell = 2\lambda$, $h\ell = 2\lambda$, and $hw = \lambda$, along with $2h + 2w + \ell = 108$. The first two equations imply $w\ell = 2\lambda = h\ell$, so $\ell > 0$ implies $h = w$. Substituting into the girth plus

length equation gives $2h + 2w + \ell = 108$; hence,

$$\ell = 108 - 4w.$$

Also, $\lambda = hw = w^2$. Substituting the latter two equations into $w\ell = 2\lambda$ gives

$$w(108 - 4w) = 2w^2;$$

hence, $0 = w(6w - 108)$. This implies $18 = w = h$ and $\ell = 36$. The fact that $\lambda = 18^2$ is irrelevant.

The critical points of the function $f(h, w, \ell) = hw\ell$ lie only on the lines $h = w = 0$, $w = \ell = 0$, and $w = \ell = 0$, as you will justify in Problem 13.2.3.15. Because $f(18, 18, 36) > 0$, it follows that the maximum of f is at $(h^\star, w^\star, \ell^\star) = (18, 18, 36)$. The most voluminous box that can be shipped this way has those dimensions and has a volume of $\frac{27}{4}\,ft^3$. ○

Theorem 13.11

(Lagrange multipliers) Suppose a constrained minimization problem

$$\left\{\begin{array}{ll} \text{Minimize} & f(\mathbf{x}) \\ \\ \text{Subject to} & f_1(\mathbf{x}) = 0 \\ & f_2(\mathbf{x}) = 0 \end{array}\right\} \tag{13.13}$$

has solution $\mathbf{x}^\star = (x^\star, y^\star, z^\star)$. If $\nabla f_1(\mathbf{x}^\star) \neq \mathbf{0}$ or $\nabla f_2(\mathbf{x}^\star) \neq \mathbf{0}$, then there exist **Lagrange multipliers** λ, μ such that

$$\nabla f(\mathbf{x}^\star) = \lambda \nabla f_1(\mathbf{x}^\star) + \mu \nabla f_2(\mathbf{x}^\star). \tag{13.14}$$

Example 13.5

Find the point on the curve $y = x^2 - x$ closest to the curve $y = \frac{1}{2}x - 6$.

Method: The square of the distance between a point (x, y) on the curve $x^2 - x - y = 0$ and a point on (X, Y) on the curve $12 - X + 2Y = 0$ is $(X - x)^2 + (Y - y)^2$. We consider X, Y, x, y to be the variables of the problem, which we state as

$$\left\{\begin{array}{ll} \text{Minimize} & (X - x)^2 + (Y - y)^2 \\ \\ \text{Subject to} & x^2 - x - y = 0 \\ & 12 - X + 2Y = 0 \end{array}\right\}. \tag{13.15}$$

Note that the distance is minimized exactly where the square of the distance is minimized.

Define $f(X, Y, x, y) \triangleq (X - x)^2 + (Y - y)^2$, $f_1(X, Y, x, y) \triangleq x^2 - x - y$, $f_2(X, Y, x, y) \triangleq 12 - X + 2Y$ and let λ, μ be the Lagrange multipliers. $[\,X \quad Y \quad x \quad y\,]^T$ is a point in $\mathbb{R}^4$, and let $\hat{\boldsymbol{I}}, \hat{\boldsymbol{J}}, \hat{\boldsymbol{\imath}}, \hat{\boldsymbol{\jmath}}$ be the corresponding unit basis vectors. So,

$$\nabla = \hat{\boldsymbol{I}}\frac{\partial}{\partial X} + \hat{\boldsymbol{J}}\frac{\partial}{\partial Y} + \hat{\boldsymbol{\imath}}\frac{\partial}{\partial x} + \hat{\boldsymbol{\jmath}}\frac{\partial}{\partial y}.$$

The candidates (X, Y, x, y) for the minimization and the Lagrange multipliers λ, μ satisfy the six equations

$$\nabla f = \lambda \nabla f_1 + \mu \nabla f_2 \quad \text{and} \quad f_1(X, Y, x, y) = f_2(X, Y, x, y) = 0,$$

that is,

$$\begin{bmatrix} 2(X-x) \\ 2(Y-y) \\ -2(X-x) \\ -2(Y-y) \end{bmatrix} = \lambda \begin{bmatrix} 0 \\ 0 \\ 2x-1 \\ -1 \end{bmatrix} + \mu \begin{bmatrix} -1 \\ 2 \\ 0 \\ 0 \end{bmatrix}, \quad x^2 - x - y = 0, \quad 12 - X + 2Y = 0.$$

The first four equations imply

$$(1)\ \ \lambda(2x-1) = -2(X-x) = \mu \quad \text{and} \quad (2)\ \ \lambda = 2(Y-y) = 2\mu;$$

hence,

$$0 = 2\mu - 2\mu = 2\lambda(2x-1) - \lambda = \lambda\big(2(2x-1) - 1\big).$$

It follows that either $\lambda = 0$ or $x = \frac{3}{4}$.

In the case when $x = \frac{3}{4}$, $y = x^2 - x = -\frac{3}{16}$ so that (X, Y, x, y) satisfies the constraint $0 = f_1(X, Y, x, y)$.

On the other hand, suppose $\lambda = 0$. Then (1) and (2) imply $X = x$ and $Y = y$, which is only possible if the two curves $x^2 - x = y$ and $y = \frac{1}{2}x - 6$ intersect. But, they don't, because $x^2 - x = y = \frac{1}{2}x - 6$ implies that x must satisfy the quadratic equation $x^2 - \frac{3}{2}x + 6 = 0$, which has no real number solution.

So, $(x, y) = (\frac{3}{4}, -\frac{3}{16})$ is the point on the curve $y = x^2 - x$ closest to the curve $y = \frac{1}{2}x - 6$. ◯

In Problem 13.2.3.13, you will continue the work of Example 13.5 to conclude that the closest approach of the two curves is where $(x, y) = (\frac{3}{4}, -\frac{3}{16})$ and $(X, Y) = \left(\frac{117}{40}, -\frac{363}{80}\right)$, and the minimum distance is $\frac{87\sqrt{5}}{40}$.

The results of Example 13.5 are illustrated in Figure 13.6.

Example 13.6

Find where the curves $xy - x + 2y = 0$ and $x - 2y + 6 = 0$ are closest to each other.

Method: $(X-x)^2 + (Y-y)^2$ is the square of the distance between a point (x, y) on the curve $xy - x + 2y = 0$ and a point on (X, Y) on the curve $X - 2Y + 6 = 0$. We consider X, Y, x, y to be the variables of the problem, which we state as

$$\left\{ \begin{array}{ll} \text{Minimize} & (X-x)^2 + (Y-y)^2 \\ \\ \text{Subject to} & xy - x + 2y = 0 \\ & X - 2Y + 6 = 0 \end{array} \right\}. \tag{13.16}$$

Define $f(X, Y, x, y) \triangleq (X-x)^2 + (Y-y)^2$, $f_1(X, Y, x, y) \triangleq xy - x + 2y$, $f_2(X, Y, x, y) \triangleq X - 2Y + 6$ and let λ, μ be the Lagrange multipliers.

As in Example 13.5, $[\, X \quad Y \quad x \quad y \,]^T$ is a point in $\mathbb{R}^4$ and

$$\nabla = \hat{\boldsymbol{I}} \frac{\partial}{\partial X} + \hat{\boldsymbol{J}} \frac{\partial}{\partial Y} + \hat{\boldsymbol{\imath}} \frac{\partial}{\partial x} + \hat{\boldsymbol{\jmath}} \frac{\partial}{\partial y}. \tag{13.17}$$

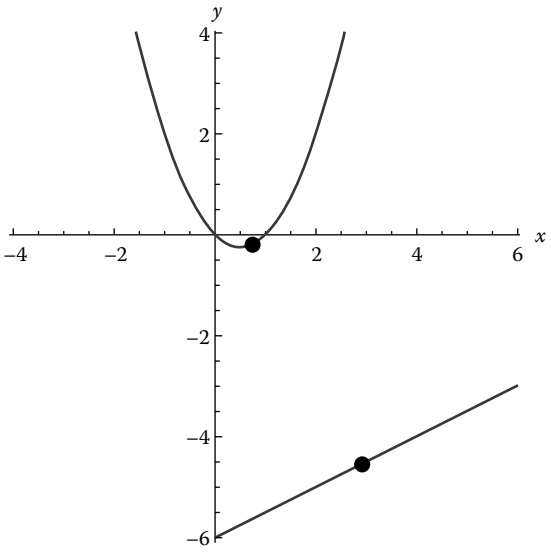


FIGURE 13.6
Example 13.5.

The candidates (X, Y, x, y) for the minimization and the Lagrange multipliers satisfy the six equations

$$\nabla f = \lambda \nabla f_1 + \mu \nabla f_2 \quad \text{and} \quad f_1(X, Y, x, y) = f_2(X, Y, x, y) = 0,$$

that is,

$$\begin{bmatrix} 2(X-x) \\ 2(Y-y) \\ -2(X-x) \\ -2(Y-y) \end{bmatrix} = \lambda \begin{bmatrix} 0 \\ 0 \\ y-1 \\ x+2 \end{bmatrix} + \mu \begin{bmatrix} 1 \\ -2 \\ 0 \\ 0 \end{bmatrix}, \quad xy - x + 2y = 0, \quad X - 2Y + 6 = 0.$$

The first four equations imply

$$(1)\ \ \lambda(y-1) = -2(X-x) = -\mu \quad \text{and} \quad (2)\ \ \lambda(x+2) = -2(Y-y) = 2\mu;$$

hence,

$$0 = 2\mu - 2\mu = \lambda(x+2) + 2(\lambda(y-1)) = \lambda\big(2(y-1) + (x+2)\big).$$

It follows that either $\lambda = 0$ or $2(y-1) + (x+2) = 0$.

In the latter case, $x = -2y$; substitute that into $f_1 = 0$ to get $(-2y)y - (-2y) + 2y = 0$; hence, $0 = 2y(y-2)$. So, our candidates so far have either (a) $y = 0$, in which case $x = 0$, or (b) $y = 2$, in which case $x = -4$.

Substitute $x = y = 0$ into (1) and (2) to get

$$2\mu = 4X \quad \text{and} \quad 2\mu = -2Y;$$

hence, $Y = -2X$. Substitute that into $f_2 = 0$ to get $0 = X - 2Y + 6 = X - 2(-2X) + 6$; hence, $X = -\frac{6}{5}$ and thus $Y = \frac{12}{5}$. The square of the distance between the respective points $(X, Y) = (-\frac{6}{5}, \frac{12}{5})$, $(x, y) = (0, 0)$ is $f(-\frac{6}{5}, \frac{12}{5}, 0, 0) = \frac{36}{5}$.

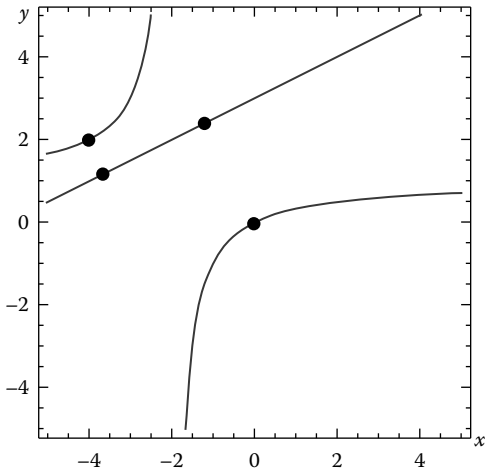


FIGURE 13.7
Example 13.6.

Substitute $x = -4, y = 2$ into (1) and (2) to get

$$2\mu = 4(X+4) = 4X + 16 \quad \text{and} \quad 2\mu = -2(Y-2) = -2Y + 4;$$

hence, $-2Y + 4 = 4X + 16$, and hence, $Y = -2X - 6$. Substitute that into $f_2 = 0$ to get $0 = X - 2Y + 6 = X - 2(-2X - 6) + 6$; hence, $X = -\frac{18}{5}$ and thus $Y = \frac{6}{5}$. The square of the distance between the respective points $(X, Y) = (-\frac{18}{5}, \frac{6}{5})$, $(x, y) = (-4, 2)$ is $f(-\frac{18}{5}, \frac{6}{5}, -4, 2) = \frac{4}{5}$.

On the other hand, suppose $\lambda = 0$. Then (1) and (2) imply $X = x$ and $Y = y$, which is only possible if the two curves $xy - x + 2y = 0$ and $x - 2y + 6 = 0$ intersect. But, they don't, because $x = 2y - 6$ substituted into the equation of the first curve gives $0 = (2y - 6)y - (2y - 6) + 2y = 2(y^2 - 3y + 3)$, which has no real solution.

So, the only *candidates* for the closest points are (a) $(X, Y) = (-\frac{6}{5}, \frac{12}{5}), (x, y) = (0, 0)$, and (b) $(X, Y) = (-\frac{18}{5}, \frac{6}{5})$, $(x, y) = (-4, 2)$. The latter gives the closest approach, and the minimum distance between the curves is $\frac{2}{\sqrt{5}}$. ○

The results of Example 13.6 are illustrated in Figure 13.7, which shows that a solution of the "Lagrange equations" (13.14) may not be a global minimizer.

13.2.2 Numerical Minimization and Steepest Descent Methods

Suppose we want to find the minimum (or maximum) value that $f = f(\mathbf{x})$ has on domain $\mathcal{D} = \mathbb{R}^n$. We will briefly discuss the ideas behind one numerical method.

Suppose the level curves of a function $f(\mathbf{x})$ are as shown in Figure 13.8a. Suppose $\mathbf{x}_0$ is our first guess for a global minimizer of f. We recall that at the point $\mathbf{x}_0$ on a level set $f(\mathbf{x}) = k$, the negative of the gradient vector, $\mathbf{g} \triangleq -\nabla f(\mathbf{x}_0)$, points in the direction of the greatest decrease of f. Let $\widehat{\mathbf{g}}$ denote the unit vector in the direction of $\mathbf{g}$. Our goal is to minimize f, so a numerical search along the normal line segment

$$\mathbf{x}(\alpha) \triangleq \mathbf{x}_0 + \alpha\,\widehat{\mathbf{g}},\ 0 \le \alpha \le \delta$$

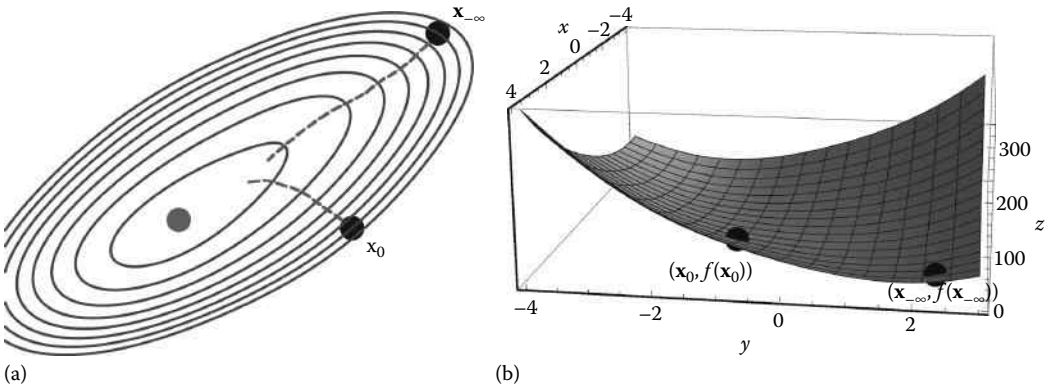


FIGURE 13.8
(a) Method of steepest descent, (b) Surface $z = f(x, y)$.

should help: Solve the single variable calculus problem

$$f\left(\mathbf{x}(\alpha^\star)\right) = \min_{0 \le \alpha \le \delta} f\left(\mathbf{x}(\alpha)\right), \tag{13.18}$$

and take as our next guess for the global minimizer $\mathbf{x}_1 \triangleq \mathbf{x}(\alpha^\star)$. By the way, the single variable calculus problem (13.18) is called a **line search**.

What we described previously is called the **method of steepest descents**. In principle, it is a reliable, cautious method. But it can be very slow because it ignores some of the geometry. Suppose that we had chosen as our initial guess the point $\mathbf{x}_{-\infty}$ also shown in Figure 13.8a. Look also at the graph of f in Figure 13.8b. It would take many iterations of the steepest descent method to get from $\mathbf{x}_{-\infty}$ to $\mathbf{x}^\star$ on a path down a long, gently sloping valley.

The **conjugate gradient** method modifies the steepest descent method by replacing the normal line's choice $\mathbf{x}(\alpha)$ with many choices of $\mathbf{x}(\alpha) = \mathbf{x}_0 + \alpha\,\widehat{\mathbf{p}}_j$, $j = 1, \ldots$. The conjugate gradient method enables us to behave more like a mountain goat than a weekend hiker.

Learn More About It

Numerical Recipes: The Art of Scientific Computing, 3rd edition, by William H. Press et al., Cambridge University Press © 2007 is a good resource for learning about the method of steepest descents and the conjugate gradient method.

13.2.3 Problems

In problems 1–3, (a) find *all* of the critical points, and (b) explain why each of the critical points is a local maximum, local minimum, or saddle point or cannot be classified by the second derivative test.

1. $f(x,y) = x^2 + xy^2 + 3y^2$
2. $f(x,y) = (2 - x^2)(1 - y^2)$
3. $f(x,y) = y^3 + 3x^2y - 3x^2 - 3y^2$
4. For $f(x,y) = xe^{-x^2}\sin^2 y$, defined only on the domain $\{(x,y) : -1 \le x \le 1, -\frac{\pi}{4} \le y \le \frac{5\pi}{4}\}$, find *all* of the critical points. Do *not* classify them into the categories of local maximum, local minimum, or saddle point.
5. Find the absolute maximum value of $f(x,y) = 4 + x - x^2 - 2y^2$ over the set of (x,y) satisfying $x^2 + y^2 = 2$.
6. Find the absolute maximum value and absolute minimum values of $f(x,y) = x^2 + 2y^2$ over the domain $0 \le 3x^2 + y^2 \le 6$.
7. Find the absolute maximum value and absolute minimum values of $f(x,y) = 2x^2 + y^2$ over the domain $0 \le x^2 + y^2 \le 4$.
8. Show that $f(x,y) \triangleq e^{-x}\sin(x+y)$ has no critical points, but that $f(x,y)$ takes on all values between $-\infty$ and ∞.
9. If a symmetric 2×2 matrix A has negative determinant, explain why the matrix has one negative and one positive eigenvalue. [Hint: $| A - \lambda I | = \begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{12} & a_{22} - \lambda \end{vmatrix} = (\lambda_1 - \lambda)(\lambda_2 - \lambda) = \lambda^2 - (a_{11} + a_{22})\lambda + |A|$.]
10. In solving the least squares solution problem in Section 2.5, we used linear algebra to conclude that if $\mathbf{x}^\star$ is the global minimizer of $f(\mathbf{x}) \triangleq ||A\mathbf{x} - \mathbf{b}||^2$, then necessarily $\mathbf{x}^\star$ solves the normal equations $A^TA\mathbf{x} = A^T\mathbf{b}$. Now, using the theory of Section 13.2, re-derive this result that a global minimum of $f(\mathbf{x})$ on domain $\mathbb{R}^n$ must satisfy the normal equations.
11. Suppose W is a real, symmetric, positive definite $n \times n$ matrix. As in Problem 2.6.3.17, define $\langle \mathbf{x}, \mathbf{y}\rangle_W \triangleq \langle W\mathbf{x}, \mathbf{y}\rangle \triangleq \mathbf{x}^T W^T\mathbf{y}$ and $|| \mathbf{x} ||_W \triangleq \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle_W}$. Suppose real, $m \times n$ matrix A and vector $\mathbf{b}$ in $\mathbb{R}^m$ are given.

    Consider the problem of finding $\mathbf{x}^\star$ that minimizes $f(\mathbf{x}) \triangleq ||A\mathbf{x} - \mathbf{b}||_W^2$. We called this a **generalized** or **weighted least squares** problem. Use the theory of Section 13.2 to re-derive the result that if $\mathbf{x}^\star$ is a global minimum of $f(\mathbf{x})$ on domain $\mathbb{R}^n$, then $\mathbf{x}^\star$ must satisfy the **generalized normal equations**

$$A^T W^T A\mathbf{x} = A^T W^T\mathbf{b}.$$

12. Derive that the function $f(x,y) \triangleq e^{-x}\sin(x+y)$ has no critical points. Also, find points at which (a) $\frac{\partial f}{\partial x} > 0$ and $\frac{\partial f}{\partial y} > 0$, (b) $\frac{\partial f}{\partial x} > 0$ and $\frac{\partial f}{\partial y} < 0$, (c) $\frac{\partial f}{\partial x} < 0$ and $\frac{\partial f}{\partial y} > 0$, (d) $\frac{\partial f}{\partial x} < 0$ and $\frac{\partial f}{\partial y} < 0$.
13. Continue the work of Example 13.5 to conclude that the closest approach of the two curves is where $(x,y) = (\frac{3}{4}, -\frac{3}{16})$ and $(X,Y) = \left(\frac{117}{40}, -\frac{363}{80}\right)$ and the minimum distance is $\frac{87\sqrt{5}}{40}$.
14. Derive that the function $k(x,y) \triangleq 5xe^y - x^5 - e^{5y}$ has a local maximum at $(x,y) = (1,0)$, but the function does not have a global maximum there despite the fact that the function has no other critical point.
15. Why do the critical points of the function $f(h,w,\ell) = hw\ell$ lie only on the lines $h = w = 0$, $w = \ell = 0$, and $\ell = h = 0$?

16. (Small project) In the paper by Aziz mentioned in Section 13.1, he also studied the case of pure radiative heat transfer from an insulated cylinder to a large medium at temperature T_∞. Define dimensionless quantities

$$\phi_0 \triangleq \frac{T_o}{T_i},\ \phi_\infty \triangleq \frac{T_\infty}{T_i},\ P \triangleq \frac{r_o}{r_i},\ B_i \triangleq \frac{\epsilon\sigma T_i^3 r_i}{k},$$

the latter being the Biot number. Given a B_i and ϕ_∞, the problem of finding the maximum heat transfer rate reduces to finding the ϕ_0 and P to

$$\left\{\begin{array}{ll}\text{Maximize} & Q \triangleq (1-\phi_0)/\ln P \\ \\ \text{Subject to} & 1-\phi_0 - B_i P \ln P\left(\phi_0^4 - \phi_\infty^4\right) = 0\end{array}\right\}.$$

Using realistic values of B_i and ϕ_∞ from, for example, Example 4 of Aziz's paper, maximize Q.

17. Suppose $\mathbf{c}$ is a constant vector. Do the calculations to explain why $\nabla\left(\mathbf{c}^T\mathbf{x}\right) = \mathrm{c}$.

13.3 Linear Programming Problems

If we are manufacturing things of various sizes and shapes, we need various raw materials, machines, and labor tasks to make them. The allocation of resources naturally leads to optimization problems.

In a **linear programming** (LP) problem, suppose we have quantities $x_1, \ldots, x_k$ of things that are needed. These quantities are assumed to be nonnegative, that is, $x_i \geq 0$ for $i = 1, \ldots, k$. We will model the demand for these quantities, or limitations on their availability, using **linear inequalities** of the form $d_1x_1 + \cdots + d_kx_k \leq b$ or $d_1x_1 + \cdots + d_kx_k \geq b$ or linear equations of the form $d_1x_1 + \cdots + d_kx_k = b$.

The last part of an LP problem is that we have a function f that is to be minimized or maximized. We assume that the objective function f is *also* linear, that is, $f(x_1, \ldots, x_k) = c_1x_1 + \cdots + c_kx_k$.

Here is an example of an LP problem:

$$\left\{\begin{array}{ll}\text{Maximize} & x + 2y \\ \\ \text{Subject to} & \begin{array}{l} x + 6y \leq 24 \\ 23x + 12y \leq 92 \\ x \geq 0,\ y \geq 0 \end{array}\end{array}\right\}. \tag{13.19}$$

In Figure 13.9, we see a drawing of the set

$$C \triangleq \{(x, y) :\ x + 6y \leq 24,\ 23x + 12y \leq 92,\ x \geq 0,\ y \geq 0\}.$$

The dotted lines are of the form $x+2y = k$, where k is a constant. A graphical solution of this LP problem is found by drawing those lines to get closer and closer until one just touches C.

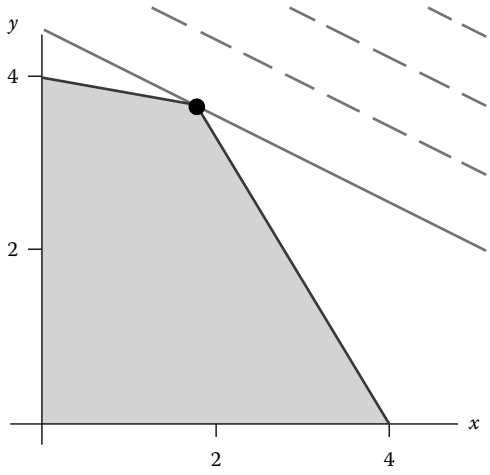


FIGURE 13.9
Convex set and lines $x + 2y = k$.

The process of getting closer and closer to C is carried out by taking the constant k smaller and smaller until at $k = k^\star$; there is a point $(x^\star, y^\star)$ in C where $2x^\star + y^\star = k^\star$.

Example 13.7

The Acme Company manufactures windows in model numbers 1 through k. Model #j requires g_j cm^2 of glass, A_j *grams* of aluminum, w_j cm^3 of wood, v_j cm^2 of vinyl plastic, and L_j hours of labor. Suppose the costs of glass, aluminum, wood, vinyl, and labor are c_g, c_A, c_w, c_v, and c_L, in dollars per appropriate unit. Suppose the company has a business plan that expects it to produce at least 800 windows per week. (If the company falls below this level of production, an investment bank may declare them in violation of their loan covenant and demand immediate repayment.) The plan also demands that they manufacture at least m_j of model #j per week, where each $m_j \geq 0$. (Some of the window models are not bought often so some of the m_j may be zero.) Assume that the company president has declared a freeze on hiring and overtime so that the total number of hours of labor available is at most 650, and assume that the company has a long term contract with a supplier that demands that it buy at least 150 kg of aluminum per week. If the Acme Company gets paid p_j dollars for each window model #j produced, formulate the problem of maximizing its profit.

Method: By the way, the amount of glass used per window by model # may take into account the average waste of glass due to breakage. Our model has several unrealistic assumptions by omission. For example, we are assuming that every week is the same and that the company sells every window it makes. In spite of this, let's formulate the problem as best we can using the given information.

Let x_j be the amount of window model #j produced per week. The minimum demand for windows, by model, translates to $x_j \geq m_j$ for $j = 1, \ldots, k$, and the business plan requires $x_1 + \cdots + x_k \geq 800$.

The total amount of glass used per week is $g_1x_1 + \cdots + g_kx_k$, and there are similar calculations for aluminum, wood, vinyl, and labor. The latter requires $L_1x_1 + \cdots + L_kx_k \leq 650$ because of the labor freeze. The total cost of the glass used in making one model #j window is $c_g g_j$, and there are similar costs for the other resources.

Our problem can be formulated as

$$\left\{\begin{array}{ll} \text{Maximize} & \sum_{j=1}^{n}\left(p_j - c_g g_j - c_A A_j - c_w w_j - c_v v_j - c_L L_j\right) x_j \\ \text{Subject to} & x_1 + \quad \cdots \quad + x_k \quad \geq 800 \\ & L_1 x_1 + \quad \cdots \quad + L_k x_k \leq 650 \\ & A_1 x_1 + \quad \cdots \quad + A_k x_k \geq 1.5 \times 10^5 \\ & \qquad x_j \quad \geq m_j,\ j = 1, \ldots, k. \end{array}\right\}. \bigcirc \tag{13.20}$$

An LP problem in "preliminary standard form" has the form

$$\left\{\begin{array}{ll} \text{Minimize} & c_1 x_1 + \cdots + c_k x_k \\ \text{Subject to} & a_{11} x_1 + \cdots + a_{1k} x_k \leq b_1 \\ & \qquad\qquad \vdots \qquad\qquad \vdots \\ & a_{m1} x_1 + \cdots + a_{mk} x_k \leq b_m \\ & \qquad x_1, \ldots, x_k \geq 0. \end{array}\right\}. \tag{13.21}$$

The letter "c" stands for the word "cost."

If our original problem has an inequality $d_1 x_1 + \cdots + d_k x_k \geq \gamma$, then we can change it to the equivalent inequality $-d_1 x_1 - \cdots - d_k x_k \leq -\gamma$. Similarly, if our original problem asked us to maximize a function $g(\mathbf{x}) = \gamma_1 x_1 + \cdots + \gamma_k x_k$, then we change it to the equivalent objective of minimizing $f(\mathbf{x}) \triangleq -g(\mathbf{x}) = -\gamma_1 x_1 - \cdots - \gamma_k x_k$. So, one theory and methods for solving (13.21) will be enough to solve many different types of LP problems.

13.3.1 Slack Variables and Standard Form

We learned how to solve linear systems of algebraic *equations* in Section 1.1. Some very clever people realized that in solving LP problems, we can take advantage of row reduction techniques by converting an inequality to an equality by introducing a **slack variable**: if $a_{11} x_1 + \cdots + a_{1k} x_k \leq b_1$, then there is a *nonnegative number* ε such that $a_{11} x_1 + \cdots + a_{1k} x_k + \varepsilon = b_1$, specifically $\varepsilon = b_1 - (a_{11} x_1 + \cdots + a_{1k} x_k)$. So, we introduce the slack variable x_{k+1}, and get the equation $a_{11} x_1 + \cdots + a_{1k} x_k + x_{k+1} = b_1$ along with the constraint that $x_{k+1} \geq 0$; the latter is the same as the nonnegativity constraint on the original variables $x_1, \ldots, x_k$.

Definition 13.9

An LP problem is in **standard form** if it is written as

$$\left\{\begin{array}{ll} \text{Minimize} & c_1 x_1 + \cdots + c_k x_k \\ \text{Subject to} & a_{11} x_1 + \cdots + a_{1k} x_k = b_1 \\ & \qquad\qquad \vdots \qquad\qquad \vdots \\ & a_{m1} x_1 + \cdots + a_{mk} x_k = b_m \\ & \qquad x_1, \ldots, x_k \geq 0. \end{array}\right\}. \tag{13.22}$$

Example 13.8

Put into standard form the LP problem

$$\left\{\begin{array}{lrcl} \text{Minimize} & 5x_1 + x_2 + 2x_3 & & \\ & & & \\ \text{Subject to} & x_1 + x_2 + x_3 & \geq & 20 \\ & 2x_1 + x_2 - x_3 & \leq & 30 \\ & 3x_1 + x_2 + x_3 & = & 60 \\ & x_1, \ldots, x_3 \geq 0 & & \end{array}\right\}.$$

Method: First, put the first inequality into the preliminary standard form, that is, $-x_1 - x_2 - x_3 \leq -20$. Next, introduce slack variables into the two inequalities. The LP problem in standard form is

$$\left\{\begin{array}{lrcl} \text{Minimize} & 5x_1 + x_2 + 2x_3 & & \\ & & & \\ \text{Subject to} & -x_1 - x_2 - x_3 + x_4 & = & -20 \\ & 2x_1 + x_2 - x_3 + x_5 & = & 30 \\ & 3x_1 + x_2 + x_3 & = & 60 \\ & x_1, \ldots, x_5 \geq 0 & & \end{array}\right\}. \bigcirc \tag{13.23}$$

Denoting

$$\mathbf{c} = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}, \quad A = \begin{bmatrix} a_{11} & . & . & . & a_{1n} \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ a_{m1} & . & . & . & a_{mn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ . \\ . \\ . \\ b_m \end{bmatrix},$$

we can rewrite (13.22), an LP in standard form, in "compact, standard form,"

$$\left\{\begin{array}{ll} \text{Minimize} & \mathbf{c}^T\mathbf{x} \\ & \\ \text{Subject to} & A\mathbf{x} = \mathbf{b}, \ \ \mathbf{x} \geq \mathbf{0} \end{array}\right\}, \tag{13.24}$$

where
$\mathbf{x}$ and $\mathbf{c}$ are in $\mathbb{R}^n$
$\mathbf{b}$ are in $\mathbb{R}^m$
A is a real, $m \times n$ matrix

Here, the notation $(x_1, \ldots, x_n) = \mathbf{x} \geq \mathbf{0}$ means $x_j \geq 0$ for $j = 1, \ldots, n$.

Definition 13.10

Suppose (13.24) is an LP in standard form. Then $\mathbf{x}$ is defined to be a

(a) **Feasible solution** if $A\mathbf{x} = \mathbf{b}$ and $\mathbf{x} \geq \mathbf{0}$;
(b) **Basic feasible solution** if $A\mathbf{x} = \mathbf{b}$, $\mathbf{x} \geq \mathbf{0}$, and no more than m of the x_j are nonzero.

Example 13.9

Find a basic feasible solution of the LP problem of Example 13.8.

Method: The augmented matrix, with variables labeled above it, is

$$\begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_5 & & \end{matrix} \\ \left[\begin{array}{rrrrr|r} -1 & -1 & -1 & 1 & 0 & -20 \\ 2 & 1 & -1 & 0 & 1 & 30 \\ 3 & 1 & 1 & 0 & 0 & 60 \end{array}\right] \end{array} \begin{array}{c} \sim \\ {\scriptstyle R_3+R_2\to R_2} \\ {\scriptstyle R_3+R_1\to R_1} \end{array} \begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_5 & \end{matrix} \\ \left[\begin{array}{rrrrr|r} 2 & 0 & 0 & 1 & 0 & 40 \\ 5 & 2 & 0 & 0 & 1 & 90 \\ 3 & 1 & 1 & 0 & 0 & 60 \end{array}\right] \end{array}.$$

So, we have a basic feasible solution $(x_1, x_2, x_3, x_4, x_5) = (0, 0, 60, 40, 90)$. ◯

Example 13.10

We want to find the proportions of wheat bran, oat flour, and rice flour to be mixed to produce emergency rations. Nutritional information* about these cereals is in Table 13.1. Assume the mixture should have, per 100 g, at least 370 kcal, 12 g of protein, and 2 g polyunsaturated fat. Set up the linear programming problem, in standard form, to minimize the grams of carbohydrates per 100 g of mixture, and find a basic feasible solution.

Method: Let x_1, x_2, and x_3 be the *proportions* of wheat bran, oat flour, and rice flour used in the mixture. It follows that $0 \le x_i \le 1$ for $i = 1, 2, 3$ and $x_1 + x_2 + x_3 = 1$.

In the mixture, the energy in *kcal* requirement means that $216x_1 + 404x_2 + 363x_3 \ge 370$. The protein requirement means that $15.55x_1 + 14.66x_2 + 7.23x_3 \ge 12$. The dietary fiber requirement means that $42.8x_1 + 6.5x_2 + 4.6x_3 \ge 12$. The polyunsaturated fat requirement means that $2.212x_1 + 3.329x_2 + 0.996x_3 \ge 2$. The amount of carbohydrates per 100 g of mixture is $f = 64.51x_1 + 65.70x_2 + 76.48x_3$.

This directly translates to

$$\left\{\begin{array}{lrcrcrcl} \text{Minimize} & 64.51x_1 & + & 65.70x_2 & + & 76.48x_3 & & \\ \text{Subject to} & 216x_1 & + & 404x_2 & + & 363x_3 & \ge & 370 \\ & 15.55x_1 & + & 14.66x_2 & + & 7.23x_3 & \ge & 12 \\ & 2.212x_1 & + & 3.329x_2 & + & 0.996x_3 & \ge & 2 \\ & x_1 & + & x_2 & + & x_3 & = & 1 \\ & x_1, \ldots, x_3 \ge 0 & & & & & & \end{array}\right\}.$$

TABLE 13.1

Nutrients per 100 g

	Wheat Bran	**Oat Flour**	**Rice Flour**
Kilocalories(kcal)	216	404	363
g Protein	15.55	14.66	7.23
g Carbs	64.51	65.70	76.48
g Dietary fiber	42.8	6.5	4.6
g Polyunsaturated fats	2.212	3.329	0.996

* From the document sr22fg20.pdf, "CerealGrains and Pasta," available from the USDA website, ⟨http://www.ars.usda.gov/ Services/docs.htm?docid=18878⟩.

The requirement $x_i \leq 1$ for $i = 1, 2, 3$ will be automatically satisfied because $x_1 + x_2 + x_3 = 1$ and $x_1, x_2, x_3 \geq 0$.

Turn all of the $\geq$ inequalities to $\leq$ inequalities and introduce slack variables to get the equivalent problem

$$\left\{\begin{array}{lrrrrrrl} \text{Minimize} & 64.51x_1 + 65.70x_2 + 76.48x_3 & & & & & & \\ \text{Subject to} & -\ 216x_1 & -\ 404x_2 & -\ 363x_3 & +x_4 & & & = -370 \\ & -15.55x_1 & -14.66x_2 & -\ 7.23x_3 & & +x_5 & & = -12 \\ & -2.212x_1 & -3.329x_2 & -0.996x_3 & & & +x_6 & = -2 \\ & x_1 & +\ x_2 & +\ x_3 & & & & = 1 \\ & x_1, \ldots, x_6 \geq 0. & & & & & & \end{array}\right\}. \bigcirc$$

In solving LP problems by hand, people have used a tableau version of the problem. In Table 13.2, the top row lists the cost coefficients $c_4, c_5, c_6, c_1, c_2, c_3$ for the materials used in the amounts $x_4, x_5, x_6, x_1, x_2, x_3$. The second row has those variable names. The next four rows give the augmented matrix for $A\mathbf{x} = \mathbf{b}$ except that the columns were permuted corresponding to listing the variables as $x_4, x_5, x_6, x_1, x_2, x_3$.

After the row operations $3.329R_4 + R_3 \to R_3$, $14.66R_4 + R_2 \to R_2$, $404R_4 + R_1 \to R_1$, we can find a basic feasible solution for x_4, x_5, x_6, x_2 as the basic variables. Permuting the columns corresponding to x_1 and x_2 gives Table 13.3. The last column is empty except for the first two entries that will have the total cost f for a known basic feasible solution. The 4×4 identity matrix in the first four columns shows that $(x_4, x_5, x_6, x_2) = (34, 2.66, 1.129, 1)$.

The total cost is

$$\mathbf{c}^T\mathbf{x} = [\,c_4 \quad c_5 \quad c_6 \quad c_2 \quad c_1 \quad c_3\,][\,x_4 \quad x_5 \quad x_6 \quad x_2 \quad x_1 \quad x_3\,]^T$$

$$= 0(34) + 0(2.66) + 0(1.129) + 65.70(1) + 64.512(0) + 76.80(0) = 65.70.$$

So, we have a basic feasible solution $(x_1, x_2, x_3, x_4, x_5, x_6) = (0, 1, 0, 34, 2.66, 1.129)$. $\bigcirc$

TABLE 13.2

Before Finding a Basic Feasible Solution

0	0	0	64.51	65.70	76.80	
x_4	x_5	x_6	x_1	x_2	x_3	y
1	0	0	−216	−404	−363	−370
0	1	0	−15.55	−14.66	−7.23	−12
0	0	1	−2.219	−3.329	−0.996	−2
0	0	0	1	1	1	1

TABLE 13.3

After Finding the First Basic Feasible Solution

0	0	0	65.70	64.51	76.80		f
x_4	x_5	x_6	x_2	x_1	x_3	y	65.70
1	0	0	0	188	41	34	
0	1	0	0	−0.89	7.43	2.66	
0	0	1	0	1.11	2.333	1.129	
0	0	0	1	1	1	1	

Note that the basic feasible solution we found is clearly suboptimal because it uses all oat flour and no wheat bran, and the latter would use fewer carbohydrates than the former. We will see how to find the optimal solution in Section 13.4.

13.3.2 Application: Structural Optimization

In *structural optimization* (Kirsch, 1993), Uri Kirsch models the collapse of structures in order to find the maximum load a structure can withstand. In the regime of **plasticity**, it is assumed that until reaching the **yield stress**, the stress is proportional to the strain, as we assumed in Section 10.4 when discussing linear elasticity. Also in the plastic regime, the bending moment of a section of material is proportional to its curvature, until the magnitude of the bending moment reaches the **fully plastic moment**, $M_{P\ell}$.

The linear theory assumes rotations are small and that the basic geometry remains the same as for the undeformed structure, for example, that a triangular strut hasn't collapsed into a straight line. So, the geometry imposes linear, static equilibrium equations relating the structure's members' bending moments, $M_1, \ldots, M_n$.

Let $\lambda \cdot P$ be the magnitude of a compressive force, where λ is a parameter that later will be maximized in order to optimize the structure. Let ℓ be a characteristic length of the structure, so $\lambda P\ell$ will be a bending moment. The optimization problem will be

$$\left\{\begin{array}{ll} \text{Maximize} & \lambda \\ \text{Subject to} & \text{static equilibrium equations in } \lambda, M_1, \ldots, M_n \\ & -M_{P\ell} \le M_j \le M_{P\ell},\ j = 1, \ldots, n \end{array}\right\}. \tag{13.25}$$

13.3.3 Problems

In problems 1–3, use row reduction to find *all* basic feasible solutions. You may use technology but write down all augmented matrices that you use and label them as to which variable you are solving for.

1. $\left\{\begin{array}{rrrrr} x_1 & -x_2 & +x_3 & & = -7 \\ 3x_1 & -4x_2 & & +x_4 & = 11 \end{array}\right\}$

2. $\left\{\begin{array}{rrrrr} x_1 & -x_2 & +x_3 & +x_4 & = 0 \\ -2x_1 & +2x_2 & -x_3 & & = 0 \\ & 3x_2 & +2x_3 & & = 4 \end{array}\right\}$

3. $\left\{\begin{array}{rrrrr} x_1 & +2x_2 & -2x_3 & +x_4 & = -6 \\ -2x_1 & -4x_2 & +5x_3 & & = 17 \\ x_1 & +2x_2 & -x_3 & & = 0 \end{array}\right\}$

In problems 4–5, (a) set up, but do not solve, the LP problem in standard form, and (b) find *all* basic feasible solutions. You may use technology but write down all augmented matrices that you use and label them as to which variable you are solving for.

4. $\begin{Bmatrix} x_1 & -x_2 & \leq 4 \\ 3x_1 & -4x_2 & \geq 5 \end{Bmatrix}$
5. $\begin{Bmatrix} x_1 & +x_2 & \leq & 2 \\ -2x_1 & -x_2 & \leq & -3 \end{Bmatrix}$

In problems 6–7, (a) set up, but do not solve, the LP problem in standard form to produce emergency rations that have the desired optimization, and (b) also find a basic feasible solution.

6. Maximize the kcal content per 100 g of mixture, with the requirements that the mixture has, per 100 g, at least 10 g of protein and 2 g of polyunsaturated fat, as well as no more than 70 g of carbohydrates.
7. Maximize the protein content in g per 100 g of mixture, with the requirements that the mixture has, per 100 g, at least 300 kcal, 10 g of dietary fiber, and 2.5 g of polyunsaturated fat.

13.4 Simplex Procedure

Here we will study Dantzig's **Simplex Procedure**, a basic but powerful method for solving LP problems. Unlike a least squares problem or the problem of minimization of a function of several variables, the difficulty is not in the function to be optimized but, instead, in a "combinatorial explosion" of candidates for the minimizer.

We saw in Section 13.2 that even with a few variables, the search for an optimum can be difficult due to searching both in the interior and on the boundary of the domain. As the number of variables increases, the difficulty increases terribly.

For an LP problem, it turns out that the work can be reduced tremendously because of the following result concerning convex geometrical analysis.

Theorem 13.12

If $\mathbf{x}^\star$ is in $\mathbb{R}^n$ and is a minimizer for LP problem (13.22) in Section 13.3, then at most m of the entries $x_j^\star$ can be nonzero.

This fundamental result reduces optimization to a search among $\binom{n}{m}$ candidates $\mathbf{x}$, because that is the number of ways of choosing m distinct objects from a menu of n distinct objects. Unfortunately, even searching the finite number $\binom{n}{m} \triangleq \frac{n!}{m!(n-m)}$ of candidates may take too much time! For example, $\binom{1000}{950} \approx 9.46046 \times 10^{84}$. This is what we mean by a combinatorial explosion of candidates.

The simplex procedure decides how to improve a basic feasible solution by replacing one of the basic variables by a new choice. This is called the **pivoting** process. The decision process first chooses a variable to leave; after that, the process chooses which variable to enter into the list of basic variables.

TABLE 13.4

Before Replacing a Basic Variable

c_{i_1}	c_{i_2}	c_{i_3}	c_{j_1}	c_{j_2}		f
x_{i_1}	x_{i_2}	x_{i_3}	x_{j_1}	x_{j_2}	$\mathbf{y}$	$y_{i_i}c_{i_1} + y_{i_2}c_{i_2} + y_{i_3}c_{i_3}$
1	0	0	α_{i_1}	β_1	y_{i_i}	
0	1	0	α_{i_2}	β_2	y_{i_2}	
0	0	1	α_{i_3}	β_3	y_{i_3}	
			$z_{j_1} - c_{j_1}$	$z_{j_2} - c_{j_2}$		

13.4.1 Unit Cost Reduction

If we move a variable into the set of basic variables, then we will, in effect, exchange specific amounts of the old basic variables for 1 unit of the new variable. If the new variable is increased from $x_j = 0$ to $x_j = 1$, then the process will increase our costs by c_j but decrease our cost by z_j, the total cost of the replaced amounts of the old quantities. We will explain how to find $z_j - c_j$, the **unit cost reduction**.

Suppose we begin with, or have arrived at, the tableau in Table 13.4. At this stage, the 3×3 identity matrix in the first three columns shows that the basic feasible solution has $(x_{i_1}, x_{i_2}, x_{i_3}) = (y_{i_1}, y_{i_2}, y_{i_3})$, all other $x_k = 0$, and the total cost is $f = \mathbf{y} \bullet [\, c_{i_1} \quad c_{i_2} \quad c_{i_3} \,]^T$. If we increase the amount of variable x_{j_1} from 0 to 1, that is, add 1 unit of x_{j_1}, at a cost of $1 \cdot c_{j_1}$, then keeping the system $A\mathbf{x} = \mathbf{b}$ satisfied implies we *must* reduce x_{i_1} by α_{i_1}, reduce x_{i_2} by α_{i_2}, and reduce x_{i_3} by α_{i_3}, changing costs by

$$z_{j_1} \triangleq \alpha_{i_1}c_{i_1} + \alpha_{i_2}c_{i_2} + \alpha_{i_3}c_{i_3}.$$

So, simultaneously increasing x_{j_1} from 0 to 1 while changing* $x_{i_1}, \ldots, x_{i_3}$ from $y_{i_1}, \ldots, y_{i_3}$ to $y_{i_1} - \alpha_{i_1}, \ldots, y_{i_3} - \alpha_{i_3}$, respectively, increases costs by c_{j_1} but also changes costs by z_{j_1}. The net unit cost saving is $z_{j_1} - c_{j_1}$.

When deciding which variable to enter into the list of basic variables, the simplex procedure says to choose the variable with the maximum positive unit reduced cost; if none of the unit reduced costs are positive, then we have already arrived at the minimizer!

Suppose we decided to move variable x_{j_1} into the list of basic variables. By Theorem 13.12, we must move out one of the old basic variables, that is, reduce it to zero, if we want our new basic feasible solution to be a minimizer. To decide which variable to move out, note that if we increase x_{j_1} from 0 to $\theta > 0$, then we will change $x_{i_1}, \ldots, x_{i_3}$ from $y_{i_1}, \ldots, y_{i_3}$ to $y_{i_1} - \theta\alpha_{i_1}, \ldots, y_{i_3} - \theta\alpha_{i_3}$, respectively. If we reduce one of the old basic variables to a negative value, then we have "gone too far" because we no longer have a feasible solution. So, we need $y_{i_1} - \theta\alpha_{i_1} \geq 0, \ldots, y_{i_3} - \theta\alpha_{i_3} \geq 0$.

If, say, $\alpha_{i_1} < 0$, then $y_{i_1} - \theta\alpha_{i_1} \geq y_{i_1} \geq 0$ for all $\theta > 0$, which imposes no restriction on θ. The actual restrictions on θ are that $y_{i_\ell} - \theta\alpha_{i_\ell} \geq 0$ for all ℓ for which $\alpha_{i_\ell} > 0$. The **minimum reduction** is

$$\theta = \min_{\text{for which } \alpha_{i_\ell} > 0} \frac{y_{i_\ell}}{\alpha_{i_\ell}}. \tag{13.26}$$

If all $\alpha_{i_\ell} \leq 0$, then we define $\theta = 0$ and the simplex procedure is complete.

* If an $\alpha < 0$, then we are increasing the corresponding variable by $|\alpha|$.

Let L be the choice of index ℓ that achieves the minimum in (13.26), that is, $\theta = \frac{y_{i_L}}{\alpha_{i_L}}$.

Assuming $\theta > 0$, to improve on our basic feasible solution we increase x_{j_1} from 0 to θ and simultaneously reduce x_{i_L} to 0; at the same time, we change the other basic variables x_{i_k} from y_{i_k} to $y_{i_k} - \theta\alpha_{i_k}$, respectively.

Example 13.11

Use the simplex procedure to find the minimum value and a minimizer of the LP problem

$$\left\{\begin{array}{lllll} \text{Minimize} & 5x_1 + x_2 + 2x_3 & & & \\ \\ \text{Subject to} & x_1 & +x_2 & +x_3 & \geq 20 \\ & 2x_1 & +x_2 & -x_3 & \leq 30 \\ & 3x_1 & +x_2 & +x_3 & = 60 \\ & & x_1, \ldots, x_3 \geq 0 & & \end{array}\right\}.$$

Method: In Section 13.3, we used slack variables x_4, x_5 to put this problem in standard form as (13.22) in Section 13.3, that is,

$$\left\{\begin{array}{lllllll} \text{Minimize} & 5x_1 + x_2 + 2x_3 & & & & & \\ \\ \text{Subject to} & -x_1 & -x_2 & -x_3 & +x_4 & = & -20 \\ & 2x_1 & +x_2 & -x_3 & +x_5 & = & 30 \\ & 3x_1 & +x_2 & +x_3 & & = & 60 \\ & x_1, \ldots, x_5 \geq 0 & & & & & \end{array}\right\}.$$

First, we will find a basic feasible solution. In **matrix-tableau form**, the problem is

$$\begin{array}{ccccccc} 5 & 1 & 2 & 0 & 0 & & f \\ x_1 & x_2 & x_3 & x_4 & x_5 & & \end{array}$$
$$\left[\begin{array}{ccccc|c} -1 & -1 & -1 & \underline{1} & 0 & -20 \\ 2 & 1 & -1 & 0 & \underline{1} & 30 \\ 3 & 1 & 1 & 0 & 0 & 60 \end{array}\right] \begin{array}{l} \\ \\ R_3 + R_2 \to R_2 \\ R_3 + R_1 \to R_1 \end{array} \sim$$
$$\begin{array}{ccccccc} 5 & 1 & 2 & 0 & 0 & & f \\ x_1 & x_2 & x_3 & x_4 & x_5 & & 120 \end{array}$$
$$\left[\begin{array}{ccccc|c} 2 & 0 & 0 & \underline{1} & 0 & 40 \\ 5 & 2 & 0 & 0 & \underline{1} & 90 \\ 3 & 1 & \underline{1} & 0 & 0 & 60 \end{array}\right] .$$
$$\begin{array}{cc} z_1 & z_2 \end{array}$$

The underlined ones in the first and second rows correspond to the basic variables x_4, x_5. After that, the elementary row operations make x_3 to be the third basic variable. After permuting the variables, the tableau is in Table 13.5, whose bottom row contains the unit costs reduction information $z_j - c_j$, which we are about to calculate. Also, later we will explain why the "5" is circled.

TABLE 13.5

After Finding the First Basic Solution

0	0	2	5	1		f
x_4	x_5	x_3	x_1	x_2	y	$120 = 0 \cdot 40 + 0 \cdot 90 + 2 \cdot 60$
1	0	0	2	0	40	
0	1	0	⑤	2	90	
0	0	1	3	1	60	
			1	1		

So far, we have a basic feasible solution $(x_1, x_2, x_3, x_4, x_5) = (0, 0, 60, 40, 90)$. The next thing to do is to decide whether we should pivot by choosing a variable to enter the list of basic variables, and *after* that, by choosing a variable to move out.

We use the "maximum unit reduced cost" criterion for choosing which variable, if any, to move in. The unit reduced cost of x_1 is

$$z_1 - c_1 = \begin{bmatrix} c_4 \\ c_5 \\ c_3 \end{bmatrix} \bullet \begin{bmatrix} \alpha_4 \\ \alpha_5 \\ \alpha_3 \end{bmatrix} - c_1 = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix} \bullet \begin{bmatrix} 2 \\ 5 \\ 3 \end{bmatrix} - 5 = 6 - 5 = 1.$$

Note that the $\boldsymbol{\alpha}$ column vector used to calculate z_1 sits under the x_1 variable in Table 13.5.

Similarly

$$z_2 - c_2 = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix} \bullet \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} - 1 = 2 - 1 = 1.$$

The maximum reduced unit cost is 1, and so we are free to choose either x_1 or x_2 to move into the list of basic variables. Let's choose x_1 to move in.

The choice to move in x_1 will, in principle, affect the choice of which variable to move out. We decide which variable to move out by analyzing the column $[\,\alpha_4 \quad \alpha_5 \quad \alpha_3\,]^T = [\,2 \quad 5 \quad 3\,]^T$ corresponding to variable x_1. We calculate that **minimum reduction** is

$$\theta = \min_{\alpha_{i_\ell} > 0} \frac{y_{i_\ell}}{\alpha_{i_\ell}} = \min\left\{\frac{y_4}{\alpha_4}, \frac{y_5}{\alpha_5}, \frac{y_3}{\alpha_3}\right\} = \min\left\{\frac{40}{2}, \frac{90}{5}, \frac{60}{3}\right\} = 18,$$

achieved at index $L = 5$. So, to improve on our basic feasible solution, we increase x_1 from 0 to $\theta = 18$ and reduce x_5 to 0; at the same time, the other basic variables x_{i_k} change from y_{i_k} to $y_{i_k} - \theta\alpha_{i_k}$, respectively.

Circle the "pivot position" ⑤ in Table 13.5 and now do row operations $0.2R_2 \to R_2$, $-2R_2 + R_1 \to R_1$, and $-3R_2 + R_3 \to R_3$ to get the tableau in Table 13.6. After that, we permute the columns, specifically by exchanging the columns corresponding to variables x_5 and x_1, to put the tableau into standard form for using unit costs reduction to discuss the next round of possibly swapping variables:

$$z_5 - c_5 = \begin{bmatrix} 0 \\ 5 \\ 2 \end{bmatrix} \bullet \begin{bmatrix} -0.4 \\ 0.2 \\ -0.6 \end{bmatrix} - 0 = -0.2 - 0 = -0.2 \quad \text{and}$$

$$z_2 - c_2 = \begin{bmatrix} 0 \\ 5 \\ 2 \end{bmatrix} \bullet \begin{bmatrix} -0.8 \\ 0.4 \\ -0.2 \end{bmatrix} - 1 = 1.6 - 1 = 0.6.$$

TABLE 13.6

After Doing Row Operations on Table 13.5

0	0	2	5	1	
x_4	x_5	x_3	x_1	x_2	f
1	−0.4	0	0	−0.8	4
0	0.2	0	1	0.4	18
0	−0.6	1	0	−0.2	6

TABLE 13.7

After Finding the Second Basic Solution

0	5	2	0	1		f
x_4	x_1	x_3	x_5	x_2	y	$102 = 0 \cdot 4 + 5 \cdot 18 + 2 \cdot 6$
1	0	0	−0.4	−0.8	4	
0	1	0	0.2	④	18	
0	0	1	−0.6	−0.2	6	

TABLE 13.8

After Finding the Third Basic Solution

0	1	2	0	5		f
x_4	x_2	x_3	x_5	x_1	y	$75 = 0 \cdot 40 + 1 \cdot 45 + 2 \cdot 15$
1	0	0	0	0.8	40	
0	1	0	0.5	2.5	45	
0	0	1	−0.5	0.2	15	
			−0.5	−2.1		

The maximum unit cost reduction is 0.6, so we move variable x_2 in. To decide which variable to move out, we calculate the minimum positive reduction, using $*$ to denote quantities not calculated because $\alpha < 0$:

$$\theta = \min_{\alpha_{i_\ell} > 0} \frac{y_{i_\ell}}{\alpha_{i_\ell}} = \min\left\{\frac{y_4}{\alpha_4}, \frac{y_1}{\alpha_1}, \frac{y_3}{\alpha_3}\right\} = \min\left\{*, \frac{18}{0.4}, *\right\} = 45,$$

which is achieved at index $L = 1$. So, to improve on our basic feasible solution, we increase x_2 from 0 to $\theta = 45$ and reduce x_1 to 0; at the same time, the other basic variables x_{i_k} change from y_{i_k} to $y_{i_k} - \theta\alpha_{i_k}$.

Circle the pivot position ④ in the tableau in Table 13.7, and do row operations $2.5R_2 \to R_2$, $0.8R_2 + R_1 \to R_1$, and $0.2R_2 + R_3 \to R_3$; after that, permute the columns to get the tableau in Table 13.8.

The unit reduced costs of x_5 and x_1 are, respectively,

$$z_5 - c_5 = \begin{bmatrix}0\\1\\2\end{bmatrix} \bullet \begin{bmatrix}0\\0.5\\-0.5\end{bmatrix} - 0 = -0.5 - 0 = -0.5 \quad \text{and}$$

$$z_1 - c_1 = \begin{bmatrix}0\\1\\2\end{bmatrix} \bullet \begin{bmatrix}0.8\\2.5\\0.2\end{bmatrix} - 5 = 2.9 - 5 = -2.1.$$

Because all of the unit cost reductions are negative, we have arrived at a minimizer! The solution is $(x_1, x_2, x_3) = (0, 45, 15)$. The slack variable values of $x_4 = 40, x_5 = 0$ are not part of the solution to the original problem but do indicate how much "wiggle room" is left in the inequalities at the optimum solution. ○

By the way, the *Mathematica* command

$$\text{FindMinimum}[\{5x_1 + x_2 + 2x_3, -x_1 - x_2 - x_3 + x_4 == -20 \,\&\&$$

$$2x_1 + x_2 - x_3 + x_5 == 30 \,\&\&\, 3x_1 + x_2 + x_3 == 60 \,\&\&$$

$$x_1 \geq 0 \,\&\&\, x_2 \geq 0 \,\&\&\, x_3 \geq 0 \,\&\&\, x_4 \geq 0 \,\&\&\, x_5 \geq 0\}, \{x_1, x_2, x_3, x_4, x_5\}]$$

gave output

$$\{75., \{x_1 \to 0., x_2 \to 45., x_3 \to 15., x_4 \to 40., x_5 \to 0.\}\}$$

that checks our conclusion for Example 13.11. While *Mathematica* can easily solve our small LP problems, we needed to work through them by hand in order to understand how the simplex procedure works.

13.4.2 Problems

In problems 1–8 solve the LP problem.

1. $$\left\{\begin{array}{ll} \text{Minimize} & x_1 + x_2 \\ \text{Subject to} & x_1 + 2x_2 \geq 3 \\ & x_1, x_2 \geq 0. \end{array}\right\}$$

2. $$\left\{\begin{array}{lll} \text{Minimize} & 2x_1 + x_2 + x_3 & \\ \text{Subject to} & x_1 + 2x_2 - x_3 & = 20 \\ & -x_1 - x_2 + 2x_3 + x_4 & = 16 \\ & 2x_1 + x_2 + 2x_3 & = 12 \\ & x_1, \ldots, x_4 \geq 0. & \end{array}\right\}$$

3. $$\left\{\begin{array}{lll} \text{Minimize} & 3x_1 + x_2 + 2x_3 & \\ \text{Subject to} & -x_1 - x_2 - x_3 + x_4 & = -10 \\ & 2x_1 + x_2 + x_3 + x_5 & = 40 \\ & 3x_1 + x_2 + x_3 & = 50 \\ & x_1, \ldots, x_5 \geq 0. & \end{array}\right\}$$

4. $$\left\{\begin{array}{lll} \text{Minimize} & 5x_1 + 2x_2 + x_3 & \\ \text{Subject to} & 2x_1 + x_2 + x_3 & = 60 \\ & 3x_1 + x_2 + x_3 & \leq 80 \\ & 2x_1 + x_2 - x_3 & \geq 40 \\ & x_1, \ldots, x_3 \geq 0. & \end{array}\right\}$$

5. $$\left\{\begin{array}{ll}\text{Minimize} & 2x_1+3x_2+x_3 \\ \text{Subject to} & \begin{array}{llll} 2x_1 & +x_2 & +x_3 & =60 \\ 3x_1 & +x_2 & +x_3 & \le 80 \\ 2x_1 & +x_2 & -x_3 & \ge 40 \end{array} \\ & x_1,\ldots,x_3 \ge 0.\end{array}\right\}$$

6. $$\left\{\begin{array}{ll}\text{Minimize} & 3x_1-2x_3 \\ \text{Subject to} & \begin{array}{llll} x_1 & +3x_2 & +4x_3 & = 80 \\ 3x_1 & -x_2 & +x_3 & \le 160 \\ x_1 & +x_2 & +x_3 & \ge 40 \end{array} \\ & x_1,\ldots,x_3 \ge 0.\end{array}\right\}$$

7. $$\left\{\begin{array}{ll}\text{Minimize} & x_1+x_2+2x_3 \\ \text{Subject to} & \begin{array}{llll} x_1 & +3x_2 & +4x_3 & \le 70 \\ 3x_1 & -x_2 & +x_3 & \le 110 \\ x_1 & +x_2 & +x_3 & \ge 40 \end{array} \\ & x_1,\ldots,x_3 \ge 0.\end{array}\right\}$$

8. $$\left\{\begin{array}{ll}\text{Minimize} & 3x_1+x_2-2x_3 \\ \text{Subject to} & \begin{array}{llll} -x_1 & +3x_2 & -x_3 & = 7 \\ -2x_2 & +x_3 & +x_4 & = 12 \\ -4x_2 & +3x_3 & +x_5 & = 10 \end{array} \\ & x_1,\ldots,x_5 \ge 0.\end{array}\right\}$$

In problems 9–11, solve a structural optimization problem of a continuous beam. [The model comes from the book by Kirsch referred to in Section 13.2.] [Hint: Substitute $\lambda P\ell = \mu$ and $x_j = M_j/M_{P\ell}$, $j = 1,2,3$.]

9. $$\left\{\begin{array}{ll}\text{Maximize} & \lambda \\ \text{Subject to} & \begin{array}{r} 4M_1+2M_2 = \lambda P\ell \\ 2M_2+4M_3 = 2\lambda P\ell \end{array} \\ & -M_{P\ell} \le M_j \le M_{P\ell},\ j=1,2,3\end{array}\right\}$$

10. $$\left\{\begin{array}{ll}\text{Maximize} & \lambda \\ \text{Subject to} & \begin{array}{r} 4M_1+2M_2 = \lambda P\ell \\ 2M_2+4M_3 = \lambda P\ell \end{array} \\ & -M_{P\ell} \le M_j \le M_{P\ell},\ j=1,2,3\end{array}\right\}$$

11. $$\left\{\begin{array}{ll}\text{Maximize} & \lambda \\ \text{Subject to} & \begin{array}{r} 4M_1+2M_2 = 5\lambda P\ell \\ 2M_2+4M_3 = \lambda P\ell \end{array} \\ & -M_{P\ell} \le M_j \le M_{P\ell},\ j=1,2,3\end{array}\right\}$$

12. Solve the LP problem of Problem 13.3.3.6.
13. Solve the LP problem of Problem 13.3.3.7.

13.5 Nonlinear Programming

Suppose $\mathbf{x}$ is in $\mathbb{R}^n$ and $f(\mathbf{x})$ and $f_i(\mathbf{x})$, $i = 1, \ldots, m$, are functions which have a common domain C in $\mathbb{R}^n$. We want to minimize the value of the objective function $f(\mathbf{x})$. The functions $f_1, \ldots, f_m$ give constraints concerning which $\mathbf{x}$ can be put into the objective function. The **nonlinear programming** problem is

$$\left\{\begin{array}{ll} \text{Minimize} & f(\mathbf{x}) \\ \text{Subject to} & f_1(\mathbf{x}) \leq 0 \\ & \vdots \\ & f_m(\mathbf{x}) \leq 0 \\ & \mathbf{x} \text{ is in } C \end{array}\right\}. \tag{13.27}$$

Definition 13.11

A set C in $\mathbb{R}^n$ is **convex** if for every pair of points $\mathbf{x}, \mathbf{y}$ in C the **line segment**

$$[\mathbf{x}, \mathbf{y}] \triangleq \{(1-t)\mathbf{x} + t\mathbf{y} : \ 0 \leq t \leq 1\}$$

is contained in C.

In $\mathbb{R}^2$, Figure 13.10 shows two examples of a convex set C.

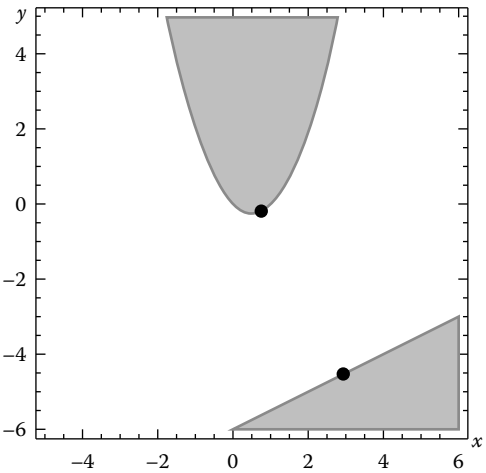


FIGURE 13.10
Example 13.12.

Definition 13.12

Suppose a function $f = f(\mathbf{x})$ is defined on a convex set C in $\mathbb{R}^n$. We say $f(\mathbf{x})$ is

(a) **Convex** if

$$f\big((1-t)\mathbf{x} + t\mathbf{y}\big) \leq (1-t)f(\mathbf{x}) + tf(\mathbf{y}), \text{ for all } 0 \leq t \leq 1, \ \mathbf{x}, \mathbf{y} \text{ in } C \tag{13.28}$$

(b) **Strongly (or strictly) convex** if

$$f\big((1-t)\mathbf{x} + t\mathbf{y}\big) < (1-t)f(\mathbf{x}) + tf(\mathbf{y}), \text{ for all } 0 < t < 1, \ \mathbf{x} \neq \mathbf{y} \text{ in } C \tag{13.29}$$

In Problem 13.1.3.9, you explained why a twice continuously differentiable function of a single variable defined on an interval is convex if, and only if, its second derivative is nonnegative on that interval. So, "convex" is what we called "concave up" for a function of a single variable in Calculus I.

Recall the definition of the Hessian matrix,

$$D^2 f(\mathbf{x}) \triangleq \left[\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) \right]_{\substack{1 \leq i \leq n \\ 1 \leq j \leq n}}.$$

Theorem 13.13

Suppose $f(\mathbf{x})$ is a function whose domain is a convex set C in $\mathbb{R}^n$ and whose gradient vector, $\nabla f(\mathbf{x})$, and Hessian matrix, $D^2 f(\mathbf{x})$, are continuous on C. Then the following are equivalent:

(a) f is convex on C.
(b) $f(\mathbf{x}_2) - f(\mathbf{x}_1) \geq \nabla f(\mathbf{x}_1) \bullet (\mathbf{x}_2 - \mathbf{x}_1)$ for all $\mathbf{x}_1, \mathbf{x}_2$ in C.
(c) $D^2 f(\mathbf{x})$ is positive semi-definite at all $\mathbf{x}$ in C.

Corollary 13.3

Suppose A is an $n \times n$ real, constant, positive semi-definite matrix and $\mathbf{c}$ is a real, constant vector in $\mathbb{R}^n$. Then $f(\mathbf{x}) \triangleq \mathbf{x}^T A \mathbf{x}$ and $g(\mathbf{x}) \triangleq \mathbf{c}^T \mathbf{x}$ are convex on all of $\mathbb{R}^n$.

Why? You will derive these results in Problem 13.5.3.14.

Recall from Section 13.3 the notation $(x_1, \ldots, x_n) = \mathbf{x} \geq \mathbf{0}$, which means $x_j \geq 0$ for $j = 1, \ldots, n$. Similarly, $(y_1, \ldots, y_m) = \mathbf{y} < \mathbf{0}$ means $y_j < 0$ for $j = 1, \ldots, m$.

Definition 13.13

Suppose $f(\mathbf{x})$ and $f_i(\mathbf{x})$, $i = 1, \ldots, m$, are functions whose common domain is a convex set C in $\mathbb{R}^n$. Denote $\mathbf{f}(\mathbf{x}) = [\, f_1(\mathbf{x}) \quad \cdots \quad f_m(\mathbf{x}) \,]^T$ in $\mathbb{R}^m$.

(a) The **feasible region** is the set $\mathcal{R} \triangleq \{\mathbf{x} \text{ in } C : \ \mathbf{f}(\mathbf{x}) \leq \mathbf{0}\}$.

(b) The **Lagrangian** is the function $F(\mathbf{x}, \boldsymbol{\lambda}) \triangleq f(\mathbf{x}) + \sum_{j=1}^{m} \lambda_j f_j(\mathbf{x}) = f(\mathbf{x}) + \boldsymbol{\lambda}\mathbf{f}(\mathbf{x})$, where $\boldsymbol{\lambda} = [\, \lambda_1 \quad \cdots \quad \lambda_m \,]^T$ is a vector of Lagrange multipliers.

(c) **Slater's condition** is that there exists an $\mathbf{x}_0$ such that $\mathbf{f}(\mathbf{x}_0) < 0$. Such an $\mathbf{x}_0$ is called a **strictly feasible solution**.

Theorem 13.14

Suppose $f(\mathbf{x})$ is a convex function whose domain is a closed, convex set C in $\mathbb{R}^n$. Then f achieves its minimum in C, that is, there exists an $\mathbf{x}^\star$ in C satisfying

$$f(\mathbf{x}^\star) \leq f(\mathbf{x}) \text{ for all } \mathbf{x} \text{ in } C$$

as long as either

(a) C is bounded* or

(b) $f(\mathbf{x})$ is both strongly convex on C and twice continuously differentiable on C.

Definition 13.14

Suppose $f(\mathbf{x})$ and $f_i(\mathbf{x})$, $i = 1, \ldots, m$, are convex functions whose common domain is a convex set C. A **convex programming** (CP) problem is

$$\left\{ \begin{array}{ll} \text{Minimize} & f(\mathbf{x}) \\ \text{Subject to} & f_1(\mathbf{x}) \leq 0 \\ & \vdots \\ & f_m(\mathbf{x}) \leq 0 \\ & \mathbf{x} \text{ is in } C \end{array} \right\}. \tag{13.30}$$

Theorem 13.15

(Karush–Kuhn–Tucker (KKT) necessary conditions) If $\mathbf{x}^\star$ is a global minimizer for the nonlinear programming problem (13.27), then there exists a vector of Lagrange multipliers

* See Section 2.10 for the definitions of "closed" and "bounded."

$\boldsymbol{\lambda}^\star = [\,\lambda_1^\star \quad \cdots \quad \lambda_m^\star\,]^T$ such that

$$\boldsymbol{\lambda}^\star \geq \mathbf{0} \quad \text{and} \quad \lambda_i^\star f_i(\mathbf{x}^\star) = 0 \text{ for } i = 1, \ldots, m, \tag{13.31}$$

and

$$\nabla f(\mathbf{x}^\star) + \left[\, \frac{\partial f_i}{\partial x_j}(\mathbf{x}^\star) \,\right]^T \boldsymbol{\lambda}^\star = \mathbf{0}, \tag{13.32}$$

in addition to the original requirements of the CP problem that

$$\mathbf{x}^\star \text{ is in } C \quad \text{and} \quad \mathbf{f}(\mathbf{x}^\star) \leq \mathbf{0}. \tag{13.33}$$

(13.32) is called the **stationarity** condition, and (13.31)'s conclusion "$\lambda_i^\star f_i(\mathbf{x}^\star) = 0$, $i = 1, \ldots, m$", is called the **complementarity** condition. As we will see, the stationarity conclusion is the same as the "vector field" conclusion of the method of Lagrange multipliers we saw in Section 13.2 for problems with *equality* constraints.

Conclusions (13.31) and (13.32) of Theorem 13.15 are also known as the KKT or Kuhn–Tucker (KT) conditions.

Theorem 13.16

(KKT sufficient conditions) Suppose that Slater's condition holds, that is, that there is a strictly feasible solution, and the functions $f(\mathbf{x})$ and $f_i(\mathbf{x})$, $i = 1, \ldots, m$, are all convex functions on a common, convex domain C. If there exists $\mathbf{x}^\star$ in C with $\mathbf{f}(\mathbf{x}^\star) \leq \mathbf{0}$ and a vector of Lagrange multipliers $\boldsymbol{\lambda}^\star$ such that the KKT conditions are satisfied, then $\mathbf{x}^\star$ is a global minimizer for the convex programming problem (13.30).

Our first example is similar to Example 13.5 in Section 13.2.

Example 13.12

Use the KT conditions to solve the CP problem of finding where the regions $x^2 - x - y \leq 0$ and $12 - X + 2Y \leq 0$ are closest to each other.

Method: $(X - x)^2 + (Y - y)^2$ is the square of the distance between a point (x, y) in the region $x^2 - x - y \leq 0$ and a point on (X, Y) in the region $12 - X + 2Y \leq 0$. We consider X, Y, x, y to be the variables of the problem, which we state as

$$\left\{ \begin{array}{ll} \text{Minimize} & (X - x)^2 + (Y - y)^2 \\ \\ \text{Subject to} & x^2 - x - y \leq 0 \\ & 12 - X + 2Y \leq 0 \end{array} \right\}. \tag{13.34}$$

Define $f(X, Y, x, y) \triangleq (X-x)^2+(Y-y)^2$, $f_1(X, Y, x, y) \triangleq x^2-x-y$, $f_2(X, Y, x, y) \triangleq 12-X+2Y$ and let $\lambda_1, \lambda_2 \geq 0$ be the Lagrange multipliers.

Note that f, f_1, and f_2 are convex functions everywhere, as we can see from using either Theorem 13.13 or Corollary 13.3.

As in Examples 13.5 and 13.6 in Section 13.2, the gradient is

$$\nabla = \hat{\boldsymbol{I}}\frac{\partial}{\partial X} + \hat{\boldsymbol{J}}\frac{\partial}{\partial Y} + \hat{\boldsymbol{\imath}}\frac{\partial}{\partial x} + \hat{\boldsymbol{\jmath}}\frac{\partial}{\partial y}.$$

The stationarity condition, (13.32), is

$$\begin{bmatrix} 2(X-x) \\ 2(Y-y) \\ -2(X-x) \\ -2(Y-y) \end{bmatrix} = \nabla f(X,Y,x,y) = -\begin{bmatrix} \frac{\partial f_1}{\partial X} & \frac{\partial f_1}{\partial Y} & \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial X} & \frac{\partial f_2}{\partial Y} & \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} \end{bmatrix}^T \boldsymbol{\lambda}$$

$$= -\lambda_1 \begin{bmatrix} 0 \\ 0 \\ 2x-1 \\ -1 \end{bmatrix} - \lambda_2 \begin{bmatrix} -1 \\ 2 \\ 0 \\ 0 \end{bmatrix}, \tag{13.35}$$

which also appeared in Example 13.5 in Section 13.2, with $\lambda = -\lambda_1$ and $\mu = -\lambda_2$.

Equation (13.35) implies

$$(1)\ \ \lambda_1(2x-1) = 2(X-x) = \lambda_2 \quad \text{and} \quad (2)\ \ \lambda_1 = -2(Y-y) = 2\lambda_2;$$

hence,

$$0 = 2\lambda_2 - 2\lambda_2 = 2\lambda_1(2x-1) - \lambda_1 = \lambda_1\big(2(2x-1)-1\big).$$

It follows that

$$\text{either } \lambda_1 = 0 \quad \text{or} \quad x = \frac{3}{4}. \tag{13.36}$$

The complementarity conditions are

$$\lambda_1(x^2 - x - y) = \lambda_2(12 - X + 2Y) = 0.$$

In effect, the complementarity conditions say that either the multiplier λ_1 is zero or the point (x, y) lies on the boundary of the region $x^2 - x - y \le 0$, and either the multiplier λ_2 is zero or the point (X, Y) lies on the boundary of the region $12 - X + 2Y \le 0$.

Combine the first complementarity condition and (13.36) to imply that

$$\lambda_1 = 0 \quad \text{or} \quad y = x^2 - x = \left(\frac{3}{4}\right)^2 - \frac{3}{4} = -\frac{3}{16}.$$

So, either $\lambda_1 = 0$ or $(x, y) = (\frac{3}{4}, -\frac{3}{16})$.

Suppose $\lambda_1 = 0$. Then (1) and (2) imply $X = x$, $Y = y$, and $\lambda_2 = 0$. The second complementarity condition is then satisfied because $\lambda_2 = 0$. The only facts we have left to work with are the feasibility requirements in (13.33), specifically $\mathbf{f}(\mathbf{x}^\star) \le \mathbf{0}$, that is, that $x^2 - x - y \le 0$ and $12 - X + 2Y \le 0$. It follows that

$$y \ge x^2 - x \quad \text{and} \quad Y \le -6 + \frac{1}{2}X. \tag{13.37}$$

But $X = x$ and $Y = y$, so $x^2 - x \le y \le -6 + \frac{1}{2}x$; hence, $0 \ge x^2 - \frac{3}{2}x + 6 = \left(x - \frac{3}{4}\right)^2 + \frac{87}{16} \ge \frac{87}{16}$, which is impossible. So, we conclude that $\lambda_1 \ne 0$.

So far, we have concluded that $(x,y) = (\frac{3}{4}, -\frac{3}{16})$ and $\lambda_1 > 0$. To find (X,Y), note that (1) and (2), along with $\lambda_1 > 0$ and the fact that $2x - 1 = 2 \cdot \frac{3}{4} - 1 \neq 0$, together imply

$$2Y + \frac{3}{8} = 2\left(Y + \frac{3}{16}\right) = 2(Y - y) = -\lambda_1 = -\frac{2(X-x)}{(2x-1)} = -\frac{2(X - \frac{3}{4})}{(2 \cdot \frac{3}{4} - 1)} = -\frac{2X - \frac{3}{2}}{\frac{1}{2}} = -4X + 3;$$

hence,

$$Y = -2X + \frac{21}{16}.$$

The second complementarity condition is that either $\lambda_2 = 0$ or $12 - X + 2Y = 0$. If $\lambda_2 = 0$ then, again, (1) and (2) imply $X = x$ and $Y = y$, leading eventually to a contradiction as in the preceding argument. If $12 - X + 2Y = 0$, then we use the fact that $Y = -2X + \frac{21}{16}$, hence

$$0 = 12 - X + 2Y = 12 - X + 2\left(-2X + \frac{21}{16}\right) = 12 - 5X + \frac{21}{8};$$

hence, $X = \frac{117}{40}$, and thus $Y = -2\frac{117}{40} + \frac{21}{16} = -\frac{363}{80}$. The closest approach of the two regions is where $(x,y) = (\frac{3}{4}, -\frac{3}{16})$, $(X,Y) = \left(\frac{117}{40}, -\frac{363}{80}\right)$, and the minimum distance is

$$\sqrt{\left(\frac{117}{40} - \frac{3}{4}\right)^2 + \left(-\frac{363}{80} + \frac{3}{16}\right)^2} = \sqrt{\frac{87^2}{40^2} + \frac{348^2}{80^2}} = \frac{87\sqrt{5}}{40}. \bigcirc$$

The results of Example 13.12 are depicted in Figure 13.10.

Example 13.13

Solve

$$\left\{\begin{array}{ll} \text{Minimize} & -\ln x - 3\ln y \\ \text{Subject to} & 2x + y \leq 4 \\ & x > 0,\ y > 0 \end{array}\right\}. \tag{13.38}$$

Method: The objective function, $f(x,y) = -\ln x - 3\ln y$, is convex on the convex set $C \triangleq \{(x,y) : x > 0, y > 0\}$ by Problems 13.1.3.10 and 13.1.3.12(a). The constraint function $f_1(x,y) = 2x + y - 4$ is also convex on C, by a result in Corollary 13.3, so we can use the KT conditions to solve the CP problem (13.38). The feasible region is shown in Figure 13.11. Let $\lambda \geq 0$ be the Lagrange multiplier.

The feasibility conditions are $2x + y - 4 \leq 0, x > 0, y > 0$. The complementarity condition is

$$\lambda \cdot (2x + y - 4) = 0. \tag{13.39}$$

The stationarity condition is

$$\begin{bmatrix} -\frac{1}{x} \\ -\frac{3}{y} \end{bmatrix} + \lambda \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{13.40}$$

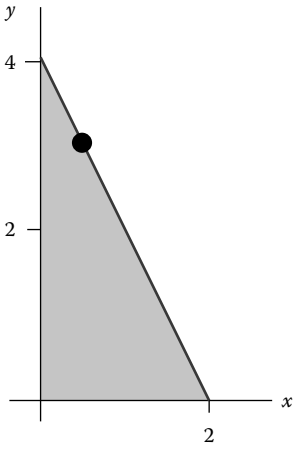


FIGURE 13.11
Example 13.13.

But, (13.40) implies $\lambda \neq 0$, because $-\frac{1}{x}$ cannot be zero for a feasible (x, y). So, the complementarity condition implies

$$2x + y - 4 = 0. \tag{13.41}$$

(13.40) also implies

$$y = \frac{3}{\lambda} \quad \text{and} \quad x = \frac{1}{2\lambda}.$$

Substitute those into (13.41) to get

$$4 = 2\left(\frac{1}{2\lambda}\right) + \left(\frac{3}{\lambda}\right) = \frac{4}{\lambda};$$

hence, $\lambda = 1$, and hence, $x = \frac{1}{2}$ and $y = 3$. The minimum value of the objective function with these constraints is $f(\frac{1}{2}, 3) = -\ln\frac{1}{2} - 3\ln 3 = -\ln\frac{27}{2}$. ○

13.5.1 Dual LP Problem

Suppose we have an LP problem in the preliminary standard form (13.21) in Section 13.3, that is,

$$\left\{\begin{array}{ll} \text{Minimize} & \mathbf{c}^T\mathbf{x} \\ \text{Subject to} & A\mathbf{x} \leq \mathbf{b} \quad \text{and} \quad \mathbf{x} \geq \mathbf{0} \end{array}\right\},$$

where
$\mathbf{x}$ and $\mathbf{c}$ are in $\mathbb{R}^k$
$\mathbf{b}$ are in $\mathbb{R}^m$
A is a real, $m \times k$ matrix.

We can refer compactly to (13.21) in Section 13.3 as the "LP problem for the 'triple'$[A, \mathbf{b}, \mathbf{c}]$."

Definition 13.15

The LP problem $[A, \mathbf{b}, \mathbf{c}]$ has **dual LP problem** $[-A^T, \mathbf{c}, -\mathbf{b}]$, that is, to find $\mathbf{y}$ in $\mathbb{R}^m$ to

$$\left\{\begin{array}{ll}\text{Minimize} & -\mathbf{b}^T\mathbf{y} \\ \text{Subject to} & -A^T\mathbf{y} \leq \mathbf{c} \quad \text{and} \quad \mathbf{y} \geq \mathbf{0}\end{array}\right\}. \tag{13.42}$$

Note that the dual LP problem (13.42) can be restated as

$$\left\{\begin{array}{ll}\text{Maximize} & \mathbf{b}^T\mathbf{y} \\ \text{Subject to} & A^T\mathbf{y} \geq -\mathbf{c} \quad \text{and} \quad \mathbf{y} \geq \mathbf{0}\end{array}\right\}.$$

Theorem 13.17

(LP duality theorem) Suppose that either the LP problem $[A, \mathbf{b}, \mathbf{c}]$ has an optimal solution $\mathbf{x}^\star$ or its dual, $[-A^T, \mathbf{c}, -\mathbf{b}]$, has an optimal solution $\mathbf{y}^\star$. Then both problems have an optimal solution and

$$\mathbf{c}^T\mathbf{x}^\star = -\mathbf{b}^T\mathbf{y}^\star.$$

Why? Define the convex set $C = \{\mathbf{x} : \mathbf{x} \geq \mathbf{0}\}$.

First, we will use the KT conditions to explain why $[A, \mathbf{b}, \mathbf{c}]$ having an optimal solution $\mathbf{x}^\star$ implies $[-A^T, \mathbf{c}, -\mathbf{b}]$ has an optimal solution $\mathbf{y}^\star$ and $\mathbf{c}^T\mathbf{x}^\star = -\mathbf{b}^T\mathbf{y}^\star$.

Define $f(\mathbf{x}) \triangleq \mathbf{c}^T\mathbf{x}$ and $\mathbf{f}(\mathbf{x}) \triangleq A\mathbf{x} - \mathbf{b}$, and define the CP problem

$$\left\{\begin{array}{ll}\text{Minimize} & f(\mathbf{x}) \\ \text{Subject to} & \mathbf{f}(\mathbf{x}) \leq \mathbf{0} \\ & \mathbf{x} \text{ is in } C\end{array}\right\}.$$

By the KT conditions, there exists a vector of Lagrange multipliers $\boldsymbol{\lambda}^\star = [\lambda_1^\star \quad \cdots \quad \lambda_m^\star]^T$ such that

$$\boldsymbol{\lambda}^\star \geq \mathbf{0} \quad \text{and} \quad \lambda_i^\star f_i(\mathbf{x}^\star) = 0 \text{ for } i = 1, \ldots, m \tag{13.43}$$

and

$$\nabla f(\mathbf{x}^\star) = -\left[\frac{\partial f_i}{\partial x_j}(\mathbf{x}^\star)\right]^T \boldsymbol{\lambda}^\star, \tag{13.44}$$

in addition to the original requirements of the CP problem that

$$\mathbf{x}^\star \text{ is in } C \quad \text{and} \quad \mathbf{f}(\mathbf{x}^\star) \leq \mathbf{0}, \quad \text{that is, that } \mathbf{x}^\star \geq \mathbf{0} \quad \text{and} \quad A\mathbf{x}^\star \leq \mathbf{b}. \tag{13.45}$$

From $f(\mathbf{x}) \triangleq \mathbf{c}^T\mathbf{x}$, we calculated in Problem 13.2.3.17 that

$$\nabla f(\mathbf{x}^\star) = \mathbf{c}.$$

From $\mathbf{f}(\mathbf{x}) \triangleq A\mathbf{x} - \mathbf{b}$, we calculate that

$$\left[\frac{\partial f_i}{\partial x_j}(\mathbf{x}^\star)\right]^T = (A)^T = A^T.$$

So, the stationarity condition is

$$\mathbf{c} = -A^T\boldsymbol{\lambda}^\star.$$

The complementarity conditions in (13.43) are

$$\lambda_i^\star \cdot \left(a_{i1}x_1^\star + \cdots + a_{ik}x_k^\star - b_i\right) = 0, \text{ for } i = 1, \ldots, m.$$

Summing up zeros over $i = 1, \ldots, m$ implies

$$\mathbf{b}^T\boldsymbol{\lambda}^\star = (A\mathbf{x}^\star)^T\boldsymbol{\lambda}^\star = (\mathbf{x}^\star)^TA^T\boldsymbol{\lambda}^\star = -(\mathbf{x}^\star)^T\mathbf{c}, \quad \text{that is,} \quad -\mathbf{b}^T\boldsymbol{\lambda}^\star = \mathbf{c}^T\mathbf{x}^\star.$$

If we *define* $\mathbf{y}_0 \triangleq \boldsymbol{\lambda}^\star$, then the KT conditions and the preceding work explain why $\mathbf{y}_0$ satisfies

$$\mathbf{y}_0 \geq \mathbf{0}, \quad -A^T\mathbf{y}_0 = \mathbf{c}, \quad \text{and} \quad -\mathbf{b}^T\mathbf{y}_0 = \mathbf{c}^T\mathbf{x}^\star.$$

This shows that the dual problem, $[-A^T, \mathbf{c}, -\mathbf{b}]$, has a feasible solution, $\mathbf{y}_0$, and therefore, there is a $\mathbf{y}^\star$ that is both feasible, that is, satisfies $\mathbf{y}^\star \geq \mathbf{0}$ and $-A^T\mathbf{y}^\star \leq \mathbf{c}$, and minimizes the dual problem's objective function, $\mathbf{b}^T\mathbf{y}$. [Note that there may be no relationship between $\mathbf{y}_0 = \boldsymbol{\lambda}^\star$ and $\mathbf{y}^\star$.]

It follows from the minimality of $\mathbf{b}^T\mathbf{y}^\star$ that

$$-\mathbf{b}^T\mathbf{y}^\star \leq -\mathbf{b}^T\mathbf{y}_0 = \mathbf{c}^T\mathbf{x}^\star. \tag{13.46}$$

But, the preceding argument can be used if we exchange the roles of the two problems: $[A, \mathbf{b}, \mathbf{c}]$ is the dual of $[-A^T, \mathbf{c}, -\mathbf{b}]$. The KT conditions for the global minimizer $\mathbf{y}^\star$ for $[-A^T, \mathbf{c}, -\mathbf{b}]$ imply there is a $\boldsymbol{\mu}^\star \geq \mathbf{0}$ for which stationarity holds, that is, $\mathbf{b} = -(-A^T)^T\boldsymbol{\mu}^\star = A\boldsymbol{\mu}^\star$ and the complementarity conditions

$$\mu_i^\star \cdot \left(-(a_{1j}y_1^\star + \cdots + a_{kj}y_k^\star) - c_j\right) = 0, \text{ for } j = 1, \ldots, k$$

hold. As before, summing up zeros over $i = 1, \ldots, m$ implies

$$\mathbf{c}^T\boldsymbol{\mu}^\star = (-A^T\mathbf{y}^\star)^T\boldsymbol{\mu}^\star = (\mathbf{y}^\star)^T(-A)\boldsymbol{\mu}^\star = -(\mathbf{y}^\star)^T\mathbf{b}, \quad \text{that is,} \quad \mathbf{c}^T\boldsymbol{\mu}^\star = -(\mathbf{y}^\star)^T\mathbf{b}.$$

If we define $\mathbf{x}_0 = \boldsymbol{\mu}^\star$, we see that $\mathbf{x}_0$ is a feasible solution of $[A, \mathbf{b}, \mathbf{c}]$, and thus the original problem's optimal solution $\mathbf{x}^\star$ satisfies

$$\mathbf{c}^T\mathbf{x}^\star \leq \mathbf{c}^T\boldsymbol{\mu}^\star = -\mathbf{b}^T\mathbf{y}^\star. \tag{13.47}$$

This and (13.46) imply that

$$\mathbf{c}^T\mathbf{x}^\star \leq \mathbf{c}^T\boldsymbol{\mu}^\star = -\mathbf{b}^T\mathbf{y}^\star \leq -\mathbf{b}^T\mathbf{y}_0 = \mathbf{c}^T\mathbf{x}^\star;$$

hence, equality holds throughout, and hence,

$$\mathbf{c}^T\mathbf{x}^\star = -\mathbf{b}^T\mathbf{y}^\star. \ \square$$

13.5.2 Application: Geometric Tolerancing

Geometric Dimensioning and **Geometric Tolerancing** are sets of *ANSI* standards for the quantitative description of machine parts and constraints on errors in their manufacture. For example, an engine design might require that the difference between a piston radius and its cylinder bore radius vary by a magnitude of at most 0.0002 in.

One basic problem is to check if a particular part's dimensions are within the margin of error. In order to do that, it may be necessary to rotate and/or translate a part to see if it fits within an allowed region. This tests the **straightness tolerancing** of the part.

For example, in Figure 13.12a, we see a curved side of a part. We want to quantify how close it is to being straight. The curviness of the side is exaggerated for convenience in illustrating the concepts involved.

After rotating the plane by an angle θ and translating the origin by (X, Y), we get the curve in Figure 13.12b. The number t_s measures how "unstraight" the side is. For example, the whole manufactured part might be rejected if $t_s > 0.01$ mm.

In order to measure the straightness of a curved side, we need to choose the rotation angle θ and the vertical translation Y so that points (x_i, y_i) on the curve are mapped to points (x_i', y_i') lying in the strip $0 \leq y' \leq t_s$. After rotation by θ and translation by (X, Y), the point (x, y) is mapped to the point

$$\begin{bmatrix} x_i' \\ y_i' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} + \begin{bmatrix} X \\ Y \end{bmatrix};$$

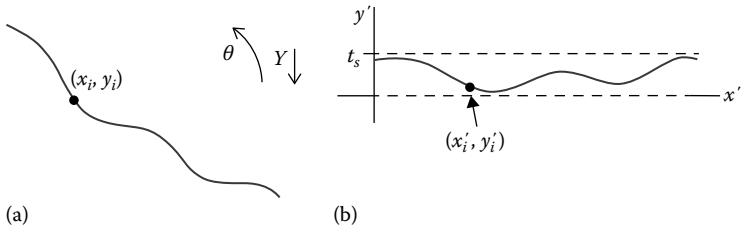


FIGURE 13.12
(a) Curved side, (b) Side after rotation and translation to check straightness.

hence, $y_i' = x_i \sin\theta + y_i \cos\theta + Y$. Testing the part involves mapping some points (x_i, y_i), $i = 1, \ldots, N$ on the original curve to the points (x_i', y_i'). The straightness of a curve can be measured by t_s, the objective function of the nonlinear programming problem (Chen and Fan, 2002)

$$\left\{\begin{array}{ll} \text{Minimize} & t_s \\ \text{Subject to} & 0 \le x_i \sin\theta + y_i \cos\theta + Y \le t_s,\ i = 1, \ldots, N \end{array}\right\}. \tag{13.48}$$

There is no sign restriction on either θ or Y. The paper referred to "Tolerance evaluation of minimum zone straightness$\cdots$" goes on to solve the nonlinear programming problem using a commonly used minimization algorithm that is available via the Excel software package.

Learn More About It

Example 13.13 is a version of an example in "A duality theorem for convex programs," W. S. Dorn, *IBM J. Res. Dev.* **4** (1960) pp. 407-413.

There are many sources to learn the justification of the KKT conditions, including *Linear Programming: Methods and Applications*, 5th edn., by Saul I. Gass, McGraw-Hill, Inc. © 1985, *Linear and Nonlinear Programming*, 3rd edn., by David G. Luenberger and Yinyu Ye, Springer © 2008, and *Elements of Applicable Functional Analysis*, by Charles W. Groetsch, Marcel Dekker, Inc. © 1980.

13.5.3 Problems

1. Use the KT conditions to find where the regions $x^2 - 2x - y \le 0$ and $8 - X + Y \le 0$ are closest to each other.
2. Use the KT conditions to find where the regions $x^2 - 2x + 1 - y \le 0$ and $6 - X + Y \le 0$ are closest to each other.
3. Solve $\left\{\begin{array}{ll} \text{Minimize} & -2\ln x - 3\ln y \\ \text{Subject to} & 2x + y \le 4 \\ & x > 0,\ y > 0 \end{array}\right\}$.
4. Solve $\left\{\begin{array}{ll} \text{Minimize} & -\ln x - 2\ln y \\ \text{Subject to} & -x + y \le 1 \\ & 2x + y \le 3 \\ & x > 0,\ y > 0 \end{array}\right\}$.
5. Given $\mathbf{x}_j$, $j = 1, \ldots, p$ in a convex set C and nonnegative real numbers λ_j satisfying $\lambda_1 + \cdots + \lambda_p = 1$, explain why $\lambda_1 \mathbf{x}_1 + \cdots + \lambda_p \mathbf{x}_p$ also must be in C.
6. Suppose $\mathbf{x}_0$ is in $\mathbb{R}^n$ and $r > 0$. Explain why each of the following are convex sets in $\mathbb{R}^n$:
   (a) The closed solid ball $\{\mathbf{x} : ||\mathbf{x} - \mathbf{x}_0|| \le r\}$
   (b) The open solid ball $\{\mathbf{x} : ||\mathbf{x} - \mathbf{x}_0|| < r\}$
   (c) A vector subspace $\mathcal{V}$.

7. Suppose $f(\mathbf{x})$ is a convex function of $\mathbf{x}$ on a convex set C and M is any real number. Explain why $S \triangleq \{\mathbf{x} : f(\mathbf{x}) \leq M\}$ is either a convex set or empty.
8. (a) Suppose $f(\mathbf{x})$ and $g(\mathbf{x})$ are convex functions defined on a common, convex set C in $\mathbb{R}^n$. Explain why $f(\mathbf{x}) + g(\mathbf{x})$ is also convex on C.
9. We know from Problem 13.1.3.12(b) that in $\mathbb{R}^1$, the composition of two convex functions need not be convex. A vector-valued function $\mathbf{g}(\mathbf{x})$ on $\mathbb{R}^n$ is **affine** if $\mathbf{g}(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ for some matrix A and $\mathbf{b}$ in $\mathbb{R}^n$. Suppose $f(\mathbf{y})$ is convex on $\mathbb{R}^m$ and $\mathbf{g}(\mathbf{x})$ is affine on $\mathbb{R}^n$ and takes on values in $\mathbb{R}^m$. Explain why $f\big(\mathbf{g}(\mathbf{x})\big)$ is convex on $\mathbb{R}^n$.
10. Suppose that C is a convex set in $\mathbb{R}^n$ and for $= 1, \ldots, m$, $f_i(\mathbf{x})$ is a convex function on common domain C. Explain why the set of feasible points, that is,

$$\{\mathbf{x} \text{ in } C : \ \mathbf{f}(\mathbf{x}) \leq \mathbf{0} \text{ and } \mathbf{x} \geq \mathbf{0}\}$$

    is a convex set in $\mathbb{R}^n$.
11. Suppose C_1 and C_2 are convex subsets of $\mathbb{R}^m$. Explain why their intersection, $C_1 \cap C_2 \triangleq \{\mathbf{x} : \ \mathbf{x}$ is in both C_1 and $C_2\}$, is also convex.
12. Explain why the results in Corollary 13.3 are true.
13. Suppose that C is a convex set in $\mathbb{R}^n$, $f(\mathbf{x})$ is a convex function on C, and $\mathbf{x}^\star$ is a local minimizer of f on C. Explain why $\mathbf{x}^\star$ must be a global minimizer of f on C. [Hint: This is a generalization of Theorem 13.4 in Section 13.1.]
14. Here you will establish the results of Corollary 13.3 without using the results of Theorem 13.13: Suppose A is an $n \times n$ real, constant, positive semi-definite matrix and $\mathbf{c}$ is a real, constant vector in $\mathbb{R}^n$. Explain why $f(\mathbf{x}) \triangleq \mathbf{x}^T A\mathbf{x}$ and $g(\mathbf{x}) \triangleq \mathbf{c}^T\mathbf{x}$ are convex on all of $\mathbb{R}^n$.
15. Suppose that A is an $n \times n$ real, constant, positive definite matrix. Explain why $f(\mathbf{x}) \triangleq \mathbf{x}^T A\mathbf{x}$ is strictly convex on all of $\mathbb{R}^n$.
16. Suppose $f(\mathbf{x})$ is strictly convex on a convex set C. Explain why f cannot have two distinct global minimizers in C.

13.6 Rayleigh–Ritz Method

Recall from Definition 2.23 in Section 2.9 that the **Rayleigh quotient** of a real, symmetric matrix A is defined by

$$\mathcal{R}_A(\mathbf{x}) \triangleq \frac{\langle \mathbf{x}, A\mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle} = \frac{\mathbf{x}^T A\mathbf{x}}{||\,\mathbf{x}\,||^2}, \quad \text{for } \mathbf{x} \neq \mathbf{0}. \tag{13.49}$$

Also, recall from Theorem 2.45 in Section 2.9 that if A is a real symmetric matrix, then the values

$$\lambda_1 \triangleq \max\{\mathcal{R}_A(\mathbf{x}) : \mathbf{x} \text{ satisfying } \mathbf{x} \neq \mathbf{0}\} \quad \text{and} \quad \lambda_n \triangleq \min\{\mathcal{R}_A(\mathbf{x}) : \mathbf{x} \text{ satisfying } \mathbf{x} \neq \mathbf{0}\}$$

exist and are eigenvalues of A, and there are eigenvectors $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(n)}$, respectively, that "achieve" the values λ_1 and λ_n, that is,

$$\lambda_1 = \mathcal{R}_A(\mathbf{x}_1) \quad \text{and} \quad \lambda_n = \mathcal{R}_A(\mathbf{x}_n).$$

Recall that $\mathcal{R}_A(\beta\mathbf{x}) = \mathcal{R}_A(\mathbf{x})$ for all nonzero scalars β. In Section 2.9, we used this to argue that we need only study the Rayleigh quotient for $||\,\mathbf{x}\,|| = 1$.

Note that if A is $n \times n$ and $1 = ||\,\mathbf{x}\,||^2 = \mathbf{x}^T\mathbf{x} = \sum_{i=1}^{n} x_j^2$ then

$$\mathcal{R}_A(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j\,.$$

It follows that we can, at least theoretically, find λ_1 and $\mathbf{x}^{(1)}$ by solving a maximization problem and similarly find λ_n and $\mathbf{x}^{(n)}$ by solving a minimization problem.

Example 13.14

Use Lagrange multipliers to explain why the minimum and the maximum of the Rayleigh quotient of a real, symmetric matrix A are eigenvalues.

Method: Our two optimization problems are of the form (13.7) in Section 13.2, specifically

$$\left\{\begin{array}{ll}\text{Maximize} & \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j \\ \\ \text{Subject to} & \displaystyle -1 + \sum_{i=1}^{n} x_i^2 = 0\end{array}\right\} \tag{13.50}$$

and

$$\left\{\begin{array}{ll}\text{Minimize} & \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j \\ \\ \text{Subject to} & \displaystyle -1 + \sum_{i=1}^{n} x_i^2 = 0\end{array}\right\}. \tag{13.51}$$

Note that (13.50) could be rewritten as to call for *minimizing* $\left(-\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j\right)$, hence could be rewritten in the form of (13.7) in Section 13.2.

Define

$$f(\mathbf{x}) \triangleq \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j.$$

For example, when $n = 2, f(x_1, x_2) = a_{11}x_1^2 + a_{12}x_1x_2 + a_{21}x_2x_1 + a_{22}x_2^2$.

So far we have not used the fact that A is symmetric. For example, when $n = 2$, symmetry of A implies $f(x_1, x_2) = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2$. Similarly, for any $n \geq 2$,

$$f(\mathbf{x}) = \sum_{i=1}^{n} a_{ii}x_i^2 + 2\sum_{j\neq i} a_{ij}x_i x_j.$$

This helps us to calculate

$$\frac{\partial f}{\partial x_i} = 2a_{ii}x_i + 2\sum_{j\neq i} a_{ij}x_j = 2(A\mathbf{x})_i,$$

that is, twice the ith component of the vector $A\mathbf{x}$.

Define $g(\mathbf{x}) = -1 + \sum_{i=1}^{n} x_i^2$. We calculate that

$$\frac{\partial g}{\partial x_i} = 2x_i = 2(\mathbf{x})_i,$$

twice the ith component of the vector $\mathbf{x}$.

Because $\nabla g = 2\mathbf{x} \neq \mathbf{0}$ for all $\mathbf{x}$ that satisfy the constraint $g(\mathbf{x}) = ||\,\mathbf{x}\,||^2 - 1 = 0$, (13.50) has Lagrange multiplier λ satisfying

$$2A\mathbf{x}^\star = \nabla f(\mathbf{x}^\star) = \lambda\,\nabla g(\mathbf{x}^\star) = 2\lambda\,\mathbf{x}^\star,$$

where $\mathbf{x}^\star$ is a maximizer. It follows that

$$A\mathbf{x}^\star = \lambda\mathbf{x}^\star.$$

Because $\mathbf{x}^\star \neq \mathbf{0}$, we conclude that $\mathbf{x}^\star$ is an eigenvector of A. We note that the maximizer $\mathbf{x}^\star$ for problem (13.50) satisfies

$$f(\mathbf{x}^\star) = (\mathbf{x}^\star)^T A\mathbf{x}^\star = (\mathbf{x}^\star)^T(\lambda\mathbf{x}^\star) = \lambda\,||\mathbf{x}^\star||^2 = \lambda\cdot 1 = \lambda.$$

But, as we saw in Theorem 2.45 in Section 2.9, both the maximum and minimum values of $\mathcal{R}_A(\mathbf{x})$ are eigenvalues of A. So, the maximum is an eigenvalue, with corresponding eigenvector $\mathbf{x}^\star$. The same reasoning explains why the minimum is an eigenvalue, with corresponding eigenvector being a minimizer. ◯

In Example 2.37 in Section 2.9 we saw that the Rayleigh quotient for the matrix

$$A = \begin{bmatrix} -4 & 2 & 3 \\ 2 & -5 & 1 \\ 3 & 1 & -8 \end{bmatrix}$$

produced approximate eigenvalues $\lambda_1 \approx -1.10996$, for the maximum eigenvalue, and $\lambda_n \approx -9.60653$, for the minimum eigenvalue.

13.6.1 Other Eigenvalues

Suppose that for a symmetric matrix A we have found λ_1, the maximum eigenvalue, and a corresponding eigenvector $\mathbf{x}^{(1)} \triangleq [x_1^{(1)} \quad \cdots \quad x_n^{(1)}]^T$ with $||\mathbf{x}^{(1)}||^2 = 1$. After that, we can look for another eigenvalue $\lambda_2 \leq \lambda_1$ and an eigenvector $\mathbf{x}$ that is orthogonal to $\mathbf{x}^{(1)}$ by using the **Rayleigh–Ritz method** of solving the maximization problem

$$\left\{\begin{array}{ll} \text{Maximize} & \mathbf{x}^T A\mathbf{x} \\ \text{Subject to} & \mathbf{x}^T\mathbf{x} - 1 = 0 \\ & \mathbf{x}^T\mathbf{x}^{(1)} = 0 \end{array}\right\}. \tag{13.52}$$

Again, we can use the method of Lagrange multipliers, except now with two constraints and thus two multipliers. Define $f(\mathbf{x}) \triangleq \mathbf{x}^T A\mathbf{x}$, $f_1(\mathbf{x}) \triangleq \mathbf{x}^T\mathbf{x}-1$, and $f_2(\mathbf{x}) \triangleq \mathbf{x}^T\mathbf{x}^{(1)}$. Symmetry of A implies that, as in Example 13.14, we have

$$\nabla f = 2A\mathbf{x}, \quad \nabla f_1 = 2\mathbf{x}, \quad \text{and} \quad \nabla f_2 = \mathbf{x}^{(1)}.$$

We conclude that there exist multipliers μ_1, μ_2 and global minimizer $\mathbf{x}^\star$ for which stationarity holds, that is,

$$2A\mathbf{x}^\star = 2\mu_1\mathbf{x}^\star + \mu_2\mathbf{x}^{(1)} \tag{13.53}$$

and

$$||\mathbf{x}^\star||^2 - 1 = 0 = \left(\mathbf{x}^\star\right)^T\mathbf{x}^{(1)}. \tag{13.54}$$

Take the dot product of (13.53) with $\mathbf{x}^\star$ and use (13.54) to get

$$2\left(\mathbf{x}^\star\right)^T A\mathbf{x}^\star = 2\mu_1||\mathbf{x}^\star||^2 + \mu_2\left(\mathbf{x}^\star\right)^T\mathbf{x}^{(1)} = 2\mu_1 \cdot 1 + \mu_2 \cdot 0 = 2\mu_1.$$

This shows that μ_1 is the global maximum value of $\mathbf{x}^T A\mathbf{x}$ subject to the constraints (13.54).

On the other hand, take the dot product of (13.53) with the unit vector $\mathbf{x}^{(1)}$ and use (13.54) to get

$$2\left(\mathbf{x}^{(1)}\right)^T A\mathbf{x}^\star = 2\mu_1\left(\mathbf{x}^{(1)}\right)^T\mathbf{x}^\star + \mu_2||\mathbf{x}^{(1)}||^2 = 2\mu_1 \cdot 0 + \mu_2 \cdot 1 = \mu_2. \tag{13.55}$$

But, symmetry of A implies

$$\left(\mathbf{x}^{(1)}\right)^T A\mathbf{x}^\star = \left(\mathbf{x}^{(1)}\right)^T A^T\mathbf{x}^\star = \left(A\mathbf{x}^{(1)}\right)^T\mathbf{x}^\star = \left(\lambda_1\mathbf{x}^{(1)}\right)^T\mathbf{x}^\star = \lambda_1\left(\mathbf{x}^{(1)}\right)^T\mathbf{x}^\star = \lambda_1 \cdot 0,$$

so (13.55) implies $\mu_2 = 0$. It follows that stationarity, (13.53), actually implies

$$2A\mathbf{x}^\star = 2\mu_1\mathbf{x}^\star + 0 \cdot \mathbf{x}^{(1)} = 2\mu_1\mathbf{x}^\star;$$

hence, $\lambda_2 \triangleq \mu_1$ is an eigenvalue of A with corresponding eigenvector $\mathbf{x}^\star$.

The last thing to notice is that $\lambda_2 \leq \lambda_1$ follows from

$$\begin{aligned}\lambda_1 &= \max\{\mathcal{R}_A(\mathbf{x}) : \mathbf{x} \text{ satisfying } ||\,\mathbf{x}\,|| = 1\} \\ &\geq \max\{\mathcal{R}_A(\mathbf{x}) : \mathbf{x} \text{ satisfying } ||\,\mathbf{x}\,|| = 1 \text{ and } \mathbf{x}^T\mathbf{x}^{(1)} = 0\} = \lambda_2.\end{aligned}$$

This is true because the maximum over a set is greater than or equal to the maximum over its subset, as we announced in Lemma 13.1 in Section 13.2.

We can continue in this way: After having found eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ and corresponding orthonormal set of eigenvectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}$, we can pose the problem

$$\left\{\begin{array}{ll} \text{Maximize} & \mathbf{x}^T A\mathbf{x} \\ \text{Subject to} & \mathbf{x}^T\mathbf{x} - 1 = 0 \\ & \mathbf{x}^T\mathbf{x}^{(1)} = 0 \\ & \vdots \\ & \mathbf{x}^T\mathbf{x}^{(k)} = 0 \end{array}\right\}. \tag{13.56}$$

In Problem 13.6.3.3, you will explain why the global minimizer of (13.56) gives an eigenvector corresponding to an eigenvalue $\lambda_{k+1} \leq \lambda_k$.

13.6.2 Hilbert Space Eigenvalue Problems

The next result for a linear operator on Hilbert space generalizes the Rayleigh method for a matrix on $\mathbb{R}^n$.

Theorem 13.18

Suppose A is a bounded, self-adjoint operator on a Hilbert space $\mathcal{H}$ and define

$$\mathcal{R}_A(\mathbf{x}) \triangleq \frac{\langle A\mathbf{x}, \mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle}.$$

Assume there exists* a specific vector $\mathbf{x}^{(1)}$ satisfying

$$\mathcal{R}_A(\mathbf{x}^{(1)}) = \lambda_1 = \max\{\mathcal{R}_A(\mathbf{x}) : \mathbf{x} \text{ satisfying } \mathbf{x} \neq \mathbf{0}\}.$$

Then $\mathbf{x}^{(1)}$ is an eigenvector of A.

Why? By Theorem 2.44(a) in Section 2.9, without loss of generality, we may assume that $||\mathbf{x}^{(1)}|| = 1$. "Perturb" $\mathbf{x}^{(1)}$ a little by considering

$$\mathbf{x} = \mathbf{x}^{(1)} + \varepsilon\mathbf{y},$$

where $\varepsilon > 0$, $||\,\mathbf{y}\,|| = 1$ and $\mathbf{y} \perp \mathbf{x}^{(1)}$; the latter and the Pythagorean theorem guarantee that

$$||\,\mathbf{x}\,||^2 = ||\mathbf{x}^{(1)}||^2 + ||\,\varepsilon\mathbf{y}\,||^2 = 1^2 + |\,\varepsilon\,|^2\,||\,\mathbf{y}\,||^2 = 1 + \varepsilon^2.$$

Because λ_1 is the maximum,

$$\lambda_1 \geq \mathcal{R}_A(\mathbf{x}) = \frac{\langle A\mathbf{x}, \mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle} = \frac{\langle A(\mathbf{x}^{(1)} + \varepsilon\mathbf{y}), \mathbf{x}^{(1)} + \varepsilon\mathbf{y}\rangle}{||\mathbf{x}^{(1)} + \varepsilon\mathbf{y}||^2} = \frac{\langle A(\mathbf{x}^{(1)} + \varepsilon\mathbf{y}), \mathbf{x}^{(1)} + \varepsilon\mathbf{y}\rangle}{1 + \varepsilon^2}. \tag{13.57}$$

* In an infinite dimensional Hilbert space, if the operator A is "compact," then an analytical argument can show that $\mathbf{x}^{(1)}$ exists.

Multiply through (13.57) by $1 + \varepsilon^2$; also, expand the numerator and use the fact that A is self-adjoint. This gives

$$(1 + \varepsilon^2)\lambda_1 \geq \langle A\mathbf{x}^{(1)}, \mathbf{x}^{(1)} \rangle + 2\varepsilon \langle A\mathbf{y}, \mathbf{x}^{(1)} \rangle + \varepsilon^2 \langle A\mathbf{y}, \mathbf{y} \rangle = \lambda_1 + 2\varepsilon \langle A\mathbf{y}, \mathbf{x}^{(1)} \rangle + \varepsilon^2 \langle A\mathbf{y}, \mathbf{y} \rangle.$$

Subtract λ_1 from both sides to get

$$\varepsilon^2 \lambda_1 \geq 2\varepsilon \langle A\mathbf{y}, \mathbf{x}^{(1)} \rangle + \varepsilon^2 \langle A\mathbf{y}, \mathbf{y} \rangle.$$

Divide both sides by ε; because it's positive, this won't change the direction of the inequality. So,

$$\varepsilon \lambda_1 \geq 2\langle A\mathbf{y}, \mathbf{x}^{(1)} \rangle + \varepsilon \langle A\mathbf{y}, \mathbf{y} \rangle.$$

Take the limit as $\varepsilon \to 0^+$ to get

$$0 \geq \langle A\mathbf{y}, \mathbf{x}^{(1)} \rangle. \tag{13.58}$$

Equation (13.58) is true as long as $||\,\mathbf{y}\,|| = 1$ and $\mathbf{y} \perp \mathbf{x}^{(1)}$. So, we can replace $\mathbf{y}$ by $-\mathbf{y}$ to get

$$0 \geq \langle A(-\mathbf{y}), \mathbf{x}^{(1)} \rangle.$$

Multiply through by (-1), which changes the direction of the inequality, so we get $0 \leq \langle A\mathbf{y}, \mathbf{x}^{(1)} \rangle$, that is,

$$\langle A\mathbf{y}, \mathbf{x}^{(1)} \rangle \geq 0. \tag{13.59}$$

Combining (13.58) and (13.59), we get

$$0 \geq \langle A\mathbf{y}, \mathbf{x}^{(1)} \rangle \geq 0.$$

Using this and self-adjointness of A, it follows that

$$0 = \langle A\mathbf{y}, \mathbf{x}^{(1)} \rangle = \langle \mathbf{y}, A\mathbf{x}^{(1)} \rangle = \langle A\mathbf{x}^{(1)}, \mathbf{y} \rangle.$$

So, every unit vector $\mathbf{y}$ that is orthogonal to $\mathbf{x}^{(1)}$ is also orthogonal to $A\mathbf{x}^{(1)}$. It follows from a Hilbert space generalization of Corollary 2.9 in Section 2.4 that $\mathbf{x}^{(1)}$ and $A\mathbf{x}^{(1)}$ are parallel; hence,

$$A\mathbf{x}^{(1)} = \alpha \mathbf{x}^{(1)}$$

for some scalar α. But we can explain why $\alpha = \lambda_1$:

$$\alpha \cdot 1 = \alpha \cdot ||\mathbf{x}^{(1)}||^2 = \alpha \langle \mathbf{x}^{(1)}, \mathbf{x}^{(1)} \rangle = \langle A\mathbf{x}^{(1)}, \mathbf{x}^{(1)} \rangle = \lambda_1 \langle \mathbf{x}^{(1)}, \mathbf{x}^{(1)} \rangle = \lambda_1.$$

So,

$$A\mathbf{x}^{(1)} = \lambda_1 \mathbf{x}^{(1)}. \; \square$$

13.6.3 Problems

1. Explain why the global maximizer of (13.56) gives an eigenvector corresponding to an eigenvalue $\lambda_{k+1} \leq \lambda_k$.

In problems 2 and 3, use the Rayleigh quotient to find approximations of the minimum and maximum eigenvalues of the 3×3 matrix found in problems 2.9.2.3 and 2.9.2.4, respectively. You may use *Mathematica* or MATLAB® as in Example 2.37 in Section 2.9. After that, use the Rayleigh–Ritz method to find an approximation of the middle eigenvalue of the matrix.

2. $A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 2 \end{bmatrix}$

3. $A = \begin{bmatrix} 2 & \sqrt{3} & 0 \\ \sqrt{3} & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix}$

Key Terms

affine function: Problem 13.5.3.9
basic feasible solution: Definition 13.10 in Section 13.3
complementarity condition: (13.31) in Section 13.5
conjugate gradient method: after (13.18) in Section 13.2
constrained optimization problem: (13.7) in Section 13.2
convex function: Definition 13.3 in Section 13.1, Definition 13.12 in Section 13.5
convex programming (CP) problem: (13.30) in Section 13.5
convex set: Definition 13.11 in Section 13.5
critical number: Definition 13.2 in Section 13.1
critical point: Definition 13.6 in Section 13.2
critical radius of insulation: Example 13.1 in Section 13.1
Dantzig's **simplex procedure**: Section 13.4
dual LP problem: Definition 13.15 in Section 13.5
feasible direction: Definition 13.4 in Section 13.2
feasible region: Definition 13.13 in Section 13.5
feasible solution: Definition 13.10 in Section 13.3
fully plastic moment: before (13.25) in Section 13.3
generalized normal equations: Problem 13.2.3.11
generalized or **weighted least squares**: Problem 13.2.3.11
geometric dimensioning, geometric : **tolerancing**: end of Section 13.5
global minimizer, global minimum (value): Definition 13.1 in Section 13.1, Definition 13.4 in Section 13.2
global maximizer: after Definition 13.1 in Section 13.1
Hessian matrix: Definition 13.8 in Section 13.2
interior: Definition 13.5 in Section 13.2
Lagrange multiplier(s): (13.8) in Section 13.2, (13.14) in Section 13.2, Definition 13.13 in Section 13.5
Lagrangian: Definition 13.13 in Section 13.5

line search after (13.18) in Section 13.2
line segment: Definition 13.11 in Section 13.5
linear inequalities: beginning of Section 13.3
Linear programming (LP) problem: beginning of Section 13.3
local (or relative) minimizer, **local minimum (value)**: Definition 13.1 in Section 13.1, Definition 13.4 in Section 13.2
matrix-tableau form: beginning of Example 13.11 in Section 13.4
method of steepest descents: after (13.18) in Section 13.2, Figure 13.8
minimum reduction: (13.26) in Section 13.4
minimization problem: (13.4) in Section 13.2
nonlinear programming problem: (13.27) in Section 13.5
objective function: (13.4) in Section 13.2
pivoting process: after Theorem 13.12 in Section 13.4
plasticity: before (13.25) in Section 13.3
Rayleigh quotient: (13.49) in Section 13.6
Rayleigh–Ritz method: (13.52) in Section 13.6
saddle point: Definition 13.7 in Section 13.2
slack variable: before Definition 13.9 in Section 13.3
Slater's condition: Definition 13.13 in Section 13.5
standard form for LP problem: Definition 13.9 in Section 13.3
stationarity: (13.32) in Section 13.5
stationary point: Definition 13.6 in Section 13.2
straightness tolerancing: end of Section 13.5
strictly feasible solution: Definition 13.13 in Section 13.5
strongly (or strictly) convex: Definition 13.3 in Section 13.1, Definition 13.12 in Section 13.5
unit cost reduction: beginning of Section 13.4.1
yield stress: before (13.25) in Section 13.3

Mathematica Commands

After Example 13.11 in Section 13.4:

$$\text{FindMinimum}[\{5x_1 + x_2 + 2x_3, -x_1 - x_2 - x_3 + x_4 == -20 \,\&\&$$
$$2x_1 + x_2 - x_3 + x_5 == 30 \,\&\& 3x_1 + x_2 + x_3 == 60 \,\&\&$$
$$x_1 \geq 0 \,\&\& x_2 \geq 0 \,\&\& x_3 \geq 0 \,\&\& x_4 \geq 0 \,\&\& x_5 \geq 0\}, \{x_1, x_2, x_3, x_4, x_5\}]$$

References

Aziz, A. The critical thickness of insulation. *Heat Transfer Engineering* 18, 61–91, 1997.

Chen, M.-C. and Fan, S.-K.S. Tolerance evaluation of minimum zone straightness using non-linear programming techniques: A spreadsheet approach. *Computers and Industrial Engineering* 43, 437–453, 2002.

Cheney, E.W. *Approximation Theory*. AMS Chelsea Publishing, New York, 1982, pp. 25–26.

Kirsch, U. *Structural Optimization: Fundamentals and Applications*. Springer-Verlag, New York, 1993.

14

Calculus of Variations

14.1 Minimization Problems

The Calculus of variations, as a systematic way of modeling and solving physical problems, is historically relevant but was also reinvigorated in the twentieth century in the subjects of both control theory and finite element methods.

Very natural contexts for Calculus of variations include engineering mechanics and electromagnetism where we use knowledge of the energy in the system. For example, it makes sense that if we deform a solid object then its new equilibrium shape should minimize its potential energy.

There are many physical problems that are modeled by a **Calculus of variations** problem of finding an "admissible" *function* $y(x)$ so as to

$$\left\{\begin{array}{ll} \text{Minimize} & J[\,y\,] \triangleq \displaystyle\int_a^b F\big(x, y(x), y'(x)\big)\,dx \\ \\ \text{Subject to} & \begin{array}{l} y(a) = y_a \\ y(b) = y_b \end{array} \end{array}\right\} \tag{14.1}$$

or subject to conditions on or including $y'(a), y'(b)$. Also, the integral may involve higher order derivatives of y. Often the problem has a natural interpretation in terms of energy in a system.

In Chapter 13 we studied minimization of a function that depends on several independent variables, that is, unknowns to be solved for. In Chapter 14 we will minimize an integral which depends on a function which is the unknown to be solved for. That is inherently a more mind boggling problem, but results from Chapter 13 will still be relevant.

A function is **admissible** if it is continuous and piecewise continuously differentiable on the interval $[\,a, b\,]$. If higher order derivatives are in the integrand then the class of admissible functions may be further restricted to involve higher order differentiability.

A **functional** is a mapping from a vector space to scalar values. For example, in (14.1) the mapping

$$y \mapsto \int_a^b F\big(x, y(x), y'(x)\big)\,dx$$

is a functional.

We will see in the rest of the chapter that there are techniques for turning a problem such as (14.1) into a differential equation with boundary conditions.

Here are some examples.

Example 14.1

Suppose a beam has rest state $y(x) \equiv 0$ for $0 \leq x \leq L$. A linear theory for the vertical deformation $y = y(x)$ of the beam subject to a constant external force field $f(x)$ in the vertical direction assumes the total potential energy stored in the beam is

$$J = J[\,y\,] = \int_0^L \left(\frac{1}{2}\,EI(y'')^2 + yf(x)\right) dx.$$

The first term in the integral is the strain energy. Assume that the beam is simply supported at the ends, so $u(0) = u(L) = 0$. Physically, it makes sense that the total potential energy should be at a minimum, and we will see in Section 14.2 that this works out well mathematically too. The calculus of variations problem is

$$\left\{\begin{array}{ll} \text{Minimize} & \displaystyle\int_0^L \left(\frac{1}{2}\,EI(y'')^2 + yf(x)\right) dx \\ \\ \text{Subject to} & y(0) = y(L) = 0 \end{array}\right\}. \tag{14.2}$$

We will see in Section 14.2 that by varying $y(x)$ we will arrive at the ODE-BVP

$$\left\{\begin{array}{c} -EIy'''' + f(x) = 0 \\ \\ y(0) = y(L) = 0, \quad y''(0) = y''(L) = 0 \end{array}\right\}.$$

The **natural boundary conditions** $y''(0) = y''(L) = 0$ will appear as a consequence of the minimization.

Example 14.2

Suppose a thin, "Euler–Bernoulli" rod has angular deflection $\theta = \theta(s)$, for $0 \leq s \leq L$, and is subject to a compressive end force P, as illustrated in Figure 14.1. The potential energy of the rod is

$$J[\,\theta\,] = \int_0^L \left(\frac{1}{2}\,EI\left(\frac{d\theta}{ds}\right)^2 - P(1 - \cos\theta)\right) ds.$$

The first term is the potential energy due to bending and the second term comes from the load on the rod. Assume that the rod is pinned at the ends, so $\theta'(0) = \theta'(L) = 0$,

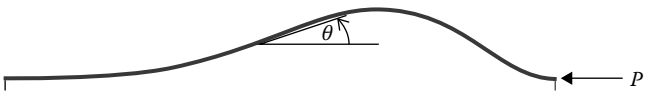


FIGURE 14.1
Example 14.2.

where $' = \dfrac{d}{ds}$. Physically, it makes sense that the total potential energy should be at a minimum, and we will see in Section 14.2 that this works out well mathematically too. The calculus of variations problem is

$$\left\{ \begin{array}{ll} \text{Minimize} & \displaystyle\int_0^L \left(\frac{1}{2} EI \left(\frac{d\theta}{ds} \right)^2 - P(1 - \cos\theta) \right) dx \\ \\ \text{Subject to} & \theta'(0) = \theta'(L) = 0 \end{array} \right\}. \tag{14.3}$$

We will see in Section 14.2 that by varying $\theta(s)$ we will arrive at the nonlinear ODE-BVP

$$\left\{ \begin{array}{l} EI\theta'' + P\sin\theta = 0 \\ \\ \theta'(0) = \theta'(L) = 0 \end{array} \right\}.$$

Example 14.3

Suppose a thin membrane has displacement $z = u(x, y, t)$ for (x, y) in some nice set $\mathcal{D}$ in $\mathbb{R}^2$. Suppose the edge of the membrane is kept at zero displacement, that is, $u = 0$, on the boundary of $\mathcal{D}$, and assume the force of gravity is negligible. By assuming a solution in the form

$$u(x, y, t) = e^{i\omega t}\phi(x, y),$$

we are led to the eigenvalue problem for the Laplacian given by

$$\left\{ \begin{array}{l} 0 = \dfrac{\partial^2\phi}{\partial x^2} + \dfrac{\partial^2\phi}{\partial y^2} + \omega^2\phi \\ \\ 0 \equiv \phi\big|_{(x,y) \text{ on boundary of } \mathcal{D}} \end{array} \right\}.$$

We will see in Section 14.4 that this problem is equivalent to the calculus of variations problem

$$\left\{ \begin{array}{ll} \text{Minimize} & \displaystyle\iint_{\mathcal{D}} \left(\left(\frac{\partial\phi}{\partial x} \right)^2 + \left(\frac{\partial\phi}{\partial y} \right)^2 \right) dx\, dy \\ \\ \text{Subject to} & \displaystyle\iint_{\mathcal{D}} |\phi(x, y)|^2\, dx\, dy = 1 \end{array} \right\}. \tag{14.4}$$

Example 14.3 is a generalization of work on the Rayleigh Quotient in Section 9.5.

Example 14.4

(The **brachistochrone problem**) Suppose a bead starts at rest and then, under the influence of the force of gravity, slides (not rolls!) without friction down the curve $z = f(x)$, as shown in Figure 14.2. Set up a calculus of variations problem whose solution would give the curve that will minimize the time it takes for the bead to reach the origin.

Method: Recall from Section 7.2 that the arclength, s, along the curve satisfies

$$\frac{ds}{dx} = \sqrt{1 + (f'(x))^2}.$$

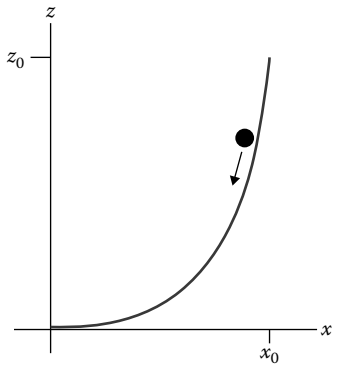


FIGURE 14.2
Example 14.4.

hence

$$ds = \sqrt{1 + (f'(x))^2}\, dx. \tag{14.5}$$

Suppose the bead has mass m and that during its slide its position vector is

$$\mathbf{r} = \mathbf{r}(t) = x(t)\,\hat{\boldsymbol{\imath}} + z(t)\,\hat{\boldsymbol{k}}.$$

The total energy, that is, sum of the kinetic and potential energies, is

$$E = \frac{1}{2}mv^2 + mgz.$$

This is conserved, so

$$\big(v(t)\big)^2 + 2gz(t) \equiv 0^2 + 2gz_0, \tag{14.6}$$

after recalling that the bead starts from rest at position $(x, z) = (x_0, z_0)$. But, (14.6) implies that the velocity of the particle is

$$0 < \frac{ds}{dt} = ||\dot{\mathbf{r}}(t)|| = |v(t)| = \sqrt{2g\,(z_0 - z(t))},$$

and $z(t) = f(x(t))$, so

$$ds = \sqrt{2g}\sqrt{z_0 - f(x(t))}\, dt,$$

hence

$$\frac{dt}{ds} = \frac{1}{\sqrt{2g}\sqrt{z_0 - f(x(t))}}.$$

It follows from (14.5) that

$$dt = \frac{dt}{ds}\, ds = \frac{1}{\sqrt{2g}\sqrt{z_0 - f(x)}} \sqrt{1 + (f'(x))^2}\, dx.$$

So, the total time required for the journey is

$$t = \int_{x_0}^{0} \frac{1}{\sqrt{2g}\sqrt{z_0 - f(x)}} \sqrt{1 + (f'(x))^2}\, dx. \tag{14.7}$$

The calculus of variations problem is to minimize

$$J[f] \triangleq \int_{x_0}^{0} \frac{1}{\sqrt{2g}\sqrt{z_0 - f(x)}} \sqrt{1 + (f'(x))^2}\, dx \tag{14.8}$$

subject to $f(x_0) = z_0$ and $f'(x)$ is continuously differentiable on $[0, x_0]$. □

It turns out that the curve that gives the shortest *time* is not a straight line from (x_0, z_0) to $(0, 0)$ despite the fact that that would give the shortest *path*. Why not? For one thing, the velocity and the speed of the bead are not constant during its journey.

14.1.1 The Rayleigh–Ritz Method

Suppose that abstractly we have a problem

$$Au = f \quad \text{and} \quad u \text{ is in } \mathcal{V}. \tag{14.9}$$

Then we will see that we can use the corresponding calculus of variations problem of minimizing

$$J[u] \triangleq \langle Au, u\rangle - 2\langle u, f\rangle \tag{14.10}$$

to find an approximate solution. Moreover, in some problems the approximate minimum value of $J[u]$ has physical significance.

For example, in Section 12.5 the problem $\big(p(x)y'(x)\big)' + q(x)y(x) = f(x)$, $y(0) = y(L) = 0$ was rewritten in the form (14.9) using a differential operator A defined by

$$(Au)(x) \triangleq -\big(p(x)u'(x)\big)' + q(x)u(x)$$

on functions $u = u(x)$ in the space $\mathcal{V} \triangleq C_0^2[0, L]$ defined by

u is in $\mathcal{V}$ if $u(x)$ is twice continuously differentiable on $[0, L]$ and $u(0) = u(L) = 0$.

The Rayleigh–Ritz method we used in Section 9.7 for eigenvalue problems can be used to find an approximate solution for a problem in the form (14.9): If u is a minimizer for the problem

$$J[y] = \int_0^L \Big(p(x)\big(y'(x)\big)^2 + q(x)\big(y(x)\big)^2 - 2y(x)f(x)\Big)\, dx$$

and satisfies the BCs $u(0) = u(L) = 0$ then it should satisfy the ODE $\big(p(x)u'(x)\big)' + q(x)u(x) = f(x)$.

Example 14.5

Use the Rayleigh–Ritz method to find an approximate solution of

$$\left\{\begin{array}{l} -\big(\varrho(x)y'\big)' = f(x) \\ y(0) = y(1) = 0 \end{array}\right\}, \tag{14.11}$$

where mass density $\varrho(x) \triangleq 3+\cos(\pi x)$ and load $f(x) = 0.2x$. This models a loaded string that is fixed at the ends $x = 0$ and $x = 1$.

Method: It makes sense to look for an approximate solution that also satisfies the boundary conditions, for example, in the form

$$y(x) = c_1 x(1-x) + c_2 x^2(1-x). \tag{14.12}$$

In Section 9.7 we referred to this as using a family of "trial" functions.

Using *Mathematica*^{TM} we found that

$$\begin{aligned} f(c_1,c_2) &\triangleq \int_0^1 \Big(\varrho(x)\big(y'(x)\big)^2 - 2y(x)f(x)\Big)\,dx \\ &= \int_0^1 \Big(\big(3+\cos(\pi x)\big)\big(c_1(1-2x)+c_2(2x-3x^2)\big)^2 \\ &\quad - 2\big(c_1x(1-x)+c_2x^2(1-x)\big)0.2x\Big)dx \\ &= c_1^2 + c_1\left(-\frac{1}{30}+c_2+\frac{144\,c_2}{\pi^4}-\frac{16\,c_2}{\pi^2}\right) \\ &\quad + \frac{c_2}{50}\left(-1+\frac{20\,c_2(180-20\pi^2+\pi^4)}{\pi^4}\right) \end{aligned} \tag{14.13}$$

and the approximate $\min_{c_1,c_2} f(c_1,c_2)$ is achieved at $c_1 = 0.008216334900156332$, $c_2 = 0.019716988798084503$. In Figure 14.3 we show the approximate solution of the ODE-BVP, namely,

$$y^\star(x) \triangleq 0.008216334900156332x(1-x) + 0.019716988798084503x^2(1-x)$$

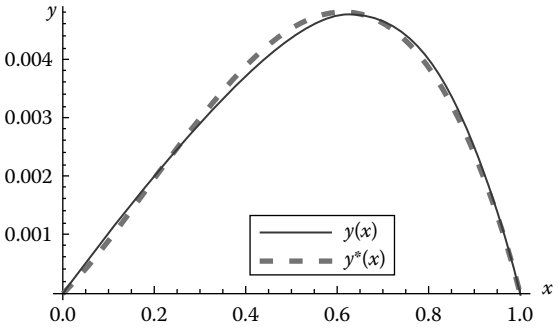


FIGURE 14.3
Example 14.5.

as a dashed graph and also *Mathematica*'s approximate solution $y(x)$ in the solid, graph.

The minimization in the Rayleigh–Ritz method was done by the command

$$\text{FindMinimum}\left[c_1^2 + c_1\left(-\frac{1}{30} + c_2 + \frac{144\,c_2}{\pi^4} - \frac{16\,c_2}{\pi^2}\right) + \frac{c_2}{50}\left(-1 + \frac{20\,c_2(180 - 20\pi^2 + \pi^4)}{\pi^4}\right), c_1, c_2\right]$$

and *Mathematica*'s approximate solving of the ODE-BVP was done using

$$\text{s} = \text{NDSolve}[\{y''[x] == \frac{1}{3 + \text{Cos}[\pi x])}\left(y'[x] * \pi * \text{Sin}[\pi x] - \frac{x}{5}\right),$$

$$y[0] == 0, y[1] == 0\}, \{y\}, x]//\text{Quiet}$$

followed by

$$\text{b=Plot}[\text{Evaluate}[\{y[x]\}/.s], \{x, 0, 1\}, \text{PlotStyle} \to \{\text{Blue, Thick}\},$$

$$\text{LabelStyle} \to \text{Directive}[\text{FontSize} \to 16],$$

$$\text{AxesLabel} \to \{x, y\}, \text{AxesStyle} \to \text{Thickness}[0.00315]].$$

14.1.2 Problems

1. Use the Rayleigh–Ritz method to find an approximate solution of

$$\left\{\begin{array}{c} -\big((3 + \cos x)y'\big)' = 0.1x \\ y'(0) = y(\pi) = 0 \end{array}\right\}$$

by using an approximate solution of the form $y(x) = c_1x^2(\pi - x) + c_2 \cos \frac{x}{2}$.

In problems 2 and 3, use the Rayleigh–Ritz method to find an approximate solution of the ODE-BVP.

2. $$\left\{\begin{array}{c} -\frac{d}{dr}\left[r\frac{du}{dr}\right] - u = r^3 \\ |\,u(0^+)\,| < \infty,\ u(1) = 0 \end{array}\right\}$$

3. $$\left\{\begin{array}{c} -\frac{d}{dr}\left[r\frac{du}{dr}\right] - 4r^2u = r^3 \\ |\,u(0^+)\,| < \infty,\ u(1) = 0 \end{array}\right\}$$

14.2 Necessary Conditions

Example 14.6

Suppose we have a twice continuously differentiable solution function $\theta_0 = \theta_0(s)$ for (14.3) in Section 14.1, that is, $J[\,\theta_0\,]$ is the minimum of

$$J[\,\theta\,] \triangleq \int_0^L \left(\frac{1}{2} EI\left(\frac{d\theta}{ds}\right)^2 - P(1-\cos\theta)\right) ds$$

subject to the requirements that $\theta\,'(0) = \theta\,'(L) = 0$. Explain why $\theta_0(s)$ must satisfy the nonlinear ODE

$$EI\theta\,'' + P\sin\theta = 0, \quad 0 < s < L. \tag{14.14}$$

Method: Let $\varepsilon_0 > 0$. Consider a family of functions given by

$$\theta(s) \triangleq \theta_0(s) + \varepsilon\phi(s), \quad -\varepsilon_0 \le \varepsilon \le \varepsilon_0, \tag{14.15}$$

where $\phi(s)$ is continuously differentiable on $(0, L)$ and $\phi\,'(0) = \phi\,'(L) = 0$. The boundary conditions on $\phi(s)$ are imposed so that $\theta\,'(0) = \theta\,'(L) = 0$.

The values of $J[\,\theta\,]$ on that family give a function of the single variable ε:

$$f(\varepsilon) \triangleq J[\,\theta_0(s) + \varepsilon\phi(s)] = \int_0^L \left(\frac{1}{2} EI\big(\theta\,'_0(s) + \varepsilon\phi'(s)\big)^2 - P\left(1 - \cos\left(\theta_0(s) + \varepsilon\phi(s)\right)\right)\right) ds$$

$$= \int_0^L \frac{1}{2} EI\left(\big(\theta\,'_0(s)\big)^2 + 2\varepsilon\theta\,'_0(s)\phi'(s) + \big(\varepsilon\phi'(s)\big)^2\right) ds - \int_0^L P\left(1 - \cos\left(\theta_0(s)\right)\right) ds,$$

$$- \int_0^L P\left(\sin\left(\theta_0(s)\right)\varepsilon\phi(s) \quad + \frac{1}{2}\cos\left(\theta_0(s)\right)(\varepsilon\phi(s))^2 + \cdots\right) ds,$$

for small ε, by using the MacLaurin series for the cosine function. So,

$$f(\varepsilon) = f(0) + \varepsilon \int_0^L \left(EI\theta_0'(s)\phi\,'(s) - P\sin\left(\theta_0(s)\right)\phi(s)\right) ds + \cdots.$$

We assumed that $\theta_0(s)$ is a global minimizer of $J[\,\theta\,]$, so we must have

$$\frac{df}{d\varepsilon}(0) = 0 \quad \text{and} \quad \frac{d^2f}{d\varepsilon^2}(0) \ge 0.$$

The first conclusion, the equality $f'(0) = 0$, is called "stationarity." It implies that

$$0 = f'(0) = \int_0^L \left(EI\theta_0'(s)\phi\,'(s) - P\sin\left(\theta_0(s)\right)\phi(s)\right) ds$$

$$= \int_0^L EI\theta_0'(s)\phi\,'(s)\, ds - \int_0^L P\sin\left(\theta_0(s)\right)\phi(s)\, ds.$$

Using integration by parts and the assumption that θ_0 is *twice* continuously differentiable, this gives

$$0 = f'(0) = \left[\, EI\theta_0'(s)\phi(s) \,\right]_0^L - \int_0^L EI\theta_0''(s)\phi(s)\, ds - \int_0^L P\sin\big(\theta_0(s)\big)\phi(s)\, ds.$$

We assumed that the minimizer satisfies the boundary conditions, so the first two terms are zero. This gives

$$0 = f'(0) = \int_0^L \big(-EI\theta_0''(s) - P\sin\big(\theta_0(s)\big)\big)\phi(s)\, ds.$$

Since this is true for *all* functions $\phi(s)$ that are continuously differentiable on $[\,0, L\,]$ and satisfy $\phi\,'(0) = \phi\,'(L) = 0$, the terms in the parentheses must add up to be identically zero, that is,

$$-EI\theta_0''(s) - P\sin\big(\theta_0(s)\big) \equiv 0, \quad 0 < s < L.$$

This says that *if* $\theta_0(s)$ is a global minimizer *then necessarily* it must satisfy the nonlinear ODE (14.14), that is,

$$EI\theta\,'' + P\sin\theta = 0,\ 0 < s < L. \bigcirc$$

So, analogously to our work in Chapter 10 for partial differential equations, we see that a model involving integration leads to a differential equation.

14.2.1 Euler–Lagrange Equations

The method used to apply stationarity to the minimization of

$$J[\,\theta\,] = \int_0^L \left(\frac{1}{2}\, EI\left(\frac{d\theta}{ds}\right)^2 - P(1-\cos\theta)\right) ds$$

was based on varying the function θ to be (14.15), that is,

$$\theta(s) \triangleq \theta_0(s) + \varepsilon\,\phi(s),$$

where $\phi(s)$ is continuously differentiable on $(0, L)$ and $\phi\,'(0) = \phi\,'(L) = 0$. It helps to think of the function $\varepsilon\phi(s)$ as being a small change of the function $\theta(s)$. Thinking of ε as an arbitrarily small real number, we define the "variation" of $\theta(s)$ as being $\varepsilon\,\phi(s)$, that is,

$$\delta\theta \triangleq \varepsilon\,\phi(s).$$

Even though we suppress the "(s)" from the notation of $\delta\theta$, we must keep in mind that $\delta\theta$ is a function that is free to vary, at least within the requirements that $\delta\theta(s)$ be continuously differentiable on $(0, L)$ and $\delta\theta\,'(0) = \delta\theta\,'(L) = 0$.

Similarly, the **variation** $\delta\theta'$ is a function that is free to vary, at least within the requirements that it be continuous on $(0, L)$ and have $\int_0^L \delta\theta'(s)\, ds = 0$. The latter follows from $\int_0^L \delta\theta'(s)\, ds = [\delta\theta(s)]_0^L$, which follows from a consistency assumption that

$$(\delta\theta(s))' = \delta(\theta\,')(s). \tag{14.16}$$

The idea of linear approximation was used in calculus of functions of a single variable and calculus of functions of several variables. For example, in $\mathbb{R}^3$, $f(\mathbf{r})$, that is, $f(x, y, z)$, has

$$f(\mathbf{r}_0 + \triangle\mathbf{r}) \approx f(\mathbf{r}_0) + \nabla f(\mathbf{r}_0) \bullet (\triangle\mathbf{r}) = f(\mathbf{r}_0) + \frac{\partial f}{\partial x}(\mathbf{r}_0)\triangle x + \frac{\partial f}{\partial y}(\mathbf{r}_0)\triangle y + \frac{\partial f}{\partial z}(\mathbf{r}_0)\triangle z.$$

As we saw in Chapter 9 when studying Fourier series, varying a function can be done in infinitely many "directions," not just the two directions of $\mathbb{R}^2$ or the n directions of $\mathbb{R}^n$.

Analogously, we have

$$J[\,\theta + \delta\theta\,] \approx J[\,\theta\,] + \delta J.$$

Examining the terms in the integral, we have

$$\delta\left(\frac{1}{2}\,EI\left(\frac{d\theta}{ds}\right)^2\right) = \big(EI\theta'(s)\big)\cdot\delta\theta\,'$$

and

$$\delta\left(-P\big(1-\cos\theta(s)\big)\right) = \big(-P\sin\theta(s)\big)\cdot\delta\theta,$$

so

$$\delta J = \int_0^L \big((EI\theta\,'(s))\cdot\delta\theta\,' - (P\sin\theta(s))\cdot\delta\theta\big)\,ds.$$

Stationarity requires $0 = \delta J$. After use of integration by parts, this becomes

$$0 = \delta J = \big[(EI\theta\,'(s))\cdot\delta\theta\big]_0^L - \int_0^L \Big(\big(EI\theta\,'(s)\big)'\cdot\delta\theta - \big(P\sin\theta(s)\big)\cdot\delta\theta\Big)\,ds$$

$$= EI\theta\,'(L)\cdot\delta\theta(L) - EI\theta\,'(0)\cdot\delta\theta(0) - \int_0^L \big(EI\theta\,''(s) + P\sin\theta(s)\big)\cdot\delta\theta\,ds.$$

Because of the boundary conditions satisfied by $\theta(s)$, we get

$$0 = \int_0^L \big(EI\theta\,''(s) + P\sin\theta(s)\big)\cdot\delta\theta\,ds.$$

Because the variation $\delta\theta(s)$ is arbitrary except for the requirements that it be continuously differentiable on $(0, L)$ and satisfy $\delta\theta\,'(0) = \delta\theta\,'(L) = 0$, we conclude that

$$EI\theta\,''(s) + P\sin\theta(s) \equiv 0, \quad 0 < s < L.$$

Note that

$$\delta J = \frac{\partial J}{\partial \varepsilon}[\,\theta + \varepsilon\phi\,]|_{\varepsilon=0}\,. \tag{14.17}$$

Let's generalize this "variational" method.

Theorem 14.1

Suppose $F = F(x, y, y')$ is continuously differentiable and suppose $y = y_0(x)$ solves (14.1), in Section 14.1, that is,

$$\left\{\begin{array}{ll} \text{Minimize} & J[\,y\,] \triangleq \int_a^b F\big(x, y(x), y'(x)\big)\,dx \\ \\ \text{Subject to} & \begin{array}{l} y(a) = y_a \\ y(b) = y_b \end{array} \end{array}\right\}.$$

Then necessarily, at $\big(x, y(x), y'(x)\big) = \big(x, y_0(x), y_0'(x)\big)$, we have the **Euler–Lagrange equation**

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\left[\,\frac{\partial F}{\partial y'}\,\right] \equiv 0,\ a < x < b. \tag{14.18}$$

Why is (14.18) true? We are allowed to vary $\delta y = \delta y(x)$ and $\delta y' = \delta y'(x)$ arbitrarily, except for requiring that δy be continuously differentiable on (a, b), $\delta y(a) = \delta y(b) = 0$, $\delta y'$ be piecewise continuous on (a, b), and $\int_a^b \delta y'(x)\,dx = 0$. The second assumption comes from the need for $y_0(x) + \delta y(x)$ to satisfy the boundary conditions $y(a) = y_a, y(b) = y_b$, and the last assumption follows from the second assumption and consistency, that is, $(\delta y)' = \delta(y')$.

Analogous to linear approximation in $\mathbb{R}^3$, we have

$$\delta F(x, y, y') = \frac{\partial F}{\partial y}(x, y, y') \cdot \delta y + \frac{\partial F}{\partial y'}(x, y, y') \cdot \delta y'.$$

There is no partial derivative of F with respect to x term because x is not being varied in the minimization process. So,

$$\delta J = \int_a^b \delta F(x, y, y')dx = \int_a^b \left(\frac{\partial F}{\partial y}\big(x, y(x), y'(x)\big) \cdot \delta y(x) + \frac{\partial F}{\partial y'}\big(x, y(x), y'(x)\big) \cdot \delta y'(x) \right) dx.$$

Concerning the last term, integration by parts gives

$$\int_a^b \frac{\partial F}{\partial y'}\big(x, y(x), y'(x)\big) \cdot \delta y'(x)\, dx =$$

$$= \left[\frac{\partial F}{\partial y'}\big(x, y(x), y'(x)\big) \cdot \delta y(x) \right]_a^b - \int_a^b \left(\frac{d}{dx} \left[\frac{\partial F}{\partial y'}\big(x, y(x), y'(x)\big) \right] \right) \cdot \delta y(x)\, dx.$$

But, $\delta y(a) = \delta y(b) = 0$, so

$$\delta J = \int_a^b \frac{\partial F}{\partial y}\big(x, y(x), y'(x)\big) \cdot \delta y(x)\, dx - \int_a^b \left(\frac{d}{dx} \left[\frac{\partial F}{\partial y'}\big(x, y(x), y'(x)\big) \right] \right) \cdot \delta y(x)\, dx.$$

Stationarity at a global minimum gives

$$0 = \delta J = \int_a^b \left(\frac{\partial F}{\partial y}\big(x, y_0(x), y_0'(x)\big) - \frac{d}{dx} \left[\frac{\partial F}{\partial y'}\big(x, y_0(x), y_0'(x)\big) \right] \right) \cdot \delta y(x)\, dx.$$

Because $\delta y(x)$ is arbitrary except for the requirements that it be continuously differentiable and satisfy $\delta y(a) = \delta y(b) = 0$, we conclude that (14.18) holds, that is,

$$\frac{\partial F}{\partial y}\big(x, y_0(x), y_0'(x)\big) - \frac{d}{dx} \left[\frac{\partial F}{\partial y'}\big(x, y_0(x), y_0'(x)\big) \right] \equiv 0, \quad a < x < b. \ \square$$

Example 14.7

Explain why the result of Example 14.6 agrees with the Euler–Lagrange equation of Theorem 14.1.

Method: In Example 14.6, $F(x, y, y') = \frac{1}{2} EI(y')^2 - P(1 - \cos y)$, where $y = \theta(s)$ and $y' = \theta'(s)$, s is the independent variable instead of x, and $' = \frac{d}{ds}$. The Euler–Lagrange equation is

$$0 = \frac{\partial F}{\partial y} - \frac{d}{ds} \left[\frac{\partial F}{\partial y'} \right] = -P \sin y - \frac{d}{ds} \left[EIy' \right] = -P \sin \theta - EIy'' = -P \sin \theta - EI\theta''.$$

This agrees with our conclusion in Example 14.6. ○

Theorem 14.1's result is powerful and enables us to avoid having to "reinvent the wheel." In the subject of Advanced Engineering Mathematics this is exactly what we want to have in our toolbox.

Thus, we also see why much of engineering's foundations in the physical sciences can be formulated on the basic principle of minimizing potential energy.

Example 14.8

Assuming $\mathcal{D}$ is a nice planar domain, find a PDE satisfied by the solution of the minimization problem

$$\left\{\begin{array}{ll}\text{Minimize} & \displaystyle\iint_{\mathcal{D}}\left(\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right)dA\\[2ex] \text{Subject to} & u=g(x,y)\text{ on }\partial\mathcal{D}\end{array}\right\}. \tag{14.19}$$

Method: Define

$$J[\,u\,]=\iint_{\mathcal{D}}\left(\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right)dA.$$

Let $u_0(x,y)$ be a solution of (14.19) and $\delta u(x,y)$ be its variation. Both u_0 and $u=u_0+\delta u$ must satisfy the Dirichlet boundary condition $u\equiv g(x,y)$ on $\partial\mathcal{D}$, so $\delta u\equiv 0$ on $\partial\mathcal{D}$. We calculate

$$\begin{aligned}J[\,u_0+\delta u\,]&=\iint_{\mathcal{D}}\left(\left(\frac{\partial u_0}{\partial x}+\frac{\partial(\delta u)}{\partial x}\right)^2+\left(\frac{\partial u_0}{\partial y}+\frac{\partial(\delta u)}{\partial y}\right)^2\right)dA\\ &=\iint_{\mathcal{D}}\left(\left(\frac{\partial u_0}{\partial x}\right)^2+2\frac{\partial u_0}{\partial x}\cdot\frac{\partial(\delta u)}{\partial x}+\left(\frac{\partial(\delta u)}{\partial x}\right)^2+\left(\frac{\partial u_0}{\partial y}\right)^2\right.\\ &\qquad\left.+2\frac{\partial u_0}{\partial y}\cdot\frac{\partial(\delta u)}{\partial y}+\left(\frac{\partial(\delta u)}{\partial y}\right)^2\right)dA,\end{aligned}$$

so

$$\delta J=2\iint_{\mathcal{D}}\left(\frac{\partial u_0}{\partial x}\cdot\frac{\partial(\delta u)}{\partial x}+\frac{\partial u_0}{\partial y}\cdot\frac{\partial(\delta u)}{\partial y}\right)dA=2\iint_{\mathcal{D}}(\boldsymbol{\nabla}u_0)\bullet(\boldsymbol{\nabla}\delta u)\,dA.$$

Corollary 6.1 in Section 6.7 implies $\boldsymbol{\nabla}u_0\bullet\boldsymbol{\nabla}\delta u=\boldsymbol{\nabla}\bullet(\delta u\,\boldsymbol{\nabla}u_0)-\delta u\nabla^2u_0$, so the divergence theorem implies

$$\delta J=2\oint_{\partial\mathcal{D}}\big(\delta u\,\boldsymbol{\nabla}u_0)\big)\bullet\widehat{\mathbf{n}}\,ds-2\iint_{\mathcal{D}}\delta u\,\nabla^2u_0\,dA=0-2\iint_{\mathcal{D}}\delta u\,\nabla^2u_0\,dA,$$

because $\delta u\equiv 0$ on $\partial\mathcal{D}$. Other than the requirement that $\delta u\equiv 0$ on $\partial\mathcal{D}$, arbitrariness of δu implies stationarity, that is, $\delta J=0$, which implies

$$\nabla^2u_0=0\text{ in }\mathcal{D},$$

that is, u_0 satisfies Laplace's equation in $\mathcal{D}$. ◯

Theorem 14.1 and Example 14.8 generalize to problems in which the minimizer is a function of two independent variables:

Corollary 14.1

Suppose $F = F\left(x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right)$ is continuously differentiable and $u = u_0(x, y)$ solves

$$\left\{\begin{array}{ll} \text{Minimize} & J[\,u\,] \triangleq \iint\limits_{\mathcal{D}} F\left(x, y, u, \dfrac{\partial u}{\partial x}, \dfrac{\partial u}{\partial y}\right) dx\, dy \\ \\ \text{Subject to} & u = g(x, y) \text{ on } \partial\mathcal{D} \end{array}\right\}.$$

Then necessarily we have the Euler–Lagrange equation

$$\left(\frac{\partial F}{\partial u} - \frac{\partial}{\partial x}\left[\frac{\partial F}{\partial(\partial u/\partial x)}\right] - \frac{\partial}{\partial y}\left[\frac{\partial F}{\partial(\partial u/\partial y)}\right]\right)\Big|_{\text{at}\left(x, y, u_0(x,y)\right)} \equiv 0, \ (x, y) \text{ in } \mathcal{D}. \tag{14.20}$$

14.2.2 Natural Boundary Condition

Example 14.9

Explain why the solution of the minimization problem

$$\left\{\begin{array}{ll} \text{Minimize} & J[\,y\,] \triangleq -\int\limits_0^L \left(\left(y'(x)\right)^2 + 2y(x)f(x)\right) dx \\ \\ \text{Subject to} & y(0) = 0 \end{array}\right\} \tag{14.21}$$

must solve the ODE-BVP

$$\left\{\begin{array}{l} y''(x) = f(x) \\ y(0) = y'(L) = 0 \end{array}\right\}. \tag{14.22}$$

The latter is a model for physical problem of the equilibrium configuration of a loaded string with free right end. Notice that the ODE-BVP includes both $y(0) = 0$, the **fixed end** boundary condition in the minimization problem, and also $y'(L) = 0$, the **free end** boundary condition. The latter is also called a **natural boundary condition**.

Method: Notice that our problem is not directly addressed by Theorem 14.1 because the minimization problem does not assume a fixed end condition at the endpoint $x = L$.

Suppose $y_0(x)$ solve the minimization problem (14.21). Within the admissible class of piecewise continuously differentiable functions $y(x)$ defined on $[\,0, L\,]$ we are free to vary both

$$y(x) = y_0(x) + \delta y(x), \quad 0 < x < L,$$

and

$$y(L) = y_0(L) + \delta y(L).$$

The two kinds of variation are independent of each other. But, in order for $y_0(x) + \delta y(x)$ to continue to satisfy the boundary condition $y(0) = 0$, we must have

$$\delta y(0) = 0. \tag{14.23}$$

We have

$$J[\,y_0+\delta y\,] = -\int_0^L \big(y_0'(x)+\delta y'(x)\big)^2 + 2\big(y_0(x)+\delta y(x)\big)f(x)\big)\,dx$$

$$= -\int_0^L \Big(\big(y_0'\big)^2 + 2y_0'(\delta y') + \big(\delta y'\big)^2 + 2y_0 f + 2(\delta y)f\Big)\,dx$$

$$= -\int_0^L \Big(\big(y_0'\big)^2 + 2y_0 f\Big)\,dx - \int_0^L \big(2y_0'(\delta y') + 2(\delta y)f\big)\,dx - \int_0^L \big(\delta y'\big)^2.$$

So, stationarity of J at y_0 requires that

$$0 = \delta J = -\int_0^L \big(2y_0'\big(\delta y'(x)\big) + 2(\delta y)f\big)\,dx.$$

Integration by parts gives that the first term can be rewritten using

$$-\int_0^L \big(2y_0'\,\delta y'(x)\,dx = -\big[\,2y_0'(x)\big(\delta y(x)\big)\,\big]_0^L + \int_0^L 2y_0''\,\delta y(x)\,dx.$$

Using this and (14.23), that is, $\delta y(0)=0$, we conclude that stationarity of J at y_0 requires that

$$0 = -2y_0'(L)\delta y(L) + 2\int_0^L \big(y_0''(x) - f(x)\big)\delta y(x)\,dx.$$

Varying $\delta y(x)$, $0 < x < L$, and $\delta y(L)$ independently implies that, respectively, both

$$y_0'(L) = 0$$

and

$$\big(y_0''(x) - f(x)\big) \equiv 0, \quad 0 < x < L.$$

Putting together (a) the fixed end BC $y_0(0) = 0$, (b) the "natural boundary condition" $y_0'(L) = 0$, and (c) the ODE $y_0''(x) - f(x) \equiv 0$, we conclude that the solution of the calculus of variations problem, y_0, satisfies the ODE-BVP (14.22). ◯

14.2.3 Hamilton's Principle

Problems of dynamics can be stated as Hamilton's equations in terms of convenient state variables, and Hamilton's principle implies that the solution is a stationary point of a functional.

For example, the motion of a mass on the end of a rod free to move in a plane is constrained to lie on a circle. This is called a **pendulum**. We studied the potential energy in this system in Problem 7.2.5.21, as an application of line integrals.

If a system has total potential energy V and kinetic energy T, the **Lagrangian** L is defined by

$$L \triangleq T - V.$$

The total **action** during a time interval $[t_1, t_2]$ is defined to be

$$I \triangleq \int_{t_1}^{t_2} L\, dt.$$

Hamilton's principle states that the motion makes the action stationary, that is, has

$$\delta I = 0.$$

Example 14.10

Study Hamilton's principle for the motion of a mass m on an undamped nonlinear pendulum of length ℓ. As in Figure 7.17, define φ to be the angle with respect to vertical, with $\varphi = 0$ for the pendulum normally at rest and $\varphi = \pi$ for the inverted pendulum. Then the total potential energy is $V = mgz = mg\ell(1 - \cos\varphi)$ and the kinetic energy is $T = \frac{1}{2}\, m(\ell\dot{\varphi})^2$.

Method: The action is

$$I = \int_{t_1}^{t_2} \left(\frac{1}{2}\, mg(\ell\dot{\varphi})^2 - mg\ell(1 - \cos\varphi) \right) dt,$$

and, by integration by parts, its variation is

$$\begin{aligned}
\delta I &= \int_{t_1}^{t_2} m\ell^2\dot{\varphi}(\delta\dot{\varphi})\, dt + \int_{t_1}^{t_2} mg\ell(-\sin\varphi)(\delta\varphi)\, dt \\
&= \Big[m\ell^2\dot{\varphi}(\delta\varphi) \Big]_{t_1}^{t_2} - \int_{t_1}^{t_2} m\ell^2\ddot{\varphi}(\delta\varphi)\, dt + \int_{t_1}^{t_2} mg\ell(-\sin\varphi)(\delta\varphi)\, dt \\
&= m\ell^2\dot{\varphi}(t_2)\delta\varphi(t_2) - m\ell^2\dot{\varphi}(t_1)\delta\varphi(t_1) - \int_{t_1}^{t_2} m\ell\big(\ell\ddot{\varphi} + g\sin\varphi\big)(\delta\varphi)\, dt\,.
\end{aligned}$$

Among all motions that start and end at fixed endpoints $\varphi(t_1) = \varphi_1$, $\varphi(t_2) = \varphi_2$, the variations $\delta\varphi$ satisfy $\delta\varphi(t_1) = \delta\varphi(t_2) = 0$, hence

$$\delta I = -\int_{t_1}^{t_2} m\ell\big(\ell\ddot{\varphi} + g\sin\varphi\big)(\delta\varphi)\, dt.$$

So, stationarity implies

$$\big(\ell\ddot{\varphi} + g\sin\varphi\big) \equiv 0, \quad t_1 < t < t_2.$$

This is the familiar equation of motion of an undamped nonlinear pendulum. ○

14.2.4 Hamilton's Principle for Continuous Media

Recall from Section 10.6 the potential and kinetic energies of linear vibrations of a string satisfying boundary conditions, that is,

$$PE = \int_0^L \frac{1}{2} T_0 \left(\frac{\partial u}{\partial x} \right)^2 dx$$

and

$$KE = \int_0^L \frac{1}{2} \varrho_0 \left(\frac{\partial u}{\partial t} \right)^2 dx,$$

where
the constant tension is T_0
the constant mass density is ϱ_0

Again, the Lagrangian is $L \triangleq KE - PE$ and the action is $I \triangleq \int_{t_1}^{t_2} L \, dt$, that is,

$$I[\, u \,] = \int_{t_1}^{t_2} \int_0^L \frac{1}{2} \left(\varrho_0 \left(\frac{\partial u}{\partial t} \right)^2 - T_0 \left(\frac{\partial u}{\partial x} \right)^2 \right) dx \, dt.$$

Hamilton's principle of least action gives

$$0 = \delta I = \int_{t_1}^{t_2} \int_0^L \left(\varrho_0 \frac{\partial u}{\partial t} \cdot \left(\frac{\partial (\delta u)}{\partial t} \right) - T_0 \frac{\partial u}{\partial x} \cdot \left(\frac{\partial (\delta u)}{\partial x} \right) \right) dx \, dt.$$

Integration by parts with respect to t for the first term and integration by parts with respect to x for the second term gives

$$0 = \delta I = \int_{t_1}^{t_2} \int_0^L \left(- \frac{\partial}{\partial t} \left[\varrho_0 \frac{\partial u}{\partial t} \right] + \frac{\partial}{\partial x} \left[T_0 \frac{\partial u}{\partial x} \right] \right) \cdot (\delta u) \, dx \, dt.$$

This being true for all admissible variations δu gives us the one dimensional wave equation PDE

$$\frac{\partial}{\partial t} \left[\varrho_0 \frac{\partial u}{\partial t} \right] = \frac{\partial}{\partial x} \left[T_0 \frac{\partial u}{\partial x} \right],$$

that is,

$$\varrho_0 \frac{\partial^2 u}{\partial t^2} = T_0 \frac{\partial^2 u}{\partial x^2}.$$

The results can be generalized to the case when the tension and the mass density are not constant.

Learn More About It

Calculus of Variations, with Applications to Physics and Engineering, by Robert Weinstock, Dover Publications, Inc. © 1974, is an excellent reference. It includes Problem 14.2.5.8.

14.2.5 Problems

In problems 1–4 find the ODE- or PDE -BVP satisfied by a solution of the calculus of variations problem.

1. Minimize $-\int_0^L \left(p(x)\left(y'(x)\right)^2 + 2y(x)f(x)\right) dx$
   Subject to $y(0) = y(L) = 0$
2. Minimize $\int_0^L \left(-\left(y'(x)\right)^2 + x^2\left(y(x)\right)^2\right) dx$
   Subject to $y(0) = y(L) = 0$.
   [Note: This is an example of a boundary value problem modeling a cantilever beam clamped at the end $x = L$, as in Problem 9.7.1.7.]
3. Minimize $\iint_{\mathcal{D}} \left(\left(\partial u/\partial x\right)^2 + \left(\partial u/\partial y\right)^2 - 2uf(x,y)\right) dA$
   Subject to $u \equiv g(x,y)$ on $\partial\mathcal{D}$
4. Minimize $\iint_{\mathcal{D}} \left(\left(\partial u/\partial x\right)^2 + \left(\partial u/\partial y\right)^2 - 2uf(x,y)\right) dA$
   Subject to $\partial u/\partial n \equiv 0$ on $\partial\mathcal{D}$
5. Express the solution of

$$\left\{\begin{array}{ll} \text{Minimize} & \int_1^3 \left(-x\left(y'\right)^2 + xy^2\right) dx \\ \\ \text{Subject to} & y(1) = 0,\ y(3) = -1 \end{array}\right\}$$

   in terms of Bessel functions of order zero.
6. Use the Euler–Lagrange ODE necessary condition to explain why there is no admissible, that is, continuous and piecewise continuously differentiable on $[-1, 1]$, function to solve

$$\left\{\begin{array}{ll} \text{Minimize} & \int_{-1}^1 x^2\left(y'\right)^2 dx \\ \\ \text{Subject to} & y(-1) = -1,\ y(1) = 1 \end{array}\right\}.$$

7. Find the first variation of

$$J[\,y,v\,] \triangleq \int_a^b F(x,y(x),y'(x),v(x),v'(x))\,dx,$$

where y, v are continuously differentiable and satisfy BCs $y(a) = y_a$, $y(b) = y_b$, $v(a) = v_a$, $v(b) = v_b$.

8. Study Hamilton's principle for the motion of a mass m on an undamped **spherical pendulum** of length ℓ. Let φ be the angle with respect to vertical, with $\varphi = 0$ for the pendulum normally at rest and $\varphi = \pi$ for the inverted pendulum, and define θ to be the usual angle in the xy-plane. Then the total potential energy is $V = mgz = mg\ell(1-\cos\varphi)$ and the kinetic energy is $T = \frac{1}{2}\,m\big((\ell\dot{\varphi})^2 + (\ell(\sin\varphi)\,\dot{\theta})^2\big)$.

9. You will explain why the minimizer of (14.2) in Section 14.1, that is,

$$\left\{\begin{array}{ll} \text{Minimize} & \displaystyle\int_0^L \left(\frac{1}{2}\,EI(y'')^2 + yf(x)\right)dx \\[2ex] \text{Subject to} & y(0) = y(L) = 0 \end{array}\right\}$$

must satisfy the fourth order ODE-BVP

$$(\star) \quad \left\{\begin{array}{c} EIy'''' + f(x) = 0 \\[2ex] y(0) = y(L) = 0, \quad y''(0) = y''(L) = 0 \end{array}\right\}.$$

The **natural boundary conditions** $y''(0) = y''(L) = 0$ will appear as a consequence of the minimization.

(a) Define $J[\,y\,] \triangleq \int_0^L \left(\frac{1}{2}\,EI(y'')^2 + yf(x)\right)dx$ and explain why at a minimizer $y_0(x)$ the variation of J satisfies

$$0 = \delta J = \int_0^L \big(EI(y_0''(x)\,\delta y''(x) + f(x)\,\delta y(x)\big)\,dx.$$

(b) In the result of part (a), integrate twice by parts and use the fact that $\delta y(0) = \delta y(L) = 0$ (Why is that true?) to get

$$0 = \int_0^L \big(EI(y_0''''(x) + f(x)\big)\delta y(x)\,dx + EIy_0''(L)\,\delta y'(L) - EIy_0''(0)\,\delta y'(0).$$

(c) From the result of part (b), argue that y_0 satisfies $(\star)$ by varying independently (1) the function $\delta y(x)$ for $0 < x < L$, (2) the number $\delta y'(L)$, and (3) the number $\delta y'(0)$.

10. Partially generalize the result of Problem 14.2.5.9 to explain why the minimizer y_0 of

$$\left\{\begin{array}{ll}\text{Minimize} & \displaystyle\int_0^L F\left(x, y, y', y''\right) dx \\ \\ \text{Subject to} & y(0) = y(L) = 0\end{array}\right\}$$

at $\left(x, y, y', y''\right) = \left(x, y_0(x), y_0'(x), y_0''(x)\right)$ must satisfy the ODE, for $0 < x < L$,

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\left[\frac{\partial F}{\partial y'}\right] + \frac{d^2}{dx^2}\left[\frac{\partial F}{\partial y''}\right] \equiv 0.$$

11. Theorem 14.1 generalizes to problems in which the minimizer is a vector valued function $\mathbf{y}$: Suppose $F = F(x, \mathbf{y}, \mathbf{y}')$ is continuously differentiable and suppose $\mathbf{y} = \mathbf{y}_0(x)$ is in $\mathbf{R}^n$ and solves (14.1) in Section 14.1, that is,

$$\left\{\begin{array}{ll}\text{Minimize} & \displaystyle J[\,\mathbf{y}\,] \triangleq \int_a^b F\big(x, \mathbf{y}(x), \mathbf{y}'(x)\big)\, dx \\ \\ \text{Subject to} & \mathbf{y}(a) = \mathbf{y}_a \\ & \mathbf{y}(b) = \mathbf{y}_b.\end{array}\right\}.$$

Find the necessary conditions satisfied by $\mathbf{y}_0(x)$.

12. Study Hamilton's principle for the motion of the three springs and two masses system in Example 5.5 in Section 5.1. The action is

$$I = \frac{1}{2}\int_{t_1}^{t_2} \left(m_1\dot{x}_1^2 + m_2\dot{x}_2^2 - k_1x_1^2 - k_2(x_2 - x_1)^2 - k_3x_2^2\right) dt.$$

14.3 Problems with Constraints

We may need to find the minimum of a functional $J[\,y\,]$ over all functions y that satisfy an **integral constraint**

$$K[\,y\,] \triangleq \int_a^b G\big(x, y(x), y'(x)\big)\, dx = c. \tag{14.24}$$

Such a minimization problem is called an **isoperimetric problem** in honor of an historic problem: Find the curve of minimum length that encloses a given area. We will mention that historic problem in Example 14.12.

Theorem 14.2

Suppose $F = F(x, y, y')$ is continuously differentiable and suppose $y = y_0(x)$ solves

$$\left\{\begin{array}{ll} \text{Minimize} & J[\,y\,] \triangleq \displaystyle\int_a^b F\big(x, y(x), y'(x)\big)\,dx \\ \\ \text{Subject to} & K[\,y\,] \triangleq \displaystyle\int_a^b G\big(x, y(x), y'(x)\big)\,dx = c \\ & \qquad\qquad\qquad\qquad y(a) = y_a \\ & \qquad\qquad\qquad\qquad y(b) = y_b \end{array}\right\}, \tag{14.25}$$

where c is a constant. Then necessarily there is a constant Lagrange multiplier λ such that

$$\delta(J - \lambda K) = \frac{\partial(F - \lambda G)}{\partial y} - \frac{d}{dx}\left[\frac{\partial(F - \lambda G)}{\partial y'}\right] \equiv 0, \quad a < x < b, \tag{14.26}$$

as long as the problem is not degenerate in the sense that at the minimizer

$$\frac{\partial G}{\partial y} - \frac{d}{dx}\left[\frac{\partial G}{\partial y'}\right] \not\equiv 0, \quad a < x < b. \tag{14.27}$$

Why? Instead of using a single parameter family of variations as in (14.15) in Section 14.2, we use a two parameter family of variations

$$y(\varepsilon_1, \varepsilon_2)(x) \triangleq y_0(x) + \varepsilon_1\phi_1(x) + \varepsilon_2\phi_2(x), \quad -\varepsilon_0 \le \varepsilon_1, \quad \varepsilon_2 \le \varepsilon_0 \tag{14.28}$$

so that we can have $y(x)$ satisfy the integral constraint (14.24) *and* still have enough freedom of variation to be able to imply a necessary condition. Here $\phi_1(x), \phi_2(x)$ are linearly independent functions, in that sense that $0 \equiv c_1\phi_1(x) + c_2\phi_2(x)$ on the interval $[\,a, b\,]$ for constants c_1, c_2 only if $c_1 = c_2 = 0$, and also $\phi_j(a) = \phi_j(b) = 0$ for $j = 1, 2$ so that $y(\varepsilon_1, \varepsilon_2)(x)$ satisfies the boundary conditions for all $\varepsilon_1, \varepsilon_2$.

Define functions of the two real variables $\varepsilon_1, \varepsilon_2$ by

$$f(\varepsilon_1, \varepsilon_2) \triangleq J\big[\,y(\varepsilon_1, \varepsilon_2)\big] \quad \text{and} \quad g(\varepsilon_1, \varepsilon_2) \triangleq K\big[\,y(\varepsilon_1, \varepsilon_2)\big].$$

The minimization problem

$$\left\{\begin{array}{ll} \text{Minimize} & f(\varepsilon_1, \varepsilon_2) \\ \\ \text{Subject to} & g(\varepsilon_1, \varepsilon_2) = c \end{array}\right\} \tag{14.29}$$

is covered by Corollary 13.2 in Section 13.2, where the vector $\boldsymbol{\varepsilon} \triangleq \overrightarrow{(\varepsilon_1, \varepsilon_2)}$ is in $\mathbb{R}^2$. Because $f(0, 0) = J[\,y_0\,]$ and we assumed y_0 is a global minimizer for (14.25), we know that

$(\varepsilon_1, \varepsilon_2) = (0, 0)$ is a global minimizer for (14.29) in $\mathbb{R}^2$. So, there exists a constant Lagrange multiplier λ so that

$$\nabla_{\boldsymbol{\varepsilon}}(f - \lambda g)\Big|_{(\varepsilon_1,\varepsilon_2)=(0,0)} = \mathbf{0}. \tag{14.30}$$

In Problem 14.3.2.5 you will explain why (14.30) implies

$$0 = \int_a^b \left(\frac{\partial(F - \lambda G)}{\partial y}\big(x, y_0(x), y_0'(x)\big) - \frac{d}{dx}\left[\frac{\partial(F - \lambda G)}{\partial y'}\big(x, y_0(x), y_0'(x)\big) \right] \right) \phi_1(x)\, dx \tag{14.31}$$

and

$$0 = \int_a^b \left(\frac{\partial(F - \lambda G)}{\partial y}\big(x, y_0(x), y_0'(x)\big) - \frac{d}{dx}\left[\frac{\partial(F - \lambda G)}{\partial y'}\big(x, y_0(x), y_0'(x)\big) \right] \right) \phi_2(x)\, dx. \tag{14.32}$$

(14.31) and (14.32) redundantly give the same requirement that

$$\frac{\partial(F - \lambda G)}{\partial y}\big(x, y_0(x), y_0'(x)\big) - \frac{d}{dx}\left[\frac{\partial(F - \lambda G)}{\partial y'}\big(x, y_0(x), y_0'(x)\big) \right] \equiv 0,$$

that is, (14.26), as long as (14.27) is true. □

The technical requirement (14.27) is analogous to the technical requirement $\nabla g(\mathbf{x}^\star) \neq \mathbf{0}$ of Theorem 13.10 in Section 13.2.

By combining the calculation of variation implicit in Corollary 14.1 in Section 14.2 with the result of Theorem 14.2 we have

Corollary 14.2

Suppose $F = F\left(x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right)$ is continuously differentiable and $u = u_0(x, y)$ solves

$$\left\{ \begin{array}{ll} \text{Minimize} & J[\,u\,] \triangleq \displaystyle\iint_{\mathcal{D}} F\left(x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right) dx\, dy \\ \text{Subject to} & K[\,u\,] \triangleq \displaystyle\iint_{\mathcal{D}} G\left(x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right) dx\, dy = c \\ & \qquad\qquad u = g(x, y) \text{ on } \partial\mathcal{D} \end{array} \right\},$$

where c is a constant. Then necessarily u satisfies

$$\left(\frac{\partial(F - \lambda G)}{\partial u} - \frac{\partial}{\partial x}\left[\frac{\partial(F - \lambda G)}{\partial(\partial u/\partial x)} \right] - \frac{\partial}{\partial y}\left[\frac{\partial(F - \lambda G)}{\partial(\partial u/\partial y)} \right] \right)\Big|_{\text{at}\left(x, y, u_0(x,y)\right)} \equiv 0,\ (x, y) \text{ in } \mathcal{D} \tag{14.33}$$

for some constant λ, as long as

$$\left(\frac{\partial G}{\partial u} - \frac{\partial}{\partial x}\left[\frac{\partial G}{\partial(\partial u/\partial x)}\right] - \frac{\partial}{\partial y}\left[\frac{\partial G}{\partial(\partial u/\partial y)}\right]\right)\Big|_{\text{at}\left(x,y,u_0(x,y)\right)} \not\equiv 0, \ (x,y) \text{ in } \mathcal{D}.$$

Theorem 14.3

Suppose $F = F(x, y, y')$ is continuously differentiable and suppose $y = y_0(x)$ solves

$$\left\{\begin{array}{ll} \text{Minimize} & J[\,y,z\,] \triangleq \displaystyle\int_a^b F(x, y(x), y'(x), z(x), z'(x))\,dx \\ \\ \text{Subject to} & G(x,y,z) = c \\ & y(a) = y_a \\ & y(b) = y_b \\ & z(a) = z_a \\ & z(b) = z_b \end{array}\right\},$$

where c is a constant. Then necessarily there is a Lagrange multiplier function $\lambda = \lambda(x)$ such that

$$\frac{\partial(F - \lambda(x)G)}{\partial y} - \frac{d}{dx}\left[\frac{\partial(F - \lambda(x)G)}{\partial y'}\right] \equiv 0, \quad a < x < b, \tag{14.34}$$

and

$$\frac{\partial(F - \lambda(x)G)}{\partial z} - \frac{d}{dx}\left[\frac{\partial(F - \lambda(x)G)}{\partial z'}\right] \equiv 0, \quad a < x < b, \tag{14.35}$$

unless the problem is degenerate in the sense that at the minimizer

$$\frac{\partial G}{\partial y}(x, y(x), z(x)) = \frac{\partial G}{\partial z}(x, y(x), z(x)) \equiv 0, \quad a < x < b. \tag{14.36}$$

Why? Please see the explanation in the book by Weinstock mentioned in the "Learn More About It" below.

The technical requirement (14.36) is analogous to the technical requirement $\nabla g(\mathbf{x}_0) \neq \mathbf{0}$ of Theorem 13.10 in Section 13.2.

14.3.1 Differential Equation Constraints

If we are to minimize an integral functional subject to constraints expressed as differential equations, we can move the constraints into the functional using Lagrange multipliers that are functions of the independent variable.

Example 14.11

The problem to find the optimum path of a single stage rocket to put a satellite into orbit can be expressed as a calculus of variations problem (Thomson, 1986) of

$$\left\{\begin{array}{ll} \text{Maximize} & J[\,\phi\,] \triangleq u_0 + \displaystyle\int_0^T \frac{F(t)}{m}\,\cos\phi\,dt \\ \\ \text{Subject to} & \dot{w} - \frac{F(t)}{m}\,\sin\phi + g = 0 \\ & \dot{y} - w = 0 \\ & y(0) = y_0, \phi(0) = \theta_0 \\ & u_0 = v_0\cos\phi_0 \\ & w_0 = v_0\sin\phi_0 \\ & w(T) = 0, y(T) = Y \end{array}\right\},$$

where $\phi = \phi(t)$ is the "thrust attitude," that is, the rocket's angle of inclination with respect to horizontal. Also, T is the unspecified time when the rocket's fuel runs out, $y(t)$ is the altitude, g is the acceleration due to gravity (assumed to differ negligibly from a constant), $F(t)$ is the magnitude of the propulsive force of the rocket, $m(t)$ is the rocket's mass, and $u(t)$ and $w(t)$ are the rocket's horizontal and vertical velocities, respectively. The same techniques that derive Theorem 14.3 say that there are Lagrange multiplier *functions* $\lambda_1(t)$ and $\lambda_2(t)$ such that the constraints can be moved into the functional to reformulate the problem as

$$\left\{\begin{array}{ll} \text{Maximize} & J_2[\phi] \triangleq u_0 + \displaystyle\int_0^T \left(\frac{F(t)}{m}\,\cos\phi + \lambda_1(t)(\dot{w} - \frac{F}{m}\,\sin\phi + g) + \lambda_2(t)(\dot{y} - w)\right) dt \\ \\ \text{Subject to} & y(0) = y_0, \phi(0) = \theta_0 \\ & u_0 = v_0\cos\phi_0 \\ & w_0 = v_0\sin\phi_0 \\ & w(T) = 0, y(T) = Y \end{array}\right\}.$$

Thomson derives that the optimum thrust attitude is given implicitly by

$$\tan\phi(t) = \left(1 - \frac{t}{T}\right)\tan\theta_0.$$

Example 14.12

Perhaps the first isoperimetric problem in history was to find the "nice," that is, simple, closed, piecewise smooth, plane curve of length $L = 2\pi$ that encloses the greatest area. Given a nice curve

$$\mathcal{C}:\ \mathbf{r}(t) = x(t)\,\hat{\boldsymbol{\imath}} + y(t)\,\hat{\boldsymbol{\jmath}}, \quad 0 \le t \le b,$$

its length is

$$L[\,x, y\,] = \int_0^b \sqrt{\dot{x}^2 + \dot{y}^2}\,dt$$

and the area it encloses is

$$A[\,x,y\,] = \int_0^b x\frac{dy}{dt}\,dt\,.$$

So, the problem is to

$$\left\{\begin{array}{ll}\text{Maximize} & A[\,x,y\,] \\ \text{Subject to} & L[\,x,y\,] = 2\pi\end{array}\right\}.$$

It turns out* that the maximum area is π and it is achieved by a circle of radius one.

Learn More About It

Calculus of Variations, with Applications to Physics and Engineering, by Robert Weinstock, Dover Publications, Inc. © 1974, is an excellent reference. It includes a derivation of Theorem 14.3 in its Sections 4 and 5, "Restrictions Imposed through Finite or Differential Equations." Another good reference is Section IV-7 of *Methods of Mathematical Physics, Volume I*, by R. Courant and D. Hilbert, John Wiley & Sons, © 1989.

The derivation for the isoperimetric problem of Example 14.12 can be found in Section III.2 of *Elements of Applicable to Functional Analysis*, by Charles W. Groetsch, Marcel Dekker, Inc. © 1980.

14.3.2 Problems

In problems 1–4 use Theorem 14.2 or Corollary 14.2 find a BVP satisfied by the minimizer.

1. Minimize $\int_0^L \left((u')^2 - q(x)u^2\right) dx$
   Subject to $\int_0^L |u(x)|^2 dx = 1$ and $u(0) = u'(L) = 0$
2. Minimize $\int_0^L \left((u')^2 - q(x)u^2\right) dx$
   Subject to $\int_0^L \sigma(x)|u(x)|^2\,dx = 1$ and $u(0) = u'(L) = 0$
3. Minimize $\iint_{\mathcal{D}} \left(\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 - f(x,y)\,u^2\right) dA$
   Subject to $\iint_{\mathcal{D}} |u(x,y)|^2\,dA = 1$ and $u \equiv g(x,y)$ on $\partial\mathcal{D}$
4. Minimize $\iint_{\mathcal{D}} \left(\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 - u^4\right) dA$
   Subject to $\iint_{\mathcal{D}} |u(x,y)|^2 dA = 1$ and $u \equiv 0$ on $\partial\mathcal{D}$
5. Explain why (14.30) implies (14.31) and (14.32).

* It is convenient to reparametrize the curve by arclength, expand $x(t)$ and $y(t)$ in Fourier series, and use a Parseval Theorem.

14.4 Eigenvalue Problems

Suppose we have an eigenvalue problem

$$A\mathbf{x} = \lambda\mathbf{x}, \quad \text{where} \quad \mathbf{x} \neq \mathbf{0}, \tag{14.37}$$

where $\mathbf{x}$ is in an inner product space $(\mathcal{V}, \langle \cdot, \cdot \rangle)$. As we have seen in Sections 2.9 and 13.6, one of the forms of the Rayleigh–Ritz method for finding an eigenvalue(s) of a square matrix A effectively replaces Problem 14.3.2.7 with

$$\left\{ \begin{array}{ll} \text{Minimize} & \langle A\mathbf{x}, \mathbf{x} \rangle \\ \text{Subject to} & \langle \mathbf{x}, \mathbf{x} \rangle = 1 \end{array} \right\}. \tag{14.38}$$

14.4.1 An ODE-BVP

Given a finite interval $[\, a, b \,]$, define a norm on functions $y(x)$ by

$$||y||_s^2 \triangleq \int_a^b \left| \frac{d^s y}{dx^s} \right|^2 dx$$

and define a normed vector space $\mathcal{V}^s(a, b)$ to consist of all s-times piecewise continuously differentiable functions $y(x)$ on $[\, a, b \,]$, which satisfy

$$||y||_s^2 < \infty \quad \textit{and} \quad \text{the boundary conditions } y(a) = y'(b) = 0.$$

For functions defined on $[\, a, b \,]$ define also an inner product by

$$\langle y, z \rangle \triangleq \int_a^b y(x)\, z(x)\, dx,$$

at least for functions that are square integrable on $[\, a, b \,]$, that is, that satisfy

$$\infty > \langle y, y \rangle = ||y||_0^2 = \int_a^b |y(x)|^2\, dx.$$

Example 14.13

Consider the eigenvalue problem

$$\left\{ \begin{array}{l} y''(x) + q(x)y = \lambda y(x),\ a < x < b \\ y(a) = y'(b) = 0 \end{array} \right\}. \tag{14.39}$$

Turn (14.39) into a calculus of variations problem.

Method: Define the operator $A = \frac{d^2}{dx^2} + q(x)$ on $\mathcal{V}^2(a,b)$, which consists of all twice piecewise continuously differentiable functions $y(x)$ on $[\,a,b\,]$ which satisfy the boundary conditions $y(a) = y'(b) = 0$, by

$$Ay \triangleq \frac{d^2}{dx^2}[\,y\,] + q(x)y$$

and define

$$J[\,y\,] \triangleq \langle Ay, y\rangle = \int_a^b \big(y''(x) + q(x)y\big)\,y(x)\,dx = \int_a^b y''(x)\,y(x)\,dx + \int_a^b q(x)y(x)\,y(x)\,dx.$$

Using integration by parts and the boundary conditions we have

$$J[y] = \Big[y'(x)\,y(x)\Big]_a^b - \int_a^b \big(y'(x)\big)^2\,dx + \int_a^b q(x)\big(y(x)\big)^2 dx$$

$$= 0 + 0 + \int_a^b \Big(-\big(y'(x)\big)^2 + q(x)\big(y(x)\big)^2\Big)\,dx.$$

Here, the calculus of variations problem (14.38) is

$$\left\{\begin{array}{l} \text{Minimize } J[\,y\,] \triangleq \displaystyle\int_a^b \Big(-\big(y'(x)\big)^2 + q(x)\big(y(x)\big)^2\Big)\,dx \\[2ex] \text{Subject to } K[\,y\,] \triangleq \displaystyle\int_a^b \big(y(x)\big)^2 dx = 1 \end{array}\right\}, \tag{14.40}$$

where y is in $\mathcal{V}^2(a,b)$, hence satisfies the boundary conditions $y(a) = y'(b) = 0$.

Define

$$F\big(x, y(x), y'(x)\big) = -\big(y'(x)\big)^2 + q(x)\big(y(x)\big)^2$$

and

$$G\big(x, y(x), y'(x)\big) = \big(y(x)\big)^2.$$

Theorem 14.2 in Section 14.3 enables us to conclude that there exists a Lagrange multiplier λ such that

$$2q(x)y - 2\lambda y + \frac{d}{dx}[\,2y'\,] = \frac{\partial(F - \lambda G)}{\partial y} - \frac{d}{dx}\left[\frac{\partial(F - \lambda G)}{\partial y'}\right] \equiv 0, \quad a < x < b,$$

as long as at the minimizer we have

$$2y(x) - (0) = \frac{\partial G}{\partial y} - \frac{d}{dx}\left[\frac{\partial G}{\partial y'}\right] \not\equiv 0, \quad a < x < b. \tag{14.41}$$

Because $y(x)$ satisfies $\int_a^b \big(y(x)\big)^2 dx = 1$, it is impossible to have $2y(x) \equiv 0$ on $a < x < b$. So, condition (14.41) is automatically true at a solution of the calculus of variations problem.

So, the ODE-BVP eigenvalue problem (14.39), that is, $y''(x) + q(x)y - \lambda y(x) = 0, a < x < b$, $y(a) = y'(b) = 0$, does indeed correspond to the calculus of variations problem (14.40). ◯

Because $y(x)$ satisfies $Ay = y''(x) + q(x)y = \lambda y$ and $y \not\equiv 0$ for $a < x < b$, we can refer to λ as an eigenvalue of the differential operator A on the space $\mathcal{V}(a, b)$. Note that the two boundary conditions, $y(a) = y'(b) = 0$, are part of the definition of the space $\mathcal{V}^2(a, b)$.

14.4.2 An Eigenvalue Problem for the Laplacian

Similarly, given a nice spatial region $\mathcal{D}$ in $\mathbb{R}^2$, define a norm on functions $\phi(x, y)$ by

$$||\phi||_s^2 \triangleq \iint_{\mathcal{D}} \left(\left(\frac{\partial^s \phi}{\partial x^s} \right)^2 + \left(\frac{\partial^s \phi}{\partial y^s} \right)^2 \right) dx\, dy.$$

Define a normed vector space $\mathcal{W}$ to consist of s-times continuously differentiable functions $\phi(x, y)$ on $\mathcal{D}$ for which

$$||\phi||_s^2 < \infty$$

and which satisfy the boundary condition

$$0 \equiv \phi\big|_{(x,y) \text{ on boundary of } \mathcal{D}}.$$

Unfortunately, this space is not a *complete* normed vector space, as defined in Section 2.10. But there is a concept called "completion of a metric space" that considers "functions" ψ that are limits, in the norm $||\cdot||_s$, of sequences of functions in $\mathcal{W}$. This completion is called a **Sobolev** (Lusternik and Sobolev 1974) space and is denoted by $H_0^s(\mathcal{D})$. The superscript^s refers to the squares of the s order derivatives in the norm and the subscript $_0$ refers to the zero Dirichlet boundary condition. For functions defined on $\mathcal{D}$, define also an inner product by

$$\langle \phi, \psi \rangle \triangleq \iint_{\mathcal{D}} \phi(x, y)\, \psi(x, y)\, dx\, dy,$$

at least for functions that are square integrable on $\mathcal{D}$, that is, that satisfy

$$\infty > \langle \phi, \phi \rangle = ||\phi||_0^2 = \iint_{\mathcal{D}} |\phi(x, y)|^2\, dx\, dy.$$

Example 14.14

We have seen in Example 14.3 in Section 14.1 that vibration of a membrane over a planar region $\mathcal{D}$ can lead to the eigenvalue problem

$$\left\{ \begin{array}{l} 0 = \dfrac{\partial^2 \phi}{\partial x^2} + \dfrac{\partial^2 \phi}{\partial y^2} + \lambda \phi, \ (x, y) \text{ in } \mathcal{D} \\ \\ 0 \equiv \phi\big|_{(x,y) \text{ on boundary of } \mathcal{D}} \end{array} \right\}, \tag{14.42}$$

assuming the boundary of the membrane has displacement $u = 0$. Look for a solution in the Sobolev space $H_0^2(\mathcal{D})$ by turning (14.42) into a calculus of variations problem.

Method: Define the operator $A = -\Delta$ on $H_0^2(\mathcal{D})$ by

$$A\phi \triangleq -\left(\frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2}\right) = -\boldsymbol{\nabla} \bullet \boldsymbol{\nabla}\phi.$$

Consider the corresponding calculus of variations problem (14.38). We have

$$\langle A\phi, \phi\rangle = -\iint\limits_{\mathcal{D}} \big(\boldsymbol{\nabla} \bullet \boldsymbol{\nabla}\phi(x,y)\big)\phi(x,y)\,dx\,dy.$$

Use the result of Corollary 6.1 in Section 6.7 and the Divergence Theorem to conclude that

$$\begin{aligned}\langle A\phi, \phi\rangle &= -\iint_{\mathcal{D}} \boldsymbol{\nabla} \bullet \big(\phi\,\boldsymbol{\nabla}\phi\big)\,dx\,dy + \iint\limits_{\mathcal{D}} \big(\boldsymbol{\nabla}\phi\big) \bullet \big(\boldsymbol{\nabla}\phi\big)\,dx\,dy \\ &= -\oint_{\partial\mathcal{D}} \phi \cdot \big(\widehat{\mathbf{n}} \bullet \boldsymbol{\nabla}\phi\big)\,ds + \iint\limits_{\mathcal{D}} \big(\boldsymbol{\nabla}\phi\big) \bullet \big(\boldsymbol{\nabla}\phi\big)\,dx\,dy \\ &= 0 + \iint\limits_{\mathcal{D}} \left(\left(\frac{\partial\phi}{\partial x}\right)^2 + \left(\frac{\partial\phi}{\partial y}\right)^2\right) dx\,dy,\end{aligned}$$

after using the boundary condition $0 \equiv \phi\big|_{(x,y) \text{ on boundary of } \mathcal{D}}$. So, the calculus of variations problem is

$$\left\{\begin{array}{ll}\text{Minimize} & \displaystyle\iint\limits_{\mathcal{D}} \left(\left(\frac{\partial\phi}{\partial x}\right)^2 + \left(\frac{\partial\phi}{\partial y}\right)^2\right) dx\,dy \\ \text{Subject to} & \displaystyle\iint\limits_{\mathcal{D}} |\phi(x,y)|^2\,dx\,dy = 1 \quad \text{and} \quad \phi \equiv 0 \text{ on } \partial\mathcal{D}\end{array}\right\}. \tag{14.43}$$

14.4.3 Sturm–Liouville Problem

Assume $s(x) > 0$ on the interval (a, b). Assume $\epsilon_0, \epsilon_1, \gamma_0, \gamma_1$ are given constants (scalars). We will assume always that either $\epsilon_0 \neq 0$ or $\epsilon_1 \neq 0$ and either $\gamma_0 \neq 0$ or $\gamma_1 \neq 0$. Below we will look for minimizers $X(x)$ that are continuous and whose derivative is piecewise continuous on the interval (a, b).

Example 14.15

The minimizer of

$$\left\{\begin{array}{ll}\text{Minimize} & J[\,X\,] \triangleq -\Big[p(x)X(x)X'(x)\Big]_a^b + \displaystyle\int_a^b \Big(p(x)\left(X'(x)\right)^2 - q(x)\left(X(x)\right)^2\Big)\,dx \\ \text{Subject to} & K[\,X\,] \triangleq \displaystyle\int_a^b s(x)(X(x))^2\,dx = 1 \\ & \epsilon_0 X(a) - \epsilon_1 X'(a) = 0 \\ & \gamma_0 X(b) + \gamma_1 X'(b) = 0\end{array}\right\}$$

solves the Sturm–Liouville eigenvalue problem consisting of ODE (9.63) in Section 9.6, that is,

$$\big(p(x)X'(x)\big)' + (\lambda s(x) + q(x))X(x) = 0,\ a < x < b,$$

and boundary conditions (9.64) in Section 9.6, that is,

$$\left\{\begin{array}{l} \epsilon_0 X(a) - \epsilon_1 X'(a) = 0 \\ \\ \gamma_0 X(b) + \gamma_1 X'(b) = 0 \end{array}\right\}.$$

An explanation for this is in Section 8.1 of the book by Weinstock mentioned in Section 14.3's "Learn More About It".

Note that the technical requirement (14.27) in Section 14.3 is automatically verified because

$$\frac{\partial G}{\partial X} - \frac{d}{dx}\left[\frac{\partial G}{\partial X'}\right] = 2s(x)X(x) \not\equiv 0$$

because $s(x) > 0$ on the interval (a, b), and a minimizer $X(x)$ is continuous on (a, b) and satisfies $\int_a^b s(x)(X(x))^2\,dx = 1$.

Learn More About It

Good expositions of Sobolev spaces are in Section 2.5 of *Elements of Functional Analysis*, by L. A. Lusternik and V. J. Sobolev, Hindustan Publishing Co., Halsted Press/John Wiley & Sons, Inc., © 1974.

14.4.4 Problems

In problems 1–4, find a calculus of variations problem whose solution is an eigenfunction for the given eigenvalue problem.

1. $\left\{\begin{array}{l} 0 = \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} + \lambda u,\ (x, y) \text{ in } \mathcal{D} \\ \\ 0 \equiv \frac{\partial u}{\partial n} \text{ on } \partial\mathcal{D} \end{array}\right\}$

2. $\left\{\begin{array}{l} 0 = \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} + f(x, y)u + \lambda u,\ (x, y) \text{ in } \mathcal{D} \\ \\ 0 \equiv u \text{ on } \partial\mathcal{D} \end{array}\right\}$

3. $\left\{\begin{array}{l} 0 = \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} + f(x, y)u + \lambda\sigma(x, y)u,\ (x, y) \text{ in } \mathcal{D} \\ \\ 0 \equiv u \text{ on } \partial\mathcal{D} \end{array}\right\}$

4. $$\left\{\begin{array}{l} 0 = \dfrac{\partial^2 u}{\partial x^2} + 2\dfrac{\partial^2 u}{\partial y^2} + f(x,y)u + \lambda u, \ (x,y) \text{ in } \mathcal{D} \\ \\ 0 \equiv u \text{ on } \partial\mathcal{D} \end{array}\right\}$$

In problems 5 and 6, use the result of Example 14.15 to find an ODE-BVP eigenvalue problem satisfied by the minimizer.

5. Minimize $\displaystyle\int_0^L \Big((u')^2 - q(x)u^2\Big)dx$
   Subject to $\displaystyle\int_0^L u^2\,dx = 1$ and $u(0) = u(L) = 0$
6. Minimize $\displaystyle\int_0^L \Big((u')^2 - q(x)u^2\Big)dx$
   Subject to $\displaystyle\int_0^L \sigma(x)u^2 dx = 1$ and $u(0) = u'(L) = 0$

14.5 Short Take: Finite Element Methods

If we combine variational methods with the approximation of a solution using a basis of splines, which we learned about in Section 8.9, then we have "Finite Element Methods (FEM)."

Abstractly, the Rayleigh–Ritz method for a boundary value problem uses the variational formulation

$$\left\{\begin{array}{ll} \text{Minimize} & J[\,u\,] \triangleq \langle u, Au\rangle - 2\langle u, f\rangle \\ \\ \text{Subject to} & u = \sum_{n=1}^{M} c_n \phi_n \end{array}\right\}, \tag{14.44}$$

where the "trial functions" ϕ_n are chosen appropriately. For example, if the operator A has boundary conditions built into its definition then the trial functions should satisfy the boundary conditions.

For example, if $A = -\frac{d^2}{dx^2} + q(x)$ is defined to include the boundary conditions $y(0) = y(L) = 0$ then we want the trial functions to satisfy $\phi_n(0) = \phi_n(L) = 0$, for $n = 1, \ldots, M$.

Recall from Section 8.8 that the uniform tent basis functions were defined in (8.86) in Section 8.9, that is,

$$T_j(x) \triangleq \left\{\begin{array}{ll} \left(\frac{x - x_j}{h} + 1\right), & x_{j-1} \le x \le x_j \\ \\ \left(1 - \frac{x - x_j}{h}\right), & x_j \le x \le x_{j+1} \\ \\ 0, & |x - x_j| \ge h \end{array}\right\} = \phi\left(\frac{x - x_j}{h}\right),$$

for $j = 0, \ldots, N+1$. If $h = \frac{L}{N}$ and $x_j = jh$ for $j = -1, \ldots, N+1$, then for $n = 1, \ldots, N-1 \triangleq M$ the uniform tent basis spline functions $T_1(x), \ldots, T_{N-1}(x)$ will satisfy the boundary conditions $y(0) = y(L) = 0$.

Example 14.16

Use the uniform tent basis functions and the Rayleigh–Ritz method to approximate the solution of the ODE-BVP

$$\left\{\begin{array}{l} -y''(x) + xy = 1 + x,\ 0 < x < 1 \\ \\ y(0) = y(1) = 0 \end{array}\right\}. \tag{14.45}$$

Method: Define the linear differential operator A by

$$Ay \triangleq \Big(-\frac{d^2}{dx^2} + x\Big)[\,y\,] = -y'' + xy$$

on functions that satisfy the boundary conditions $y(0) = y(1) = 0$. Using integration by parts and the boundary conditions we calculate

$$\begin{aligned} \langle Ay, y\rangle &= \int_0^1 \big(-y'' + xy\big)y\,dx = \Big[-y'(x)y(x)\Big]_0^1 + \int_0^1 (y')y'\,dx + \int_0^1 (xy)y\,dx \\ &= \int_0^1 (y')y'\,dx + \int_0^1 (xy)y\,dx. \end{aligned}$$

So, the Rayleigh–Ritz method using the uniform tent basis functions is the equivalent optimization problem

$$\left\{\begin{array}{ll} \text{Minimize} & J[\,y\,] \triangleq \displaystyle\int_0^1 \big((y')^2 + xy^2 - 2(1+x)y\big)\,dx \\ \\ \text{Subject to} & y(x) = \sum_{n=1}^{M} c_n T_n(x) \end{array}\right\},$$

where

$M = N - 1$

$h = \frac{1}{N}$

We used *Mathematica* (see the commands below) to solve for the constants $c_1, \ldots, c_7$ in the special case $N = 8$, that is, grid size $h = 0.125$, and got the approximate dashed solution graph in Figure 14.4. Figure 14.4 also displays *Mathematica*'s approximate solution graph in a solid curve.

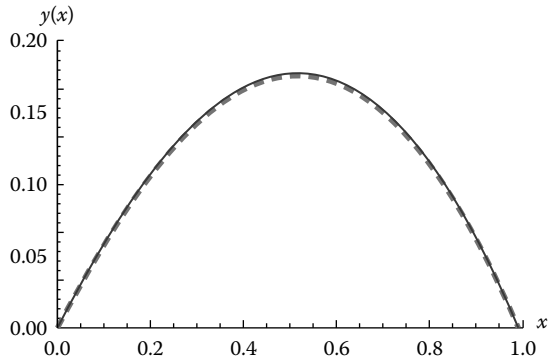


FIGURE 14.4
Example 14.16.

14.5.1 *Mathematica* Commands

After defining h = 0.125, the basic tent function $T_1(x)$ in Example 14.16 was defined by

$$\text{T1[t\_]:=Piecewise[\{\{}\frac{\text{t}}{\text{h}}\text{, 0< t} \le \text{h\}, \{1 -} \frac{\text{t} - \text{h}}{\text{h}}\text{, h< t} \le \text{2h\}, \{0, -1 < \text{t} \le 0\}, \{0, 2\text{h} < \text{t} \le 1\}\}].}$$

Noting that $T_2(x) = T_1(x - h)$, $T_3(x) = T_1(x - 2h)$, etc., we defined

$$\text{y[x\_, c1\_, c2\_, c3\_, c4\_, c5\_, c6\_, c7\_] := c1} * \text{T1[x]} + \text{c2} * \text{T1[x} - \text{h]} + \text{c3} * \text{T1[x} - \text{2h]}$$
$$+\text{c4} * \text{T1[x} - \text{3h]} + \text{c5} * \text{T1[x} - \text{4h]} + \text{c6} * \text{T1[x} - \text{5h]} + \text{c7} * \text{T1[x} - \text{6h]}$$

and the functional by

$$\text{J[c1\_, c2\_, c3\_, c4\_, c5\_, c6\_, c7\_]}$$
$$:=\text{Evaluate}\Big[\int_0^1 \big((\text{D[y[x, c1, c2, c3, c4, c5, c6, c7], x]})^2 + \text{x (y[x, c1, c2, c3, c4, c5, c6, c7]})^2$$
$$-2 * (1 + \text{x}) * \text{y[x, c1, c2, c3, c4, c5, c6, c7]}\big)\,\text{dx}\Big].$$

Then we used

$$\text{FindMinimum[J[c1, c2, c3, c4, c5, c6, c7], \{\{c1, .1\}, \{c2, 0.1\}, \{c3, 0.1\}, \{c4, .1\}, \{c5, 0.1\},}$$
$$\text{\{c6, 0.1\}, \{c7, 0.1\}\}].}$$

to get output

$$\{-0.176893, \{\text{c1} \to 0.0722715, \text{c2} \to 0.127121, \text{c3} \to 0.162938, \text{c4} \to 0.178214,$$
$$\text{c5} \to 0.171418, \text{c6} \to 0.140859, \text{c7} \to 0.0845438\}\}.$$

We substituted those values into y to get a function

$$\text{y[x, 0.07227152517438032', 0.1271210999916466', 0.16293835769807058',}$$
$$\text{0.17821421372289542', 0.17141750813321335', 0.14085942298213375',}$$
$$\text{0.08454383690012605']}$$

whose graph we plotted as the dashed curve in Figure 14.4.

To get and plot *Mathematica*'s approximate solution, we used the command

$$\text{s = NDSolve[\{x''[t] == tx[t] - 1 - t, x[0] == 0,x[1] == 0\}, \{x\}, \{t, 0, 1\}]}$$

followed by

```
bb = Plot[Evaluate[x[t]/.s], {t, 0, 1}, PlotStyle → {Blue, Thick},
  AxesLabel → {Text[Style["x", Italic, 18]], Text[Style["y(x)", Italic, 18]]}]
```

to get the solid curve in Figure 14.4.

Example 14.17

Use cubic spline uniform basis functions and the Rayleigh–Ritz method to approximate the solution of the ODE-BVP

$$\left\{\begin{array}{l} -y''(x) + x^2 y = x,\ 0 < x < 1 \\ \\ y(0) = y'(1) = 0 \end{array}\right\}. \tag{14.46}$$

Method: Define the linear differential operator A by

$$Ay \triangleq \Big(-\frac{d^2}{dx^2} + x\Big)[\,y\,] = -y'' + x^2 y$$

on functions that satisfy the boundary conditions $y(0) = y'(1) = 0$. Using integration by parts and the boundary conditions we calculate

$$\begin{aligned} \langle Ay, y\rangle &= \int_0^1 \big(-y'' + xy\big) y\, dx = \Big[-y'(x)y(x)\Big]_0^1 + \int_0^1 (y')y'\, dx + \int_0^1 (x^2 y) y\, dx \\ &= \int_0^1 (y')y'\, dx + \int_0^1 (x^2 y) y\, dx. \end{aligned}$$

The instructions said to use cubic spline uniform basis functions because having a boundary condition involving the derivative of the solution, for example, at $x = 1$, contraindicates use of tent spline uniform basis functions, which lack smoothness.

Let $h = \frac{1}{N}$ be the grid size, and define $x_j = jh$, for $j = -1, 0, \ldots, N+1$. Recall that the cubic spline uniform basis functions $C_j(x)$ were defined in (8.90) in Section 8.9, for $j = -1, 0, \ldots, N+1$. Using the data in (8.93) in Section 8.9, specifically

$$\begin{gathered} C_j(x_{j-2}) = 0, \quad C_j(x_{j-1}) = \frac{1}{6}, \quad C_j(x_j) = \frac{2}{3}, C_j(x_{j+1}) = \frac{1}{6}, \quad C_j(x_{j+2}) = 0, \\ C_j{}'(x_{j-2}) = 0, \quad C_j{}'(x_{j-1}) = \frac{1}{2h}, \quad C_j{}'(x_j) = 0, C_j{}'(x_{j+1}) = -\frac{1}{2h}, \quad C_j{}'(x_{j+2}) = 0, \end{gathered}$$

we see that an approximate solution $y(x) \triangleq \sum_{j=-1}^{N+1} c_j C_j(x)$ will satisfy the boundary condition $0 = y(0)$ if, and only if, $\frac{1}{6}c_{-1} + \frac{2}{3}c_0 + \frac{1}{6}c_1 = 0$, that is, $c_{-1} + 4c_0 + c_1 = 0$. Also, $y(x)$ will satisfy the boundary condition $0 = y'(1) = y'(x_N)$ if, and only if, $-\frac{1}{2h}c_{N-1} + \frac{1}{2h}c_{N+1} = 0$, that is, $-c_{N-1} + c_{N+1} = 0$.

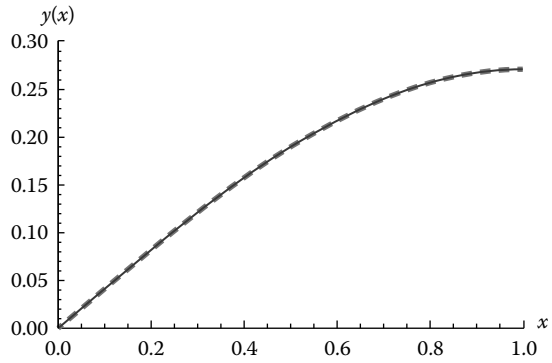


FIGURE 14.5
Example 14.17.

So, the Rayleigh–Ritz method using the cubic spline uniform basis functions is the equivalent optimization problem

$$\left\{\begin{array}{ll} \text{Minimize} & J[\,y\,] \triangleq \displaystyle\int_0^1 \left((y')^2 + x^2y^2 - 2xy\right) dx \\ \\ \text{Subject to} & y(x) = \displaystyle\sum_{j=-1}^{N+1} c_j C_j(x), c_{-1} + 4c_0 + c_1 = 0, \quad \text{and} \quad -c_{N-1} + c_{N+1} = 0 \end{array}\right\}.$$

Note that it is *not* true that each of the individual test functions $C_j(x)$ satisfies the boundary conditions $y(0) = y'(1) = 0$. But, all test functions $\sum_{j=-1}^{N+1} c_j C_j(x)$ *do* satisfy the boundary conditions as long as the coefficients $c_{-1}, c_0, \ldots, c_{N+1}$ satisfy the constraints $c_{-1} + 4c_0 + c_1 = 0$ and $-c_{N-1} + c_{N+1} = 0$.

We used *Mathematica* (see the commands below) to solve for the constants $c_{-1}, c_0, \ldots, c_3$ in the special case $N = 3$, that is, grid size $h = \frac{1}{3}$, and got the approximate dashed solution graph in Figure 14.5. Figure 14.5 also displays *Mathematica*'s approximate solution graph in a solid curve. ○

To create the cubic spline B3[x] $\triangleq \psi(x)$ defined in (8.88) in Section 8.9, we used *Mathematica* commands

$$\text{BB[x\_]} := \text{Piecewise}\left[\left\{\left\{\frac{3x^3 - 6x^2 + 4}{6}, 0 < x \le 1\right\}, \left\{\frac{(2-x)^3}{6}, 1 < x < 2\right\}\right\}\right]$$

$$\text{B3[x\_]} := \text{Piecewise}[\{\{\text{BB[x]}, 0 < x < 2\}, \{\text{BB}[-x], -2 < x < 0\}, \{0, x \le -2\}, \{0, x \ge 2\}\}]$$

After that we used *Mathematica* commands

$$h = \frac{1}{3}; \text{CM1[x\_]} := \text{B3}\left[\frac{x+h}{h}\right]; \text{C0[x\_]} := \text{B3}\left[\frac{x}{h}\right]; \text{C1[x\_]} := \text{B3}\left[\frac{x-h}{h}\right];$$

$$\text{C2[x\_]} := \text{B3}\left[\frac{x-2h}{h}\right]; \text{C3[x\_]} := \text{B3}\left[\frac{x-3h}{h}\right]; \text{C4[x\_]} := \text{B3}\left[\frac{x-4h}{h}\right];$$

```
y[x_, cm1_, c0_, c1_, c2_, c3_, c4_] :=
  cm1 CM1[x] + c0 C0[x] + c1 C1[x] + c2 C2[x] + c3 C3[x] + c4 C4[x]

f[cm1_, c0_, c1_, c2_, c3_, c4_] := Evaluate[ ∫_0^1 ((D[y[x, cm1, c0, c1, c2, c3, c4], x])^2
  + x^2 * (y[x, cm1, c0, c1, c2, c3, c4])^2 - 2x * y[x, cm1, c0, c1, c2, c3, c4]) dx ]

FindMinimum[{f[cm1, c0, c1, c2, c3, c4], {-c2 + c4 == 0, cm1 + 4c0 + c1 == 0}},
  {{cm1, .1}, {c0, .1}, {c1, .1}, {c2, 0.1}, {c3, .1}, {c4, .1}}]
```

14.5.2 Rayleigh–Ritz, Galerkin, and Least Squares

Suppose A is a linear differential operator on a vector space that includes in its definition boundary conditions. A could be an ordinary differential operator or a partial differential operator, such as the Laplacian.

The Galerkin method is, abstractly, as in (12.47) in Section 12.5, to find $c_1, \ldots, c_M$ that solves

$$\sum_{j=1}^{M} \langle \phi_i, A\phi_j \rangle c_j = \langle f, \phi_i \rangle, \; i = 1, \ldots, M, \tag{14.47}$$

which can be rewritten as (12.48) in Section 12.5, that is,

$$\mathrm{A}\mathbf{c} = \mathbf{f}, \tag{14.48}$$

where $\mathbf{c} = [c_1 \quad \ldots \quad c_M]^T$ and

$$\mathbf{f} = \left[\langle \phi_1, f \rangle \quad \ldots \quad \langle \phi_M, f \rangle \right]^T \tag{14.49}$$

are vectors in $\mathbb{R}^M$, and A is the $M \times M$ matrix defined by (12.49) in Section 12.5, that is,

$$\mathrm{A} \triangleq \left[\langle \phi_i, A\phi_j \rangle \right]_{\substack{1 \le i \le M \\ 1 \le j \le M}}. \tag{14.50}$$

Definition 14.1

A linear operator A on an inner product space is **symmetric** if

$$\langle Au, v \rangle = \langle u, Av \rangle \quad \text{for all } u, v.$$

The definition of "symmetric" is similar to Definition 2.36 in Section 12.10 of "self-adjoint" for a bounded linear operator on a Hilbert space.

For example, the differential operator in Example 14.16 is symmetric, because for all functions u and v that satisfy the boundary conditions, that is, satisfy $u(0) = u(1) = 0$, $v(0) = v(1) = 0$, and are sufficiently differentiable, integration by parts twice gives

$$\langle Au, v\rangle = \int_0^1 (-u''(x) + xu)v\,dx = \Big[-u'(x)v(x)\Big]_0^1 + \int_0^1 \big(u'(x)v'(x) + xu(x)v(x)\big)\,dx$$

$$= -0 + 0 + \big[\,u(x)v'(x)\,\big]_0^1 + \int_0^1 \big(-u(x)v''(x) + xu(x)v(x)\big)\,dx = 0 - 0 + \langle u, Av\rangle.$$

Theorem 14.4

If A is linear and symmetric, then the Galerkin method and the Rayleigh–Ritz method given in (14.44) give the same approximate solution of $Au = f$.

Why? It will suffice to show that $u = \sum_{n=1}^M c_n\phi_n$ minimizes $J[\,u\,] \triangleq \langle u, Au\rangle - 2\langle u, f\rangle$ if, and only if, $\mathbf{c} = [c_1 \quad \ldots \quad c_M]^T$ satisfies (14.48).

We calculate

$$q(\mathbf{c}) \triangleq J\left[\sum_{n=1}^M c_n\phi_n\right] = \left\langle \sum_{i=1}^M c_i\phi_i,\ A\sum_{j=1}^M c_j\phi_j\right\rangle - 2\left\langle \sum_{i=1}^M c_i\phi_i,\ f\right\rangle$$

$$= \sum_{i=1}^M\sum_{j=1}^M c_ic_j\langle\phi_i, A\phi_j\rangle - 2\sum_{i=1}^M c_i\langle\phi_i, f\rangle = \mathbf{c}^T\mathrm{A}\mathbf{c} - 2\mathbf{c}^T\mathbf{f},$$

where A and $\mathbf{f}$ are as in (14.50) and (14.49), respectively.

Because we assumed A is symmetric, the matrix A is symmetric, that is, $\mathrm{A}^T = \mathrm{A}$.

Using symmetry of A, (Fermat's) Theorem 13.6 in Section 13.2, that is, stationarity, states that a minimizer $\mathbf{c}$ for $j(\mathbf{c})$ must satisfy

$$\mathbf{0} = \frac{1}{2}\,\mathbf{0} = \frac{1}{2}\,\frac{\partial q}{\partial \mathbf{c}} = \mathrm{A}\mathbf{c} - \mathbf{f},$$

which is (14.48) from Galerkin's method. So, the minimizer $u = \sum_{n=1}^M c_n\phi_n$ for the Rayleigh–Ritz method is the same as the solution given by Galerkin's method. □

The Rayleigh–Ritz method was given in (14.44) and used the functional

$$J[\,u\,] \triangleq \langle u,\ Au\rangle - 2\langle u, f\rangle.$$

The **least squares method** is to, instead, find constants $c_1, \ldots, c_M$ to

$$\left\{\begin{array}{ll}\text{Minimize} & J_2[\,u\,] \triangleq ||Au - f||^2 = \langle Au - f, Au - f\rangle \\ \\ \text{Subject to} & u = \sum_{n=1}^M c_n\phi_n\end{array}\right\}. \tag{14.51}$$

In Problem 14.5.4.6 you will explain why the least squares method gives the same solution as a *generalization* of Galerkin's method, assuming A is a symmetric linear operator.

14.5.3 Finite Elements for PDEs

For PDEs we can use surface splines and triangular surface patches, which were mentioned in Section 8.9, analogously to the way we used tent functions to approximate the solution of an ODE earlier in this section. There are a lot of sophisticated computational science problems associated with decomposing a planar or spatial region into triangles or tetrahedra, or into rectangles or boxes. But after that hard work, the idea of using the Rayleigh–Ritz or Galerkin method is abstractly the same as for ODEs.

Like many other chapters or even sections of this book, the topic of finite elements for PDEs deserves a shelf of books to discuss either the technicalities of practical implementation or the mathematical theory that justifies the hope of convergence to a solution.

Learn More About It

Among the many good references for finite element methods for PDEs are *Finite Elements for Electrical Engineers*, 3rd ed., by Peter P. Silvester and Ronald l. Ferrari, Cambridge University Press, 1996, *Introduction to Finite Elements in Engineering*, 3rd ed., by Tirupathi R. Chandrupatla and Ashok D. Belegundu, Prentice Hall, Inc., 2002, and *An Analysis of the Finite Element Method*, 2nd ed., by Gilbert R. Strang and George J. Fix, Wellesley-Cambridge, 2008.

14.5.4 Problems

In problems 1 and 2, use the uniform tent basis functions and the Rayleigh–Ritz method to approximate the solution of the ODE-BVP, using the given value of N.

1. $\left\{\begin{array}{l} -y''(x) + x^2 y = x,\ 0 < x < 1 \\ \\ y(0) = y(1) = 0 \end{array}\right\}, N = 4$

2. $\left\{\begin{array}{l} -y''(x) + y\cos(\pi x) = x,\ 0 < x < 1 \\ \\ y(0) = y(1) = 0 \end{array}\right\}, N = 5$

In problems 3 and 4, use the cubic spline uniform basis functions and the Rayleigh–Ritz method to approximate the solution of the ODE-BVP, using the given value of N.

3. $\left\{\begin{array}{l} -y''(x) + xy = 1 + x,\ 0 < x < 1 \\ \\ y(0) = y'(1) = 0 \end{array}\right\}, N = 2$

4. $\left\{\begin{array}{l} -y''(x) + xy = 1 + x,\ 0 < x < 1 \\ \\ y'(0) = y(1) = 0 \end{array}\right\}, N = 2$

5. For the *nonlinear* ODE-BVP $\left\{\begin{array}{l} -y''(x) + y - y^3 = x,\ 0 < x < 1 \\ \\ y(0) = y(1) = 0 \end{array}\right\}$,
   (a) explain why its solution is a minimizer for the problem

$$\left\{\begin{array}{ll} \text{Minimize} & J[\,y\,] \\ \\ \text{Subject to} & y(0) = y(1) = 0 \end{array}\right\},$$

   where

$$(\star\star)[y] \triangleq \int_0^1 \left(\frac{1}{2}\,(y')^2 + \frac{1}{2}\,y^2 - \frac{1}{4}\,y^4 - xy\right) dx.$$

   (b) Use the uniform tent basis functions and the Rayleigh–Ritz method to approximate the solution of the ODE-BVP, using $N = 4$.

6. You will explain why the least squares method gives the same solution as a *generalization* of Galerkin's method, assuming A is a symmetric linear operator.
   (a) Use $||Au - f||^2 = \langle Au - f,\ Au - f\rangle$, $u = \sum_{n=1}^{M} c_n\phi_n$, and symmetry of A to explain why

$$||Au - f||^2 = \sum_{i=1}^{M}\sum_{j=1}^{M} c_i c_j \langle A\phi_i, A\phi_j\rangle - 2\sum_{i=1}^{M} c_i\langle A\phi_i, f\rangle + ||f||^2.$$

   (b) Define $\Psi_j \triangleq A\phi_j$, $\widehat{\mathrm{A}} \triangleq \left[\langle \Psi_i,\ A\phi_j\rangle\right]_{\substack{1\,\le\, i\,\le\, M \\ 1\,\le\, j\,\le\, M}}$, and $\widehat{\mathbf{f}} \triangleq \left[\langle \Psi_1,\ f\rangle \cdots \langle \Psi_M,\ f\rangle\right]^T$.
   Explain why the vector $\mathbf{c} = \left[c_1 \cdots c_M\right]^T$ minimizes $||Au - f||^2$ if, and only if, $\mathbf{c}$ solves the system of equations

$$\widehat{\mathrm{A}}\mathbf{c} = \widehat{\mathbf{f}},$$

   which is a generalization of Galerkin's method found in (12.53) in Section 12.5.

Key Terms

action: before Example 14.10 in Section 14.2
admissible function: after (14.1) in Section 14.1
brachistochrone problem: Example 14.4 in Section 14.1
calculus of variations: before (14.1) in Section 14.2
Euler–Lagrange equation: (14.18) in Section 14.2
fixed end boundary condition: after (14.23) in Section 14.2
free end boundary condition: after (14.23) in Section 14.2
functional: after (14.1) in Section 14.1
Hamilton's principle: before Example 14.10 in Section 14.2
integral constraint: (14.24) in Section 14.3
isoperimetric problem: after (14.24) in Section 14.3

least squares method: (14.46) in Section 14.5
natural boundary condition: after (14.2) in Section 14.1, after (14.23) in Section 14.2
pendulum: before Example 14.10 in Section 14.2
Rayleigh–Ritz method: after (14.10) in Section 14.1
Sobolev space: before (14.42) in Section 14.4
spherical pendulum: Problem 14.2.5.8
symmetric: Definition 14.1 in Section 14.5
variation: before (14.16) in Section 14.2

Mathematica Commands

After Example 14.5 in Section 14.1:

$$\text{s} = \text{NDSolve}[\{\text{y}''[\text{x}] == \frac{1}{3+\text{Cos}[\pi\,\text{x}]}\left(\text{y}'[\text{x}]*\pi*\text{Sin}[\pi\,\text{x}] - \frac{\text{x}}{5}\right),$$
$$\text{y}[0] == 0, \text{y}[1] == 0\}, \{\text{y}\}, \text{x}]//\text{Quiet}$$
$$\text{b} = \text{Plot}[\text{Evaluate}[\{\text{y}[\text{x}]\}/.\text{s}], \{\text{x}, 0, 1\}, \text{PlotStyle} \to \{\text{Blue}, \text{Thick}\},$$
$$\text{LabelStyle} \to \text{Directive}[\text{FontSize} \to 16], \text{AxesLabel} \to \{\text{x}, \text{y}\},$$
$$\text{AxesStyle} \to \text{Thickness}[0.00315]].$$

In Section 14.5:

$$\text{T1}[\text{t\_}]:=\text{Piecewise}[\{\{\frac{\text{t}}{\text{h}}, 0 < \text{t} \le \text{h}\}, \{1 - \frac{\text{t}-\text{h}}{\text{h}}, \text{h} < \text{t} \le 2\text{h}\}, \{0, -1 < \text{t} \le 0\}, \{0, 2\text{h} < \text{t} \le 1\}\}]$$

$$\text{y}[\text{x\_}, \text{c1\_}, \text{c2\_}, \text{c3\_}, \text{c4\_}, \text{c5\_}, \text{c6\_}, \text{c7\_}] := \text{c1}*\text{T1}[\text{x}] + \text{c2}*\text{T1}[\text{x}-\text{h}] + \text{c3}*\text{T1}[\text{x}-2\text{h}]$$
$$+\text{c4}*\text{T1}[\text{x}-3\text{h}] + \text{c5}*\text{T1}[\text{x}-4\text{h}] + \text{c6}*\text{T1}[\text{x}-5\text{h}] + \text{c7}*\text{T1}[\text{x}-6\text{h}]$$

$$\text{J}[\text{c1\_}, \text{c2\_}, \text{c3\_}, \text{c4\_}, \text{c5\_}, \text{c6\_}, \text{c7\_}]$$
$$:=\text{Evaluate}\Big[\int_0^1 \big((\text{D}[\text{y}[\text{x}, \text{c1}, \text{c2}, \text{c3}, \text{c4}, \text{c5}, \text{c6}, \text{c7}], \text{x}])^2 + \text{x}\,(\text{y}[\text{x}, \text{c1}, \text{c2}, \text{c3}, \text{c4}, \text{c5}, \text{c6}, \text{c7}])^2$$
$$-2*(1+\text{x})*\text{y}[\text{x}, \text{c1}, \text{c2}, \text{c3}, \text{c4}, \text{c5}, \text{c6}, \text{c7}]\big)\,\text{dx}\Big]$$

$$\text{FindMinimum}[\text{J}[\text{c1}, \text{c2}, \text{c3}, \text{c4}, \text{c5}, \text{c6}, \text{c7}], \{\{\text{c1}, .1\}, \{\text{c2}, 0.1\}, \{\text{c3}, 0.1\}, \{\text{c4}, .1\}, \{\text{c5}, 0.1\},$$
$$\{\text{c6}, 0.1\}, \{\text{c7}, 0.1\}\}]$$

$$\text{s} = \text{NDSolve}[\{\text{x}''[\text{t}] == \text{tx}[\text{t}] - 1 - \text{t}, \text{x}[0] == 0, \text{x}[1] == 0\}, \{\text{x}\}, \{\text{t}, 0, 1\}]$$

$$\text{bb}=\text{Plot}[\text{Evaluate}[\text{x}[\text{t}]/.\text{s}], \{\text{t}, 0, 1\}, \text{PlotStyle}\to\{\text{Blue}, \text{Thick}\},$$
$$\text{AxesLabel} \to \{\text{Text}[\text{Style}[''\text{x}'', \text{Italic}, 18]], \text{Text}[\text{Style}[''\text{y(x)}'', \text{Italic}, 18]]\}]$$

Further *Mathematica* commands are found after Example 14.17 in Section 14.5.

References

Lusternik, L.A. and Sobolev, V.J. *Elements of Functional Analysis*, Section 2.5. John Wiley & Sons, Inc., New York, 1974.

Thomson, W.T. *Introduction to Space Dynamics*, Section 8.3. Dover Publications, Inc., Mineola, NY, 1986.

15

Functions of a Complex Variable

15.1 Complex Numbers, Roots, and Functions

We begin by briefly reviewing complex numbers and their properties.

In Section 2.1, we defined a number z to be **complex** if $z = x + iy$, where x and y are real and $i \triangleq \sqrt{-1}$. Given a complex number $z = x + iy$, where x and y are real, we call x the **real part** of z and write $x = \mathcal{R}e(z)$, and we call y the **imaginary part** of z and write $y = \mathcal{I}m(z)$. Both $\mathcal{R}e(z)$ and $\mathcal{I}m(z)$ are real numbers. If $\mathcal{R}e(z) = 0$, we say z is **imaginary**; if $\mathcal{I}m(z) = 0$, we say z is **real**. From now on, if we write $a > 0$ or say that "a is positive," then we are implicitly assuming that a is real.

Given a complex number $z = x + iy$, where x and y are real, we denote $\bar{z} = x - iy$ and call it the **complex conjugate** of z. We have that

$$\mathcal{R}e(z) = \frac{z + \bar{z}}{2} \quad \text{and} \quad \mathcal{I}m(z) = \frac{z - \bar{z}}{2i}.$$

Note that

$$z\bar{z} = (x + iy)(x - iy) = x^2 - (iy)^2 = x^2 + y^2,$$

which is the square of the distance in the xy-plane between the points (x, y) and $(0, 0)$. Figure 15.1 shows a complex number z, its complex conjugate $\bar{z}$, and the quantity

$$|z| \triangleq \sqrt{x^2 + y^2},$$

which is called the **modulus** or **absolute value** of $x+iy$, assuming x and y are real. Note that if x is real, then the modulus of x is just the absolute value of x. Also, note that $\big|\, |z| \,\big| = |z|$, that is, the modulus of the modulus of z is just the modulus of z.

We have the important geometric property that $|z-w|$ is the distance between the points z and w in the complex plane.

15.1.1 Polar Forms

To solve difference equations in Section 4.6, we used the **polar form** of a complex number:

$$x + iy = r\cos\theta + ir\sin\theta. \tag{15.1}$$

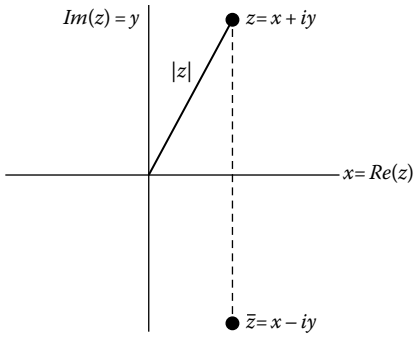


FIGURE 15.1
Complex numbers.

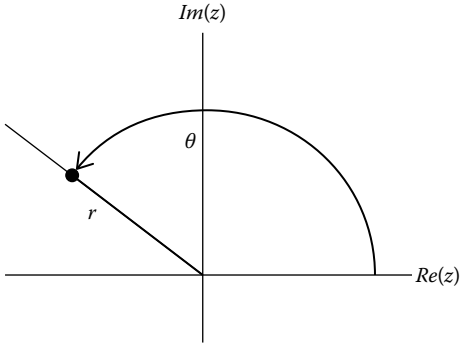


FIGURE 15.2
Complex numbers and the polar form.

This is just another example of using polar coordinates, as shown in Figure 15.2. Note that

$$r = \sqrt{x^2 + y^2} = |z| \quad \text{and} \quad \tan\theta = \frac{y}{x}, \quad \text{if} \quad x \neq 0. \tag{15.2}$$

To solve harmonic oscillator problems in Section 3.3, we used

$$\textbf{Euler's formula}: \quad e^{i\theta} \triangleq \cos\theta + i\sin\theta\,. \tag{15.3}$$

Combining (15.1) through (15.3), we have

$$z = x + iy = r(\cos\theta + i\sin\theta) = re^{i\theta} = |z|\,e^{i\theta}, \tag{15.4}$$

where

- $r = |z| \geq 0$
- θ is real
- $-\pi < \theta \leq \pi$

The **polar exponential form** of z is $|z|\,e^{i\theta}$, *where* θ is real and $-\pi < \theta \leq \pi$. Familiar laws of exponents also hold for positive integer powers of $z = re^{i\theta}$:

$$z^n = \left(re^{i\theta}\right)^n = r^n\left(e^{i\theta}\right)^n = r^n e^{in\theta} = r^n\left(\cos n\theta + i\sin n\theta\right)$$

$$= |z|^n\left(\cos n\theta + i\sin n\theta\right). \tag{15.5}$$

Note that $n\theta$ may not be in the interval $(-\pi, \pi]$, so the polar exponential form of z^n may not be $|z|^n e^{in\theta}$.

In the special case that $r = 1$, we have

$$\textbf{DeMoivre's theorem}: \quad (\cos\theta + i\sin\theta)^n = \cos n\theta + i\sin n\theta.$$

For all integers k,

$$e^{i2\pi k} = 1, \quad e^{ik\pi} = (-1)^k, \quad \text{and} \quad e^{i\theta + i2\pi k} = e^{i\theta}.$$

Further,

Theorem 15.1

For real numbers α and β,

$$e^{i\alpha} = e^{i\beta} \quad \Longleftrightarrow \quad \alpha + 2\pi k = \beta \quad \text{for some integer } k. \tag{15.6}$$

Also,

$$e^{-i\theta} = \cos(-\theta) + i\sin(-\theta) = \cos\theta - i\sin\theta.$$

Example 15.1

Find the exact polar exponential and polar forms of the numbers

(a) $-\sqrt{3} + i$, (b) $(-\sqrt{3} + i)^5$, (c) $(-\sqrt{3} + i)(1 - i)$, (d) $\frac{-\sqrt{3}+i}{1-i}$.

Method:

(a) Figure 15.3a shows the number $-\sqrt{3} + i$ in the complex plane, so we see that $r = \sqrt{(-3)^2 + (1)^2} = 2$ and the angle $\theta = \frac{5\pi}{6}$. So, in the polar form $-\sqrt{3} + i = 2\cos\frac{5\pi}{6} + i2\sin\frac{5\pi}{6}$ and in the polar exponential form $-\sqrt{3} + i = 2e^{i5\pi/6}$.

(b) Using the result of part (a), we have

$$(-\sqrt{3} + i)^5 = \left(2e^{i5\pi/6}\right)^5 = 2^5 e^{i25\pi/6} = 32e^{i25\pi/6} = 32e^{i\pi/6}$$

as the polar exponential form. It follows that its polar form is

$$(-\sqrt{3} + i)^5 = 32\cos\frac{\pi}{6} + i32\sin\frac{\pi}{6}.$$

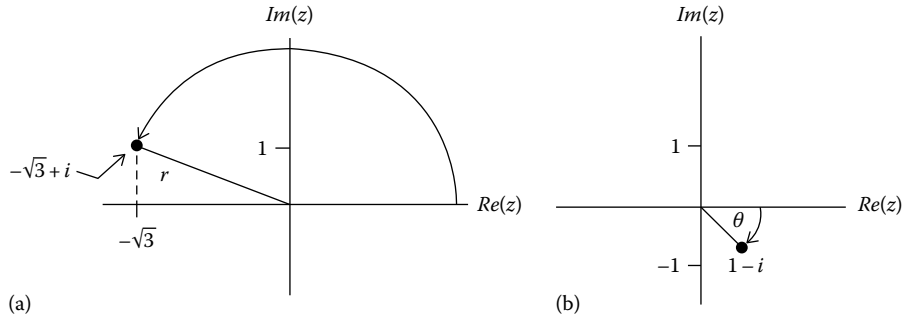


FIGURE 15.3
Example 15.1: (a) $-\sqrt{3}+i$ and (b) $1-i$.

(c) Figure 15.3b shows the number $1-i$ in the complex plane, so we see that $r = \sqrt{(1)^2+(-1)^2} = \sqrt{2}$ and the angle $\theta = -\frac{\pi}{4}$. So, $1-i = \sqrt{2}\,e^{-i\pi/4}$. Using this and a result from part (a), we have

$$(-\sqrt{3}+i)(1-i) = \left(2e^{i5\pi/6}\right)\left(\sqrt{2}e^{-i\pi/4}\right) = 2\sqrt{2}\,e^{i\left(\frac{5\pi}{6}-\frac{\pi}{4}\right)} = 2\sqrt{2}\,e^{i7\pi/12}$$

is the polar exponential form. The polar form is $2\sqrt{2}\cos\frac{7\pi}{12} + i2\sqrt{2}\sin\frac{7\pi}{12}$.

(d) Using results from parts (a) and (c),

$$\frac{-\sqrt{3}+i}{1-i} = \frac{2e^{i5\pi/6}}{\sqrt{2}e^{-i\pi/4}} = \sqrt{2}\,e^{i\frac{5\pi}{6}-\left(-i\frac{\pi}{4}\right)} = \sqrt{2}\,e^{i13\pi/12} = \sqrt{2}\,e^{-i11\pi/12}$$

is the polar exponential form, and the polar form is $\sqrt{2}\cos\left(-\frac{11\pi}{12}\right) + i\sqrt{2}\sin\left(-\frac{11\pi}{12}\right)$. ○

Example 15.2

Explain why $\cos 3\theta = \cos^3\theta - 3\cos\theta\sin^2\theta$.

Method: $\cos 3\theta = \mathcal{R}e\left(e^{i3\theta}\right) = \mathcal{R}e\left(\left(e^{i\theta}\right)^3\right) = \mathcal{R}e\left((\cos\theta + i\sin\theta)^3\right)$

$= \mathcal{R}e\left(\cos^3\theta + i3\cos^2\theta\sin\theta - 3\cos\theta\sin^2\theta - i\sin^3\theta\right) = \cos^3\theta - 3\cos\theta\sin^2\theta$. ○

Because we have

$$e^{i\theta} = \cos\theta + i\sin\theta \quad \text{and} \quad e^{-i\theta} = \cos(-\theta) + i\sin(-\theta) = \cos\theta - i\sin\theta, \tag{15.7}$$

it follows that

$$e^{-i\theta} = \overline{e^{i\theta}}.$$

Using that, we have

$$\cos\theta = \mathcal{R}e\left(e^{i\theta}\right) = \frac{e^{i\theta}+e^{-i\theta}}{2} \tag{15.8}$$

and

$$\sin\theta = \mathcal{I}m\left(e^{i\theta}\right) = \frac{e^{i\theta} - e^{-i\theta}}{2i} = i \cdot \frac{-e^{i\theta} + e^{-i\theta}}{2}. \tag{15.9}$$

15.1.2 Roots

We are familiar with the fact that the equation $z^2 = 9$ has two solutions, $z = \pm 3 = \pm\sqrt{9}$. Now we will see how this fits into the general calculation of roots that are complex numbers.

Example 15.3

For each of the given equations, find all exact solutions in the form $a + ib$:

(a) $z^2 = -5$,

(b) $z^2 = -1 + i\sqrt{3}$,

(c) $z^3 = -8$,

(d) $z^4 = 81$.

Method: Each of problems (a), . . . , (d) asks us to solve an equation of the form $z^n = w$ where z is the unknown and $w \neq 0$ is given. Our solution method is to use the polar exponential forms of both $z = re^{i\theta}$ and $w = \rho e^{i\omega}$, so the equation we are to solve is

$$(\star) \quad \rho e^{i\omega} = w = z^n = (re^{i\theta})^n = r^n e^{in\theta}.$$

Taking the modulus of both sides and noting that θ is real gives us

$$\rho = \left|\rho e^{i\omega}\right| = \left|r^n e^{in\theta}\right| = r^n,$$

where $\rho = |w| > 0$ and $r = |z|$ are both real. This implies that $r = \rho^{1/n}$. Using $\rho = r^n$ and $(\star)$ implies that we need to have

$$e^{i\omega} = e^{in\theta}.$$

Using the result of Theorem 15.1, we get $\omega + 2\pi k = n\theta$ for some integer k for all solutions $z = re^{i\theta}$.

(a) Using polar exponential forms $z = re^{i\theta}$ and $w = -5 = 5e^{i\pi}$, we see that the equation $z^2 = -5$ is equivalent to $5e^{i\pi} = -5 = z^2 = \left(re^{i\theta}\right)^2 = r^2 e^{i2\theta}$. As in the general discussion of the method earlier, we need to solve (1) $5 = |w| = r^2$ for $r = |z|$ and (2) $e^{i\pi} = e^{i2\theta}$ for real θ. This gives us $r = \sqrt{5}$, which, by definition, is the unique nonnegative number whose square is 5.

By (15.6), (2) is true if, and only if, $\pi + 2\pi k = 2\theta$, for some integer k. This is equivalent to $\theta = \frac{\pi}{2} + \pi k$. So, the solutions are

$$z = re^{i\theta} = \sqrt{5}e^{i\left(\frac{\pi}{2} + \pi k\right)}, \quad k = \textit{any integer}.$$

At first glance, we seem to have infinitely many solutions, and we know from previous experience with high school algebra that there should be only two solutions, $z = \pm i\sqrt{5}$. We rewrite the solutions as

$$z = \sqrt{5} \cdot e^{i\frac{\pi}{2}} e^{i\pi k} = \sqrt{5} \cdot i \cdot (-1)^k.$$

Since there are only two possibilities for $(-1)^k$, what looks like infinitely many solutions does reduce to two solutions, $z = \pm i\sqrt{5}$.

(b) Using polar exponential forms $z = re^{i\theta}$ and $w = -1 + i\sqrt{3} = 2\,e^{i\frac{2\pi}{3}}$, we see that the equation $z^2 = -1 + i\sqrt{3}$ is equivalent to $2\,e^{i\frac{2\pi}{3}} = -1 + i\sqrt{3} = z^2 = \left(re^{i\theta}\right)^2 = r^2\,e^{i2\theta}$. As in the general discussion of the method earlier, we need to solve (1) $2 = |w| = r^2$ for $r = |z|$ and (2) $e^{i\frac{2\pi}{3}} = e^{i2\theta}$ for real θ. This gives us $r = \sqrt{2}$.

By (15.6), (2) is true if, and only if, $\frac{2\pi}{3} + 2\pi k = 2\theta$, that is, $\theta = \frac{\pi}{3} + \pi k$, for some integer k. So, the solutions are

$$z = re^{i\theta} = \sqrt{2}\,e^{i\left(\frac{\pi}{3}+\pi k\right)}, \quad k = \textit{any integer}.$$

Again, what appears to be infinitely many solutions reduces to two solutions:

$$z = \sqrt{2}\cdot e^{i\pi/3}e^{i\pi k} = \sqrt{2}\cdot\left(-\frac{1}{2} + i\,\frac{\sqrt{3}}{2}\right)\cdot(-1)^k = \pm\left(-\frac{1}{\sqrt{2}} + i\,\frac{\sqrt{3}}{\sqrt{2}}\right).$$

The set of all solutions is $\left\{-\frac{1}{\sqrt{2}} + i\sqrt{\frac{3}{2}}, -\frac{1}{\sqrt{2}} - i\sqrt{\frac{3}{2}}\right\}$.

(c) Using polar exponential forms $z = re^{i\theta}$ and $w = -1 + i\sqrt{3} = 2\,e^{i\frac{2\pi}{3}}$, we see that the equation $z^3 = -8$ is equivalent to $8\,e^{i\pi} = -8 = z^3 = \left(re^{i\theta}\right)^3 = r^3\,e^{i3\theta}$. This gives both (1) $8 = r^3$, which tells us $r = 2$, and (2) $e^{i\pi} = e^{i3\theta}$.

By (15.6), (2) is equivalent to $\pi + 2\pi k = 3\theta$, that is, $\theta = \frac{\pi}{3} + \frac{2\pi}{3}k$, for some integer k. So, the solutions are

$$z = re^{i\theta} = 2\,e^{i\left(\frac{\pi}{3}+\frac{2\pi}{3}k\right)}, \quad k = \textit{any integer}.$$

What appears to be infinitely many solutions reduces to three solutions:

$$z_0 = 2\,e^{i\left(\frac{\pi}{3}+0\right)} = 2\left(\tfrac{1}{2} + i\,\tfrac{\sqrt{3}}{2}\right) = 1 + i\sqrt{3},$$

$$z_1 = 2\,e^{i\left(\frac{\pi}{3}+\frac{2\pi}{3}\right)} = 2\,e^{i\pi} = -2,$$

and

$$z_2 = 2\,e^{i\left(\frac{\pi}{3}+\frac{4\pi}{3}\right)} = 2\,e^{i\frac{5\pi}{3}} = 2\left(\frac{1}{2} - i\,\frac{\sqrt{3}}{2}\right) = 1 - i\sqrt{3}.$$

The set of all solutions is $\{1 + i\sqrt{3},\ -2,\ 1 - i\sqrt{3}\}$ and is displayed in Figure 15.4a.

(d) Using polar exponential forms $z = re^{i\theta}$ and $w = 81 = 81\,e^{i\cdot 0}$, we see that the equation $z^3 = 81$ is equivalent to $81\,e^{i\cdot 0} = 81 = z^4 = \left(re^{i\theta}\right)^4 = r^4\,e^{i4\theta}$. This gives both (1) $81 = r^4$, which tells us $r = 3$, and (2) $e^{i\cdot 0} = e^{i4\theta}$.

By (15.6), (2) is equivalent to $0 + 2\pi k = 4\theta$, that is, $\theta = \frac{\pi}{2}k$, for some integer k. So, the solutions are

$$z = re^{i\theta} = 3\,e^{i\frac{\pi}{2}k}, \quad k = \textit{any integer}.$$

What appears to be infinitely many solutions reduces to four solutions:

$$z_0 = 3\,e^{i0} = 3,$$

$$z_1 = 3\,e^{i\frac{\pi}{2}} = i3,$$

$$z_2 = 3\,e^{i\pi} = -3,$$

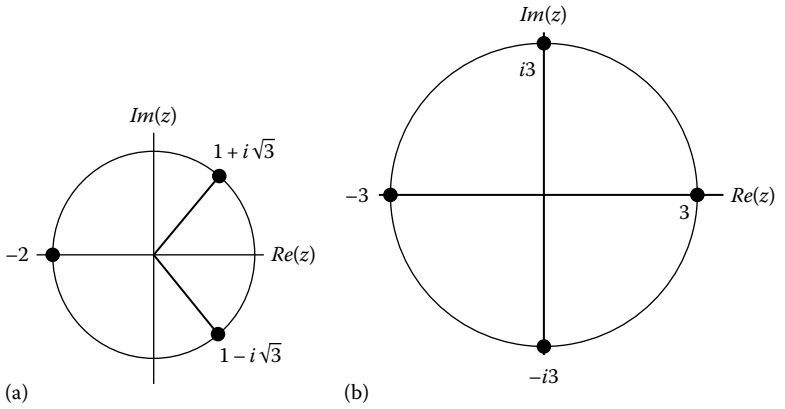


FIGURE 15.4
Example 15.3 (a) and (b).

and

$$z_3 = 3e^{i\frac{3\pi}{2}} = -i3.$$

The set of all solutions is $\{3,\ i3,\ -3,\ -i3\}$ and is displayed in Figure 15.4b. ◯

Even though the instructions asked for the solutions in the form $z = a + ib$, it helped to use the polar exponential form of z.

It is not an accident that the solutions of $z^3 = w$ are the vertices of an equilateral triangle, if $w \neq 0$. You will explain why in Problem 15.1.4.11.

By the way, we can check that the solutions in part (c) agree with what we know from high school algebra: the solutions of

$$0 = z^3 + 8 = (z+2)(z^2 - 2z + 4)$$

are $z = -2$ and, by the quadratic formula, $z = \frac{2 \pm \sqrt{(-2)^2 - 4 \cdot 1 \cdot 4}}{2 \cdot 1} = \frac{2 \pm \sqrt{-12}}{2} = \frac{2}{2} \pm \frac{i2\sqrt{3}}{2} = 1 \pm i\sqrt{3}$, as we found earlier.

In general, if w is any complex number and n is an integer ≥ 2, then we can solve $z^n = w$ by using polar exponential forms $w = \rho e^{i\alpha}$ and $z = re^{i\theta}$:

$$\rho e^{i\alpha} = w = z^n = \left(re^{i\theta}\right)^n = r^n e^{in\theta}$$

has solutions

$$z = \rho^{1/n} e^{i(\alpha + 2\pi k)/n} = \rho^{1/n} e^{i\alpha}\, e^{i2\pi k/n} = \rho^{1/n} e^{i\alpha} \left(e^{i2\pi/n}\right)^k, \quad k = 0, 1, \ldots, n-1.$$

Define the **principal nth root of unity** by

$$\omega_n = e^{i2\pi/n}$$

because, first, $\omega_n^n = 1$ and, second, *all* of the roots of 1 are given by

$$1,\ \omega_n,\ \omega_n^2,\ \ldots,\ \omega_n^{n-1}.$$

Further, all of the solutions of $z^n = w$ are given by

$$\rho^{1/n} e^{i\alpha/n},\quad \rho^{1/n} e^{i\alpha/n}\omega_n\ ,\ldots,\quad \rho^{1/n} e^{i\alpha/n}\omega_n^{n-1}.$$

15.1.3 Functions

A **function of a complex variable** is a "machine" that gives a single complex number output, $f(z)$, for every allowed complex number input z, as depicted in Figure 15.5. The word **mapping** is synonymous with the word "function."

Given a function f and a set of allowed inputs, A, the **image** of A under f is the set

$$\{f(z) :\ z \text{ in } A\}.$$

Example 15.4

Under the mapping $f(z) \triangleq 1/z$, find the image of the set $\{z :\ |z| = 2\}$.

Method: Define $w = u + iv$ by $w = f(z)$, where $u = \mathcal{R}e\big(f(z)\big)$ and $v = \mathcal{I}m\big(f(z)\big)$. Define set $A = \{z :\ |z| = 2\}$. For every z in A, the corresponding image point $w = f(z) = 1/z$ has $1/w = z$ satisfy

$$2 = |z| = \left|\frac{1}{w}\right| = \frac{|1|}{|w|} = \frac{1}{|w|},$$

so $|w| = \frac{1}{2}$. So, the image set $f(A)$ satisfies

$$f(A) \subseteq \left\{w :\ |w| = \frac{1}{2}\right\}.$$

Is $f(A) = \{w :\ |w| = \frac{1}{2}\}$? That is, does the image of A include *all* of the points in that circle in the w-plane? For any w satisfying $|w| = \frac{1}{2}$, if we let $z = \frac{1}{w}$, then

$$f(z) = \frac{1}{z} = \frac{1}{\frac{1}{w}} = w$$

and $|z| = \frac{1}{|w|} = \frac{1}{1/2} = 2$ so that z is in A. So, $f(A)$ includes all of the points of the circle $|w| = \frac{1}{2}$. That fact, combined with the **inclusion** $f(A) \subseteq \{w :\ |w| = \frac{1}{2}\}$, gives us the conclusion that $f(A) = \{w :\ |w| = \frac{1}{2}\}$. ◯

Geometrically, the set A, that is, $\{z :\ |z| = 2\}$, is the circle in the z-plane of radius 2 whose center is at the origin. Similarly, the set $f(A) = \{w :\ |w| = \frac{1}{2}\}$ is the circle in the w-plane of radius $\frac{1}{2}$ whose center is at the origin.

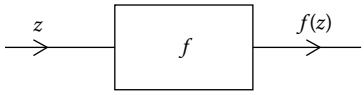


FIGURE 15.5
Function as a machine.

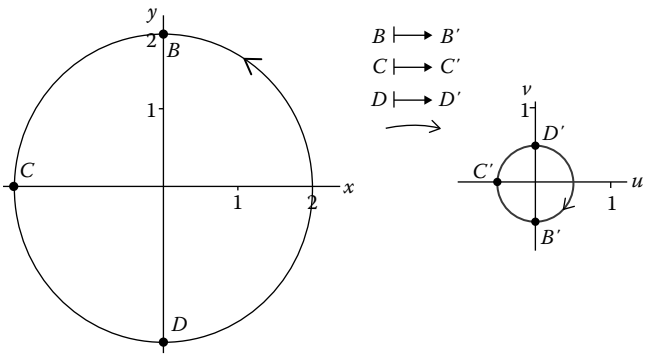


FIGURE 15.6
Mapping of a circle by the inversion mapping.

By the way, implicit in our way of solving Example 15.4 was our use of the inverse function $f^{-1}(w)$ because

$$w = f(z) = \frac{1}{z} \iff z = f^{-1}(w) = \frac{1}{w}.$$

While we cannot draw a graph of $f(z)$ versus z for z in A, we can show the effect of the mapping f on A using a transformation picture like the ones we showed in Section 6.6 for linear transformations of the plane. The transformation picture is shown in Figure 15.6.

The mapping $f(z) = 1/z$ is called the **inversion mapping**. It is possible to show that whenever A is a circle in the z-plane that does not pass through the origin, $z = 0$, then $f(A)$ is a circle in the w-plane.

Under the inversion mapping, what is the image of a circle A that *does* pass through the origin? The inversion mapping is undefined at the origin, so we should omit it from A by defining $\widehat{A} = \{z \neq 0 : z \text{ is in } A\}$ and then finding the image of $\widehat{A}$ under the inversion mapping.

Example 15.5

Find the image under the inversion mapping of the set A, which is the circle in the z-plane of radius two whose center is at $-i2$.

Method: Define $w = u + iv$ by $w = f(z)$, where $u = \mathcal{R}e\big(f(z)\big)$ and $v = \mathcal{I}m\big(f(z)\big)$. The set A consists of $z = x + iy$ whose distance from the point $z_0 \triangleq -i2$ is two, that is, $2 = |x + iy - (0 - i2)| = \sqrt{(x-0)^2 + (y-(-2))^2}$. So,

$$A = \{z : |z + i2| = 2\},$$

and every z in A is of the form

$$z = -i2 + 2e^{it}$$

for some real t. We define

$$\widehat{A} = \{z \neq 0 : z \text{ is in } A\} = \left\{-i2 + 2e^{it} : \frac{\pi}{2} < t < \frac{5\pi}{2}\right\}.$$

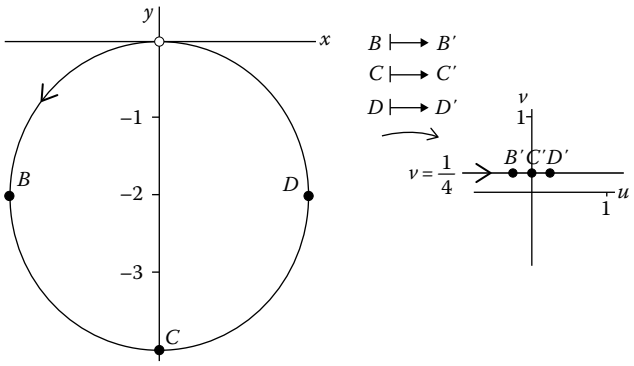


FIGURE 15.7
Example 15.5.

It follows that for every z in $\widehat{A}$,

$$u + iv = w = f(z) = \frac{1}{-2i + 2e^{it}} = \frac{1}{-2i + 2\cos t + i2\sin t} = \frac{1}{2} \cdot \frac{1}{\cos t + i(-1 + \sin t)}.$$

Rationalizing the denominator gives

$$u + iv = \frac{1}{2} \cdot \frac{\cos t - i(-1 + \sin t)}{\cos^2 t + (-1 + \sin t)^2} = \frac{1}{2} \cdot \frac{\cos t - i(-1 + \sin t)}{\cos^2 t + (1 - 2\sin t + \sin^2 t)}$$

$$= \frac{1}{2} \cdot \frac{\cos t + i(1 - \sin t)}{2 - 2\sin t} = \frac{1}{4}\left(\frac{\cos t}{1 - \sin t} + i\,\frac{1 - \sin t}{1 - \sin t}\right) = \frac{\cos t}{4(1 - \sin t)} + i\,\frac{1}{4}.$$

So, $f(\widehat{A}) \subseteq \{u + iv : v = \frac{1}{4}\}$; the latter is a horizontal line in the w-plane. The ultimate results are illustrated in Figure 15.7.

Is $f(\widehat{A}) = \{u + iv : v = \frac{1}{4}\}$? That is, does the image of $\widehat{A}$ include *all* of the points in that line in the w-plane? For any $w = u + iv$ satisfying $\mathcal{I}m(w) = \frac{1}{4}$, we need to show that there is at least one value of t in the interval $\frac{\pi}{2} < t < \frac{5\pi}{2}$ for which $g(t) \triangleq \frac{\cos t}{4(1-\sin t)} = u$. We calculate

$$\lim_{t \to \frac{\pi}{2}+} \frac{\cos t}{4(1 - \sin t)} = \lim_{t \to \frac{\pi}{2}+} \frac{(1 + \sin t)\cos t}{4(1 + \sin t)(1 - \sin t)} = \lim_{t \to \frac{\pi}{2}+} \frac{(1 + \sin t)}{4\cos t} = \lim_{t \to \frac{\pi}{2}+} \frac{\approx 2}{\approx 0^-} = -\infty, \tag{15.10}$$

and similarly

$$\lim_{t \to \frac{5\pi}{2}-} g(t) = \lim_{t \to \frac{5\pi}{2}-} \frac{\cos t}{4(1 - \sin t)} = \infty. \tag{15.11}$$

Because $g(t)$ is continuous for $\frac{\pi}{2} < t < \frac{5\pi}{2}$, (15.10) and (15.11) imply $g(t)$ takes on all values in the interval $(-\infty, \infty)$. This concludes the explanation why $f(\widehat{A}) = \{u + iv : v = \frac{1}{4}\}$, a horizontal line in the w-plane. ○

In general, if A is a circle passing through the origin in the z-plane, then the image of $\widehat{A}$ under the inversion mapping is a line in the w-plane. We will explain this in Chapter 16 when we need this general result.

15.1.4 Problems

1. Why is $\overline{wz} = \overline{w}\,\overline{z}$ for all complex numbers $z \neq 0$ and w?
2. Why is $\overline{\left(\frac{w}{z}\right)} = \frac{\overline{w}}{\overline{z}}$ for all complex numbers w, z?
3. Explain why $z_1\bar{z}_2 - z_2\bar{z}_1$ is imaginary for all complex numbers z_1, z_2.
4. Why is $|z - w|$ the distance between the points z and w in the complex plane?
5. Explain why the only complex number z for which $\bar{z} = 2 - \dfrac{1}{z}$ is $z = 1$. [Hint: Multiply both sides by z.]
6. Find the exact polar exponential form of (a) $\frac{-1+i}{\sqrt{3}-i}$, (b) $(-1+i)^3$, and (c) $(\sqrt{3}-i)^7$.
7. Find the exact polar exponential and $x + iy$ forms of (a) $(-\sqrt{3} - i)^2$, (b) $\left(\frac{3}{5} - i\,\frac{4}{5}\right)^2$, and (c) $\frac{-\sqrt{3}-i}{1+i}$.
8. Explain why $\sin 4\theta \equiv 4\cos\theta\,\sin\theta(\cos^2\theta - \sin^2\theta)$ is an identity.
9. Explain why (a) $\sin 3\theta \equiv 3\cos^2\theta\sin\theta - \sin^3\theta$ and (b) $\cos 4\theta \equiv \cos^4\theta - 6\cos^2\theta\sin^2\theta + \sin^4\theta$ are identities.
10. The number π is **irrational**, that is, cannot be written as a ratio of integers. Use this fact to explain why $z \triangleq e^i$ has $z^n \neq 1$ for all integers n.
11. Suppose $\rho > 0$ and α is real. Explain why the cube roots of any number $0 \neq \rho e^{i\alpha}$ are the vertices of an equilateral triangle. What is the length of each of its sides?
12. Suppose $\rho > 0$ and α is real. Explain why the sixth roots of any number $\rho e^{i\alpha}$ are the vertices of an equilateral hexagon. What is the length of each of its sides?
13. Find the exact solutions of the equations. Give your final conclusions in both polar exponential and $x + iy$ forms. (a) $z^2 = i4$, (b) $z^2 = -2 - i2\sqrt{3}$, (c) $z^3 = 27$, (d) $z^3 = -2 - i2$, (e) $z^3 = -4 + i4\sqrt{3}$, (f) $z^4 = -81$.
14. Solve $z^3 = -1 + i$ and express the solution(s) in polar exponential form.
15. Solve $z^3 = -1 - i\sqrt{3}$ and express the solution(s) in polar exponential form.
16. Find the exact solutions of the equations. Give your final conclusions in both polar exponential and $x + iy$ forms:
    (a) $z^2 + 4z + 8 = 0$,
    (b) $z^3 - z^2 + z = 0$.
17. Find all solutions of the equations:
    (a) $\dfrac{1-z}{1+z} = z$,
    (b) $(z-i)^3 = z^3$.
18. For each of (a), (b), and (c), describe in words involving geometric concepts the set of points and draw it in the complex plane:
    (a) $|z + 1 - i\sqrt{3}| = 1$,
    (b) $|z - 1| = |z - i2|$,
    (c) $|z - i| = |z + 1|$.

19. Find the image of the given sets under the inversion mapping. Draw an appropriate transformation picture and give the image in set notation:
    (a) $\{z : |z| = 3\}$,
    (b) $\{z : |z - 1| = 1\}$,
    (c) $\{z : |z + 2| = 2\}$,
    (d) $\{z : |2z + i| = 1\}$.
20. Find the image of the given sets under the mapping $f(z) = z^2$. Draw an appropriate transformation picture and give the image in set notation. (a) The half of the circle $|z| = 2$ with $\mathcal{R}e(z) \geq 0$; (b) the half of the circle $|z| = 2$ with $\mathcal{I}m(z) \leq 0$; (c) the quarter *disk* $\{z = x + iy : 0 \leq x^2 + y^2 \leq 4, x \geq 0, y \geq 0\}$; (d) the line $\mathcal{I}m(z) = -1$, (e) the line $\mathcal{R}e(z) = 1$, (f) the line $\mathcal{I}m(z) = \beta$, where β is an unspecified real constant, (g) the line $\mathcal{R}e(z) = \alpha$, where α is an unspecified real constant, (h) the line $\mathcal{R}e(z) = \mathcal{I}m(z)$.
21. Find the image of the given sets under the mapping $f(z) = iz - 3$. Draw an appropriate transformation picture and give the image in set notation. (a) $\{z : |z| = 1\}$, (b) $\{z : \mathcal{R}e(z) = 2\}$, (c) $\{z : \mathcal{I}m(z) = -1\}$, (d) $\{z : \mathcal{R}e(z) = -\mathcal{I}m(z)\}$.
22. Find the exact value of $(-\sqrt{3}+i)^5$ in the form $x+iy$, where x, y are real. Do evaluate exact trigonometric function values.
23. Suppose a and R are unspecified real numbers. Using the formula $|x + iy| = \sqrt{x^2 + y^2}$, explain why, for all real θ, both (a) $|a + Re^{i\theta}| = |a + Re^{-i\theta}|$ and (b) $|a + Re^{i\theta}| = |ae^{i\theta} + R|$.

15.2 Derivative and the Cauchy–Riemann Equations

Just as for a function of a single real variable, there are concepts of limit, continuity, and derivative for a function of a single complex variable. Such functions usually take on complex values. For example, if $z = x + iy$ where x and y are real, then the function defined by $f(z) = z^2$ can also be expressed as

$$f(z) \triangleq z^2 = (x + iy)^2 = (x^2 - y^2) + i2xy \triangleq u(x, y) + iv(x, y),$$

whose real part is $u(x, y) = \mathcal{R}e\big(f(z)\big) = x^2 - y^2$ and whose imaginary part is $v(x, y) = \mathcal{I}m\big(f(z)\big) = 2xy$.

So, a function of a single complex variable $z = x + iy$ can be expressed in terms of two real-valued functions of the two real variables x, y. Our work with functions of a single complex variable will have a lot in common with the multivariable calculus study of vector fields and thus will be richer than Calculus I study of a function of a single real variable.

For example, the function of a single real variable x given by $g(x) = |x|^2$ is differentiable at all x, but the function of a single complex variable z given by $f(z) = |z|^2$ is differentiable *only* at $z = 0$, as we will see in Example 15.12.

Example 15.6

Express $f(z) \triangleq 1/z$ in terms of two functions of two real variables.

Method: Let $z = x + iy$ where x and y are real. Using rationalization of the denominator,

$$f(z) \triangleq \frac{1}{z} = \frac{1}{x+iy} = \frac{(x-iy)}{(x+iy)(x-iy)} = \frac{x-iy}{x^2+y^2} = \frac{x}{x^2+y^2} + i\,\frac{-y}{x^2+y^2} \triangleq u(x,y) + iv(x,y),$$

whose real part is $u(x,y) = \dfrac{x}{x^2+y^2}$ and whose imaginary part is $v(x,y) = -\dfrac{y}{x^2+y^2}$. ○

An **open disk** around the complex number z_0 is a set of the form

$$\mathcal{D}_r(z_0) \triangleq \{z : |z - z_0| < r\},$$

where $r > 0$. The **punctured disk** of positive radius r about z_0 is the set $\{z : 0 < |z-z_0| < r\}$.

Intuitively, the complex output values $w \triangleq f(z)$ tend to a limit w_0 if $f(z)$ gets closer and closer to w_0 as the input values z get closer and closer to z_0. Because modulus measures distance between two complex numbers, having such a limit means that

$$|f(z) - w_0| \quad \text{goes to zero} \quad \text{as } |z - z_0| \quad \text{goes to zero.}$$

Here is a precise definition.*

Definition 15.1

Suppose the domain of $f(z)$ includes a punctured disk about z_0. Suppose that for all $\varepsilon > 0$ there is a $\delta > 0$ such that $0 < |z - z_0| < \delta$ implies $|f(z) - w_0| < \varepsilon$. Then we say that the **limit of** $f(z)$, **as** z **approaches** z_0, **exists and equals** w_0, and we write

$$\lim_{z \to z_0} f(z) = w_0 \tag{15.12}$$

or

$$f(z) \to w_0 \quad \text{as} \quad z \to z_0. \tag{15.13}$$

In the definition, given a function f, the $\delta > 0$ may depend on either or both of ε and z_0.

The existence of the limit in (15.12) means that no matter how small an open disk (of radius ε) is drawn around w_0, there is a sufficiently small, positive radius δ such that f maps the punctured disk $\{z : 0 < |z - z_0| < \delta\}$ into the disk $\{w : |w - w_0| < \varepsilon\}$. This is depicted in Figure 15.8, where complex numbers $w = u + iv$ with u and v being real.

* There is a more general definition of limit when the domain of $f(z)$ does not include a punctured disk about z_0: we say z_0 is a **limit point** of a set A if for all $r > 0$, there is a point $z_r \neq z_0$ in the intersection $A \cap \mathcal{D}_r(z_0)$. Suppose z_0 is a limit point of the domain of $f(z)$ and for all $\varepsilon > 0$ there exists a $\delta > 0$ such that z in both the domain of f and the punctured disk of radius δ about z_0 implies $|f(z) - w_0| < \varepsilon$. Then we say that the limit of $f(z)$, as z approaches z_0, exists and equals w_0, and we write $\lim_{z \to z_0} f(z) = w_0$ or $f(z) \to w_0$ as $z \to z_0$. This more general definition is used when we consider boundary values, for example, for partial differential equations (PDEs) rewritten as problems in the complex plane.

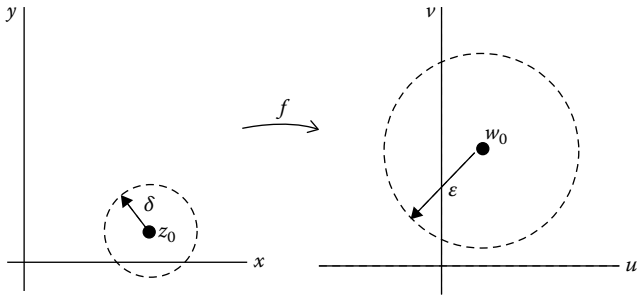


FIGURE 15.8
Limit in terms of ε and δ.

Theorem 15.2

Suppose $\lim_{z\to z_0} f(z)$ and $\lim_{z\to z_0} g(z)$ exist and c is a complex constant. Then there exist

(a) $\lim_{z\to z_0}\big(f(z)+g(z)\big) = \lim_{z\to z_0} f(z) + \lim_{z\to z_0} g(z)$.
(b) $\lim_{z\to z_0}\big(cf(z)\big) = c\,\lim_{z\to z_0} f(z)$.
(c) $\lim_{z\to z_0}\big(f(z)g(z)\big) = \big(\lim_{z\to z_0} f(z)\big)\big(\lim_{z\to z_0} g(z)\big)$.
(d) $\lim\limits_{z\to z_0} \dfrac{g(z)}{h(z)} = \dfrac{\lim_{z\to z_0} g(z)}{\lim_{z\to z_0} h(z)}$, as long as there exists $\lim_{z\to z_0} h(z) \neq 0$.

Example 15.7

Explain why $\lim_{z\to i} \frac{3(z^2+1)}{z-i} = i6$.

Method: Let $z_0 = i$ and define $f(z) \triangleq \dfrac{3(z^2+1)}{z-i}$ for $z \neq i$. For $z \neq i$,

$$f(z) = \frac{3(z^2+1)}{(z-i)} = \frac{3(z+i)\cancel{(z-i)}}{\cancel{(z-i)}} = 3(z+i).$$

The problem's narrative gave us the clue that $f(z)$ should be approaching $i6 = w_0$ as $z \to i$. For $z \neq i$, we calculate

$$|f(z) - w_0| = |3(z+i) - i6| = |3z - 3i| = |3(z-i)| = |3|\,|z-i| = 3\,|z-i|.$$

So, for any $\varepsilon > 0$, we can choose $\delta = \frac{1}{3}\,\varepsilon$ to see that $0 < |z-i| < \delta$ implies $|f(z) - i6| < \varepsilon$, as we desired. ◯

The results of Example 15.7 are illustrated in Figure 15.9.

In Example 15.7, the limit exists as $z \to i$ even though $f(z)$ is undefined at $z = i$. When we say that the complex numbers $f(z)$ approach $i6$, it is a completely separate issue whether f does or does not take on the value $i6$, or any value at all, at $z = i$.

In general, for $\lim_{z\to z_0} f(z) = w_0$ to exist, it must be true that the complex numbers $f(z)$ approach w_0 for *all* paths of complex numbers z approaching z_0.

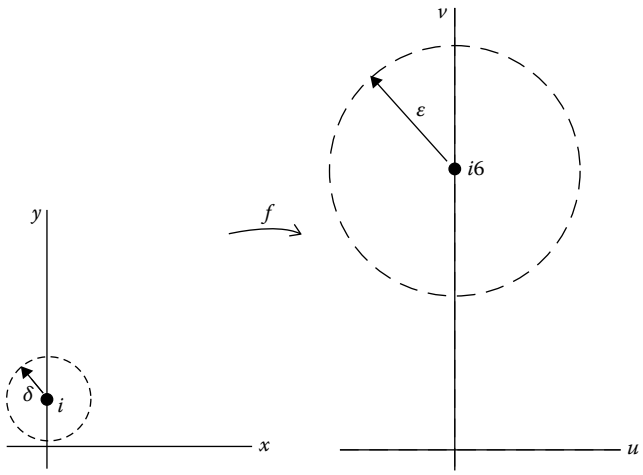


FIGURE 15.9
Example 15.7.

Definition 15.2

A function f is **continuous** at z_0 if f is defined on a sufficiently small open disk around z_0 and $\lim_{z\to z_0} f(z) = f(z_0)$.

Definition 15.3

A function f whose domain is $\mathcal{D}$ has an **extension** $\widetilde{f}$ if both (a) the domain of $\widetilde{f}$ is a set $\widetilde{\mathcal{D}}$ that contains $\mathcal{D}$ and (b) $f(z) = \widetilde{f}(z)$ for all z in $\mathcal{D}$.

Example 15.8

Because $f(z) \triangleq \dfrac{3(z^2+1)}{z-i}$ is undefined at $z = i$, f is not continuous at i. Find an extension of f that is continuous at i.

Method: The natural domain of f is $\mathcal{D} = \{z : z \neq i\}$. Since we want $\widetilde{f}(z) = f(z)$ for all $z \neq i$ and we know from Example 15.7 that $\lim_{z\to i} f(z) = i6$, we should try

$$\widetilde{f}(z) \triangleq \left\{ \begin{array}{ll} \dfrac{3(z^2+1)}{z-i}, & \text{if } z \neq i \\ \\ i6, & \text{if } z = i \end{array} \right\}.$$

We constructed $\widetilde{f}$ to agree with f on $\mathcal{D}$ and the domain of $\widetilde{f}$ is the whole z-plane, so $\widetilde{f}$ is an extension of f.

Because $\widetilde{f}$ is defined at $z = i$ and on an open disk around $z = i$, and

$$\lim_{z\to i} f(z) = \lim_{z\to i} \widetilde{f}(z) = i6,$$

we conclude that $\widetilde{f}$ is continuous at $z = i$ and thus is as desired. ○

By a **polynomial in** z of **degree** n, we will mean a function of the form $p(z) = a_0 + a_1 z + \cdots + a_n z^n$, where $a_0, \ldots, a_n$ are constant complex numbers and $a_n \neq 0$.

Theorem 15.3

(a) If $f(z)$ is a polynomial, then $f(z)$ is continuous at every z.
(b) If $p(z)$ and $q(z)$ are polynomials, then $f(z) \triangleq p(z)/q(z)$ is continuous at every z where $q(z) \neq 0$.
(c) If $f(z)$ is a polynomial defined on a punctured disk $\{z : 0 < |z - z_0| < \delta\}$ whose radius is positive, then there exists $\lim_{z \to z_0} f(z)$.

15.2.1 Derivatives

The definition of the derivative of a function of a single complex variable is analogous to the definition of derivative of a function of a single real variable.

Throughout the rest of this section, we will assume that a function $f(z)$ is defined at least for all z in some open disk about a point z_0, that is, the domain of f includes a disk $\{z : |z - z_0| < \delta\}$ for some $\delta > 0$, possibly very small.

Definition 15.4

The **derivative** of f at z_0 exists and is given by

$$f'(z_0) \triangleq \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0},$$

if the limit exists. If it does, then we say that f is **differentiable at** z_0.

An alternative notation for the limit in the definition of the derivative is

$$f'(z) = \lim_{\triangle z \to 0} \frac{f(z + \triangle z) - f(z)}{\triangle z},$$

if the limit exists.

We begin with an example and some properties of derivatives that are familiar to us from our Calculus I courses.

Example 15.9

Assume a is an unspecified complex number. Explain why $f(z) \triangleq 3(z^2 + a)$ is differentiable everywhere and find a formula for its derivative as a function of z.

Method: For any z_0,

$$f'(z_0) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} = \lim_{z \to z_0} \frac{3(z^2 + \cancel{a}) - 3(z_0^2 + \cancel{a})}{z - z_0} = \lim_{z \to z_0} \frac{3(z^2 - z_0^2)}{z - z_0},$$

so, using Theorem 15.3(c), there exists

$$f'(z_0) = \lim_{z \to z_0} \frac{3(z + z_0)\cancel{(z - z_0)}}{\cancel{(z - z_0)}} = \lim_{z \to z_0} 3(z + z_0) = 3(z_0 + z_0) = 6z_0.$$

As a function of z, the derivative is $f'(z) = 6z$. ○

Theorem 15.4

(a) If n is a positive integer, then z^n is differentiable everywhere and $\left(z^n\right)' = nz^{n-1}$.

(b) If $f(z)$ is a polynomial, then $f(z)$ is differentiable everywhere.

(c) If $p(z)$ and $q(z)$ are polynomials, then $f(z) \triangleq p(z)/q(z)$ is differentiable at every z for which $q(z) \neq 0$.

Theorem 15.5

If $f(z)$ is differentiable at z_0, then $f(z)$ is continuous at z_0.

Theorem 15.6

Suppose $f(z)$ and $g(z)$ are differentiable at z and c is a complex constant. Then the following derivatives at z exist:

(a) $\left(f(z) + g(z)\right)' = f'(z) + g'(z)$.

(b) $\left(cf(z)\right)' = cf'(z)$.

(c) $\left(f(z)g(z)\right)' = f'(z)g(z) + f(z)g'(z)$.

(d) $\dfrac{d}{dz}\left[\dfrac{f(z)}{g(z)}\right] = \dfrac{f'(z)g(z) - f(z)g'(z)}{\left(g(z)\right)^2}$, as long as $g(z) \neq 0$.

(e) $(f \circ g)'(z) = \dfrac{d}{dz}\left[f\left(g(z)\right)\right] = f'(g(z))g'(z)$, as long as f is differentiable at $g(z)$.

The result of Theorem 15.6(c) is the product rule, the result of Theorem 15.6(d) is the quotient rule, and the result of Theorem 15.6(e) is the chain rule for $(f \circ g)(z) = f\left(g(z)\right)$, the composition of two functions of complex variables.

Example 15.10

At what value(s) of z is $(z + i)/(z - i)$ differentiable? Find its derivative where it exists.

Method: The denominator is zero only at $z = i$, so the derivative of $(z + i)/(z - i)$ fails to exists at $z = i$. At every other z in the complex plane, the derivative exists and is given by

$$\left(\frac{z+i}{z-i}\right)' = \frac{(z+i)'(z-i) - (z+i)(z-i)'}{(z-i)^2} = \frac{(z-i) - (z+i)}{(z-i)^2} = -\frac{2i}{(z-i)^2}. \text{ ○}$$

15.2.2 Cauchy–Riemann Equations

Here we begin to see how the results for a function $f(z)$ of one complex variable $z = x + iy$ are richer than the results for a function of one real variable or even the results for a planar vector field, that is, a vector valued function of two real variables (x, y).

Theorem 15.7

Suppose $f(z) = u(x, y) + iv(x, y)$, where $z = x + iy$, and x, y, u, v are real. If $f(z)$ is differentiable at $z_0 = x_0 + iy_0$, then $\frac{\partial u}{\partial x}$, $\frac{\partial u}{\partial y}$, $\frac{\partial v}{\partial x}$, $\frac{\partial v}{\partial y}$ all exist at (x_0, y_0) and the **Cauchy–Riemann equations** hold at (x_0, y_0), that is,

$$\left\{ \begin{array}{l} \dfrac{\partial u}{\partial x}(x_0, y_0) = \dfrac{\partial v}{\partial y}(x_0, y_0) \\ \\ \dfrac{\partial u}{\partial y}(x_0, y_0) = -\dfrac{\partial v}{\partial x}(x_0, y_0) \end{array} \right\}. \tag{15.14}$$

Further

$$f'(z_0) = \frac{\partial u}{\partial x}(x_0, y_0) + i\,\frac{\partial v}{\partial x}(x_0, y_0) \tag{15.15}$$

and

$$f'(z_0) = \frac{\partial v}{\partial y}(x_0, y_0) - i\,\frac{\partial u}{\partial y}(x_0, y_0)\,. \tag{15.16}$$

Why? First, we see that the Cauchy–Riemann equations, (15.14), follow from comparing (15.15) with (15.16). So, it will suffice to show that both (15.15) and (15.16) are true and that the four partial derivatives that appear in those equations exist.

Assume $f(z)$ is differentiable at $z_0 = x_0 + iy_0$. Then, by definition,

$$f'(z_0) = f'(x_0 + iy_0) = \lim_{x+iy \to x_0+iy_0} \frac{f(x + iy) - f(x_0 + iy_0)}{(x + iy) - (x_0 + iy_0)} \tag{15.17}$$

exists and is the same complex number, no matter what path in the z-plane along which we choose to have $x + iy$ approach $x_0 + iy_0$.

In particular, along horizontal paths with $y \equiv y_0$, we calculate

$$f'(z_0) = \lim_{\substack{x \to x_0 \\ y \equiv y_0}} \frac{f(x + iy_0) - f(x_0 + iy_0)}{(x + iy_0) - (x_0 + iy_0)} = \lim_{x \to x_0} \frac{u(x, y_0) + iv(x, y_0) - \big(u(x_0, y_0) + iv(x_0, y_0)\big)}{x - x_0},$$

that is,

$$(\star)\ f'(z_0) = \lim_{x \to x_0} \left(\frac{u(x, y_0) - u(x_0, y_0)}{x - x_0} + i\,\frac{v(x, y_0) - v(x_0, y_0)}{x - x_0} \right).$$

In order for the limit of a complex-valued expression to exist, the limits of both the real and imaginary parts must exist. So, if $f'(z_0)$ exists, then there exist both

$$\lim_{x\to x_0} \frac{u(x,y_0)-u(x_0,y_0)}{x-x_0} = \frac{\partial u}{\partial x}(x_0,y_0)$$

and

$$\lim_{x\to x_0} \frac{v(x,y_0)-v(x_0,y_0)}{x-x_0} = \frac{\partial v}{\partial x}(x_0,y_0).$$

So, from the existence of $f'(z_0)$ and $(\star)$, it follows that (15.15) is true, that is,

$$f'(z_0) = \frac{\partial u}{\partial x}(x_0,y_0) \; + i\,\frac{\partial v}{\partial x}(x_0,y_0)\,.$$

Along vertical paths with $x \equiv x_0$, we calculate

$$f'(z_0) = \lim_{\substack{x\equiv x_0\\ y\to y_0}} \frac{f(x_0+iy)-f(x_0+iy_0)}{(x_0+iy)-(x_0+iy_0)} = \lim_{y\to y_0} \frac{u(x_0,y)+iv(x_0,y)-\big(u(x_0,y_0)+iv(x_0,y_0)\big)}{i(y-y_0)},$$

that is,

$$(\star\star)\;\; f'(z_0) = \lim_{y\to y_0}\left(\frac{u(x_0,y)-u(x_0,y_0)}{i(y-y_0)} + i\,\frac{v(x_0,y)-v(x_0,y_0)}{i(y-y_0)}\right).$$

In order for the limit of a complex-valued expression to exist, the limits of both the real and imaginary parts must exist. So, if $f'(z_0)$ exists, then there exist both

$$\lim_{y\to y_0} \frac{v(x_0,y)-v(x_0,y_0)}{y-y_0} = \frac{\partial v}{\partial y}(x_0,y_0)$$

and

$$-\lim_{y\to y_0} \frac{u(x_0,y)-u(x_0,y_0)}{x-x_0} = -\frac{\partial u}{\partial y}(x_0,y_0).$$

So, from the existence of $f'(z_0)$ and $(\star\star)$, it follows that (15.16) is true, that is,

$$f'(z_0) = \frac{\partial v}{\partial y}(x_0,y_0) \; - i\,\frac{\partial u}{\partial y}(x_0,y_0)\,.$$

Equating (15.15) and (15.16) establishes the rest of the theorem. □

So, differentiability of $f = u + iv$ implies that the Cauchy–Riemann equations, relating partial derivatives of u and v, are satisfied. This suggests the possibility that a function f satisfying the Cauchy–Riemann equations would be differentiable. Unfortunately, satisfying the Cauchy–Riemann equations, while a necessary consequence of differentiability, does not, by itself, guarantee differentiability. We do have a partially satisfying result:

Theorem 15.8

Suppose $f(z) = u(x, y) + iv(x, y)$, where $z = x + iy$, and x, y, u, v are real. If $\frac{\partial u}{\partial x}$, $\frac{\partial u}{\partial y}$, $\frac{\partial v}{\partial x}$, $\frac{\partial v}{\partial y}$ are all continuous at (x_0, y_0) and the Cauchy–Riemann equations, (15.14), hold at (x_0, y_0), then $f(z)$ is differentiable at $z_0 = x_0 + iy_0$ and $f'(z_0)$ is given by (15.15) and by (15.16).

We may refer to the properties that "$\frac{\partial u}{\partial x}$, $\frac{\partial u}{\partial y}$, $\frac{\partial v}{\partial x}$, $\frac{\partial v}{\partial y}$ are all continuous at (x_0, y_0)" as the "technical hypotheses."

Example 15.11

At what value(s) of z is $f(z) \triangleq 2xy + i(y^2 - x^2)$ differentiable? Find a formula for $f'(z)$ where it exists.

Method: $u(x, y) = 2xy$ and $v(x, y) = y^2 - x^2$ are polynomials in real variables x and y, so $\frac{\partial u}{\partial x} = 2y$, $\frac{\partial u}{\partial y} = 2x$, $\frac{\partial v}{\partial x} = -2x$, and $\frac{\partial v}{\partial y} = 2y$ exist and are continuous everywhere. Thus, the technical hypotheses of Theorem 15.8 are satisfied. This implies that the Cauchy–Riemann equations are "decisive" in this example, in the sense that $f'(x_0 + iy_0)$ exists if, and only if, the Cauchy–Riemann equations are satisfied at (x_0, y_0).

For any $x + iy$, we verify the Cauchy–Riemann equations:

$$\frac{\partial u}{\partial x}(x, y) - \frac{\partial v}{\partial y}(x, y) = 2y - 2y \equiv 0 \quad \text{and} \quad \frac{\partial u}{\partial y}(x, y) + \frac{\partial v}{\partial x}(x, y) = 2x + (-2x) \equiv 0.$$

So, f is differentiable everywhere. By (15.15), $f'(z) = \frac{\partial u}{\partial x} + i\,\frac{\partial v}{\partial x} = 2y - i2x = -i2z$. ◯

Example 15.12

At what value(s) of z is $f(z) \triangleq |z|^2$ differentiable? Find a formula for $f'(z)$ where it exists.

Method: $f(z) = x^2 + y^2$, so $u(x, y) = x^2 + y^2$ and $v(x, y) \equiv 0$ are polynomials in real variables x and y, so $\frac{\partial u}{\partial x} = 2x$, $\frac{\partial u}{\partial y} = 2y$, $\frac{\partial v}{\partial x} = \frac{\partial v}{\partial y} \equiv 0$ exist and are continuous everywhere. So the technical hypotheses of Theorem 15.8 are satisfied. This implies that the Cauchy–Riemann equations are decisive in this example. For any $x + iy$, we have

$$\frac{\partial u}{\partial x}(x, y) - \frac{\partial v}{\partial y}(x, y) = 2x - 0 = 2x \quad \text{and} \quad \frac{\partial u}{\partial y}(x, y) + \frac{\partial v}{\partial x}(x, y) = 2y + 0 = 2y,$$

so f is differentiable only where both $2x = 0$ and $2y = 0$. So, f is differentiable *only* at the origin, $z = 0$, where (15.15) gives $f'(0) = \frac{\partial u}{\partial x}(0, 0) + i\,\frac{\partial v}{\partial x}(0, 0) = 0 + i\,0 = 0$. ◯

Example 15.13

At what value(s) of z is $f(z) \triangleq y^2 - x^2 + x + i(-2xy + y^2)$ differentiable? Find a formula for $f'(z)$ where it exists.

Method: $u(x,y) = y^2 - x^2 + x$ and $v(x,y) = -2xy + y^2$ are polynomials in real variables x and y, so $\frac{\partial u}{\partial x} = -2x + 1$, $\frac{\partial u}{\partial y} = 2y$, $\frac{\partial v}{\partial x} = -2y$, $\frac{\partial v}{\partial y} = -2x + 2y$ exist and are continuous everywhere. So, the technical hypotheses of Theorem 15.8 are satisfied. This implies that the Cauchy–Riemann equations are decisive in this example.

For any $x + iy$, we have

$$\frac{\partial u}{\partial x}(x,y) - \frac{\partial v}{\partial y}(x,y) = -2x + 1 - (-2x + 2y) = 1 - 2y \quad \text{and}$$

$$\frac{\partial u}{\partial y}(x,y) + \frac{\partial v}{\partial x}(x,y) = 2y + (-2y) \equiv 0,$$

so f is differentiable only if $1 - 2y = 0$. So, f is differentiable *only* on the line $y = \frac{1}{2}$, where (15.15) gives $f'(x + i\frac{1}{2}) = \frac{\partial u}{\partial x}(x, \frac{1}{2}) + i\frac{\partial v}{\partial x}(x, \frac{1}{2}) = (1 - 2x) + i(-2 \cdot \frac{1}{2}) = 1 - 2x - i = 1 - 2z$. ◯

In the appendix of this section, we will study two examples to further understand complications involving the Cauchy–Riemann equations.

15.2.3 Orthogonal Families of Curves and an Application to Fluid Flow

Suppose we have a function of two real variables $\Phi = \Phi(x,y)$ and a point $P = (x_0, y_0)$ that lies on a level curve $\{(x,y) : \Phi(x,y) = k_1\}$; hence, $\Phi(x_0, y_0) = k_1$. Recall from Section 6.5 that the vector

$$\nabla\Phi(x_0,y_0) = \frac{\partial \Phi}{\partial x}(x_0,y_0)\hat{\boldsymbol{\imath}} + \frac{\partial \Phi}{\partial y}(x_0,y_0)\hat{\boldsymbol{\jmath}}$$

is normal to the level curve at point P, as long as $\nabla\Phi(x_0,y_0) \neq 0$.

As usual, write $z = x + iy$. Suppose further that $\Phi(x,y)$ is the real part of a function $f(z)$ that is differentiable at $z_0 = x_0 + iy_0$, that is, there is a function $\Psi(x,y)$ such that

$$f(z) \triangleq \Phi(x,y) + i\Psi(x,y)$$

is differentiable at $z_0 = x_0 + iy_0$. By the Cauchy–Riemann equations, at $(x,y) = (x_0, y_0)$,

$$\nabla\Phi \bullet \nabla\Psi = \left(\frac{\partial \Phi}{\partial x}\hat{\boldsymbol{\imath}} + \frac{\partial \Phi}{\partial y}\hat{\boldsymbol{\jmath}}\right) \bullet \left(\frac{\partial \Psi}{\partial x}\hat{\boldsymbol{\imath}} + \frac{\partial \Psi}{\partial y}\hat{\boldsymbol{\jmath}}\right) = \underline{\frac{\partial \Phi}{\partial x}}\,\frac{\partial \Psi}{\partial x} + \underbrace{\frac{\partial \Phi}{\partial y}}\,\frac{\partial \Psi}{\partial y}$$

$$= \underline{\frac{\partial \Psi}{\partial y}}\,\frac{\partial \Psi}{\partial x} - \underbrace{\frac{\partial \Psi}{\partial x}}\,\frac{\partial \Psi}{\partial y} \equiv 0\,. \tag{15.18}$$

Let $k_1 = \Phi(x_0,y_0)$ and $k_2 = \Psi(x_0,y_0)$, so the level curves $\Phi(x,y) = k_1$ and $\Psi(x,y) = k_2$ intersect at the point (x_0,y_0). [We've dropped the set notation here for brevity.] By (15.18), those two level curves intersect orthogonally, as long as $\nabla\Phi(x_0,y_0) \neq \mathbf{0}$.

The two *families* of level "curves"

$$\mathcal{F}_1 = \{\Phi(x,y) = k_1 : k_1 \text{ in } \mathbb{R}\}, \quad \mathcal{F}_2 = \{\Psi(x,y) = k_2 : k_2 \text{ in } \mathbb{R}\}$$

are **orthogonal families**, that is, at all points where a level curve $\Phi(x,y) = k_1$ from family $\mathcal{F}_1$ intersects a level curve $\Psi(x,y) = k_2$ from family $\mathcal{F}_2$, these two level curves intersect orthogonally.

Example 15.14

Fluids that have **potential flow** have particle trajectories $\begin{bmatrix} \dot{x}(t) \\ \dot{y}(t) \end{bmatrix}$ whose velocities are given by

$$\begin{bmatrix} \dot{x}(t) \\ \\ \dot{y}(t) \end{bmatrix} = \begin{bmatrix} \dfrac{\partial \Phi}{\partial x}(x,y) \\ \\ \dfrac{\partial \Phi}{\partial y}(x,y) \end{bmatrix}\Bigg|_{(x,y) = (x(t),y(t))} = \nabla \Phi(x,y)\Big|_{(x,y) = (x(t),y(t))}$$

for some scalar potential function $\Phi(x,y)$. An **equipotential** curve is a level set $\Phi(x,y) = k_1$, for some constant k_1. In 2D fluid flow, we denote the velocity by

$$\mathbf{v} = \dot{x}(t)\hat{\boldsymbol{\imath}} + \dot{y}(t)\hat{\boldsymbol{\jmath}} = \nabla \Phi(x,y).$$

The fluid is **incompressible** if $\nabla \bullet \mathbf{v} = 0$, in which case

$$0 = \nabla \bullet \mathbf{v} = \nabla \bullet \big(\nabla \Phi(x,y)\big) = \frac{\partial}{\partial x}\left[\frac{\partial \Phi}{\partial x}\right] + \frac{\partial}{\partial y}\left[\frac{\partial \Phi}{\partial y}\right];$$

hence,

$$\frac{\partial^2 \Phi}{\partial x^2} + \frac{\partial^2 \Phi}{\partial y^2} = 0,$$

that is, $\Phi(x,y)$ satisfies Laplace's equation. We will have more to say about solutions of Laplace's equation when we discuss "harmonic functions" in Section 15.3 and when we use complex variables to solve Laplace's equation in Chapter 16. ○

As usual, we write $z = x + iy$. Suppose further that a given potential function $\Phi(x,y)$ is the real part of a function $f(z)$ that is continuously differentiable, that is, there is a function $\Psi(x,y)$ such that

$$f(z) \triangleq \Phi(x,y) + i\Psi(x,y)$$

is *continuously* differentiable; hence, Φ, Ψ satisfy the Cauchy–Riemann equations and all of their first partial derivatives are continuous functions. As we saw, the equipotential curves $\Phi(x,y)$ are orthogonal to the **streamlines**, defined to be the level sets $\Psi(x,y) = k_2$.

Let us understand why the name "streamlines" makes sense: because (15.18) is true and $\nabla \Phi = \mathbf{v}$,

$$0 = \mathbf{v} \bullet \nabla \Psi.$$

Thus, automatically, a nonzero vector $\nabla \Psi$ is normal to the streamline, that is, to the level curve, $\Psi(x,y) = k_2$, and (15.18) says that $\nabla \Psi$ is orthogonal to the velocity vector, $\mathbf{v}$. This implies that at a point where there is a tangent vector, a nonzero velocity vector is tangent to the streamline, by the result of Problem 15.2.5.15.

15.2.4 Appendix

Here we will explore further the technical hypotheses or conclusions involved in differentiability of a function $f(z)$, where, as usual, $z = x + iy$, x, y being real. We will assume in the two examples in the succeding text that $f(z) = u(x, y) + iv(x, y)$ and that $u(x, y), v(x, y)$ have their first partial derivatives with respect to x and y existing at (x_0, y_0).

Example 15.15

In Problem 15.2.5.14* you will explain why the function

$$f(z) \triangleq \begin{Bmatrix} x^2 \sin\left(\frac{1}{x}\right) + i\,x\,y \sin\left(\frac{1}{x}\right), & \text{if } x \neq 0 \\ 0, & \text{if } x = 0 \end{Bmatrix}$$

is differentiable at $z_0 = 0$ even though

$$\frac{\partial u}{\partial x}(x, y) = \begin{Bmatrix} 2x \sin\left(\frac{1}{x}\right) - \cos\left(\frac{1}{x}\right), & \text{if } x \neq 0 \\ 0, & \text{if } x = 0 \end{Bmatrix}$$

is not continuous at $(x, y) = (0, 0)$. So, $f(z)$ being differentiable at a point $z_0 = x_0 + iy_0$ does not imply that the first partial derivatives of u, v must be continuous at (x_0, y_0).

Example 15.16

Define

$$f(z) \triangleq \begin{Bmatrix} \bar{z}^3/|z|^2, & z \neq 0 \\ 0, & z = 0 \end{Bmatrix}.$$

In Problem 15.2.5.16, you will explain why f is not differentiable at the point $z_0 = 0$ even though u, v satisfy the Cauchy–Riemann equations at $(0, 0)$. So, u, v satisfying the Cauchy–Riemann equations at a point z_0 are not enough to imply that $f = u + iv$ is differentiable at z_0.

Learn More About It

An explanation of Theorem 15.8 can be found in Section 2.1 of *Complex Variables*, 2nd ed., by Stephen D. Fisher, Dover Publications © 1999.

15.2.5 Problems

1. Evaluate the limit, if it exists:
   (a) $\lim_{z \to -1} \dfrac{3(z^2 - 1)}{z + 1}$,

* This example was created by my colleague, Professor James T. Vance, Jr.

(b) $\lim_{z\to -1} \dfrac{2(z^3+1)}{z+1}$,

(c) $\lim_{z\to -1} \dfrac{2(z^3+1)}{3(z^2-1)}$.

2. Evaluate the limit, if it exists:

(a) $\lim_{z\to -i} \dfrac{3(z^2+1)}{z+i}$,

(b) $\lim_{z\to -i} \dfrac{z^4+1}{z+i}$,

(c) $\lim_{z\to -i} \dfrac{z^4+1}{3(z^2+1)}$.

In problems 3 and 4, for the given function use results from problems 1 and 2, respectively, to find an extension that is continuous everywhere.

3. (a) $f(z) = \dfrac{3(z^2-1)}{z+1}$,

(b) $g(z) = \dfrac{2(z^3+1)}{z+1}$.

4. $f(z) = \dfrac{3(z^2+1)}{z+i}$.

5. For the given function, find the derivative where it exists:

(a) $f(z) = iz^2 + 2z - (1+i)$,

(b) $g(z) = \pi(2z-i)^2$,

(c) $h(z) = (z^2+1)(z^3+2z+4)$,

(d) $k(z) = \dfrac{z+i2}{z-i}$,

(e) $\ell(z) = \dfrac{1}{z^2 - i2z - 4}$,

(f) $m(z) = \dfrac{\pi^2}{z^2(z-1)}$.

In problems 6–12, a function $f(z)$ is given. For each, (a) where is f differentiable? (b) where is f not differentiable? and (c) find a formula for $f'(z)$, where it exists, in terms of z. As usual, we write $z = x + iy$, where x and y are real.

6. $f(z) = |z|^2$.
7. $f(z) = (-1+2x)y + i(x - x^2 + y^2)$.
8. $f(z) = (1+2x)y + i(x - x^2 + y^2)$.
9. $f(z) = (-1-2x)y + i(x - x^2 - y^2)$.
10. $f(z) = 3x^2 + y^2 + i2xy$.
11. $f(z) = z\,\mathcal{R}e(z)$.
12. $f(z) = x^2 + 3y^2 + i2xy$.

13. Suppose $f(z) = (x - iy)(2 - x^2 - y^2)$. Explain why f is differentiable only at the points on the circle $x^2 + y^2 = 1$.
14. Define

$$f(z) \triangleq \begin{Bmatrix} x^2 \sin\left(\frac{1}{x}\right) + ixy \sin\left(\frac{1}{x}\right), & \text{if } x \neq 0 \\ 0, & \text{if } x = 0 \end{Bmatrix}.$$

    (a) Explain why $f(z)$ is differentiable at $z_0 = 0$.
    (b) Explain why

$$\frac{\partial u}{\partial x}(x, y) = \begin{Bmatrix} 2x \sin\left(\frac{1}{x}\right) - \cos\left(\frac{1}{x}\right), & \text{if } x \neq 0 \\ 0, & \text{if } x = 0 \end{Bmatrix}$$

    is not continuous at $(x, y) = (0, 0)$.
    (c) Why do these results not contradict Theorem 15.7?
15. Explain why $\nabla\Psi$ both being orthogonal to a nonzero vector $\mathbf{v}$ and normal to the curve $\Psi = k_2$ in $\mathbb{R}^2$ implies that $\mathbf{v}$ is tangent to the streamline $\Psi = k_2$, at a point where there is a tangent vector to this streamline.
16. Define $f(z) \triangleq \begin{Bmatrix} \bar{z}^3/|z|^2, & \text{if } z \neq 0 \\ 0, & \text{if } z = 0 \end{Bmatrix}$.

    Explain why f is not differentiable at a point $z = 0$ even though u, v satisfy the Cauchy–Riemann equations at $(0, 0)$. So, u, v satisfying the Cauchy–Riemann equations at a point z_0 are not enough to imply that $f = u + iv$ is differentiable at z_0.

15.3 Analyticity, Harmonic Function, and Harmonic Conjugate

We saw that a function $f(z)$ may be differentiable everywhere, or everywhere except at some isolated point(s), or only on a line. In principle, the set of points at which a function is differentiable could be anything. In this section, we will study the set of points of differentiability.

A set A in the complex plane is **open** if at every point z_0 in A there is *some* open disk $\mathcal{D}_r(z_0)$ contained in A. The radius r may depend on the point z_0. For example, let A be the first quadrant, that is, $A = \{x + iy : x > 0, y > 0\}$. For any $z_0 = x_0 + iy_0$ in A, if we choose $r < \min\{x_0, y_0\}$, then $\mathcal{D}_r(z_0)$ is contained in A. This is illustrated in Figure 15.10a.

To study open sets, it helps to have a property of complex numbers.

Theorem 15.9

(Triangle inequality) For all complex numbers z_1, z_2, $|z_1 + z_2| \leq |z_1| + |z_2|$.

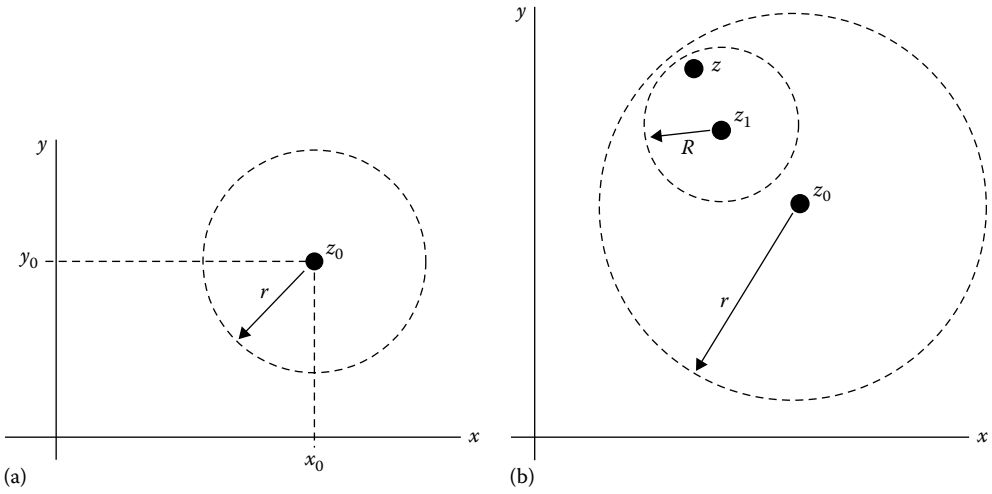


FIGURE 15.10
(a) The first quadrant is an open set. (b) An open disk is an open set.

Example 15.17

Explain why every open disk is an open set.

Method: The picture in Figure 15.10b says a thousand words: suppose z_1 is any point in an open disk $\mathcal{D}_r(z_0)$. We will explain why $\mathcal{D}_r(z_0)$ is an open set by finding an open disk $\mathcal{D}_R(z_1)$ that is contained in $\mathcal{D}_r(z_0)$.

Because z_1 is in $\mathcal{D}_r(z_0)$, $|z_1 - z_0| < r$, by definition of "open disk." Define $R = r - |z_1 - z_0|$, and choose any z in $\mathcal{D}_R(z_1)$. We will explain why z must also be in $\mathcal{D}_r(z_0)$, so we will have shown that $\mathcal{D}_R(z_1)$ is contained in $\mathcal{D}_r(z_0)$ and then we will be done.

To explain why z in $\mathcal{D}_R(z_1)$ is also in $\mathcal{D}_r(z_0)$, the triangle inequality and the definition of R allow us to calculate

$$|z - z_0| = |(z - z_1) + (z_1 - z_0)| \leq |z - z_1| + |z_1 - z_0| < R + |z_1 - z_0|$$
$$= (r - |z_1 - z_0|) + |z_1 - z_0| = r. \bigcirc$$

Definition 15.5

A function is **analytic** on an open set $\mathcal{O}$ if the function is differentiable at every point in $\mathcal{O}$.

Definition 15.6

A function is **analytic** at a point z_0 if the function is differentiable at every point in some open disk $\mathcal{D}_r(z_0)$.

Definition 15.5 is more important than Definition 15.6. In fact, a function is analytic at a point only if the function is analytic on some open disk centered at that point.

Many authors use the word "**holomorphic**" to mean the same thing as "analytic."

Example 15.18

Find where $f(z) \triangleq 2xy + i(y^2 - x^2)$ is analytic.

Method: In Example 15.11 in Section 15.2, we explained why f is analytic everywhere. So, f is analytic on the whole complex plane, $\mathbb{C}$. ◯

Example 15.19

Find where $f(z) \triangleq \dfrac{1}{z(z-i)}$ is analytic.

Method: By Theorem 15.6 in Section 15.2, f is differentiable at all z except where its denominator is zero, that is, except at $z = 0$ and $z = i$. By the result of Problem 15.3.3.24, the set $A = \{z : z \neq 0 \text{ and } z \neq i\}$ is open, so f is analytic on A. ◯

Example 15.20

Where is $f(z) \triangleq y^2 - x^2 + i(-2xy + y^2)$ analytic?

Method: In Example 15.13 in Section 15.2, we saw that f is differentiable only on the line $A \triangleq \{x + iy : y = \frac{1}{2}\}$. A contains no open disk, as illustrated in Figure 15.11: there is no point z_0 such that f is differentiable on an open disk centered at z_0 because no matter how small r is, f is not differentiable at $z_0 + \frac{i}{3}r$. So, f is analytic nowhere. ◯

Definition 15.7

A function is **entire** if the function is analytic on the whole complex plane, $\mathbb{C}$.

Theorem 15.10

Polynomials in z are entire functions.

For future reference, we state

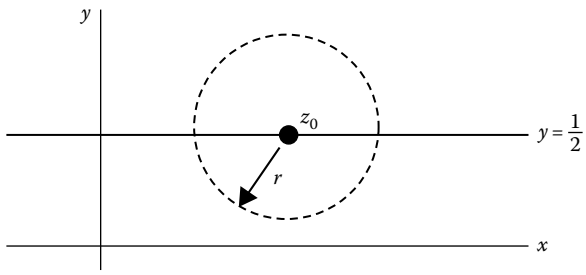


FIGURE 15.11
A function differentiable only on $y = \frac{1}{2}$.

Theorem 15.11

Suppose A is an open set in $\mathbb{C}$. If, at every point in A, $f(z) = u(x, y) + iv(x, y)$ both satisfies the Cauchy–Riemann equations and has $\frac{\partial u}{\partial x}$, $\frac{\partial u}{\partial y}$, $\frac{\partial v}{\partial x}$, $\frac{\partial v}{\partial y}$ each continuous, then f is analytic on A.

Theorem 15.12

(L'Hôpital's rule) Suppose f and g are analytic at a point z_0. If either, $f(z_0) = g(z_0) = 0$ or $\lim_{z\to z_0} |f(z)| = \lim_{z\to z_0} |g(z)| = \infty$, then

$$\lim_{z\to z_0} \frac{f(z)}{g(z)} = \lim_{z\to z_0} \frac{f'(z)}{g'(z)},$$

assuming the latter limit exists.

15.3.1 Harmonic Functions

We need further study of sets of points.

Definition 15.8

(a) A **directed line segment** is a parameterized curve $\mathcal{C} : z = z(t) = z_0 + tz_1$, where z_0 and z_1 are complex constants, t is real, and $a \le t \le b$. We say that this directed line segment connects $z(a)$ to $z(b)$, and we may notate it as $z(a) \dashrightarrow z(b)$.

(b) A **polygonal curve** connecting z_0 to z_n is either a finite sequence of directed line segments $z_0 \dashrightarrow z_1$, $z_1 \dashrightarrow z_2, \ldots, z_{n-1} \dashrightarrow z_n$ or a single directed line segment $z_0 \dashrightarrow z_n$.

Definition 15.9

(a) A set A in $\mathbb{C}$ is **connected** if for every pair of distinct points z_0, z_n in A, there is a polygonal curve $\mathcal{C}$ that satisfies all of the requirements that it lies in A, connects the point z_0 to z_n, and is **simple**, that is, does not intersect itself.

(b) A **domain** in the complex plane is a set A that is both open and connected.

The requirement that the polygonal curve $\mathcal{C}$ be simple means that no two of its directed line segments have a point in common other than the end of the k-th directed line segment being the beginning of the $k + 1$st directed line segment and, possibly, $z_0 = z_n$. Examples of domains include the sets illustrated in Figure 15.12.

An example of a set that is not a domain is the set

$$B \triangleq \{z : |z| < 1 \text{ or } |z - 2| < 1\}$$

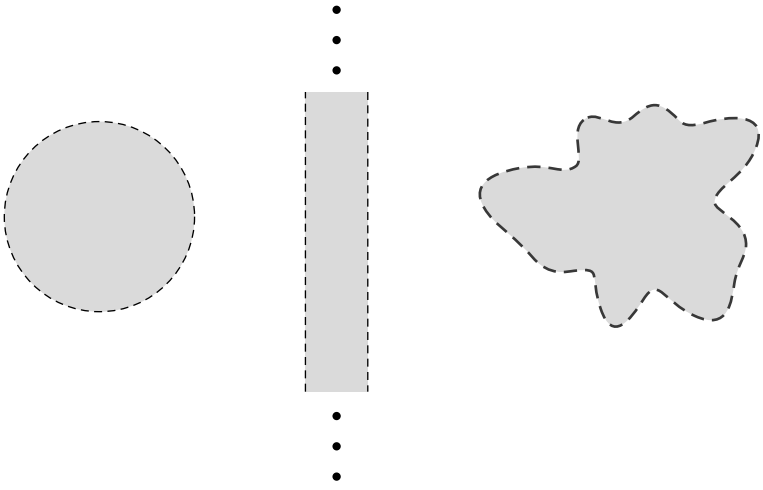


FIGURE 15.12
Three examples of domains.

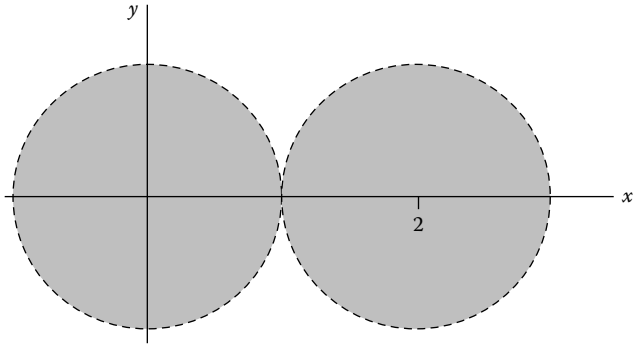


FIGURE 15.13
Set B.

shown in Figure 15.13. B is the union of two open disks that do not "communicate," that is, B is like two circular rooms with no connecting "doorway." Note that the point $z = 1$ is not in B, so there is no way to get from a point in the disk $|z| < 1$ to the disk $|z - 2| < 1$ by a polygonal curve that remains in B.

Definition 15.10

A function $\Phi(x, y)$ is **harmonic** on an open set $\mathcal{D}$ if Φ satisfies Laplace's equation,

$$\frac{\partial^2 \Phi}{\partial x^2} + \frac{\partial^2 \Phi}{\partial y^2} = 0,$$

at all points (x, y) in $\mathcal{D}$.

Clearly, harmonic functions are important in the study of partial differential equations. That is one of the reasons the next result is so valuable.

Theorem 15.13

As usual, we write $z = x + iy$, where x and y are real. If f is analytic on an open set $\mathcal{D}$, then both

$$u(x,y) \triangleq \mathcal{R}e\big(f(z)\big) \quad \text{and} \quad v(x,y) \triangleq \mathcal{I}m\big(f(z)\big)$$

are harmonic on $\mathcal{D}$.

Why? Suppose f is analytic on an open set $\mathcal{D}$ and $z = x + iy$ is in $\mathcal{D}$. Theorem 15.40 in Section 15.9 will guarantee that f' and f'' are also analytic on $\mathcal{D}$. This implies that all of the second order partial derivatives of u and v exist and are continuous at (x,y) and u and v satisfy the Cauchy–Riemann equations at z. We calculate that

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial}{\partial x}\left[\frac{\partial u}{\partial x}\right] \equiv \frac{\partial}{\partial x}\left[\frac{\partial v}{\partial y}\right] \equiv \frac{\partial}{\partial y}\left[\frac{\partial v}{\partial x}\right] \equiv \frac{\partial}{\partial y}\left[-\frac{\partial u}{\partial y}\right] = -\frac{\partial^2 u}{\partial y^2};$$

hence, u is harmonic at z. A similar calculation explains why v is harmonic at z. □

Example 15.21

Explain why $u(x,y) = 3x^2y - y^3$ is harmonic on $\mathbb{C}$.

Method: $u(x,y) = \mathcal{R}e\big(f(z)\big)$ where $f(z) \triangleq -iz^3$ is an entire function, so $u(x,y)$ is harmonic on $\mathbb{C}$. ◯

Alternatively, we could substitute $u(x,y)$ into Laplace's equation and verify that it is satisfied at all (x,y) in $\mathbb{R}^2$.

15.3.2 Harmonic Conjugate

Definition 15.11

Given a function $u(x,y)$ whose first partial derivatives are continuous on a domain $\mathcal{D}$, we say v is a **harmonic conjugate** of u if $f(z) \triangleq u(x,y) + iv(x,y)$ is analytic on $\mathcal{D}$.

Given a function u, part of the procedure for *finding* a harmonic conjugate, v, uses the same technique for finding a scalar potential function as we used in Section 6.4. In order for v to be a harmonic conjugate for the given u, the Cauchy–Riemann equations (15.14) in Section 15.2, for $f = u + iv$ demand that

$$\left\{ \begin{array}{l} \dfrac{\partial v}{\partial y} \equiv \dfrac{\partial u}{\partial x} \\ \\ \dfrac{\partial v}{\partial x} \equiv -\dfrac{\partial u}{\partial y} \end{array} \right\} \text{ on } \mathcal{D}.$$

Because u is given, we are effectively given

$$\nabla v = \frac{\partial v}{\partial x}\,\hat{\boldsymbol{\imath}} + \frac{\partial v}{\partial y}\,\hat{\boldsymbol{\jmath}} = -\frac{\partial u}{\partial y}\,\hat{\boldsymbol{\imath}} + \frac{\partial u}{\partial x}\,\hat{\boldsymbol{\jmath}}.$$

Given vector field ∇v, we are asked to find a "scalar potential function," v. Such a function v exists only if the given function u is harmonic. Why? By Clairaut's theorem, such a function $v(x, y)$ exists only if ∇v satisfies the exactness criterion, that is,

$$\frac{\partial}{\partial x}\left[\frac{\partial v}{\partial y}\right] \equiv \frac{\partial}{\partial y}\left[\frac{\partial v}{\partial x}\right];$$

hence,

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial}{\partial x}\left[\frac{\partial u}{\partial x}\right] \equiv \frac{\partial}{\partial x}\left[\frac{\partial v}{\partial y}\right] \equiv \frac{\partial}{\partial y}\left[\frac{\partial v}{\partial x}\right] \equiv \frac{\partial}{\partial y}\left[-\frac{\partial u}{\partial y}\right] = -\frac{\partial^2 u}{\partial y^2}. \tag{15.19}$$

So, for u to have a harmonic conjugate, it is *necessary* that u be harmonic on $\mathcal{D}$.

The terminology makes sense! Only a harmonic function can have a harmonic conjugate.

Example 15.22

Find a harmonic conjugate for $u(x, y) = 2 + x + x^3 - 3xy^2$, if it is possible to do so.

Method: Since the domain $\mathcal{D}$ is not specified in the problem, we should assume that it is the whole complex plane or, perhaps, the largest set in the complex plane on which u and its first and second partial derivatives are continuous. Later, we may have to specify the domain $\mathcal{D}$.

The given function u must be a real-valued function, as should be the desired harmonic conjugate function v.

We want v to satisfy the Cauchy–Riemann equations; hence,

$$\frac{\partial v}{\partial y} = \frac{\partial u}{\partial x} = \frac{\partial}{\partial x}\big[\,2 + x + x^3 - 3xy^2\,\big] = 1 + 3x^2 - 3y^2,$$

so

$$v = \int (1 + 3x^2 - 3y^2)\partial y = y + 3x^2 y - y^3 + g(x),$$

where $g(x)$ is an arbitrary function of x alone. Substitute v into the other Cauchy–Riemann equation to get

$$\begin{aligned} -6xy = \frac{\partial}{\partial y}\big[\,2 + x + x^2 - 3xy^2\,\big] = \frac{\partial u}{\partial y} &= -\frac{\partial v}{\partial x} = -\frac{\partial}{\partial x}\big[\,y + 3x^2 y - y^3 + g(x)\,\big] \\ &= -6xy - g'(x); \end{aligned}$$

hence, $g'(x) \equiv 0$. So, $v(x,y) = y + 3x^2y - y^3 + c$ is a harmonic conjugate of u, for any real constant c. ◯

Theorem 15.14

If $u = u(x,y)$ is harmonic on a domain $\mathcal{D}$ and z_0 is any point in $\mathcal{D}$, then $u(x,y)$ has a harmonic conjugate on some open ball $B_\delta(z_0)$ contained in $\mathcal{D}$.

Because Theorem 15.14 guarantees only that there is a harmonic conjugate near z_0, we may say that this is a local harmonic conjugate. For example, if $z_0 \neq 0$ is not on the negative *x-axis*, then we will see in Section 15.4 that the function $u(x,y) \triangleq \frac{1}{2}\ln(x^2+y^2)$ has a local harmonic conjugate $\arctan(y/x)$ corresponding to $f(z) = u(x,y) + i\,v(x,y) = \frac{1}{2}\ln(x^2+y^2) + \arctan(y/x) \triangleq Log(z)$.

Theorem 15.37 in Section 15.8, an improved version of Theorem 15.14, relies on concepts we will define in Section 15.8.

Example 15.23

Find a streamline function for potential flow for the velocity vector field:

$$\mathbf{v}(x,y) = \frac{dx}{dt}\,\hat{\boldsymbol{\imath}} + \frac{dy}{dt}\,\hat{\boldsymbol{\jmath}} = \left(1 + \frac{y^2 - x^2}{\left(x^2+y^2\right)^2}\right)\hat{\boldsymbol{\imath}} - \frac{2xy}{\left(x^2+y^2\right)^2}\,\hat{\boldsymbol{\jmath}}.$$

Method: This is potential flow for potential function $\Phi(x,y)$ if the velocity is

$$\mathbf{v}(x,y) = \frac{\partial\Phi}{\partial x}\,\hat{\boldsymbol{\imath}} + \frac{\partial\Phi}{\partial y}\,\hat{\boldsymbol{\jmath}}.$$

As discussed in Example 15.14 in Section 15.2, the streamline function $\Psi(x,y)$ is to be chosen so that

$$f(z) = \Phi(x,y) + i\Psi(x,y)$$

is analytic on some domain. As we saw in the present section, Ψ is a harmonic conjugate of Φ. So, by the Cauchy–Riemann equations, we want Ψ to satisfy

$$\frac{\partial\Psi}{\partial y} \equiv \frac{\partial\Phi}{\partial x} = 1 + \frac{y^2 - x^2}{\left(x^2+y^2\right)^2} \tag{15.20}$$

and

$$\frac{\partial\Psi}{\partial x} \equiv -\frac{\partial\Phi}{\partial y} = \frac{2xy}{\left(x^2+y^2\right)^2} \tag{15.21}$$

on some domain $\mathcal{D}$. From (15.21), using the substitution $w = x^2 + y^2$ and calculating $\partial w = 2x\,\partial x$, we integrate to get

$$\begin{aligned}\Psi &= \int \frac{2xy}{\left(x^2+y^2\right)^2}\,\partial x = \int \frac{2xy}{w^2}\cdot\frac{\partial w}{2x} = y\int w^{-2}\,\partial w = -\frac{y}{w} + g(y)\\ &= -\frac{y}{x^2+y^2} + g(y),\end{aligned} \tag{15.22}$$

where $g(y)$ is an arbitrary function of y alone. [We were able to pull out the "y" factor from the integral with respect to w because the original integral was with respect to x; y and x are independent variables.]

Substitute this into (15.20), the first Cauchy–Riemann equation, to get

$$1 + \frac{y^2 - x^2}{\left(x^2 + y^2\right)^2} = \frac{\partial \Phi}{\partial x} = \frac{\partial \Psi}{\partial y} = \frac{\partial}{\partial y}\left[- \frac{y}{x^2 + y^2} + g(y) \right]$$
$$= - \frac{1}{x^2 + y^2} + \frac{2y^2}{\left(x^2 + y^2\right)^2} + g'(y).$$

At first glance, it looks hopeless to find $g'(y)$ as a function of y alone, but patient algebraic manipulation gives us

$$g'(y) = 1 + \frac{y^2 - x^2}{\left(x^2 + y^2\right)^2} - \left(- \frac{1}{x^2 + y^2} + \frac{2y^2}{\left(x^2 + y^2\right)^2} \right)$$
$$= 1 + \frac{y^2 - x^2}{\left(x^2 + y^2\right)^2} - \left(\frac{-x^2 + y^2}{\left(x^2 + y^2\right)^2} \right) = 1,$$

so $g(y) = y + c$, where c is an arbitrary real constant. For convenience, choose $c = 0$ to get the streamline function:

$$\Psi(x, y) = - \frac{y}{x^2 + y^2} + y.$$

All of the earlier calculations made sense for $x^2 + y^2 > 0$, so we may take as the domain the **punctured plane** $\mathcal{D} = \{z : z \neq 0\}$. ◯

By the way, if we take instead the domain to be $\mathcal{D} = \{z : |z| \geq 1\}$, then the streamlines, that is, the level sets $\Psi(x, y) = k$ for real constants k, give the 2D **Stokes potential** fluid flow around a cylinder.

15.3.3 Problems

1. Suppose, as usual, that $z = x + iy$ where x and y are real. Where, if anywhere, does $f(z) \triangleq x - y + i(x + y)$ satisfy the Cauchy–Riemann equations? Find $f'(z)$ where it exists. Is f an entire function?
2. Define $f(z) = \frac{1}{2}\left(z + \frac{3}{z}\right)$. Use Definition 15.4 in Section 15.2 and calculations such as those done in Example 15.9 in Section 15.2 to find $f'(z)$, where it exists. Where is f analytic?
3. Define $f(z) = 1/(z - i)$. Use Definition 15.4 in Section 15.2 and calculations such as those done in Example 15.9 in Section 15.2 to find $f'(z)$, where it exists. Where is f analytic?

In problems 4–6, a function $u(x, y)$ is given. Find a harmonic conjugate.

4. $u(x, y) = (-1 + 2x)y$
5. $u(x, y) = x - 2xy$
6. $u(x, y) = 8x^3y - 8xy^3 + 1$

In problems 7–10, a function $u(x, y)$ is given. Find an entire function $f(z)$ such that $\mathcal{R}e(f(z)) = u(x, y)$. Express $f(z)$ and $f'(z)$ as functions of z. (Your final conclusions should be formulas that do not involve x or y.)

7. $u(x, y) = x - 2xy$
8. $u(x, y) = 4xy^3 - 4x^3y$
9. $u(x, y) = y^2 + x^3 - x^2 - 3xy^2 + 2$
10. $u(x, y) = 3xy^2 - y^3 + 2xy - y$
11. For the function $u(x, y) = y/x^2 + y^2$, find a domain in which it is harmonic and find a harmonic conjugate for u in that domain.
12. Derive the Cauchy–Riemann equations in polar coordinates: given $f(z) = f(x + iy) = u(x, y) + iv(x, y)$, define $U(r, \theta) \triangleq u(r\cos\theta, r\sin\theta)$, $V(r, \theta) \triangleq v(r\cos\theta, r\sin\theta)$. If $f(z)$ is differentiable at a point, then

$$\frac{\partial U}{\partial r} = \frac{1}{r}\frac{\partial V}{\partial \theta} \quad \text{and} \quad \frac{\partial V}{\partial r} = -\frac{1}{r}\frac{\partial U}{\partial \theta}$$

there. Why? Use the chain rule, as in Chapter 6, $x = r\cos\theta$, $y = r\sin\theta$, and the Cauchy–Riemann equations in x, y to get

$$\begin{aligned}\frac{\partial U}{\partial r} = \frac{\partial u}{\partial x}\cdot\frac{\partial x}{\partial r} + \frac{\partial u}{\partial y}\cdot\frac{\partial y}{\partial r} &= \cos\theta\cdot\frac{\partial u}{\partial x} + \sin\theta\cdot\frac{\partial u}{\partial y} \\ &= \cos\theta\cdot\frac{\partial v}{\partial y} - \sin\theta\cdot\frac{\partial v}{\partial x}\end{aligned}$$

and similar chain rule calculations of

$$\frac{1}{r}\frac{\partial V}{\partial \theta}, \quad \frac{\partial V}{\partial r}, \quad -\frac{1}{r}\frac{\partial U}{\partial \theta}.$$

13. Use the Cauchy–Riemann equations in polar coordinates to find a harmonic conjugate for $U(r, \theta) = v_\infty\left(r + \dfrac{a^2}{r}\right)\cos\theta$, where a and v_∞ are unspecified positive constants. This function is used in a model of potential flow past a circular cylinder.
14. Use Cauchy–Riemann equations in polar coordinates to find a harmonic conjugate for $U(r, \theta) = v_\infty\left(r + \dfrac{a^2}{r}\right)\cos\theta - \kappa\theta$, where a, v_∞, and κ are unspecified positive constants. This function is used in a model of potential flow with circulation κ past a circular cylinder.
15. For the function $u(x, y) = \frac{2xy}{(x^2+y^2)^2}$, use the Cauchy–Riemann equations in polar coordinates to find a domain in which it is harmonic and to find a harmonic conjugate for u in that domain.

16. Suppose n is a positive integer. Explain why $r^n \cos n\theta$ has harmonic conjugate $r^n \sin n\theta$.
17. Suppose n is a positive integer. Find a harmonic conjugate for $r^n \sin n\theta$.
18. Find an analytic function $f(z)$ for which $\mathcal{R}e(f(z)) = \dfrac{y}{(x+2)^2 + y^2}$.
19. Suppose $f(z) = e^{\alpha z}$, where α is a complex constant, specifically $\alpha = a + ib$, where a, b are real:
    (a) Find a formula for $f(z)$ in terms of *real* exponential functions of x and y and *real* sine and cosine trigonometric functions of x and y.
    (b) Where is $f(z)$ analytic?
20. Suppose $g(z) = u(x,y) + iv(x,y)$ is analytic on a domain $\mathcal{D}$ and define $U(x,y) = \big(u(x,y)\big)^2 - \big(v(x,y)\big)^2$. Find an analytic function $f(z)$ for which $\mathcal{R}e(f(z)) = U(x,y)$.
21. If $u(x,y)$ has harmonic conjugate $v(x,y)$, why does $v(x,y)$ have harmonic conjugate $-u(x,y)$?
22. Let $f(z) = z^3$. Use this to solve the system of ODEs
$$\begin{bmatrix} \dot{x}(t) \\ \\ \dot{y}(t) \end{bmatrix} = \begin{bmatrix} 3(x^2 - y^2) \\ \\ -6xy \end{bmatrix}.$$
23. Find solution curves in the xy-plane for the system of ODEs
$$\begin{bmatrix} \dot{x}(t) \\ \\ \dot{y}(t) \end{bmatrix} = \begin{bmatrix} 2y \\ \\ 2x \end{bmatrix}.$$
24. Suppose $\mathcal{O}$ is an open subset of the complex plane and $z_1, \ldots, z_n$ are complex numbers in $\mathcal{O}$. Explain why
$$\{z : z \text{ is in } \mathcal{O} \text{ but } z \neq z_k \text{ for } k = 1, \ldots, n\}$$
is open.

15.4 Elementary Functions

All of the commonly used functions of a real variable, such as the exponential, logarithmic, trigonometric, and power functions, have generalizations to functions of a single complex variable. Many of their properties are the same. But there are some nonroutine aspects of the generalization to functions of a complex variable, and a new type of function, $Arg(z)$, plays a crucial role.

15.4.1 Arg(z)

Given a complex number z, there are infinitely many values of the real number θ for which $z = |z|e^{i\theta}$. This defines a new notation:

$$arg(z) \triangleq \{\theta : z = |z|e^{i\theta} \quad \text{and } \theta \text{ is real}\}, \quad \text{for} \quad z \neq 0. \tag{15.23}$$

For example, $arg(i5) = \{\frac{\pi}{2} + 2\pi k : k \text{ is any integer}\}$ because $i5 = 5e^{i\pi/2}$. For any nonzero complex number z, $arg(z)$ is a *set* of infinitely many real numbers.

It will be very useful to select just one of the real numbers out of the infinite set $arg(z)$:

$$Arg(z) \triangleq \text{ the unique } \theta \text{ in } arg(z) \text{ that satisfies } -\pi < \theta \leq \pi, \text{ for } z \neq 0. \tag{15.24}$$

Example 15.24

Find the exact value of $Arg\left(\dfrac{-\sqrt{3}+i}{1-i}\right)$.

Method: As in Example 15.1(d) in Section 15.1, $\dfrac{-\sqrt{3}+i}{1-i} = \dfrac{2e^{i5\pi/6}}{\sqrt{2}e^{-i\pi/4}} = \sqrt{2}\,e^{i\left(\frac{5\pi}{6}-(-i\,\frac{\pi}{4})\right)} = \sqrt{2}\,e^{i13\pi/12}$. So,

$$arg(z) = \left\{\frac{13\pi}{12} + 2\pi k : k \text{ is any integer}\right\}.$$

Figure 15.14 shows the set $arg(z)$ written in an explicit form.

$Arg\left(\dfrac{-\sqrt{3}+i}{1-i}\right) = -\dfrac{11\pi}{12}$, using $k=-1$, is the unique θ in $arg(z)$ satisfying $-\pi < \theta \leq \pi$. ◯

As illustrated in Figure 15.15,

$$Arg(z) = \begin{Bmatrix} \cos^{-1}\left(\dfrac{x}{\sqrt{x^2+y^2}}\right), & y \geq 0 \\ -\cos^{-1}\left(\dfrac{x}{\sqrt{x^2+y^2}}\right), & y < 0 \end{Bmatrix}. \tag{15.25}$$

For example, $z = -\sqrt{3} - i = x + iy$ has

$$Arg(z) = -\cos^{-1}\left(\frac{-\sqrt{3}}{(-\sqrt{3})^2+(-1)^2}\right) = -\cos^{-1}\left(\frac{-\sqrt{3}}{2}\right) = -\frac{5\pi}{6}.$$

$$arg(z) = \left\{ \underset{\uparrow}{\ldots}, \quad \underset{k\leq -2}{-\frac{11\pi}{12}}, \quad \underset{k=-1}{\frac{13\pi}{12}}, \quad \underset{k=0}{\frac{37\pi}{12}}, \quad \underset{k=1}{\ldots}, \quad \underset{k\geq 2}{\ldots} \right\}.$$

FIGURE 15.14
Explicit set notation for $arg(z)$.

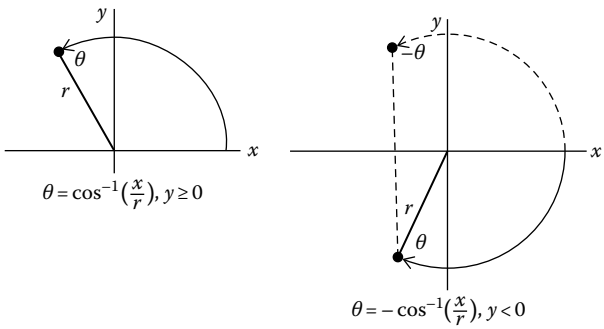


FIGURE 15.15
Arg in terms of inverse trigonometric functions.

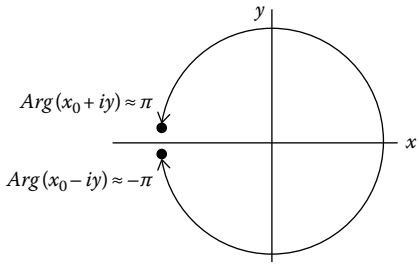


FIGURE 15.16
Discontinuity of *Arg*.

The function $Arg(z)$ is continuous at all z_0 *except* those z_0 that lie on the nonpositive real axis. Why? Regarding the lack of continuity, suppose first that $z_0 = x_0 < 0$. Then for small positive y, $Arg(x_0 + iy) \approx \pi$ and $Arg(x_0 - iy) \approx -\pi$, as illustrated in Figure 15.16. So,

$$\lim_{y \to 0^+} (Arg(x_0 + iy) = \pi \quad \text{and} \quad \lim_{y \to 0^-} (Arg(x_0 + iy) = -\pi,$$

which explains why $Arg(z)$ is discontinuous at all z of the form $z = x_0$ with $x_0 < 0$.

The $Arg(z)$ function is not even defined at $z = 0$, so $Arg(z)$ is not continuous at 0. So, $Arg(z)$ is discontinuous on the nonpositive real axis.

Formulas (15.25) and continuity of the function $\cos^{-1}$ explain why $Arg(z)$ is continuous off the positive x-axis. In Problem 15.4.6.26, you will establish the continuity of $Arg(z)$ at points on the positive real axis.

15.4.2 Exp(*z*)

Definition 15.12

Given $z = x + iy$, where x and y are real, we can define

$$exp(z) \triangleq e^x (\cos y + i \sin y). \tag{15.26}$$

Another notation for this is e^z.

Theorem 15.15

(a) e^z satisfies the laws of exponents, that is,

$$(i)\ e^{z_1+z_2} = e^{z_1}e^{z_2} \quad \text{and} \quad (ii)\ e^{z_1-z_2} = e^{z_1}/e^{z_2} \quad \text{and} \quad (iii)\ \left(e^z\right)^n = e^{nz} \tag{15.27}$$

for all complex numbers z_1, z_2, z and integers n, and

(b) e^z is an entire function, and $d/dz\left[e^z\right] = e^z$.

Why? Using trigonometric identities, you will explain (a) in Problem 15.4.6.24. For (b), write $u + iv = e^z \triangleq e^x \cos y + ie^x \sin y$, that is, $u(x, y) = e^x \cos y$ and $v(x, y) = e^x \sin y$. We can verify that the Cauchy–Riemann equations are true at all (x, y):

$$\left\{\begin{array}{l} \dfrac{\partial u}{\partial x} - \dfrac{\partial v}{\partial y} = e^x \cos y - e^x \cos y = 0 \\ \dfrac{\partial u}{\partial y} + \dfrac{\partial v}{\partial x} = -e^x \sin y + e^x \sin y = 0 \end{array}\right\}.$$

The functions $\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial v}{\partial x}, \frac{\partial v}{\partial y}$ exist and are continuous everywhere, so the technical hypotheses of Theorem 15.8 in Section 15.2 are satisfied. This establishes why e^z is differentiable everywhere and thus entire. Further, by formula (15.15) in Section 15.2,

$$\frac{d}{dz}\left[e^z\right] = \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x} = e^x \cos y + ie^x \sin y = e^z. \ \square$$

15.4.3 Log(*z*)

Given any $z \neq 0$, the set $\{w : e^w = z\}$ contains infinitely many complex numbers because $\cos y$ and $\sin y$ are periodic functions.

Example 15.25

Find the set of all exact solutions of the equation $e^w = 1 - i$.

Method: $1 - i = \sqrt{2}\, e^{i\left(-\frac{\pi}{4}\right)} = e^{\ln\sqrt{2}}\, e^{i\left(-\frac{\pi}{4}\right)} = e^{\ln\sqrt{2} + i\left(-\frac{\pi}{4}\right)}$, where $\ln x$ is the usual Calculus I function defined for positive real numbers x. The equation to solve is

$$e^w = 1 - i = e^{\ln\sqrt{2} + i\left(-\frac{\pi}{4}\right)}.$$

So, there is at least one exact solution: $w = \ln\sqrt{2} + i\left(-\frac{\pi}{4}\right)$. By (15.6) in Section 15.1, that is, $e^{i\theta - i2\pi k} = e^{i\theta}$ for all integers k, there are infinitely many solutions for w. The set of all solutions is

$$\left\{\frac{1}{2}\ln 2 - i\frac{\pi}{4} + i2\pi k : k \text{ is any integer}\right\}. \ \bigcirc$$

$$\left\{ \ldots, \underset{\substack{\uparrow \\ k \le -2}}{\frac{1}{2}\ln(2) - i\frac{9\pi}{4}}, \underset{\substack{\uparrow \\ k=-1}}{\frac{1}{2}\ln(2) - i\frac{\pi}{4}}, \underset{\substack{\uparrow \\ k=0}}{\frac{1}{2}\ln(2) + i\frac{7\pi}{4}}, \ldots \right\}.$$

FIGURE 15.17
Example 15.25.

It may help to write such a set in a more explicit form (Figure 15.17).

Definition 15.13

We define

$$log(z) \triangleq \left\{ \ln|z| + iArg(z) + i2\pi k : k \text{ is any integer} \right\}, \text{ for } z \neq 0. \tag{15.28}$$

The solution set in Example 15.25 was $log(1 - i)$.

When $z = 0$, there is no solution w for the equation $e^w = 0$, as you will see in Problem 15.4.6.19. So, $log(0)$ is undefined.

When it *is* defined, $log(z)$ is an infinite set of values w, so it would seem that we have not successfully generalized the familiar concept of logarithm of a positive real numbers, which produces a single value. In algebra and calculus of functions of a real variable, $\ln(x)$ is truly a function for $x > 0$. In fact, in this book, the natural logarithm function, ln, will allow only inputs that are positive, real numbers.

But, we have already faced a similar issue: $arg(z)$ is an infinite set but we defined $Arg(z)$, which is truly a function in that it gives a single output for every nonzero input z. In fact, this enables us to do the same for logarithms!

Definition 15.14

We define

$$Log(z) \triangleq \ln|z| + iArg(z), \quad \text{for} \quad z \neq 0. \tag{15.29}$$

So, $Log(z)$ chooses from $log(z)$ the unique element w that satisfies $-\pi < \mathcal{I}m(w) \le \pi$, as you will explain in Problem 15.4.6.22. Note that

$$log(z) = \left\{ Log(z) + i2\pi k : k \text{ is any integer} \right\}. \tag{15.30}$$

Note also that if x is a real, positive number, then $Log(x) = \ln x$, as you will explain in Problem 15.4.6.23.

Example 15.26

Find the exact values of $Log\left(-\sqrt{3}+i\right)$ and $e^{Log\left(-\sqrt{3}+i\right)}$.

Method: $Log\left(-\sqrt{3}+i\right) = \ln\left(|-\sqrt{3}+i|\right) + iArg\left(-\sqrt{3}+i\right) = \ln 2 + i\,\frac{5\pi}{6}$. Because $w = Log\left(-\sqrt{3}+i\right)$ is one of the infinitely many complex numbers that satisfy $e^w = -\sqrt{3}+i$, it follows with no effort (other than thinking!) that $e^{Log\left(-\sqrt{3}+i\right)} = -\sqrt{3}+i$. ○

Theorem 15.16

$e^{Log(z)} = z$ for all $z \neq 0$.

Theorem 15.17

$d/dz\left[Log(z)\right] = 1/z$ for all z in $\mathcal{D} \triangleq \{z : \mathcal{R}e(z) > 0 \text{ or } \mathcal{I}m(z) \neq 0\}$.

Note that $\mathcal{D}$ consists of all z in $\mathbb{C}$ *except* z on the nonpositive real axis, so we can also notate $\mathcal{D} \triangleq \{z : Arg(z) \neq \pi\}$. The set $\mathcal{D}$ is sometimes called a "cut" plane and is illustrated in Figure 15.18.

Why is Theorem 15.17 true? Choose any $z = x + iy$ where x and y are real and either $x > 0$ or $y \neq 0$. (By the way, the "or" is the "inclusive or," that is, the case where both $x > 0$ and $y \neq 0$ are included.) Denote

$$u + iv \triangleq Log(z) = \ln|z| + iArg(z) = \ln(\sqrt{x^2+y^2}) + iArg(z) = \frac{1}{2}\ln(x^2+y^2) + iArg(z).$$

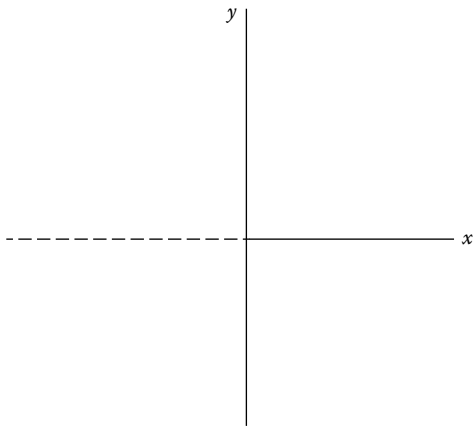


FIGURE 15.18
Cut plane.

We have

$$\frac{\partial u}{\partial x} = \frac{\partial}{\partial x}\left[\frac{1}{2}\ln(x^2+y^2)\right] = \frac{x}{x^2+y^2} \quad \text{and} \quad \frac{\partial u}{\partial y} = \frac{y}{x^2+y^2},$$

so both $\frac{\partial u}{\partial x}$ and $\frac{\partial u}{\partial y}$ are defined and continuous everywhere except at $(x,y) = (0,0)$. We will see that the tricky part of the justification is in studying the derivatives of $v(x,y) \triangleq Arg(z)$. To do that, we will use the formulas for $Arg(z)$ in (15.25).

Case (1): If (x,y) lies in the first or second quadrant or lies on the positive y-axis, that is, has $y > 0$, then $v(x,y) = Arg(z) = \cos^{-1}\left(\frac{x}{\sqrt{x^2+y^2}}\right)$, hence, $\cos v = \frac{x}{\sqrt{x^2+y^2}}$. Using implicit differentiation, we calculate that

$$-\sin v \cdot \frac{\partial v}{\partial x} = \frac{\partial}{\partial x}\big[\cos v\big] = \frac{\partial}{\partial x}\left[\frac{x}{\sqrt{x^2+y^2}}\right] = \frac{1}{\sqrt{x^2+y^2}} - \frac{x^2}{\big(x^2+y^2\big)^{3/2}} = \frac{y^2}{\big(x^2+y^2\big)^{3/2}}. \tag{15.31}$$

Because $y > 0$, hence $0 < v < \pi$,

$$\sin v = +\sqrt{1-\cos^2 v} = \sqrt{1-\frac{x^2}{x^2+y^2}} = \frac{\sqrt{y^2}}{\sqrt{x^2+y^2}} = \frac{|y|}{\sqrt{x^2+y^2}} = \frac{y}{\sqrt{x^2+y^2}}. \tag{15.32}$$

Substituting (15.32) into (15.31), and noting that $\sqrt{y^2} = |y| = y$ for $y > 0$, we get

$$\begin{aligned}\frac{\partial v}{\partial x} &= -\frac{1}{\sin v}\cdot\frac{y^2}{\big(x^2+y^2\big)^{3/2}} = -\left(\frac{y}{\sqrt{x^2+y^2}}\right)^{-1}\cdot\frac{y^2}{\big(x^2+y^2\big)^{3/2}} = -\frac{\sqrt{x^2+y^2}}{y}\cdot\frac{y^2}{\big(x^2+y^2\big)^{3/2}} \\ &= -\frac{y}{x^2+y^2} = -\frac{\partial u}{\partial y}.\end{aligned}$$

This shows that the second of the Cauchy–Riemann equations is satisfied.

Similarly,

$$\begin{aligned}\frac{\partial v}{\partial y} &= -\frac{1}{\sin v}\cdot\frac{\partial}{\partial y}\big[\cos v\big] = -\frac{1}{\sin v}\cdot\frac{\partial}{\partial y}\left[\frac{x}{\sqrt{x^2+y^2}}\right] = -\frac{1}{\sin v}\cdot\left(-\frac{xy}{\big(x^2+y^2\big)^{3/2}}\right) \\ &= -\frac{\sqrt{x^2+y^2}}{y}\cdot\left(-\frac{xy}{\big(x^2+y^2\big)^{3/2}}\right) = \frac{x}{x^2+y^2} = \frac{\partial u}{\partial x},\end{aligned}$$

which confirms the first of the Cauchy–Riemann equations. Thus, if (x,y) has $y > 0$, then the Cauchy–Riemann equations,

$$\frac{\partial v}{\partial x} = \cdots = -\frac{y}{x^2+y^2} = -\frac{\partial u}{\partial y} \quad \text{and} \quad \frac{\partial v}{\partial y} = \cdots = \frac{x}{x^2+y^2} = \frac{\partial u}{\partial x}, \tag{15.33}$$

are satisfied. Because all of the first partial derivatives $\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial v}{\partial x}, \frac{\partial v}{\partial y}$ are defined and continuous everywhere except at $(x,y) = (0,0)$, Theorem 15.8 in Section 15.2 allows us to conclude that $Log(z)$ is differentiable at such a point $z = x + iy$.

Case (2): If (x,y) lies in the third or fourth quadrant or lies on the negative y-axis, that is, has $y < 0$, then the calculations concerning derivatives of $\cos v$ work just as well for derivatives of $-\cos v$. Using the same reasoning using Theorem 15.8 in Section 15.2 as in Case (1), we conclude that $Log(z)$ is differentiable at such a point $z = x + iy$.

Case (3): If (x,y) lies on the positive x-axis, that is, has $x > 0$ and $y = 0$, you will calculate in Problem 15.4.6.25 that again (15.33) holds. By the same reasoning, using Theorem 15.8 in Section 15.2 as in Case (1), we conclude that $Log(z)$ is differentiable at such a point $z = x + iy$.

In all of these cases, formula (15.15) in Section 15.2 gives the formula for the derivative of $Log(z)$:

$$\frac{d}{dz}\big[Log(z)\big] = \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x} = \frac{x}{x^2+y^2} + i\left(-\frac{y}{x^2+y^2}\right) = \frac{x-iy}{x^2+y^2} = \frac{\bar{z}}{|z|^2} = \frac{\bar{z}}{z\bar{z}} = \frac{1}{z}.$$

This generalizes the Calculus I fact that $(\ln x)' = 1/x$ for real numbers $x > 0$. □

15.4.4 Branches of Logarithms

The function $Log(z)$ gives the unique value w in $log(z)$ that satisfies $-\pi < \mathcal{I}m(w) \leq \pi$. For any real number σ, we can define a different "branch" of $log(z)$ by

$$Log_\sigma(z) = w \text{ is the unique } w \text{ in } log(z) \text{ satisfying } -\pi + \sigma < \mathcal{I}m(w) \leq \pi + \sigma. \tag{15.34}$$

So, $Log(z) = Log_0(z)$.

The subscript σ in $Log_\sigma(z)$ does *not* denote a base for a logarithm. This is different from the notation $\ell og_2(x) = t \Leftrightarrow x = 2^t$, for real numbers $x > 0$.

The usefulness of defining such branches $Log_\sigma(z)$ is that

$$\frac{d}{dz}\big[Log_\sigma(z)\big] = \frac{1}{z}, \text{ for all } z \neq 0 \text{ for which } \sigma + \pi \text{ is not in } arg(z).$$

For example, $Log_{-\frac{\pi}{4}}(z)$ has a cut on the ray $\theta = \frac{3\pi}{4}$, which is illustrated in Figure 15.19. $Log_{-\frac{\pi}{4}}(z)$ is differentiable on $\mathcal{D} \triangleq \{z : Arg(z) \neq \frac{3\pi}{4}\}$, which is the rotation by angle $-\frac{\pi}{4}$ of the cut plane for $Log_0(z)$.

15.4.5 Power Functions

If n is an integer, $f(z) \triangleq z^n$ is easy to define and calculate. For any positive integer n, the binomial theorem gives

$$z^n = (x+iy)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k}(iy)^k,$$

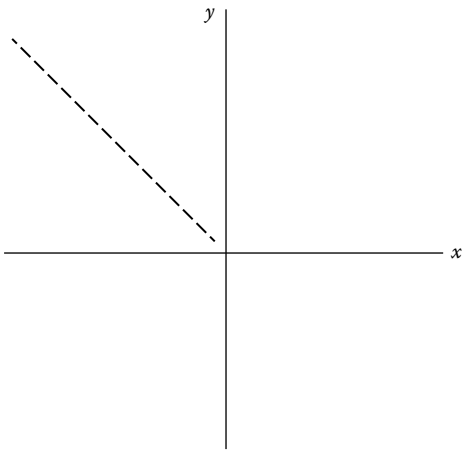


FIGURE 15.19
Another cut plane.

and for any negative integer m, we can define $z^m \triangleq \dfrac{1}{z^{-m}}$. We also want to have a definition for z^α.

First, let's recall from (15.30) that $log(z)$ is an infinite set. If α is a complex number, we can define

$$\alpha\, log(z) \triangleq \{\alpha \cdot (Log(z) + i2\pi k) : k \text{ is any integer}\},$$

and, similarly,

$$e^{\alpha log(z)} \triangleq \left\{e^{\alpha\cdot(Log(z)+i2\pi k)} : k \text{ is any integer}\right\}.$$

This set could be an infinite set, meaning that it could contain infinitely many distinct numbers.

If α is *any* complex number, we would want to define z^α in such a way that $z^\alpha = e^{log(z^\alpha)}$. This suggests defining the **power function** by

$$z^\alpha \triangleq e^{\alpha\, log(z)}, \tag{15.35}$$

which could be an infinite set.

The **principal branch**, or **principal value**, of z^α is defined to be

$$e^{\alpha\, Log(z)}$$

and will also be denoted by $z^{\backslash\alpha\backslash}$. The latter is *not* a standard notation, but we hope it will become popular.

If $\alpha = \frac{1}{n}$, we will also denote the principal value of $z^{1/n}$ by

$$\sqrt[n]{z} \triangleq e^{\frac{1}{n}Log(z)} = z^{\backslash\frac{1}{n}\backslash}.$$

Example 15.27

For each of the numbers z in the following, evaluate $z^{\backslash(1+i)\backslash}$ and z^{1+i} and express the number or set of numbers in the form $a + ib$:

$$(a)\ z = -3, \quad (b)\ z = i3, \quad (c)\ z = -\sqrt{3} + i.$$

Method: Note that

$$z^{\alpha} = e^{\alpha\, log(z)} = \left\{ e^{\alpha\,(Log(z)+i2\pi k)} : k \text{ is any integer} \right\} = \left\{ z^{\backslash\alpha\backslash} \cdot e^{\alpha(i2\pi k)} : k \text{ is any integer} \right\}. \tag{15.36}$$

(a) $Log(-3) = \ln|-3| + iArg(-3) = \ln 3 + i\pi$, so

$$\begin{aligned}(-3)^{\backslash(1+i)\backslash} &= e^{(1+i)Log(-3)} = e^{(1+i)(\ln 3+i\pi)} = e^{\ln 3}\, e^{-\pi}\, e^{i\ln 3}\, e^{i\pi} = 3 \cdot e^{-\pi} \cdot e^{i\ln 3} \cdot (-1) \\ &= -3\, e^{-\pi} \cdot e^{i\ln 3} = -3\, e^{-\pi} \cdot \big(\cos(\ln 3) + i \sin(\ln 3)\big) \\ &= -3\, e^{-\pi} \cos(\ln 3) - i3\, e^{-\pi} \sin(\ln 3).\end{aligned}$$

Using (15.36) and the earlier calculation of $(-3)^{\backslash(1+i)\backslash}$, we have

$$\begin{aligned}(-3)^{1+i} &= \left\{ (-3)^{\backslash(1+i)\backslash} \cdot e^{(1+i)i2\pi k} : k \text{ is any integer} \right\} \\ &= \left\{ (-3)^{\backslash(1+i)\backslash} \cdot e^{i2\pi k} \cdot e^{-2\pi k} : k \text{ is any integer} \right\} \\ &= \left\{ -3\, e^{-\pi} \cdot e^{i\ln 3} \cdot 1 \cdot e^{-2\pi k} : k \text{ is any integer} \right\} \\ &= \left\{ -3\, e^{-(1+2k)\pi}\, e^{i\ln 3} : k \text{ is any integer} \right\} \\ &= \left\{ -3\, e^{-(1+2k)\pi} \cos(\ln 3) - i3\, e^{-(1+2k)\pi} \sin(\ln 3) : k \text{ is any integer} \right\}.\end{aligned}$$

(b) $Log(i3) = \ln|i3| + iArg(i3) = \ln 3 + i\,\frac{\pi}{2}$, so

$$\begin{aligned}(i3)^{\backslash(1+i)\backslash} &= e^{(1+i)Log(i3)} = e^{(1+i)(\ln 3+i\,\frac{\pi}{2})} = e^{\ln 3}\, e^{-\frac{\pi}{2}}\, e^{i\ln 3}\, e^{i\pi/2} = 3 \cdot e^{-\frac{\pi}{2}} \cdot e^{i\ln 3} \cdot (i) \\ &= 3ie^{-\frac{\pi}{2}} \cdot e^{i\ln 3} = -3e^{-\frac{\pi}{2}} \sin(\ln 3) + i3e^{-\frac{\pi}{2}} \cos(\ln 3).\end{aligned}$$

Using (15.36) and the earlier calculation of $(i3)^{\backslash(1+i)\backslash}$, we have

$$\begin{aligned}(i3)^{1+i} &= \left\{ (i3)^{\backslash(1+i)\backslash} \cdot e^{(1+i)i2\pi k} : k \text{ is any integer} \right\} \\ &= \left\{ (i3)^{\backslash(1+i)\backslash} \cdot e^{i2\pi k} \cdot e^{-2\pi k} : k \text{ is any integer} \right\} \\ &= \left\{ i3\, e^{-\frac{\pi}{2}} \cdot e^{i\ln 3} \cdot e^{-2\pi k} : k \text{ is any integer} \right\} \\ &= \left\{ i3\, e^{-(\frac{1}{2}+2k)\pi}\, e^{i\ln 3} : k \text{ is any integer} \right\} \\ &= \left\{ -3\, e^{-(\frac{1}{2}+2k)\pi} \sin(\ln 3) + i3\, e^{-(\frac{1}{2}+2k)\pi} \cos(\ln 3) : k \text{ is any integer} \right\}.\end{aligned}$$

(c) $Log(-\sqrt{3}+i) = \ln|-\sqrt{3}+i| + iArg(-\sqrt{3}+i) = \ln 2 + i\,\frac{5\pi}{6}$, so

$$\begin{aligned}(-\sqrt{3}+i)^{\backslash(1+i)\backslash} &= e^{(1+i)Log(-\sqrt{3}+i)} = e^{(1+i)(\ln 2 + i\,\frac{5\pi}{6})} = e^{\ln 2}\, e^{-5\pi/6}\, e^{i\ln 2}\, e^{i5\pi/6} \\ &= 2\cdot e^{-5\pi/6}\cdot e^{i\ln 2}\cdot\left(\frac{1}{2}\cdot(-\sqrt{3}+i)\right) = e^{-5\pi/6}\cdot e^{i\ln 2}\cdot(-\sqrt{3}+i) \\ &= e^{-5\pi/6}\cdot\big(\cos(\ln 2) + i\sin(\ln 2)\big)\cdot(-\sqrt{3}+i) \\ &= -e^{-5\pi/6}\left(\sqrt{3}\,\cos(\ln 2) + \sin(\ln 2)\right) + ie^{-5\pi/6}\left(-\sqrt{3}\,\sin(\ln 2) + \cos(\ln 2)\right).\end{aligned}$$

Using (15.36) and the earlier calculation of $(-\sqrt{3}+i)^{\backslash(1+i)\backslash}$, we have

$$\begin{aligned}(-\sqrt{3}+i)^{1+i} &= \left\{(-\sqrt{3}+i)^{\backslash(1+i)\backslash}\cdot e^{(1+i)i2\pi k} : \text{any integer } k\right\} \\ &= \left\{(-\sqrt{3}+i)^{\backslash(1+i)\backslash}\cdot e^{i2\pi k}\cdot e^{-2\pi k} : \text{any integer } k\right\} \\ &= \left\{e^{-5\pi/6}\cdot e^{i\ln 2}\cdot(-\sqrt{3}+i)\cdot 1\cdot e^{-2\pi k} : k \text{ is any integer}\right\} \\ &= \left\{e^{-(\frac{5}{6}+2k)\pi}\cdot e^{i\ln 2}\cdot(-\sqrt{3}+i) : k \text{ is any integer}\right\} \\ &= \Big\{-e^{-(\frac{5}{6}+2k)\pi}\left(\sqrt{3}\cos(\ln 2) + \sin(\ln 2)\right) \\ &\quad + ie^{-(\frac{5}{6}+2k)\pi}\left(-\sqrt{3}\sin(\ln 2) + \cos(\ln 2)\right) : \text{any integer } k\Big\}. \bigcirc\end{aligned}$$

Example 15.28

For each of the numbers z in the following, evaluate (i) $\sqrt{z}$, (ii) $z^{1/2}$, and (iii) $z^{2/3}$ in "rectangular" form $a+ib$ or in polar exponential form:

$$(a)\ z = -5, \quad (b)\ z = -\sqrt{3}+i.$$

Method: Recall that polar exponential form is $re^{i\theta}$ *with* $-\pi < \theta \le \pi$.

(a) Similar to Example 15.27(a), $Log(-5) = \ln 5 + i\pi$, so

$$(i)\quad \sqrt{-5} = (-5)^{\backslash\frac{1}{2}\backslash} \triangleq e^{\frac{1}{2}Log(-5)} = e^{\frac{1}{2}(\ln 5 + i\pi)} = e^{\frac{1}{2}\ln 5}\cdot e^{i\pi/2} = e^{\ln\sqrt{5}}\cdot(i) = i\sqrt{5}$$

in rectangular form. This agrees with what we learned in a precalculus course.

On the other hand, using (15.36),

$$\begin{aligned}(-5)^{1/2} = e^{\frac{1}{2}log(-5)} &= \left\{e^{\frac{1}{2}Log(-5)}e^{\frac{1}{2}(i2\pi k)} : k \text{ is any integer}\right\} \\ &= \left\{i\sqrt{5}\cdot e^{i\pi k} :\ k \text{ is any integer}\right\} = \left\{i(-1)^k\sqrt{5} :\ k \text{ is any integer}\right\}.\end{aligned}$$

At first glance, it seems as if $(-5)^{1/2}$ is an infinite set. But since $(-1)^k = \pm 1$, we see that in "rectangular" form,

$$(ii)\qquad (-5)^{1/2} = \{\pm i\sqrt{5}\} = \{i\sqrt{5}, -i\sqrt{5}\}$$

is a finite set. We also see that, as in Example 15.3(a) in Section 15.1, $(-5)^{1/2}$ is the set of all solutions, that is, both solutions, of the equation $z^2 = -5$.

Further,

$$(-5)^{2/3} = e^{\frac{2}{3}\,log(-5)} = \left\{e^{\frac{2}{3}\,(Log(-5)+i2\pi k)} : k \text{ is any integer}\right\}$$
$$= \left\{e^{\frac{2}{3}\,(\ln 5+i(\pi+2\pi k))} : k \text{ is any integer}\right\} = \left\{\sqrt[3]{25}\cdot e^{i2\pi/3}\cdot e^{i4\pi k/3} : k \text{ is any integer}\right\}$$
$$= \left\{\sqrt[3]{25}\,e^{i2\pi/3},\ \sqrt[3]{25}\,e^{i2\pi},\ \sqrt[3]{25}\,e^{i10\pi/3}\right\}.$$

In polar exponential form,

$$(iii)\quad (-5)^{2/3} = \left\{\sqrt[3]{25}\,e^{i2\pi/3},\ \sqrt[3]{25},\ \sqrt[3]{25}e^{-i2\pi/3}\right\}.$$

(b) $Log(-\sqrt{3}+i) = \ln 2 + i\,\frac{5\pi}{6}$, so

$$\sqrt{-\sqrt{3}+i} = (-\sqrt{3}+i)^{\backslash\frac{1}{2}\backslash} \triangleq e^{\frac{1}{2}\,Log(-\sqrt{3}+i)} = e^{\frac{1}{2}\,(\ln 2+i\,\frac{5\pi}{6})} = e^{\frac{1}{2}\,\ln 2}\cdot e^{i5\pi/12}$$
$$= e^{\ln\sqrt{2}}\cdot e^{i5\pi/12}.$$

In polar exponential form,

$$(i)\qquad \sqrt{-\sqrt{3}+i} = \sqrt{2}\,e^{i5\pi/12}.$$

On the other hand, using (15.36),

$$(-\sqrt{3}+i)^{1/2} = \left\{e^{\frac{1}{2}\,Log(-\sqrt{3}+i)}e^{\frac{1}{2}\,(i2\pi k)} : k \text{ is any integer}\right\}$$
$$= \left\{\sqrt{2}\,e^{i5\pi/12}\cdot e^{i\pi k} : k \text{ is any integer}\right\}.$$

Because $e^{i\pi k} = (-1)^k$, in polar exponential form,

$$(ii)\quad (-\sqrt{3}+i)^{1/2} = \left\{\sqrt{2}e^{i5\pi/12},\ \sqrt{2}e^{-i7\pi/12}\right\},$$

which is a finite set. We also see that, as in Example 15.3(a) in Section 15.1, $(-\sqrt{3}+i)^{1/2}$ is the set of all solutions, that is, both solutions, of the equation $z^2 = -\sqrt{3}+i$.

Finally,

$$(-\sqrt{3}+i)^{2/3} = \left\{e^{\frac{2}{3}\,Log(-\sqrt{3}+i)}e^{\frac{2}{3}\,(i2\pi k)} : k \text{ is any integer}\right\}$$
$$= \left\{\sqrt[3]{4}\cdot e^{i5\pi/9}\cdot e^{i4\pi k/3} : k \text{ is any integer}\right\}$$
$$= \left\{\sqrt[3]{4}\,e^{i5\pi/9},\ \sqrt[3]{25}\,e^{i17\pi/9},\ \sqrt[3]{4}\,e^{i29\pi/9}\right\}.$$

In polar exponential form,

$$(iii)\quad (-\sqrt{3}+i)^{2/3} = \left\{\sqrt[3]{4}\,e^{i5\pi/9},\ \sqrt[3]{4}\,e^{-i\pi/9},\ \sqrt[3]{4}\,e^{-i7\pi/9}\right\}. \bigcirc$$

Example 15.29

Where is $\sqrt{z}$ differentiable, and what is its derivative there?

Method: By the chain rule,

$$\sqrt{z} \triangleq e^{\frac{1}{2}\,Log(z)}$$

is differentiable everywhere $Log(z)$ is differentiable, that is, for all z in $\mathcal{D} \triangleq \{z: Arg(z) \neq \pi\} = \{z = x + iy : x > 0 \text{ or } y \neq 0\}$. At those z,

$$\frac{d}{dz}\Big[\sqrt{z}\Big] = \frac{d}{dz}\Big[e^{\frac{1}{2}Log(z)}\Big] = e^{\frac{1}{2}Log(z)} \cdot \frac{d}{dz}\Big[\frac{1}{2}Log(z)\Big] = e^{\frac{1}{2}Log(z)} \cdot \frac{1}{2} \cdot \frac{1}{z} = \sqrt{z} \cdot \frac{1}{2z} = \frac{1}{2\sqrt{z}}.$$

This generalizes the Calculus I fact that $(\sqrt{x})' = \dfrac{1}{2\sqrt{x}}$ for real numbers $x > 0$. ◯

This tells us everything about differentiability of $\sqrt{z}$ for z in $\mathcal{D}$. But what about z not in $\mathcal{D}$? We will explain why $\sqrt{z}$ is not differentiable on the nonpositive real axis.

First, recall that in this book the natural logarithm function, ln, will allow only inputs that are positive, real numbers, such as $|z|^{1/2}$ when $z \neq 0$.

Define

$$f(z) = \sqrt{z} = \cdots = e^{\frac{1}{2}\ln|z|}\, e^{i\,Arg(z)/2} = e^{\ln(|z|^{1/2})}\, e^{i\,Arg(z)/2} = |z|^{1/2}\, e^{i\,Arg(z)/2}.$$

We recall that $Arg(z)$ is not continuous at $z = x + i0$ for $x \leq 0$. In fact, $Arg(0)$ is undefined, so $f(z)$ is not even defined at $z = 0$. Further, if $z_0 = x_0 + i0$ and $x_0 < 0$, then the facts that

$$\lim_{y \to 0^+} Arg(x_0 + iy) = \pi \quad \text{and} \quad \lim_{y \to 0^-} Arg(x_0 + iy) = -\pi$$

imply that

$$\lim_{y \to 0^+} f(x_0 + iy) = \ln(|x_0|^{1/2}) e^{i\frac{\pi}{2}} = i\sqrt{|x_0|} \quad \text{and} \quad \lim_{y \to 0^-} f(x_0 + iy) = \ln(|x_0|^{1/2}) e^{-i\frac{\pi}{2}} = -i\sqrt{|x_0|},$$

so $f(z)$ is not continuous at $z_0 = x_0 + i0$ and thus cannot be differentiable there. ◯

We recall in passing that we know that we can define for the logarithm different branches $Log_\sigma(z)$. Using them, we can define a function $f_\sigma(z) \triangleq e^{\frac{1}{2}Log_\sigma(z)}$ that is differentiable except on the ray $Arg(z) = \sigma + \pi$.

15.4.6 Problems

1. Find a complex number z for which $|z - i| = 2$ and $Arg(z) = \frac{3\pi}{4}$.
2. Find a complex number z for which $|z - i| = 3$ and $Arg(z) = \frac{5\pi}{6}$.
3. For each of the numbers z in the following, find $arg(z)$ and $Arg(z)$:
   (a) $-1 + i$,
   (b) $-\sqrt{3} - i$,
   (c) $2e^{i\frac{5\pi}{3}}$,
   (d) $\dfrac{-\sqrt{3} - i}{-1 + i}$.
4. Find the Arg of the complex numbers:
   (a) $-2\sqrt{2} - i2\sqrt{2}$,
   (b) $(1 + i)(-\sqrt{3} - i)$,
   (c) $\frac{1+i\sqrt{3}}{-1+i}$.

5. For each of (a) and (b), decide whether the given statement is true or false. If it's true, explain why; if it's false, give a specific, explicit counterexample:
   (a) $Log(z^2) = 2\,Log(z)$ for all $z \neq 0$.
   (b) $Log(\sqrt{z}) = \frac{1}{2}\,Log(z)$, where $\sqrt{z} = z^{\backslash 1/2 \backslash}$, that is, the principal value of $z^{1/2}$, for all $z \neq 0$.
6. For which value(s) of z is $Log\left(\dfrac{1}{z}\right) = -Log(z)$?
7. For which value(s) of z is (a) $Log\left(e^z\right) = z$, and (b) $e^{Log(z)} = z$?
8. Solve the equation
$$e^{(1-\frac{z}{2})} = -1 + i.$$
9. Solve the equation $\dfrac{e^z}{e^z + 1} = i$.
10. We know that $Log(z_1 z_2)$ may not equal $Log(z_1) + Log(z_2)$. Explain why
$$Log(z_1 z_2) + i2k\pi = Log(z_1) + Log(z_2),$$
where k must be either $-1, 0$, or 1.
11. Define the principal square root function by $\sqrt{z}$, that is, $z^{\backslash 1/2 \backslash} \triangleq \sqrt{|z|}\, e^{i\,Arg(z)/2}$. Find a number z and an integer n for which $Arg\left(\left(z^n\right)^{1/2}\right) \neq n\,Arg\left(z^{1/2}\right)$.
12. Define a function
$$f(z) \triangleq |z|^{\alpha^2} e^{i(3\alpha - 2)Arg(z)}.$$
Find all real values of the constant α for which $f(z)$ satisfies the Cauchy–Riemann equations for $z \neq 0$. For each such value of α, write down the corresponding functions $f'(z)$.
13. Evaluate $(2i)^i$.
14. Find the exact value(s) of $(-1 - i)^{1/3}$ and express it or them in polar exponential form.
15. Evaluate $\sqrt[3]{-\sqrt{3} + i}$ and express your final conclusion in polar exponential form.
16. Evaluate $(-\sqrt{3} + i)^{1/3}$ and express your final conclusion in polar exponential form.
17. Where is $\sqrt[3]{z}$ differentiable?
18. Where is $\sqrt[3]{z^2}$ differentiable?
19. Why is there no solution of the equation $e^w = 0$?
20. Define $f(z) = Arg(z)$ and $g(z) = \left(f(z)\right)^2$. Where is $g(z)$ differentiable? At such z, find a formula for $g'(z)$.
21. Why does the power function defined in (15.35) give only a single value when the exponent α is an integer?
22. Why does $Log(z)$ choose from $log(z)$ the unique element w satisfying $-\pi < \mathcal{I}m(w) \leq \pi$?
23. If x is a real, positive number, explain why $Log(x) = \ln x$.

24. Using trigonometric identities, explain why e^z satisfies the laws of exponents, (15.27), that is, (*i*) $e^{z_1+z_2} = e^{z_1}e^{z_2}$, (*ii*) $e^{z_1-z_2} = e^{z_1}/e^{z_2}$, and (*iii*) $\left(e^z\right)^n = e^{nz}$ for all complex numbers z_1, z_2, z and integers n.
25. Establish (15.33) in Case (3) of the explanation of Theorem 15.17: that is, if (x, y) lies on the positive x-axis, that is, has $x > 0$ and $y = 0$, explain why $\frac{\partial v}{\partial x} = \cdots = -\frac{y}{x^2+y^2} = -\frac{\partial u}{\partial y}$ and $\frac{\partial v}{\partial y} = \cdots = \frac{x}{x^2+y^2} = \frac{\partial u}{\partial x}$. [Caution: $\sqrt{y^2} = |y|$.]
26. Explain why $Arg(z)$ is continuous on the positive real axis.

15.5 Trigonometric Functions

Based on Euler's formula for $e^{i\theta}$ and our previous experience with the sine and cosine functions of a real variable as solutions of ODEs, for all complex numbers z, we define

$$\cos z \triangleq \frac{e^{iz} + e^{-iz}}{2} \quad \text{and} \quad \sin z \triangleq \frac{e^{iz} + e^{-iz}}{i2}. \tag{15.37}$$

Theorem 15.18

For all complex numbers z_1 and z_2,

$$\cos(z_1 + z_2) = \cos z_1 \cos z_2 - \sin z_1 \sin z_2 \tag{15.38}$$

and

$$\sin(z_1 + z_2) = \sin z_1 \cos z_2 + \cos z_1 \sin z_2\,. \tag{15.39}$$

Why? In justifying an identity, it is usually easier to start on the more complicated side and use facts and/or algebraic manipulations to get to the other side. For example, starting on the right-hand side of (15.38), we have

$$\begin{aligned}
\cos z_1 \cos z_2 - \sin z_1 \sin z_2 &= \frac{e^{iz_1} + e^{-iz_1}}{2} \cdot \frac{e^{iz_2} + e^{-iz_2}}{2} - \frac{e^{iz_1} - e^{-iz_1}}{i2} \cdot \frac{e^{iz_2} - e^{-iz_2}}{i2} \\
&= \frac{1}{4}\left(e^{i(z_1+z_2)} + \cancel{e^{i(-z_1+z_2)}} + \cancel{e^{i(z_1-z_2)}} + e^{-i(z_1+z_2)}\right) \\
&\quad + \frac{1}{4}\left(e^{i(z_1+z_2)} - \cancel{e^{i(-z_1+z_2)}} - \cancel{e^{i(z_1-z_2)}} + e^{-i(z_1+z_2)}\right) \\
&= \frac{1}{4} \cdot 2 \cdot \left(e^{i(z_1+z_2)} + e^{-i(z_1+z_2)}\right) = \frac{e^{i(z_1+z_2)} + e^{-i(z_1+z_2)}}{2} \\
&\triangleq \cos(z_1 + z_2),
\end{aligned}$$

thus establishing (15.38). You will establish (15.39) in Problem 15.5.1.13 using similar work. □

Example 15.30

Find the exact value of $\sin(-\sqrt{3}+i)$.

Method: It's actually easier to *not* use the polar exponential form of $-\sqrt{3}+i$, and it's easier to use (15.39) rather than the definition of $\sin(-\sqrt{3}+i)$ given by (15.37).

First, it helps to get two more formulas: for any real number y, using (15.37), we have

$$\cos(iy) = \frac{e^{i(iy)} + e^{-i(iy)}}{2} = \frac{e^{-y} + e^{y}}{2} \triangleq \cosh y \tag{15.40}$$

and

$$\sin(iy) = \frac{e^{i(iy)} - e^{-i(iy)}}{i2} = \frac{e^{-y} - e^{y}}{i2} = -i \cdot \frac{e^{-y} - e^{y}}{2} = i \cdot \frac{e^{y} - e^{-y}}{2} \triangleq i \sinh y\,. \tag{15.41}$$

So,

$$\begin{aligned}\sin(-\sqrt{3}+i) &= \sin(-\sqrt{3})\cos(i) + \cos(-\sqrt{3})\sin(i)\\ &= -\sin(\sqrt{3})\cosh(1) + i\cos(\sqrt{3})\sinh(1).\end{aligned}$$ ○

Example 15.31

Solve the equations (a) $2\sin z = -1$, and (b) $\cos z = i3$.

Method: (a) For $z = x + iy$, where x, y are real, the equation is

$$-\frac{1}{2} + i0 = \sin(x+iy) = \sin x \cos(iy) + \cos x \sin(iy) = \sin x \cosh y + i \cos x \sinh y.$$

Separating the real and imaginary parts gives the system of equations

$$\left\{\begin{array}{ll}(1) & -\frac{1}{2} = \sin x \cosh y\\ (2) & 0 = \cos x \sinh y\end{array}\right\}.$$

Equation (2) is easier to solve than the first because 0 is on the left-hand side: $0 = \cos x \sinh y \iff$ (i) $0 = \cos x$ or (ii) $0 = \sinh y \iff$ (i) $x = (n - \frac{1}{2})\pi$ for some integer n or (ii) $y = 0$.

Substitute each of these possibilities (i), (ii) into the equation (1): First, using (i), hence $x = (n - \frac{1}{2})\pi$,

$$(1)(\text{i}) \quad -\frac{1}{2} = \sin x \cosh y = \sin\left(\left(n - \frac{1}{2}\right)\pi\right)\cosh y = (-1)^{n+1}\cosh y$$

and the fact that $\cosh(y) \geq 1$ for all real y shows that there is no solution of (1)(i) for y. So far, there is no solution.

Second, using (ii), hence $y = 0$,

$$(1)(\text{ii}) \quad -\frac{1}{2} = \sin x \cosh y = \sin x \cosh 0 = \sin x \cdot 1 = \sin x,$$

which gives $x = \left(-\frac{\pi}{2} \pm \frac{\pi}{3}\right) + 2k\pi$ for any integer k.

So, the set of solutions is

$$\left\{-\frac{\pi}{2} \pm \frac{\pi}{3} + 2k\pi + i0 : k \text{ is any integer}\right\}.$$

These solutions are exactly the ones we would find in a trigonometry course.
(b) For $z = x + iy$, where x, y are real, the equation is

$$i3 = \cos(x + iy) = \cos x \cos(iy) - \sin x \sin(iy) = \cos x \cosh y - i \sin x \sinh y.$$

Separating the real and imaginary parts gives the system of equation

$$\left\{ \begin{array}{lll} (1) & 0 = & \cos x \cosh y \\ (2) & 3 = & -\sin x \sinh y \end{array} \right\}.$$

Equation (1) is easier to solve than (2): $0 = \cos x \cosh y \iff$ (i) $0 = \cos x$ or (ii) $0 = \cosh y$. But (ii) is impossible to satisfy because $\cosh y \geq 1$ for all y. The solutions of (1) have $x = (n - \frac{1}{2})\pi$.

Substitute $x = (n - \frac{1}{2})\pi$ into the second equation (2):

$$3 = -\sin x \sinh y = -\sin\left(\left(n - \frac{1}{2}\right)\pi\right) \sinh y = (-1)^n \sinh y.$$

For $n = even = 2k$, (2) becomes $\sinh y = 3$, whose solution would be $y = arcsinh(3)$ if we had defined such a function. Instead, there is a solution technique of independent interest:

$$3 = \sinh y \triangleq \frac{e^y - e^{-y}}{2} \iff 6 = e^y - e^{-y} \iff 6e^y = e^y \cdot (e^y - e^{-y}) = (e^y)^2 - 1,$$

so substituting $w = e^y$ turns that equation into $6w = w^2 - 1$, that is, $w^2 - 6w - 1 = 0$, whose solutions are

$$e^y = w = \frac{6 \pm \sqrt{40}}{2} = 3 \pm \sqrt{10}.$$

This gives us only $y = \ln\left(3 + \sqrt{10}\right)$, because $e^y > 0$ for any real y, hence, e^y cannot equal $3 - \sqrt{10} < 0$.

So far, our only solutions are

$$z = x + iy = \left(2k - \frac{1}{2}\right)\pi + i \ln(3 + \sqrt{10}).$$

For $n = odd = 2\ell - 1$, (2) becomes $\sinh y = -3$, whose solutions are found using

$$-3 = \sinh y \triangleq \frac{e^y - e^{-y}}{2} \iff -6 = e^y - e^{-y} \iff -6e^y = e^y \cdot (e^y - e^{-y}) = (e^y)^2 - 1,$$

so substituting $w = e^y$ turns that equation into $-6w = w^2 - 1$, that is, $w^2 + 6w - 1 = 0$, whose solutions are

$$e^y = w = \frac{-6 \pm \sqrt{40}}{2} = -3 \pm \sqrt{10}.$$

Again, this gives us only $y = \ln\left(-3 + \sqrt{10}\right)$.

Putting everything together, the set of solutions is

$$\left\{\left(2k-\frac{1}{2}\right)\pi + i\ln(3+\sqrt{10}) : \text{ any integer } k\right\}$$
$$\cup \left\{\left(2\ell-1-\frac{1}{2}\right)\pi + i\ln(-3+\sqrt{10}) : \text{any integer } \ell\right\}. \bigcirc$$

The $\cup$ symbol, that is, **set union**, means that the solutions consist of all z that are either in the first set or in the second set.

Notice that we found *all* of the *exact* solutions.

Example 15.31(a) shows that solving a trigonometric equation for the complex variable z may turn out to only give real solutions that we are familiar with from solving a trigonometric equation for a real variable.

But, as in Example 15.31(b), solving a trigonometric equation for a complex variable z may give infinitely many complex solutions even though the corresponding real trigonometric equation has no solution, such as $\sin x = 2$, as you will see in Problem 15.5.1.2 in the following.

15.5.1 Problems

1. Express $\sin\left(\frac{\pi+i4}{4}\right)$ in the form $a+ib$.
2. Find the solutions of $\sin z = 2$. Note that $\sin x = 2$, the corresponding trigonometric equation for real variable x, has no solution!
3. Find the solutions of $\cos z = 2$. Note that $\cos x = 2$, the corresponding trigonometric equation for real variable x, has no solution!

In problems 4–11, solve the equation, that is, find all solutions of the equation.

4. $2\cos z = -1$,
5. $\sin z = i3$,
6. $\cos z = -\cosh\pi$.
7. $\sin z = -i\sinh\frac{\pi}{2}$.
8. $\sin z = -i\sinh 3$.
9. $\cos z = i\sinh 3$.
10. $\cos z = -\cosh\pi$.
11. $\sin z = -\cos z$.
12. For complex numbers $z = x+iy$, where x, y are real, define

$$\cosh(z) \triangleq \frac{e^z + e^{-z}}{2}.$$

    (a) Explain why $\cosh(z) = \cosh x\,\cos y + i\sinh x\,\sin y$.
    (b) Find all solutions of $\cosh z = i\sinh z$.
    (c) Solve $\cosh z = -e$.
13. Establish (15.39).

14. Suppose b is a real number and $|b| \leq 1$. Explain why the equation $\sin z = b$ has only real solutions for z.
15. Suppose b is a real number and $|b| \leq 1$. Explain why the equation $\cos z = b$ has only real solutions for z.
16. Suppose b is a real number and $|b| \leq 1$. Explain why the equation $\sinh z = b$ has only imaginary solutions for z.
17. Suppose b is a real number and $|b| \leq 1$. Explain why the equation $\cosh z = b$ has only imaginary solutions for z.

15.6 Taylor and Laurent Series

Similar to the definition in calculus of real variables,

$$\sum_{j=0}^{\infty} a_j (z - z_0)^j \tag{15.42}$$

is a **power series about** z_0. The a_j's can be complex constants. As in calculus, convergence of such an infinite series at particular values of z is an issue. Also, similar to calculus, study of series begins with study of series of complex constants $\sum_{j=0}^{\infty} c_j$.

Some very useful background information is from precalculus: if r is a real number, then the sum of a **geometric series** is given by

$$\sum_{j=0}^{\infty} r^j = 1 + r + r^2 + r^3 + \cdots = \frac{1}{1-r}, \text{ as long as } |r| < 1\,. \tag{15.43}$$

Definition 15.15

(a) A series of complex numbers $\sum_{j=0}^{\infty} c_j$ **converges absolutely** if the corresponding real series $\sum_{j=0}^{\infty} |c_j|$ converges.

(b) A power series (15.42) **converges absolutely** at a specific z_1 if the complex series $\sum_{j=0}^{\infty} c_j$ converges absolutely, where $c_j \triangleq a_j (z_1 - z_0)^j$.

(c) A series $\sum_{j=0}^{\infty} c_j$ **diverges** if the sequence of partial sums $\{s_N\}_{N=0}^{\infty}$ does not converge (to a finite number), where $s_N \triangleq \sum_{j=0}^{N} c_j$.

(d) A power series (15.42) **diverges** at z_1 if the complex series $\sum_{j=0}^{\infty} c_j$ diverges, where $c_j \triangleq a_j (z_1 - z_0)^j$.

Example 15.32

For which value(s) of the complex constant α is the geometric series $\sum_{j=0}^{\infty} \alpha^j$ absolutely convergent?

Method: Define $c_j = \alpha^j$. We calculate that $|c_j| = |\alpha^j| = |\alpha|^j$, so we see that $\sum_{j=0}^{\infty} c_j$ is a geometric series that converges absolutely if, and only if, $|\alpha| < 1$. So our original series, $\sum_{j=0}^{\infty} \alpha^j$, converges absolutely if, and only if, $|\alpha| < 1$. ◯

Example 15.33

Does the series $\sum_{j=0}^{\infty} e^{(-0.1+i)j}$ converge absolutely?

Method: Define $c_j = e^{(-0.1+i)j} = \left(e^{-0.1j}\, e^{ij}\right)^j = \alpha^j$, where $\alpha = e^{-0.1}\, e^{i}$, and note that $|\alpha| = |e^{-0.1}| \cdot |e^{i}| = e^{-0.1} \cdot 1 < 1$. So, by Example 15.32, $\sum_{j=0}^{\infty} e^{(-0.1+i)j}$ is an absolutely convergent geometric series. ◯

Theorem 15.19

Suppose $\sum_{j=0}^{\infty} c_j$ and $\sum_{j=0}^{\infty} b_j$ are absolutely convergent series and $\sum_{j=0}^{\infty} h_j$ is not absolutely convergent. Then for all complex constants β and γ,

(a) $\sum_{j=0}^{\infty}(\beta c_j + \gamma b_j)$ is absolutely convergent.
(b) $\sum_{j=0}^{\infty}(\beta c_j + \gamma h_j)$ is not absolutely convergent unless $\gamma = 0$.

Why? (a) follows from the triangle inequality for complex numbers because $|\beta c_j + \gamma b_j| \leq |\beta c_j| + |\gamma b_j| = |\beta|\,|c_j| + |\gamma|\,|b_j|$. (b) Assume contrarily that $\gamma \neq 0$ and $\sum_{j=0}^{\infty}(\beta c_j + \gamma h_j)$ is absolutely convergent. Then the result of part (a) would imply that

$$\sum_{j=0}^{\infty} h_j = \sum_{j=0}^{\infty} \frac{1}{\gamma}\left((\beta c_j + \gamma h_j) + (-\beta c_j)\right)$$

is absolutely convergent, which would imply that $\sum_{j=0}^{\infty} h_j$ is absolutely convergent, giving a contradiction. □

Theorem 15.20

Suppose there exists

$$\alpha \triangleq \lim_{j\to\infty} \left|\frac{a_{j+1}}{a_j}\right|. \tag{15.44}$$

(a) If $0 < \alpha < \infty$, then the power series (15.42), that is, $\sum_{j=0}^{\infty} a_j\,(z - z_0)^j$, converges absolutely for $|z - z_0| < R \triangleq \dfrac{1}{\alpha}$ and diverges for $|z - z_0| > R$.
(b) If $\alpha = 0$, then the power series (15.42) converges absolutely at all z.

Why? (a) The absolute convergence follows from the absolute convergence of geometric series along with an (unstated) "comparison theorem" based on a similar result from

calculus of real variables. Similar reasoning explains divergence in part (a) and absolute convergence in part (b). □

We note that it is possible for a power series (15.42) to converge only at z_0.

Definition 15.16

(a) A finite number R is called the **radius of convergence** for (15.42) if the power series both (a) converges absolutely for $|z - z_0| < R$ and (b) diverges for $|z - z_0| > R$.
(b) The radius of convergence is defined to be $R = \infty$, and we say the radius of convergence is infinite, if the power series (15.42) converges absolutely at all z.
(c) The radius of convergence is defined to be $R = 0$ if the power series (15.42) converges only at z_0.

We need one further concept of convergence

Definition 15.17

Let z_0 be a complex number and $r > 0$. A power series (15.42) **converges uniformly** on a *closed* disk $\overline{\mathcal{D}}_r(z_0) \triangleq \{z : |z - z_0 \leq r\}$ if for all $\varepsilon > 0$, there exists an integer N, possibly dependent on ε, such that

$$\text{for all } n \geq N \text{ and for all } z \text{ in } \overline{\mathcal{D}}_r(z_0), \left| \sum_{j=n+1}^{\infty} a_j (z - z_0)^j \right| < \varepsilon . \tag{15.45}$$

For uniform convergence, N is not allowed to depend on z in $\overline{\mathcal{D}}_r(z_0)$.

Intuitively, $\sum_{j=N+1}^{\infty} a_j (z - z_0)^j$ is a "tail" of the series and we hope that the tail is unimportant. One way to remember this is that "the tail does not wag the dog." In applications of mathematics to engineering, usually computer software will **truncate** a series at a finite number of terms; the *reliability* of the truncation process may depend on a mathematical theorem that guarantees uniform convergence.

Theorem 15.21

Suppose a power series (15.42) has a *positive* radius of convergence, R, possibly infinite. Then

(a) The power series converges uniformly in every closed disk of finite radius that is contained in the open disk $\mathcal{D}_R(z_0) = \{z : |z - z_0| < R\}$.
(b) The power series defines a function

$$f(z) \triangleq \sum_{j=0}^{\infty} a_j (z - z_0)^j \tag{15.46}$$

that is analytic on the open disk $\mathcal{D}_R(z_0)$.

Example 15.34

Find the radius of convergence of the series $\sum_{j=0}^{\infty} \left(-\frac{z}{2}\right)^j$, and find a simple formula for an analytic function, which equals the sum of the series.

Method: The power series is $\sum_{j=0}^{\infty} \left(-\frac{1}{2}\right)^j z^j$, so $a_j = \left(-\frac{1}{2}\right)^j$, and we calculate

$$\lim_{j\to\infty} \left|\frac{a_{j+1}}{a_j}\right| = \lim_{j\to\infty} \left|\frac{\left(-\frac{1}{2}\right)^{j+1}}{\left(-\frac{1}{2}\right)^j}\right| = \lim_{j\to\infty} \left|-\frac{1}{2}\right| = \frac{1}{2} = \alpha,$$

so $R = \frac{1}{\alpha} = 2$ is the radius of convergence.

Alternatively, we note that $\sum_{j=0}^{\infty} \left(-\frac{z}{2}\right)^j$ is a geometric series that converges for $\left|-\frac{z}{2}\right| = \frac{|z|}{|2|} = \frac{|z|}{2} < 1$, that is, for $|z| < 2$, and diverges for $\left|-\frac{z}{2}\right| > 1$, that is, for $|z| > 2$.

The conclusions about the analyticity of $f(z) \triangleq \sum_{j=0}^{\infty} \left(-\frac{z}{2}\right)^j$ follow from Theorem 15.21.

The formula for $f(z)$ follows from the fact that the sum of a geometric series $\sum_{j=0}^{\infty} r^j$ is $\frac{1}{1-r}$, as long as $|r| < 1$. Here, $r = -\frac{z}{2}$ has $|r| < 1$ as long as $|z - 0| < 2$, in which case

$$f(z) = \frac{1}{1 - \left(-\frac{z}{2}\right)} = \frac{1}{1 + \frac{z}{2}} = \frac{1}{1 + \frac{z}{2}} \cdot \frac{2}{2} = \frac{2}{\left(1 + \frac{z}{2}\right) \cdot 2} = \frac{2}{2 + z}. \bigcirc$$

The fact that the function $f(z) = \frac{2}{z+2}$ is undefined at $z = -2$ reassures us that the radius of convergence being $R = 2$ makes sense. On the other hand, while the function $f(z)$ is defined at $z = 2$, that point is not within the radius of convergence, because the sequence of partial sums for

$$\sum_{j=0}^{\infty} \left(-\frac{2}{2}\right)^j = \sum_{j=0}^{\infty} (-1)^j$$

is $\{s_N\}_{j=0}^{\infty} = 1, 0, 1, 0, \ldots$, which does not converge due to oscillation.

15.6.1 Taylor Series

Earlier, we explained how a power series can define an analytic function. The reverse process also works! From a function analytic on an open disk, we can get a power series that equals the function on that disk. The process is familiar to us from calculus of a real variable.

If $f(z)$ is analytic on an open disk about z_0, Theorem 15.40 in Section 15.9, a consequence of Cauchy's integral formula, will guarantee that *all* of its derivatives $f', f'', f^{(3)}, \ldots, f^{(n)}, \ldots$

are defined, continuous, and differentiable on that open disk about z_0. So, we can calculate the numbers

$$a_j = \frac{1}{j!} f^{(j)}(z_0), \quad \text{for} \quad j = 0, 1, 2, \ldots . \tag{15.47}$$

Theorem 15.22

(Taylor series) Suppose $f(z)$ is analytic on an open disk $\mathcal{D}$ centered at z_0. Then at all z in $\mathcal{D}$,

$$f(z) = \sum_{j=0}^{\infty} a_j (z - z_0)^j = \sum_{j=0}^{\infty} \frac{1}{j!} f^{(j)}(z_0) (z - z_0)^j. \tag{15.48}$$

Also, the convergence in (15.48) is absolute inside the open disk $\mathcal{D}$ and uniform on any finite closed disk contained in $\mathcal{D}$.

Example 15.35

Find the Taylor series centered at $z_0 = 0$ for the function $f(z) \triangleq e^{3z}$.

Method: We calculate $f'(z) = 3e^{3z}$, $f''(z) = 9e^{3z}, \ldots, f^{(j)} = 3^j e^{3z}$, so $a_j = f^{(j)}(0) = 3^j e^{3 \cdot 0} = 3^j$. Let $z_0 = 0$. Theorem 15.22 on Taylor series implies that

$$f(z) = \sum_{j=0}^{\infty} \frac{1}{j!} f^{(j)}(0) (z - 0)^j = \sum_{j=0}^{\infty} \frac{1}{j!} 3^j z^j = \sum_{j=0}^{\infty} \frac{(3z)^j}{j!} . \tag{15.49}$$

As for convergence of the series in (15.48), $f(z) = e^{3z}$ is analytic on the whole complex plane $\mathbb{C}$, so Theorem 15.22 guarantees that the convergence is absolute everywhere and uniform on any finite closed disk. ◯

The Taylor series for e^{3z} agrees with the Taylor series for e^{3x} with real variable x replaced by complex variable z. This is not a coincidence!

In general, we define a function $g(x)$ of the real variable x to be **real analytic** on an interval $a < x < b$ if for some function $f(z)$ we have that $g(x) = f(x + i0)$ and $f(z)$ is analytic on some set in the complex plane that is open and contains the points $z = x + i0$ for $a < x < b$.

15.6.2 Laurent Series

There can be more than one way to expand the same function into a power series, if we include negative powers.

Example 15.36

Expand $\dfrac{1}{1 - z}$ into (a) a power series that converges inside an open disk, and (b) a series in nonpositive powers of z that converges outside an open disk.

Method:

(a) We could use the Taylor series found in (15.48) but it's easier to use geometric series and possibly manipulations of geometric series: first, we have

$$f(z) \triangleq \frac{1}{1-z} = \sum_{j=0}^{\infty} z^j, \quad \text{converges for } |z| < 1 \tag{15.50}$$

just by substituting $r = z$ into the formula for geometric series, (15.43). The power series for $f(z)$ is defined and analytic on the set $\{z : |z| < 1\}$, that is, inside an open disk.

(b) Instead of writing $\dfrac{1}{1-z}$ in powers of z, we can write

$$g(z) \triangleq \frac{1}{1-z} = \frac{1}{-z\left(1-\frac{1}{z}\right)} = \frac{1}{-z} \cdot \frac{1}{1-\frac{1}{z}} = -\frac{1}{z} \cdot \left(1 + \frac{1}{z} + \left(\frac{1}{z}\right)^2 + \left(\frac{1}{z}\right)^3 + \cdots\right)$$

$$= -\frac{1}{z} - \left(\frac{1}{z}\right)^2 - \left(\frac{1}{z}\right)^3 - \left(\frac{1}{z}\right)^4 - \cdots,$$

where we substituted $r = 1/z$ into the geometric series formula (15.43). The geometric series for $\dfrac{1}{1-\frac{1}{z}}$ converges for $1 > |r| = \left|\dfrac{1}{z}\right| = \dfrac{1}{|z|}$, that is, for $|z| > 1$. So,

$$g(z) \triangleq \frac{1}{1-z} = -\sum_{j=1}^{\infty} z^{-j} \quad \text{converges for } |z| > 1\,. \tag{15.51}$$

The series in nonpositive powers of z for $g(z)$ is defined and analytic on the set $\{z : |z| > 1\}$, that is, *outside* an open disk. ◯

It is true that we were being a little fussy by giving different names to the functions f and g. While both of them give a formula for $\dfrac{1}{1-z}$, the two functions have different domains, so giving them different names helps to separate them in our minds.

Neither the series for $f(z)$ given in (15.50), nor the series for $g(z)$ given in (15.51), were asked to converge at $z = 1$. This is good because $\dfrac{1}{1-z}$ is undefined at $z = 1$. As for other points z on the circle $|z| = 1$, such as $z = -1$, neither series converges there even though the original, given function, $\dfrac{1}{1-z}$ is defined at all $z \neq 1$.

Partial differential equations in Sections 11.5 and 11.6, specifically Problem 11.5.3.2 and Example 11.22 in Section 11.6, modeled physical problems that take place outside of a disk or a ball.

Definition 15.18

An **open annulus** is a set of the form

$$\mathcal{A}_{r,R}(z_0) \triangleq \{z : r < |z - z_0| < R\}, \tag{15.52}$$

where $R > r \geq 0$; $R = \infty$ is allowed.

When $r = 0$, the "annulus" $\mathcal{A}_{0,R}(z_0)$ is actually a punctured disk, and when $R = \infty$, the "annulus" $\mathcal{A}_{0,\infty}(z_0)$ is actually the punctured plane $\mathbb{C}\backslash\{z_0\} \triangleq \{z{:}0 < |z - z_0| < \infty\}$.

Definition 15.19

A **Laurent series** is an expansion of the form

$$f(z) \triangleq \sum_{j=1}^{\infty} a_{-j} (z - z_0)^{-j} + \sum_{j=0}^{\infty} a_j (z - z_0)^j . \tag{15.53}$$

Theorem 15.23

(Laurent series) If a Laurent series (15.53) converges absolutely for all z in an open annulus $\mathcal{A}_{r,R}(z_0)$ (or in some open disk $\mathcal{D}_R(z_0)$), then the series converges uniformly in any closed annulus contained in $\mathcal{A}_{r,R}(z_0)$, that is, in any set

$$\{z : \widetilde{r} \le |z - z_0| \le \widetilde{R}\}, \text{ where } r < \widetilde{r} < \widetilde{R} < R$$

(or, respectively, in any closed disk $\{z : |z - z_0| \le \widetilde{R}\}$, where $0 < \widetilde{R} < R$).

Example 15.37

Find a Laurent series for $\dfrac{1}{3 - z}$ in each of several significantly different open sets that are annuli or disks.

Method: A trivial way to get a Laurent series in several different open sets that are annuli or disks would be to get a Laurent series in one such set and then use the same Laurent series in some subsets of the first such set.

Instead, we will get sets that are significantly different from each other because this will help later in this section and chapter.

First,

$$\frac{1}{3 - z} = \frac{1}{3\left(1 - \dfrac{z}{3}\right)} = \frac{1}{3} \cdot \frac{1}{1 - \dfrac{z}{3}} = \frac{1}{3} \cdot \sum_{j=0}^{\infty} \left(\frac{z}{3}\right)^j = \sum_{j=0}^{\infty} 3^{-(j+1)} z^j, \text{ for } |z| < 3,$$

because $\left|\frac{z}{3}\right| < 1 \iff |z| < 3$.

Second,

$$\frac{1}{3 - z} = \frac{1}{-z\left(1 - \frac{3}{z}\right)} = -\frac{1}{z} \cdot \frac{1}{1 - \frac{3}{z}} = -\frac{1}{z} \cdot \sum_{j=0}^{\infty} \left(\frac{3}{z}\right)^j = \sum_{j=1}^{\infty} -(3^{j-1}) z^{-j}, \text{ for } |z| > 3,$$

because $\left|\frac{3}{z}\right| < 1 \iff |z| > 3$.

Third, we can get a Laurent series in an annulus centered at a $z_0 \neq 0$, for example,

$$\frac{1}{3 - z} = \frac{1}{1 - (z - 2)} = \sum_{j=0}^{\infty} (z - 2)^j , \text{ for } |z - 2| < 1.$$

Each of the three corresponding sets, $\mathcal{D}_3(0) = \{z : |z| < 3\}$, $\mathcal{A}_{3,\infty}(0) = \{z : |z| > 3\}$, and $\mathcal{D}_1(2) = \{z : |z-2| < 1\}$, is an open annulus or open disk. ◯

Example 15.38

Find Laurent series for $f(z) = \dfrac{1}{z+1} - \dfrac{2}{z-2}$ in each of several significantly different open sets that are annuli or disks.

Method: We were not told what z_0 should be, so we might as well try to use $z_0 = 0$, for convenience. Each of the two terms in $f(z)$ can be written in *many* choices of Laurent series, as we saw in Example 15.37. But when we add the two terms together to get $f(z)$, we need both of the two Laurent series to converge in the same disk or annulus. To solve this problem, it will help to be systematic!

We will get sets that are significantly different from each other because this will help later in this section and chapter.

The first term, $\dfrac{1}{z+1}$, can be written as either

$$\frac{1}{z+1} = \frac{1}{1-(-z)} = \sum_{j=0}^{\infty}(-z)^j = \sum_{j=0}^{\infty}(-1)^j z^j, \text{ for } |z| < 1, \tag{15.54}$$

because $|-z| < 1 \iff |z| < 1$, or

$$\frac{1}{z+1} = \frac{1}{z}\cdot\frac{1}{1-\left(-\frac{1}{z}\right)} = \frac{1}{z}\cdot\sum_{k=0}^{\infty}\left(-\frac{1}{z}\right)^k = \sum_{j=1}^{\infty}(-1)^{j-1}z^{-j}, \text{ for } |z| > 1, \tag{15.55}$$

because $\left|-\dfrac{1}{z}\right| < 1 \iff |z| > 1$.

The second term, $-\dfrac{2}{z-2}$, can be written as either

$$\frac{-2}{z-2} = \frac{2}{2-z} = \frac{\cancel{2}}{\cancel{2}\left(1-\frac{z}{2}\right)} = \sum_{j=0}^{\infty}\left(\frac{z}{2}\right)^j, \text{ for } |z| < 2, \tag{15.56}$$

because $\left|\frac{z}{2}\right| < 1 \iff |z| < 2$, or

$$\frac{-2}{z-2} = \frac{-2}{z}\cdot\frac{1}{1-\left(\frac{2}{z}\right)} = \frac{-2}{z}\cdot\sum_{k=0}^{\infty}\left(\frac{2}{z}\right)^k = \sum_{j=1}^{\infty}-\left(2^j\right)z^{-j}, \text{ for } |z| > 2, \tag{15.57}$$

because $\left|\frac{2}{z}\right| < 1 \iff |z| > 2$.

If we combine (15.54) and (15.56), we can write $f(z)$ as the sum of two terms, each of which converges for $|z| < 1$. If we combine (15.55) and (15.56), we can write $f(z)$ as the sum of two terms, each of which converges for $1 < |z| < 2$. If we combine (15.55) and (15.57), we can write $f(z)$ as the sum of two terms, each of which converges for $|z| > 2$. Trying to combine (15.54) and (15.57) does not work because there is no z that satisfies both $|z| < 1$ and $|z| > 2$.

Combining (15.54) and (15.56), we get

$$f(z) = \frac{1}{z+1} - \frac{2}{z-2} = \sum_{j=0}^{\infty} \left((-1)^j + \frac{1}{2^j} \right) z^j, \text{ for } |z| < 1,$$

that is, in the disk $\mathcal{D}_1(0)$.

On the other hand, combining (15.55) and (15.56), we get

$$f(z) = \frac{1}{z+1} - \frac{2}{z-2} = \sum_{j=1}^{\infty} (-1)^{j-1} z^{-j} + \sum_{j=0}^{\infty} \left(\frac{z}{2} \right)^j, \text{ for } 1 < |z| < 2,$$

that is, in the annulus $\mathcal{A}_{1,2}(0)$.

A third correct final conclusion comes from combining (15.55) and (15.57) to get

$$f(z) = \frac{1}{z+1} - \frac{2}{z-2} = \sum_{j=1}^{\infty} \left((-1)^{j-1} - 2^j \right) z^{-j}, \text{ for } |z| > 2,$$

that is, in the annulus $\mathcal{A}_{2,\infty}(0)$. ◯

In Example 15.38 we were able to express a function $f(z)$ as a Laurent series in each of the sets $\mathcal{D}_1(0)$, $\mathcal{A}_{1,2}(0)$, and $\mathcal{A}_{2,\infty}(0)$. This gives us a way to work with the function everywhere except on the circles $z = 1$ and $z = 2$. The three domains are shown in Figure 15.20. ◯

Example 15.39

Find a Laurent series for $f(z) = \sin z / z^3$.

Method: The Taylor series expansion for $\sin z$ about $z_0 = 0$ is the same as the Maclaurin series for $\sin x$ except with the real variable x replaced by the complex variable z. So,

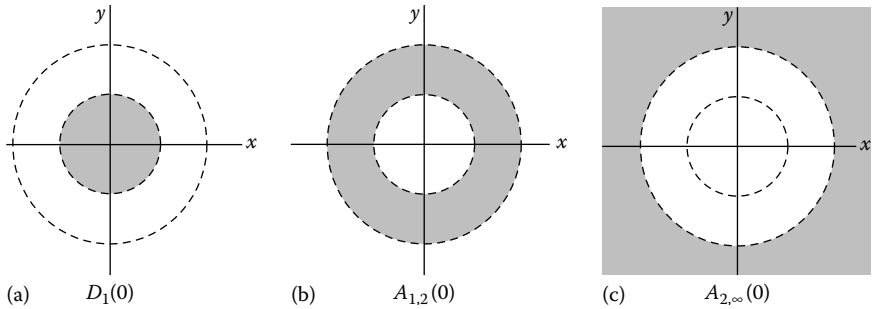


FIGURE 15.20
Example 15.38's Laurent series domains.

$$f(z) = \frac{1}{z^3} \cdot \sin z = \frac{1}{z^3} \cdot \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} z^{2k+1} = \frac{1}{z^3} \cdot \left(z - \frac{z^3}{3!} + \frac{z^5}{5!} \pm \cdots \right)$$

$$= \frac{1}{1!} z^{-2} - \frac{1}{3!} z^0 + \frac{1}{5!} z^2 - \frac{1}{7!} z^4 \pm \cdots = z^{-2} + \sum_{\ell=0}^{\infty} \frac{(-1)^{\ell+1}}{(2\ell+3)!} z^{2\ell}$$

is its Laurent series about $z_0 = 0$.

The Laurent series is undefined at $z = 0$. Because the Maclaurin series for $\sin z$ converges for all z—this follows from the ratio test, that is, Theorem 15.20–we see that the Laurent series for $f(z)$ about $z_0 = 0$ converges for all z with $|z| > 0$, that is, in the annulus $\mathcal{A}_{0,\infty}(0)$, a punctured plane. ◯

Example 15.40

Find a Laurent series for $f(z) = \dfrac{\sin z}{z}$.

Method: The Taylor series expansion for $\sin z$ about $z_0 = 0$ is the same as the Maclaurin series for $\sin x$ except with the real variable x replaced by the complex variable z. So,

$$f(z) = \frac{1}{z} \cdot \sin z = \frac{1}{z} \cdot \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} z^{2k+1} = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} z^{2k} = \frac{1}{1!} z^0 - \frac{1}{3!} z^2 + \frac{1}{5!} z^4 + \cdots$$

$$= \sum_{\ell=0}^{\infty} \frac{(-1)^{\ell}}{(2\ell+1)!} z^{2\ell}, \text{ for } 0 < |z| < \infty$$

is its Laurent series.

In fact, this Laurent series can also be called a power series because it only has nonnegative powers of $(z - z_0)$.

The Laurent series is also defined at $z = 0$. By the ratio test, that is, Theorem 15.20, we see that the Laurent series for $f(z)$ about $z_0 = 0$ converges for all z. The original function $f(z) = \frac{\sin z}{z}$ wasn't even defined at $z = 0$: nevertheless, this Laurent series is defined everywhere. ◯

We'll see in Section 15.7 that Example 15.40 illustrates the concept of a "removable" singularity.

Example 15.41

Find a Laurent series for $f(z) = \dfrac{\sin^2 z}{z}$ and its domain of convergence.

Method: It is possible to find the product of two Taylor series, but in this problem, it is easier and more effective to use the trigonometric identity

$$\sin^2 z = \frac{1 - \cos 2z}{2}.$$

Using the Maclaurin series for $\cos x$ and replacing x by $2z$, we have that

$$f(z) = \frac{1}{z} \cdot \sin^2 z = \frac{1}{z} \cdot \frac{1 - \cos 2z}{2} = \frac{1}{2z} \left(1 - \sum_{k=0}^{\infty} (-1)^k \frac{(2z)^{2k}}{(2k)!} \right) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{2^{2k-1} z^{2k-1}}{(2k)!} \tag{15.58}$$

is a Laurent series for $f(z)$. As for the Maclaurin series for $\cos x$, the ratio test shows that the Maclaurin series for $\cos z$ converges for all z. So, the Laurent series (15.58) converges for $|z| < \infty$, that is, the whole plane. While the original function is undefined at $z = 0$, its Laurent series is a continuous extension of $f(z)$.

In fact, this Laurent series can also be called a power series because it only has non-negative powers of $(z - z_0)$. ◯

15.6.3 Product of Taylor Series

If

$$f(z) = \sum_{j=0}^{\infty} a_j (z - z_0)^j \quad \text{and} \quad g(z) = \sum_{j=0}^{\infty} b_j (z - z_0)^j$$

both converge for $|z - z_0| < R$, then

$$f(z)g(z) = \sum_{j=0}^{\infty} \left(\sum_{\ell=0}^{j} a_\ell b_{j-\ell} \right) (z - z_0)^j$$

converges for $|z - z_0| < R$. This makes sense, for example, for $z_0 = 0$, because

$$\begin{aligned} f(z)g(z) &= \Big(a_0 + a_1 z + a_2 z^2 + \cdots\Big)\Big(b_0 + b_1 z + b_2 z^2 + \cdots\Big) \\ &= a_0 b_0 + (a_0 b_1 + a_1 b_0) z + (a_0 b_2 + a_1 b_1 + a_2 b_0) z^2 + \cdots . \end{aligned}$$

For example, we could calculate that

$$e^{3z} e^z = \left(\sum_{j=0}^{\infty} \frac{(3z)^j}{j!} \right) \left(\sum_{j=0}^{\infty} \frac{z^j}{j!} \right) = \sum_{j=0}^{\infty} \left(\sum_{\ell=0}^{j} \frac{3^\ell}{\ell!} \cdot \frac{1}{(j-\ell)!} \right) z^j, \tag{15.59}$$

but it is a lot easier to write

$$e^{3z} e^z = e^{4z} = \sum_{j=0}^{\infty} \frac{(4z)^j}{j!}.$$

By the way, $\displaystyle\sum_{\ell=0}^{j} a_\ell b_{j-\ell}$ is called a **discrete convolution** of the sequences of coefficients. We saw this also in Section 4.7 when studying the z-transform and in Section 9.5 when studying the discrete Fourier transform.

15.6.4 Problems

1. Find a Laurent series for $f(z) = \frac{3}{z} + \frac{1}{z-2}$ in each of the domains (a) $0 < |z| < 2$ and (b) $2 < |z|$.

2. Find a Laurent series for $f(z) = \frac{3}{z-1} + \frac{1}{z-2}$ in each of the domains (a) $1 < |z-1|$ and (b) $0 < |z-1| < 1$.
3. Find a Laurent series for $f(z) = \frac{1}{z-1} + \frac{1}{z+2}$ in each of the domains (a) $1 < |z| < 2$ and (b) $0 < |z| < 1$.

In each of problems 4–7, find a Laurent series for the given function in two domains that are disjoint, that is, have no points in common. It may help to use partial fractions.

4. $f(z) = \frac{1}{z+1} + \frac{1}{z-2}$
5. $f(z) = \frac{1}{z(z-1)}$
6. $f(z) = -\frac{5}{(z+2)(z-3)}$
7. $f(z) = -\frac{z+5}{(z+2)(2z+1)}$

In each of problems 8–10, for the given function, find a Laurent series and a domain on which it converges.

8. $f(z) = \dfrac{\cos^2 z}{z}$
9. $f(z) = z\, e^{1/z}$
10. $f(z) = \dfrac{\arcsin^2 z}{z}$. [Hint: The Maclaurin series for $\arcsin x$ is $x + \frac{1}{2}\frac{x^3}{3} + \frac{1\cdot 3}{2\cdot 4}\frac{x^5}{5} + \frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}\frac{x^7}{7} + \cdots$.]
11. A consequence of (15.59) is the identity

$$(\star) \qquad \frac{4^j}{j!} = \sum_{\ell=0}^{j} \frac{3^\ell}{\ell!} \cdot \frac{1}{(j-\ell)!},$$

for all integers $j \geq 0$. Explain $(\star)$ by applying the binomial theorem, $(x+y)^j = \sum_{\ell=0}^{j} \frac{j!}{\ell!(j-\ell)!} x^\ell y^{j-\ell}$, in the particular case of $x = 3$ and $y = 1$, that is, using the binomial theorem on the right-hand side of $4^j = (3+1)^j$. Next, generalize this to give an explanation of why the law of exponents (15.27)(i) in Section 15.4, that is, $e^{z_1+z_2} = e^{z_1}e^{z_2}$, is true.

15.7 Zeros and Poles

Definition 15.20

Suppose $f(z)$ is analytic at z_0. We say $f(z)$ has a **zero of order** m at z_0 if

$$f(z_0) = f'(z_0) = \cdots = f^{(m-1)}(z_0) = 0 \quad \text{and} \quad f^{(m)}(z_0) \neq 0\,. \tag{15.60}$$

Note that Theorem 15.40 in Section 15.9 in the following will justify that $f, f', \ldots, f^{(m)}$ exist at z_0.

Example 15.42

Find the zero(s) and their order(s) for $f(z) \triangleq (z - z_0)^2$.

Method:

$$f(z_0) = (z_0 - z_0)^2 = 0$$

$$f'(z) = 2(z - z_0), \quad f'(z_0) = 2(z_0 - z_0) = 0$$

$$f''(z) = 2, \qquad f''(z_0) = 2 \neq 0,$$

so $f(z)$ has a zero of order two at z_0. That is the only point at which $f(z) = 0$, so z_0 is the only zero of $f(z)$. ◯

Example 15.43

Find the order of the zero of $f(z) \triangleq (\sin z - z)$ at $z = 0$.

Method: The "at $z = 0$" tells us that $z_0 = 0$. We calculate

$$f(0) = \sin 0 - 0 = 0$$

$$f'(z) = \cos z - 1, \quad f'(0) = \cos 0 - 1 = 0$$

$$f''(z) = -\sin z, \qquad f''(0) = -\sin 0 = 0$$

$$f'''(z) = -\cos z, \qquad f'''(0) = -\cos 0 = -1 \neq 0,$$

so $f(z)$ has a zero of order three at $z = 0$. ◯

Theorem 15.24

(A "preparation theorem") Suppose $f(z)$ is analytic at z_0. Then $f(z)$ has a zero of order m at z_0 if, and only if,

$$f(z) = (z - z_0)^m g(z)$$

for some function $g(z)$ that both is analytic at z_0 and has $g(z_0) = \dfrac{f^{(m)}(z_0)}{m!} \neq 0$.

Why? Given that $f(z)$ has a zero of order m at z_0, the Taylor series expansion of $f(z)$ gives

$$f(z) = f(z_0) + \frac{f'(z_0)}{1!}(z - z_0) + \cdots + \frac{f^{(m-1)}(z_0)}{(m-1)!}(z - z_0)^{m-1} + \frac{f^{(m)}(z_0)}{m!}(z - z_0)^m + \cdots,$$

so knowing that $f(z)$ has a zero of order m at z_0 implies

$$\begin{aligned} f(z) &= 0 + \frac{0}{1!}(z-z_0) + \cdots + \frac{0}{(m-1)!}(z-z_0)^{m-1} + \frac{f^{(m)}(z_0)}{m!}(z-z_0)^m \\ &\quad + \frac{f^{(m+1)}(z_0)}{(m+1)!}(z-z_0)^{m+1} + \cdots \\ &= (z-z_0)^m \left(\frac{f^{(m)}(z_0)}{m!} + \frac{f^{(m+1)}(z_0)}{(m+1)!}(z-z_0) + \cdots \right) = (z-z_0)^m g(z), \end{aligned}$$

where

$$g(z) \triangleq \frac{f^{(m)}(z_0)}{m!} + \frac{f^{(m+1)}(z_0)}{(m+1)!}(z-z_0) + \cdots$$

converges and satisfies

$$g(z_0) = \frac{f^{(m)}(z_0)}{m!} + \frac{f^{(m+1)}(z_0)}{(m+1)!}(0) + \cdots = \frac{f^{(m)}(z_0)}{m!} \neq 0. \square$$

15.7.1 Singularities

Here we will study the ways a function can fail to be analytic at z_0.

Definition 15.21

$f(z)$ has an **isolated singularity** at z_0 if $f(z)$ is not analytic at z_0 but is analytic on some punctured disk $\mathcal{A}_{0,r}(z_0) = \{z : 0 < |z - z_0| < r\}$.

Example 15.44

By Theorem 15.4 in Section 15.2, $f(z) \triangleq \dfrac{1}{(z-1)^2}$ is not analytic at $z = 1$ but is analytic for $0 < |z-1| < \infty$, so $f(z)$ has an isolated singularity at $z = 1$. ◯

Example 15.45

$f(z) \triangleq \frac{\sin z}{z}$ is not analytic at $z = 0$ because it is undefined at $z = 0$. But, by Theorem 15.6(d) in Section 15.2, $f(z)$ is analytic for $0 < |z| < \infty$, so $f(z)$ has an isolated singularity at $z = 0$. ◯

Example 15.46

$f(z) \triangleq e^{-1/z}$ is not analytic at $z = 0$ because it is undefined there. But $f(z)$ is analytic for $0 < |z| < \infty$, because $f(z) = g(h(z))$, where $h(z) \triangleq -\dfrac{1}{z}$ is analytic everywhere except at $z = 0$, and $g(z) \triangleq e^z$ is analytic everywhere. So $f(z)$ has an isolated singularity at $z = 0$. ◯

Definition 15.22

Suppose $f(z)$ has an isolated singularity at z_0 and, in some punctured disk $\mathcal{A}_{0,r}(z_0)$, $f(z)$ has Laurent series

$$f(z) \triangleq \sum_{j=1}^{\infty} a_{-j}\,(z - z_0)^{-j} + \sum_{j=0}^{\infty} a_j\,(z - z_0)^j.$$

(a) If $a_{-j} = 0$, for all $j \geq 1$, then we say $f(z)$ has a **removable singularity** at z_0.
(b) If $a_{-j} \neq 0$ for infinitely many $j \geq 1$, then we say $f(z)$ has an **essential singularity** at z_0.
(c) If $a_{-j} = 0$ for all $j > m \geq 1$ *and* $a_{-m} \neq 0$, then we say $f(z)$ has a **pole of order** m at z_0.

Example 15.47

$f(z) \triangleq \dfrac{\sin z}{z}$ has what kind of singularity at $z_0 = 0$?

Method: From Example 15.45 we know $f(z)$ has an isolated singularity at $z_0 = 0$. From Example 15.40 in Section 15.6, the Laurent series of $f(z)$ about $z_0 = 0$ is

$$f(z) = \frac{1}{1!}z^0 - \frac{1}{3!}z^2 + \frac{1}{5!}z^4 + \cdots, \text{ for } 0 < |z| < \infty.$$

By definition, $f(z)$ has a removable singularity at $z_0 = 0$. ◯

The phrase "removable singularity" suggests that there exists some sort of "removal process." Indeed, for Example 15.47, we can define an extension of $f(z)$ by

$$\widetilde{f}(z) \triangleq \begin{Bmatrix} f(z), & \text{if } z \neq 0 \\ 1, & \text{if } z = 0 \end{Bmatrix}.$$

We choose $\widetilde{f}(0) = 1$ because L'Hôpital's rule, that is, Theorem 15.12 in Section 15.3, calculates that

$$\lim_{z\to 0} f(z) = \lim_{z\to 0} \frac{\sin z}{z} = \lim_{z\to 0} \frac{(\sin z)'}{(z)'} = \lim_{z\to 0} \frac{\cos z}{1} = 1.$$

Theorem 15.25

(An "extension theorem" or "singularity removal process") Suppose $f(z)$ has a removable singularity at z_0 and a Laurent series (15.53) in Section 15.6 for z in some punctured disk $\mathcal{A}_{0,r}(z_0)$. Then

$$\widetilde{f}(z) \triangleq \begin{cases} f(z), & \text{if } z \neq z_0 \\ a_0, & \text{if } z = z_0 \end{cases}$$

defines an extension of $f(z)$ that is analytic on the disk $\mathcal{D}_r(z_0)$.

Why? The reasoning is similar to what we used when removing the singularity of $\dfrac{\sin z}{z}$ at $z = 0$. □

Example 15.48

$f(z) \triangleq e^{-1/z}$ has what kind of singularity at $z_0 = 0$?

Method: From Example 15.46, we know $f(z)$ has an isolated singularity at $z = 0$. Using the Maclaurin series of e^z and replacing z by $-\dfrac{1}{z}$, we see that $f(z)$ has a Laurent series

$$f(z) = \exp\left(-\frac{1}{z}\right) = \sum_{n=0}^{\infty} \frac{1}{n!}\left(-\frac{1}{z}\right)^n = 1 + \sum_{j=0}^{\infty} \frac{(-1)^j}{j!} z^{-j}.$$

Because $a_{-j} = \frac{(-1)^j}{j!} \neq 0$ for infinitely many $j \geq 1$, in fact all $j \geq 1$, $f(z)$ has an essential singularity at $z_0 = 0$. ◯

Example 15.49

$f(z) \triangleq \dfrac{1}{(z-1)^2}$ has what kind of singularity at $z = 1$?

Method: From Example 15.44, we know $f(z)$ has an isolated singularity at $z = 1$. This function "is its own Laurent series," that is,

$$f(z) = \cdots + 0 \cdot (z-1)^{-4} + 0 \cdot (z-1)^{-3} + 1(z-1)^{-2} + 0 \cdot (z-1)^{-1} + 0 \cdot (z-1)^0 + 0 \cdot (z-1)^1 + \cdots,$$

for all $z \neq 1$. $f(z)$ fits the definition of having a pole of order two at $z_0 = 1$. ◯

It is often easier to find the location of a pole and its order by studying the zeros of the denominator and numerator of a function. Intuitively, the "algebra of poles" is the reciprocal of the "algebra of zeros."

Theorem 15.26

Suppose m is a positive integer. Any one of the following "data" is enough to imply that $f(z)$ has a pole of order m at z_0:

(a) $f(z) = \dfrac{g(z)}{(z - z_0)^m}$, where $g(z)$ is analytic at z_0 and has $g(z_0) \neq 0$.

(b) $f(z) = \dfrac{g(z)}{h(z)}$, where $h(z)$ has a zero of order m at z_0, $g(z_0) \neq 0$, and both $g(z)$ and $h(z)$ are analytic at z_0.

(c) $f(z) = f_1(z) + f_2(z)$, where $f_1(z)$ has a pole of order m at z_0 and $f_2(z)$ is analytic at z_0 or has a pole of order less than m at z_0.

Why? (a) The assumption that $g(z)$ is analytic at z_0 guarantees that $g(z)$ has Taylor series given as in (15.48) in Section 15.6, that is,

$$g(z) = \sum_{j=0}^{\infty} c_j (z - z_0)^j$$

that converges in some open disk $\mathcal{D}_r(z_0)$. We also assumed that $c_0 = g(z_0) \neq 0$. So,

$$\begin{aligned} f(z) = \frac{g(z)}{(z - z_0)^m} &= (z - z_0)^{-m} \left(c_0 + c_1(z - z_0) + c_2 (z - z_0)^2 + \cdots \right) \\ &= c_0(z - z_0)^{-m} + c_1(z - z_0)^{-m+1} + c_2 (z - z_0)^{-m+2} + \cdots . \end{aligned}$$

According to Definition 15.22(c), $f(z)$ has a pole of order m at z_0.

(b) and (c) will be in your homework Problems 15.7.2.11 and 15.7.2.12, respectively. □

Example 15.50

Find the order of the pole $f(z) \triangleq \dfrac{z - 5}{(\sin z)^2}$ has at $z = 0$.

Method: Let $h(z) \triangleq (\sin z)^2$. In order to use Theorem 15.26(b), we will need to find the order of the zero of $h(z)$ at $z = 0$. We calculate

$$\begin{aligned} & & h(0) &= \sin^2 0 = 0 \\ h'(z) &= 2 \sin z \cos z = \sin 2z, & h'(0) &= \sin(2 \cdot 0) = 0 \\ h''(z) &= 2 \cos 2z, & h''(0) &= 2 \cos 0 = 2 \neq 0, \end{aligned}$$

so $h(z)$ has a zero of order two at $z = 0$. Because $g(z) \triangleq z - 5$ is analytic at $z = 0$ and has $g(0) = -5 \neq 0$, Theorem 15.26(b) implies that the original function, $f(z) \triangleq \dfrac{g(z)}{h(z)}$, has a pole of order two at $z = 0$. ◯

Theorem 15.27

(A second "preparation theorem") Suppose $f(z)$ has an isolated singularity at z_0. Then $f(z)$ has a pole of order m at z_0 if, and only if,

$$f(z) = (z - z_0)^{-m} k(z)$$

for some function $k(z)$ that is both analytic at z_0 and has $k(z_0) \neq 0$.

Why? By definition, $f(z)$ has a pole of order m at z_0 when for z in some punctured disk $\mathcal{A}_{0,r}(z_0)$, it has a Laurent series (15.53) in Section 15.6 with $a_{-j} = 0$ for all $j > m$ *and* $a_{-m} \neq 0$. In "explicit" form,

$$\begin{aligned} f(z) &= \cdots + 0 + \frac{a_{-m}}{(z-z_0)^m} + \frac{a_{-m+1}}{(z-z_0)^{m-1}} + \cdots + a_0 + a_1(z-z_0) + a_2(z-z_0)^2 + \cdots \\ &= (z-z_0)^{-m}\Big(a_{-m} + a_{-m+1}(z-z_0) + a_{-m+2}(z-z_0)^2 + \cdots\Big) = (z-z_0)^{-m}k(z). \end{aligned}$$

Note that $k(z_0) = a_{-m} + a_{-m+1}(z_0 - z_0) + \cdots = a_{-m} + 0 + \cdots = a_{-m} \neq 0$. □

To make Theorem 15.26 more useful, it helps to have corresponding results about zeros.

Theorem 15.28

Suppose $m > 0$ and n are integers:

(a) If $g(z)$ and $h(z)$ are analytic at z_0, $g(z)$ has a zero of order m at z_0, and $h(z)$ has a zero of order n at z_0, then the product $f(z) \triangleq g(z)h(z)$ has a zero of order $(m+n)$ at z_0.

(b) If $g(z)$ has a pole of order m at z_0 and $h(z)$ is analytic and has a zero of order n at z_0, then
   (i) if $m > n$, then the product $f(z) \triangleq g(z)h(z)$ has a pole of order $(m-n)$ at z_0;
   (ii) if $m = n$, then $f(z) \triangleq g(z)h(z)$ has a removable singularity at z_0;
   (iii) if $m < n$, then $f(z)$ has a removable singularity at z_0 and an extension of $f(z) \triangleq g(z)h(z)$ has a zero of order $(n-m)$ at z_0.

Why? For example, for (b)(i), if $g(z)$ has a pole of order m at z_0, then the second "preparation" Theorem 15.27 implies that $g(z) = (z-z_0)^{-m}k(z)$ for some function $k(z)$ that both is analytic at z_0 and has $k(z_0) \neq 0$. Also, if $h(z)$ is analytic and has a zero of order n at z_0, then the first preparation Theorem 15.24 implies that $h(z) = (z-z_0)^n\ell(z)$ for some function $\ell(z)$ that is analytic at z_0 and has $\ell(z_0) \neq 0$. It follows that

$$f(z) = g(z)h(z) = (z-z_0)^{-m}k(z) \cdot (z-z_0)^n\ell(z) = (z-z_0)^{-(m-n)} \cdot k(z)\ell(z).$$

Because $k(z_0)\ell(z_0) \neq 0$, the second preparation theorem implies that $f(z)$ has a pole of order $(m-n)$ at z_0, as we wanted to explain.

As for explaining the other results in Theorem 15.28, that is left to you in Problems 15.7.2.13 through 15.7.2.15. □

Corollary 15.1

If $g(z)$ has a pole of order m at z_0, then

$$\widetilde{\frac{1}{g(z)}} \triangleq \begin{cases} \dfrac{1}{g(z)}, & z \neq z_0 \\ 0, & z = z_0 \end{cases}$$

has a zero of order m at z_0.

Why? You will explain this in Problem 15.7.2.16. □

15.7.2 Problems

1. Find a function that has a pole of order one at each of $z = \pm i$ and a pole of order three at $z = -1$.

In problems 2 and 3, what is the order of the pole of the given function $f(z)$ at the given point z_0?

2. $f(z) = \dfrac{1}{e^z - 1 - z}$ at $z_0 = 0$?
3. $f(z) = \dfrac{1}{(e^z - 1)^2}$ at $z_0 = 0$?

In problems 4–6, find and classify the isolated singularities of the given function.

4. $f(z) = \dfrac{1}{z^2(z+1)}$.
5. $f(z) = \dfrac{1}{e^z + 1}$.
6. $f(z) = \dfrac{z}{z^2 + 4} + \dfrac{1}{z}$.
7. Construct a function $f(z)$ that has a zero of order one at each of $z = \pm i$ and a pole of order three at $z = -1$.
8. Explain why $f'(z)$ has a zero of order $m - 1$ at z_0 if $f(z)$ has a zero of order m at z_0, assuming m is an integer ≥ 2. What happens if $m = 1$?
9. Explain why $f'(z)$ has a pole of order $m + 1$ at z_0 if $f(z)$ has a pole of order m at z_0, assuming m is a positive integer.
10. Suppose that $R > 0$ and for $k = 1, ..., 4$ each of the functions $f_k(z)$ has a zero of order two at z_0, is analytic at z_0, and is nonzero for $0 < |z - z_0| < R$. Define

$$h(z) = \det\left(\begin{bmatrix} f_1(z) & f_2(z) \\ f_3(z) & f_4(z) \end{bmatrix}\right)$$

and $H(z) = \dfrac{f_4(z)}{h(z)}$. Investigate the nature of the singularity of $H(z)$ at z_0.

11. Explain why Theorem 15.26(b) is true.
12. Explain why Theorem 15.26(c) is true.
13. Explain why Theorem 15.28(a) is true.
14. Explain why Theorem 15.28(b)(ii) is true.
15. Explain why Theorem 15.28(b)(iii) is true.
16. Explain why Corollary 15.1 is true.

15.8 Complex Integration and Cauchy's Integral Theorem

In the complex plane, line integrals of functions of a complex variable have properties that are wonderful and more amazing than we would have expected from our experience with line integrals of vector fields in $\mathbb{R}^2$ or $\mathbb{R}^3$.

Definition 15.23

In the complex plane,

$$\mathcal{C} : z = z(t) = x(t) + iy(t),\ a \leq t \leq b$$

is a **parametrized curve** if $z(t)$ is continuous on interval $[a, b]$. In this case, $z(a)$ is the **initial point** and $z(b)$ is the **terminal point** of $\mathcal{C}$. A parametrized curve is **piecewise smooth** if $x(t)$ and $y(t)$ are piecewise smooth functions of the real variable t in the interval $[a, b]$. A parametrized curve is **closed** if $z(a) = z(b)$. A parametrized curve is **simple** if $z(s) \neq z(t)$ for all s, t with $a \leq s < t \leq b$ except possibly $z(a) = z(b)$.

Recall from Section 7.2 our definition that a simple, closed, parametrized curve in the xy-plane is positively oriented if, while standing in the xy-plane with head in the $\hat{k}$ direction, the inside of the curve is on our left as we travel on the curve in the direction of its parametrization. While this definition is admittedly imprecise, it will suffice for our needs. We will say $\mathcal{C} : z = z(t) = x(t) + iy(t),\ a \leq t \leq b$, a simple, closed, parametrized curve in the complex plane, is **positively oriented** if $\mathcal{C} : \mathbf{r}(t) = x(t)\hat{\boldsymbol{\imath}} + y(t)\hat{\boldsymbol{\jmath}}$, the corresponding curve in $\mathbb{R}^2$, is positively oriented.

We may abuse the definitions by referring to $\mathcal{C}$ as both (1) a parametrized curve, $\mathcal{C} : z = z(t),\ a \leq t \leq b$, and (2) the set $\{z(t) : a \leq t \leq b\}$, that is, the set consisting of all of the points $z(t),\ a \leq t \leq b$.

Definition 15.24

In the complex plane, a **contour** is a chain of parametrized, piecewise smooth curves $\mathcal{C}_1, \ldots, \mathcal{C}_N$ where the initial point of $\mathcal{C}_{k+1}$ is the terminal point of $\mathcal{C}_k : z = z_k(t),\ a_k \leq t \leq b_k$, for $k = 1, \ldots, N-1$. We denote such a contour by $\mathcal{C}_1 + \cdots + \mathcal{C}_N$. A contour is **simple** if

for $k \neq \ell$, $\mathcal{C}_k$ and $\mathcal{C}_\ell$ have no points in common except that the initial point of $\mathcal{C}_{k+1}$ is the terminal point of $\mathcal{C}_k$ and, possibly, $z_1(a_1) = z_N(b_N)$. If the latter is true, then we say the contour is **closed**.

A simple, closed, contour in the complex plane will be called **positively oriented** if $\mathcal{C}_1 + \cdots + \mathcal{C}_N$, its chain of parametrized curves, $\mathcal{C}_k : z = z_k(t)$, $a_k \leq t \leq b_k$, $k = 1, \ldots, N$, is positively oriented when rewritten as a single parametrized curve:

$$z(t) = \left\{ \begin{array}{rl} z_1(t), & \text{if } a_1 \leq t \leq b_1 \\ z_2(t + a_2 - b_1), & \text{if } b_1 \leq t \leq b_1 + b_2 - a_2 \\ \vdots & \\ z_N\left(t + \sum_{k=1}^{N-1}(a_{k+1} - b_k)\right), & \text{if } b_1 + \sum_{k=2}^{N-1}(b_k - a_k) \leq t \leq b_1 + \sum_{k=2}^{N}(b_k - a_k) \end{array} \right\}.$$

An example of a contour is depicted in Figure 15.21.

We may abuse the definitions by referring to $\mathcal{C}$ as both (1) a contour, that is, a chain of parametrized curves, $\mathcal{C}_k : z = z_k(t)$, $a_k \leq t \leq b_k$, $k = 1, \ldots, N$, and (2) the set union of all of the points $z_k(t)$, $a_k \leq t \leq b_k$, $k = 1, \ldots, N$.

We will define a line integral in the complex plane analogous to the line integral of a vector field over a curve in $\mathbb{R}^2$. Given a parametrized curve $\mathcal{C} : z = z(t) = x(t) + iy(t)$, $a \leq t \leq b$, partition the interval $[a, b]$ into $a = t_0 < t_1 < \cdots < t_n = b$. This gives an approximation of $\mathcal{C}$ by a polygonal curve* consisting of directed line segments $z(t_0) \dashrightarrow z(t_1) \dashrightarrow \ldots \dashrightarrow z(t_n)$, as shown in Figure 15.22.

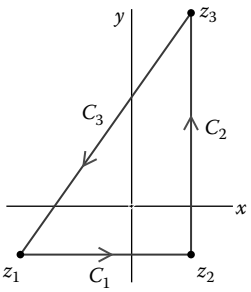


FIGURE 15.21
Contour $\mathcal{C}_1 + \mathcal{C}_2 + \mathcal{C}_3$.

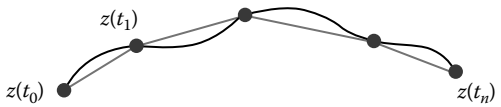


FIGURE 15.22
Polygonal curve approximation.

* See Definition 15.8(b) in Section 15.3.

Define $\triangle z_k = z(t_k) - z(t_{k-1})$. Given a function $f(z)$, we can form the Riemann sum $\sum_{k=1}^{n} f(z(t_k)) \triangle z_k$. If we can take its limit as both $n \to \infty$ and $\left(\max_{1 \le k \le n} |\triangle z_k|\right) \to 0$, then we can define the line integral of a function of a complex variable over a curve in the complex plane by

$$\int_{\mathcal{C}} f(z)\, dz \triangleq \lim_{n \to \infty} \sum_{k=1}^{n} f(z(t_k)) \triangle z_k \,. \tag{15.61}$$

Because both z and $f(z)$ have real and imaginary components, the concept defined in (15.61) is like that of a line integral of a vector field rather than of a scalar field.

Theorem 15.29

Suppose $\mathcal{C}: z = z(t) = x(t) + iy(t)$, $a \le t \le b$, is a piecewise smooth, parametrized curve and $f(z(t))$ is a continuous function of t in the interval $[a, b]$. Then there exists $\int_{\mathcal{C}} f(z)\, dz$ as defined in (15.61) and it can be calculated using

$$\int_{\mathcal{C}} f(z)\, dz = \int_{a}^{b} f(z(t)) \frac{dz}{dt}(t)\, dt \,. \tag{15.62}$$

Similar to the result of Problem 7.2.5.22, the line integral of a function of a complex variable over a curve in the complex plane does not depend on the parametrization of the curve.

Corollary 15.2

Suppose the function $f(z)$ is continuous at every point on a contour $\mathcal{C}_1 + \cdots + \mathcal{C}_N$. Then there exists

$$\int_{\mathcal{C}_1 + \cdots + \mathcal{C}_N} f(z)\, dz \triangleq \int_{\mathcal{C}_1} f(z)\, dz + \cdots + \int_{\mathcal{C}_N} f(z)\, dz \,. \tag{15.63}$$

Example 15.51

Evaluate $\int_{\mathcal{C}} \frac{1}{z-1}\, dz$ where the circle $\mathcal{C}$ is shown in Figure 15.23.

Method: The circle $\mathcal{C}$ can be parametrized by

$$\mathcal{C}: z = z(t) = (1 + 2\cos t) + i2\sin t = 1 + 2e^{it},\ 0 \le t \le 2\pi,$$

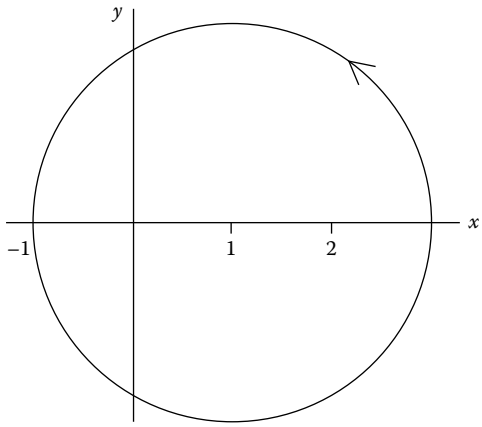


FIGURE 15.23
Example 15.51.

so (15.62) calculates that

$$\int_{\mathcal{C}} \frac{1}{z-1}\,dz = \int_0^{2\pi} \frac{1}{z(t)-1}\,z'(t)\,dt = \int_0^{2\pi} \frac{1}{\cancel{1}+2\cos t+i2\sin t-\cancel{1}}(-2\sin t+i2\cos t)\,dt$$

$$= \int_0^{2\pi} \frac{-2\sin t+i2\cos t}{2\cos t+i2\sin t}\,dt = \int_0^{2\pi} \frac{i(2\cos t+i2\sin t)}{2\cos t+i2\sin t}\,dt = \int_0^{2\pi} i\,dt = 2\pi i. \bigcirc$$

By the way, a streamlined version of the earlier calculation is

$$\int_{\mathcal{C}} \frac{1}{z-1}\,dz = \int_0^{2\pi} \frac{1}{z(t)-1}\,z'(t)\,dt = \int_0^{2\pi} \frac{1}{(\cancel{1}+2e^{it})-\cancel{1}}\,(i2e^{it})dt = \int_0^{2\pi} \frac{1}{\cancel{2e^{it}}}\,(i\cancel{2e^{it}})dt$$

$$= \int_0^{2\pi} i\,dt = 2\pi i.$$

We will see that Example 15.51 and related calculations are very useful.

Theorem 15.30

$$\int_{\mathcal{C}} \frac{1}{(z-z_0)^m}\,dz = \begin{Bmatrix} 2\pi i, & m=1 \\ 0, & m\neq 1 \end{Bmatrix}, \tag{15.64}$$

where m is any integer and $\mathcal{C}$ is any positively oriented circle centered at z_0.

Why? You will establish the case $m = 1$ in Problem 15.8.3.8; it's similar to the calculation of Example 15.51. For integer $m \neq 1$ and circle $\mathcal{C}: z(t) = z_0 + re^{it}$, $0 \leq t \leq 2\pi$,

$$\int_{\mathcal{C}} \frac{1}{(z-z_0)^m}\,dz = \int_0^{2\pi} \frac{1}{\left(\cancel{z_0} + re^{it} - \cancel{z_0}\right)^m}\left(ire^{it}\right)dt = \int_0^{2\pi} r^{-m}e^{it(-m)}\,(ir)e^{it}dt$$

$$= i\,r^{1-m}\int_0^{2\pi} e^{it(1-m)}\,dt = i\,r^{1-m}\left[\frac{e^{it(1-m)}}{i(1-m)}\right]_0^{2\pi}$$

$$= i\,r^{1-m}\left(\cancel{\frac{1}{i(1-m)}} - \cancel{\frac{1}{i(1-m)}}\right) = 0.\ \square$$

Example 15.52

Let $\mathcal{C}$ be the straight line from $z = 1$ to $z = i$. Evaluate $\int_{\mathcal{C}} \overline{z}\,dz$.

Method: $\mathcal{C}$ can be parametrized by $z = z(t) = 1 + (i-1)t$, $0 \le t \le 1$. On $\mathcal{C}$, $z'(t) = i - 1$ and $\overline{z(t)} = \overline{1 - t + it} = 1 + (-i-1)t$, so, by (15.62),

$$\int_{\mathcal{C}} \overline{z}\,dz \triangleq \int_0^1 \overline{z(t)}\,z'(t)\,dt = \int_0^1 (1 + (-i-1)t)\,(i-1)\,dt$$

$$= \int_0^1 (i - 1 + 2t)\,dt = \left[\,(i-1)t + t^2\,\right]_0^1 = (i-1) + 1 = i.\ \bigcirc$$

Definition 15.25

Given a parametrized curve $\mathcal{C}:\ z = z(t), a \le t \le b$, the **opposite** curve is the parametrized curve defined by $-\mathcal{C}:\ z = z(a + b - t), a \le t \le b$.

Theorem 15.31

If $\int_{\mathcal{C}} f(z)\,dz$ exists, then so does the integral over the opposite parametrized curve, and

$$\int_{-\mathcal{C}} f(z)\,dz = -\int_{\mathcal{C}} f(z)\,dz\,. \tag{15.65}$$

So, in this respect, integration on a parametrized curve in the complex plane behaves like the line integral of a vector field.

15.8.1 Integration on a Closed Contour

We will study integration on closed contours and learn about deep connections with singularities of functions. First, recall the definition of "connected" from Section 15.3. Next, we define a set A in the complex plane to be **bounded** if there is an $M > 0$ such that $|z| \le M$ for all z in A. For example, the square $\{x + iy:\ -1 \le x \le 1, -1 \le y \le 1\}$ is bounded with $M = \sqrt{2}$.

First, we need a major theorem of pure mathematics.

Theorem 15.32

(Jordan curve theorem) A simple closed contour $\mathcal{C}$ separates the complex plane into two open connected sets, $\mathcal{U}_i$ and $\mathcal{U}_o$ with the properties that (i) $\mathbb{C}$ equals the disjoint union $\mathcal{U}_i \cup \mathcal{C} \cup \mathcal{U}_o$, (ii) $\mathcal{U}_i$ is bounded, and (iii) $\mathcal{U}_o$ is not bounded.

We call the bounded set $\mathcal{U}_i$ the **interior** of $\mathcal{C}$, we say $\mathcal{U}_i$ is **inside** $\mathcal{C}$, and we call the unbounded set $\mathcal{U}_o$ the **exterior** of $\mathcal{C}$.

Example 15.53

$\mathcal{C}: z = 1 + 2e^{it}, 0 \le t \le 2\pi$ has positive orientation, and the interior of $\mathcal{C}$ is the open disk $\mathcal{D}_2(1)$, as shown in Figure 15.23. ○

Recall from Definition 15.9 in Section 15.3 that a "domain" is an open, connected set.

Definition 15.26

A function $f(z)$ has an **antiderivative** $F(z)$ on a domain $\mathcal{D}$ if $F'(z) = f(z)$ for all z in $\mathcal{D}$.

Theorem 15.33

Suppose $f(z)$ has an antiderivative $F(z)$ on a domain $\mathcal{D}$ and $\mathcal{C}: z = z(t)$, $a \le t \le b$ is a contour contained in $\mathcal{D}$. Then

$$\int_{\mathcal{C}} f(z)\,dz = F(z(b)) - F(z(a))\,. \tag{15.66}$$

Why? The explanation is the same as for the fundamental theorem of line integrals, Theorem 7.8 in Section 7.2.

Example 15.54

Evaluate $\int_{\mathcal{C}} \frac{1}{z}\,dz$ where $\mathcal{C}$ is shown in Figure 15.24a.

Method: Recall from Theorem 15.17 that $d/dz\left[Log(z)\right] = 1/z$ for all z in any domain $\mathcal{D}$ as long as *no* point on the nonpositive real axis is in $\mathcal{D}$. An example of such a domain $\mathcal{D}$ is the interior of the dashed ellipse illustrated in Figure 15.24b. So, Theorem 15.33 allows us to calculate

$$\int_{\mathcal{C}} \frac{1}{z}\,dz = Log(i3) - Log(2) = \left(\ln 3 + i\,\frac{\pi}{2}\right) - \left(\ln 2 + i0\right) = \ln\left(\frac{3}{2}\right) + i\,\frac{\pi}{2}. \text{ ○}$$

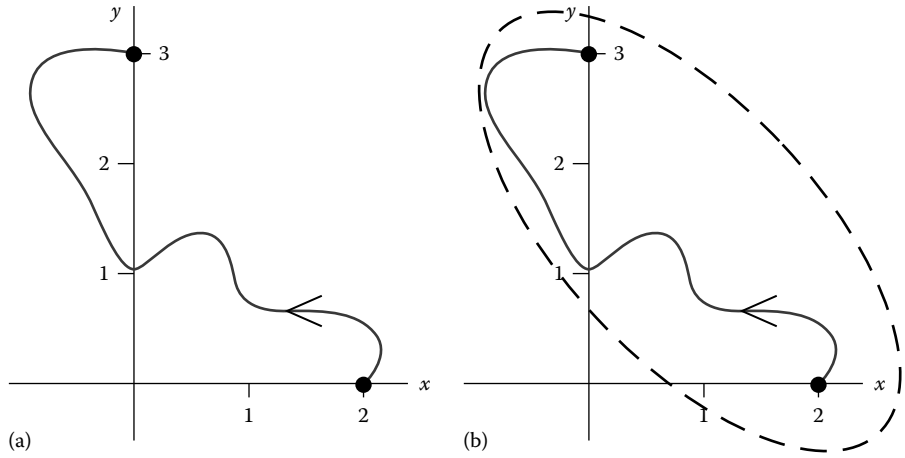


FIGURE 15.24
(a) $\mathcal{C}$, (b) $\mathcal{C}$ in $\mathcal{D}$.

Corollary 15.3

Suppose $f(z)$ has an antiderivative $F(z)$ on a domain $\mathcal{D}$ and $\mathcal{C}$ is a simple *closed* contour contained in $\mathcal{D}$. Then $\int_{\mathcal{C}} f(z)\,dz = 0$.

Why? Because $F(z)$ is an antiderivative for $f(z)$ on $\mathcal{D}$, $\int_{\mathcal{C}} f(z)\,dz = F(z(b)) - F(z(a)) = F(z(a)) - F(z(a)) = 0$, where $\mathcal{C}$ has the same initial as terminal point $z(b) = z(a)$. □

Theorem 15.34

(Path independence) Suppose $f(z)$ has an anti-derivative $F(z)$ on a domain $\mathcal{D}$, z_I and z_T are in $\mathcal{D}$, and contained in $\mathcal{D}$ are any two contours $\mathcal{C}_1$ and $\mathcal{C}_2$ both having initial point z_I and terminal point z_T. Then

$$\int_{\mathcal{C}_2} f(z)\,dz = \int_{\mathcal{C}_1} f(z)\,dz\,. \tag{15.67}$$

Why? By Theorem 15.33,

$$\int_{\mathcal{C}_2} f(z)\,dz = F(z_T) - F(z_I) = \int_{\mathcal{C}_1} f(z)\,dz.\ \square$$

15.8.2 Cauchy–Goursat Integral Theorem

Here we will learn about a basic but powerful result for calculating contour integrals. First, we need a definition.

Definition 15.27

A **deformation** in domain $\mathcal{D}$ of a contour $\mathcal{C}_0$ to a contour $\mathcal{C}_1$, denoted by $\mathcal{C}_0 \leftrightsquigarrow \mathcal{C}_1$ or $z(t,s) : \mathcal{C}_0 \leftrightsquigarrow \mathcal{C}_1$, is a map

$$z = z(t,s), \quad a \le t \le b, \quad 0 \le s \le 1$$

that satisfies

$$\left\{ \begin{array}{l} z_0(t) \triangleq z(t,0),\ a \le t \le b,\ \text{parametrizes } \mathcal{C}_0 \\ z_1(t) \triangleq z(t,1),\ a \le t \le b,\ \text{parametrizes } \mathcal{C}_1 \\ z(t,s) \text{ is continuous in } (t,s) \text{ for } a \le t \le b,\ 0 \le s \le 1 \\ z(t,s) \text{ remains in } \mathcal{D} \text{ for all } a \le t \le b,\ 0 \le s \le 1 \end{array} \right\}.$$

We say $\mathcal{C}_0$ **can be deformed in** $\mathcal{D}$ **to** $\mathcal{C}_1$ if there exists a deformation $\mathcal{C}_0 \leftrightsquigarrow \mathcal{C}_1$ in $\mathcal{D}$.

Figure 15.25 illustrates a deformation with the arrows and dotted lines depicting the process of deforming $\mathcal{C}_0$ to $\mathcal{C}_1$. This specific deformation is

$$z(t,s) = \frac{1}{\sqrt{2}} \left((2-s)^2 \cos t + (2-s) \sin t + i \left(-(2-s)^2 \cos t + (2-s) \sin t \right) \right), 0 \le t \le 2\pi.$$

Note that if there exists a deformation $z(t,s) : \mathcal{C}_0 \leftrightsquigarrow \mathcal{C}_1$ in $\mathcal{D}$, then there exists a deformation $z(t, 1-s) : \mathcal{C}_1 \leftrightsquigarrow \mathcal{C}_0$ in $\mathcal{D}$.

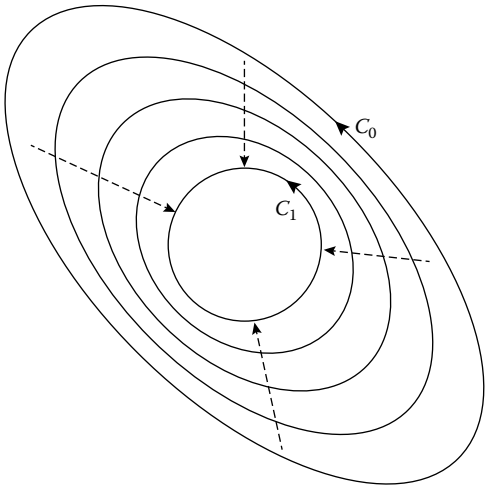


FIGURE 15.25
Deformation of $\mathcal{C}_0$ to $\mathcal{C}_1$.

Theorem 15.35

(Deformation theorem) Suppose $f(z)$ is analytic on a domain $\mathcal{D}$ and contour $\mathcal{C}_0$ can be deformed in $\mathcal{D}$ to contour $\mathcal{C}_1$. Then

$$\int_{\mathcal{C}_0} f(z)\,dz = \int_{\mathcal{C}_1} f(z)\,dz\,. \tag{15.68}$$

Definition 15.28

A domain $\mathcal{D}$ is **simply connected** if every simple, closed contour contained in $\mathcal{D}$ can be deformed in $\mathcal{D}$ to a single point.

Every open disk is simply connected, but every annulus and punctured disk is not simply connected.

Theorem 15.36

(Cauchy–Goursat integral theorem) Suppose $f(z)$ is analytic on a simply connected domain $\mathcal{D}$ and $\mathcal{C}$ is a simple, closed contour contained in $\mathcal{D}$. Then

$$\int_{\mathcal{C}} f(z)\,dz = 0.$$

Why? Given a simple closed contour $\mathcal{C}$ contained in simply connected domain $\mathcal{D}$ where $f(z)$ is analytic, find a deformation $\mathcal{C} \leftrightsquigarrow \{z_0\}$ in $\mathcal{D}$, that is, deformation to the parametrized curve $\mathcal{C}_1 : z = z(t) \equiv z_0$, $a \le t \le b$. By the Deformation Theorem 15.35,

$$\int_{\mathcal{C}} f(z)\,dz = \int_{\{z_0\}} f(z)\,dz = \int_a^b f\,(z(t)) \cdot 0\,dt = 0. \;\square$$

Example 15.55

Evaluate $\oint_{|z-2|=1} \dfrac{1}{z}\,dz$.

Method: Recall that the arrow direction in the symbol $\oint$ indicates that the parametrization of the curve is positively oriented. Define a parametrized curve by $\mathcal{C} : z = z(t) = 2 + e^{it}$, $0 \le t \le 2\pi$, and a domain by $\mathcal{D} \triangleq \mathcal{D}_{1+\varepsilon}(2)$. Then $f(z) \triangleq \dfrac{1}{z}$ is analytic in $\mathcal{D}$ as long as $\varepsilon < 1$, and $\mathcal{C}$ is contained in $\mathcal{D}$ as long as $\varepsilon > 0$. Choose $\varepsilon = \frac{1}{2}$, for example, to see that Theorem 15.36 implies that $\int_{\mathcal{C}} f(z)\,dz = 0$. $\bigcirc$

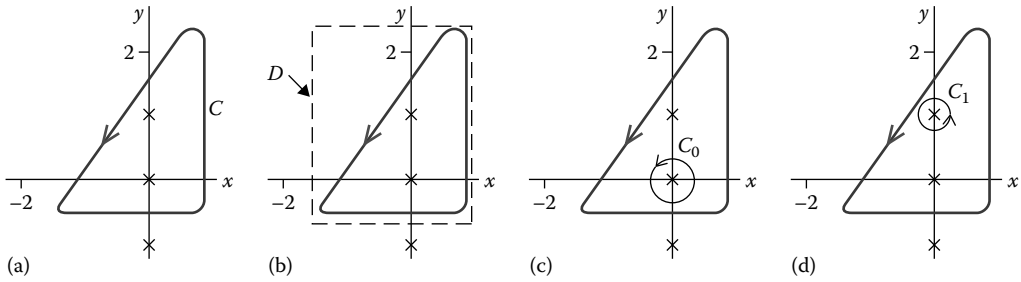


FIGURE 15.26
Example 15.56: using theorems.

Example 15.56

Evaluate $\oint_{\mathcal{C}} \dfrac{2z+i}{z^3+z}\,dz$, where $\mathcal{C}$ is shown in Figure 15.26a.

Method: The ×'s in Figure 15.26a show the location of the singularities of $f(z) \triangleq \dfrac{2z+i}{z^3+z}$; everywhere else $f(z)$ is analytic. We will use a combination of Theorems 15.30, 15.35, and 15.36 to evaluate the integral.

First, it helps to get the partial fractions expansion of $f(z)$. Because the denominator is $z^3+z = z(z^2+1) = z(z-i)(z+i)$, the correct form of the expansion is

$$\frac{2z+i}{z^3+z} = \frac{2z+i}{z^3+z} = \frac{A}{z} + \frac{B}{z-i} + \frac{C}{z+i},$$

where A, B, C are constants to be determined. As discussed in Appendix A, we can calculate A, B, C, specifically that

$$f(z) = \frac{i}{z} + \frac{-\frac{i3}{2}}{z-i} + \frac{\frac{i}{2}}{z+i},$$

so

$$\oint_{\mathcal{C}} \frac{2z+i}{z^3+z}\,dz = i\oint_{\mathcal{C}} \frac{1}{z}\,dz - \frac{i3}{2}\oint_{\mathcal{C}} \frac{1}{z-i}\,dz + \frac{i}{2}\oint_{\mathcal{C}} \frac{1}{z+i}\,dz.$$

The third term is zero, by Theorem 15.36, because $f(z) = \frac{1}{z+i}$ is analytic on the domain $\mathcal{D}$ shown in Figure 15.26b.

The first term can be calculated by the deformation $\mathcal{C} \leftrightsquigarrow \mathcal{C}_0$ in the same domain $\mathcal{D}$—see Figure 15.26c:

$$i\oint_{\mathcal{C}} \frac{1}{z}\,dz = i\oint_{\mathcal{C}_0} \frac{1}{z}\,dz = i\cdot 2\pi i = -2\pi,$$

using Theorem 15.30.

The second term can be calculated by the deformation $\mathcal{C} \leftrightsquigarrow \mathcal{C}_1$ in the same domain $\mathcal{D}$–(see Figure 15.26d):

$$-\frac{i3}{2}\oint_{\mathcal{C}} \frac{1}{z-i}\,dz = -\frac{i3}{2}\oint_{\mathcal{C}_1} \frac{1}{z-i}\,dz = -\frac{i3}{2}\cdot 2\pi i = 3\pi.$$

Finally, we conclude that

$$\oint_{\mathcal{C}} \frac{2z+i}{z^3+z}\,dz = \oint_{\mathcal{C}} \left(i\frac{i}{z} - \frac{i3}{2}\cdot\frac{1}{z-i} + \frac{i}{2}\cdot\frac{1}{z+i}\right)dz = -2\pi + 3\pi + 0 = \pi. \bigcirc$$

As we promised in Section 15.3, we have a result that also relies on the concept of simple connectedness.

Theorem 15.37

If $u = u(x, y)$ is harmonic on a simply connected domain $\mathcal{D}$, then $u(x, y)$ has a harmonic conjugate on $\mathcal{D}$.

15.8.3 Problems

In problems 1–4, evaluate the integral.

1. $\displaystyle\oint_{|z|=3} \overline{z}\, dz$
2. $\displaystyle\oint_{|z-i|=1} \overline{z}\, dz$
3. $\displaystyle\oint_{|z|=\frac{3}{2}} \frac{z}{(z-i)(z-2i)}\, dz$
4. $\displaystyle\oint_{|z|=\frac{5}{2}} \frac{z}{(z-i)(z-2i)}\, dz$
5. Evaluate the integral, where the parametrized curve $\mathcal{C}$ is shown in Figure 15.27b. [Hint: Use a suitable branch of the logarithm.]

$$\oint_{\mathcal{C}} \frac{1}{z}\, dz$$

6. Evaluate the integral, where the parametrized curve $\mathcal{C}$ is shown in Figure 15.27a:

$$\oint_{\mathcal{C}} \frac{1}{z}\, dz.$$

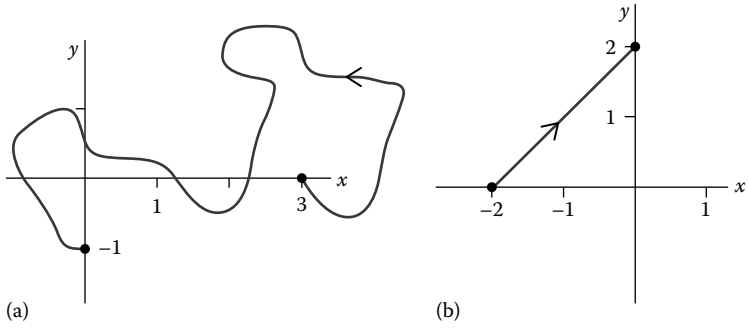


FIGURE 15.27
(a) Problem 15.8.3.6, (b) Problem 15.8.3.5.

7. Theorem 15.34 on path independence in the complex plane is analogous to the fundamental theorem of line integrals, Theorem 7.8 in Section 7.2, in $\mathbb{R}^2$ or $\mathbb{R}^3$. Theorem 7.8 in Section 7.2 applies when a vector field $\mathbf{F}$ is exact. It turns out that if we study path independence in the complex $z = x + iy$ plane by rewriting a contour integral in terms of real variables (x, y), then we will see in this problem that the Cauchy–Riemann equations naturally appear:
   (a) Given a function $f(z)$, write $f(z) = u(x, y) + iv(x, y)$ and $z = x + iy$ and thus

$$\int_{\mathcal{C}} f(z)\,dz = \int_{\mathcal{C}} \big(u(x,y) + iv(x,y)\big)\; d(x+iy)$$
$$= \cdots = \int_{\mathcal{C}} (u\,dx - v\,dy) + i\int_{\mathcal{C}} (v dx + u dy).$$

   Fill in those details.
   (b) Use the result of part (a) to argue that in order to have path independence of the contour integral $\int_{\mathcal{C}} f(z)\,dz$, we need to have path independence of both $\int_{\mathcal{C}} (u\,dx - v\,dy)$ and $\int_{\mathcal{C}} (v\,dx + u\,dy)$.
   (c) Argue that path independence of $\int_{\mathcal{C}} (udx - vdy)$ implies that the vector field $\mathbf{F} = u\,\hat{\boldsymbol{\imath}} - v\,\hat{\boldsymbol{\jmath}}$ must be exact. Argue similarly why the vector field $\mathbf{G} = v\,\hat{\boldsymbol{\imath}} + u\,\hat{\boldsymbol{\jmath}}$ must be exact.
   (d) Use the results of part (c) to argue that path independence of the contour integral $\int_{\mathcal{C}} f(z)\,dz$ implies that u, v must satisfy the Cauchy–Riemann equations.
   (e) Study Example 15.54 and Problems 7.2.5.19 and 7.2.5.20 and discuss the relative advantages and/or disadvantages of applying path independence of contour integrals in the complex plane versus applying path independence of line integrals in $\mathbb{R}^2$.
8. Establish the case $m = 1$ of Theorem 15.30; it's similar to the calculation of Example 15.51.

15.9 Cauchy's Integral Formulas and Residues

Theorem 15.38

(Cauchy's integral formula) Suppose $f(z)$ is analytic on a simply connected domain $\mathcal{D}$ that contains both a simple, closed, positively oriented contour $\mathcal{C}$ and the interior of $\mathcal{C}$. If z_0 is inside $\mathcal{C}$, then

$$\int_{\mathcal{C}} \frac{f(z)}{z - z_0}\,dz = 2\pi i\, f(z_0). \tag{15.69}$$

Suppose z is inside $\mathcal{C}$. If we replace z by ζ and then replace z_0 by z, we get

$$\frac{1}{2\pi i}\int_{\mathcal{C}} \frac{f(\zeta)}{\zeta - z}\, d\zeta = f(z)\,. \tag{15.70}$$

Example 15.57

Evaluate $\oint_{\mathcal{C}} \dfrac{\sin(z^2)}{z - i2}\, dz$, where $\mathcal{C}$ is (a) the circle $|z| = 3$ oriented positively and (b) the circle $|z| = 1$ oriented positively.

Method:

(a) By Cauchy's integral formula with $f(z) \triangleq \sin(z^2)$ and $\mathcal{D} = \mathbb{C}$,

$$\oint_{|z|=3} \frac{\sin(z^2)}{z - i2}\, dz = 2\pi i \cdot f(i2) = 2\pi i \cdot \sin\left((i2)^2\right) = 2\pi i \cdot \sin(-4) = -i2\pi \sin 4.$$

(b) By the Cauchy–Goursat Theorem 15.36 with, for example, $\mathcal{D} = \mathcal{D}_{1.5}(0)$,

$$\oint_{|z|=1} \frac{\sin(z^2)}{z - i2}\, dz = 0$$

because the function $g(z) \triangleq \dfrac{\sin(z^2)}{z - i2}$ is analytic on $\mathcal{D}_{1.5}(0)$. ○

Example 15.58

Evaluate $\oint_{|z+i|=1.5} \dfrac{z}{z^2 + 1}\, dz$.

Method: $\dfrac{z}{z^2+1} = \dfrac{z}{(z-i)(z+i)}$ is analytic everywhere on and inside the circle $|z + i| = 1.5$, except at $z = -i$. Figure 15.28 shows the contour $|z + i| = 1.5$ oriented positively, with the $\times$'s marking the poles of the function $\frac{z}{z^2+1}$. Define $f(z) \triangleq \frac{z}{z-i}$, which is analytic on $\mathcal{D} = \mathcal{D}_{1.6}(-i)$, for example. By Cauchy's integral formula on $\mathcal{D}$, we calculate that

$$\oint_{|z+i|=1.5} \frac{z}{z^2+1}\, dz = \oint_{|z+i|=1.5} \frac{f(z)}{z+i}\, dz = 2\pi i \cdot f(-i) = 2\pi i \cdot \frac{-i}{-i-i} = i\pi\,. \text{ ○}$$

Example 15.59

Evaluate $\oint_{|z|=R} \dfrac{e^{i\omega z}}{z^2+1}\, dz$, where ω is an unspecified constant and the constant $R > 1$.

Method: Partial fractions give

$$\frac{1}{z^2+1} = \frac{A}{z-i} + \frac{B}{z+i} = \ldots = \frac{-i/2}{z-i} + \frac{i/2}{z+i}.$$

Define $f(z) = e^{i\omega z}$. Partial fractions and then Cauchy's integral formula give

$$\oint_{|z|=R} \frac{e^{i\omega z}}{z^2+1}\, dz = -\frac{i}{2}\oint_{|z|=R} \frac{f(z)}{z-i}\, dz + \frac{i}{2}\oint_{|z|=R} \frac{f(z)}{z+i}\, dz = -\frac{i}{2}\cdot 2\pi i f(i) + \frac{i}{2}\cdot 2\pi i f(-i)$$

$$= \pi f(i) - \pi f(-i) = \pi\left(e^{i\omega i} - e^{i\omega(-i)}\right) = 2\pi \cdot \frac{e^{-\omega} - e^{\omega}}{2} = -2\pi \sinh\omega. \text{ ○}$$

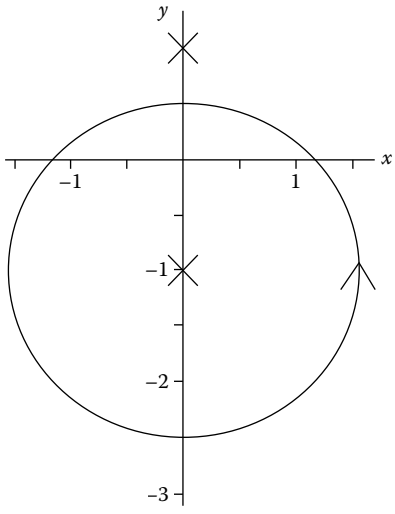


FIGURE 15.28
Example 15.58.

15.9.1 Use of a Dumb-Bell Contour

We present an alternative method for calculating integrals by using "dumb-bell" contours, as shown in Figure 15.29. Here, $\mathcal{C}_0$ is the circle $|z| = \frac{1}{3}$, $\mathcal{C}_1$ is the circle $|z-i| = \frac{1}{4}$, and they are "joined" by the two line segments (a) $L:\ z = z(t) = i\left(\frac{1}{3} + \frac{5t}{12}\right)$, $0 \le t \le 1$, and (b) $-L$. Note that the opposite parametrized curves $\pm L$ have opposite direction arrows in Figure 15.29.

Defining the contour $\mathcal{C} \triangleq \mathcal{C}_0 + L + \mathcal{C}_1 + (-L)$, if $f(z)$ is analytic on a simply connected domain $\mathcal{D}$, which contains $\mathcal{C}$, then

$$\begin{aligned}\oint_{\mathcal{C}} f(z)\,dz &= \int_{\mathcal{C}_0} f(z)\,dz + \int_{L} f(z)\,dz + \int_{\mathcal{C}_1} f(z)\,dz + \int_{-L} f(z)\,dz \\ &= \int_{\mathcal{C}_0} f(z)\,dz + \cancel{\int_{L} f(z)\,dz} + \int_{\mathcal{C}_1} f(z)\,dz - \cancel{\int_{L} f(z)\,dz};\end{aligned}$$

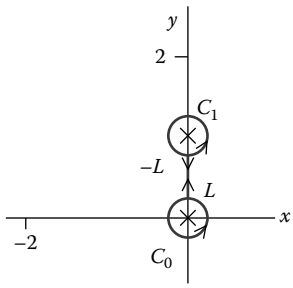


FIGURE 15.29
Dumb-bell contour.

hence,

$$\oint_{\mathcal{C}} f(z)\,dz = \int_{\mathcal{C}_0} f(z)\,dz + \int_{\mathcal{C}_1} f(z)\,dz\,. \tag{15.71}$$

The latter is the sum of integrals over two disjoint circles.

Example 15.60

(Example 15.56 in Section 15.8 again) Evaluate $\oint_{\mathcal{C}} \dfrac{2z+i}{z^3+z}\,dz$, where $\mathcal{C}$ is shown in Figure 15.26a.

Method: Using the dumbbell contour shown in Figure 15.29, (15.71) and Cauchy's integral formula allows us to calculate

$$\oint_{\mathcal{C}} \frac{2z+i}{z^3+z}\,dz = \int_{\mathcal{C}_0} \frac{2z+i}{z^3+z}\,dz + \int_{\mathcal{C}_1} \frac{2z+i}{z^3+z}\,dz = \int_{\mathcal{C}_0} \frac{\dfrac{2z+i}{z^2+1}}{z}\,dz + \int_{\mathcal{C}_1} \frac{\dfrac{2z+i}{z(z+i)}}{z-i}\,dz$$

$$= 2\pi i\left(\frac{2z+i}{z^2+1}\Big|_{at\ z=0}\right) + 2\pi i\left(\frac{2z+i}{z(z+i)}\Big|_{at\ z=i}\right)$$

$$= 2\pi i\left(\frac{i}{1}\right) + 2\pi i\left(\frac{2i+i}{i(i+i)}\right) = -2\pi + 3\pi = \pi\,. \bigcirc$$

One advantage of this method is that we didn't have to use a partial fractions expansion.

Theorem 15.39

(Cauchy's integral formula in general) Suppose $f(z)$ is analytic on a simply connected domain $\mathcal{D}$ that contains a simple, closed, positively oriented contour $\mathcal{C}$ and its interior. If z_0 is inside $\mathcal{C}$ and m is a nonnegative integer, then

$$\int_{\mathcal{C}} \frac{f(z)}{(z-z_0)^{m+1}}\,dz = \frac{2\pi i}{m!} f^{(m)}(z_0). \tag{15.72}$$

Suppose z is inside $\mathcal{C}$. If we replace z by ζ and then replace z_0 by z, we get

$$\frac{m!}{2\pi i}\int_{\mathcal{C}} \frac{f(\zeta)}{(\zeta - z)^{m+1}}\,d\zeta = f^{(m)}(z)\,. \tag{15.73}$$

When $m = 0$, (15.72) reduces to (15.69) and (15.73) reduces to (15.70). An important consequence of (15.73) and thus a consequence of Theorem 15.39 are the following.

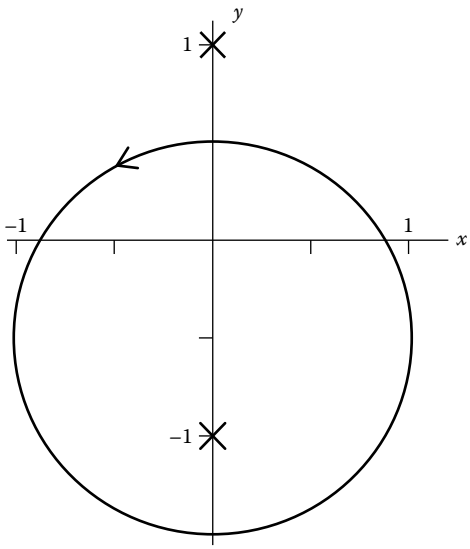


FIGURE 15.30
Example 15.61.

Theorem 15.40

If $f(z)$ is analytic at z_0, then so are $f', f'', \ldots$.

Example 15.61

Evaluate $\oint_{\mathcal{C}} \dfrac{\sin 2z}{\left(z^2+1\right)^2}\,dz$, where $\mathcal{C}:\ z = z(t) = -\frac{i}{2} + e^{it},\ 0 \le t \le 2\pi$.

Method: $\dfrac{\sin 2z}{\left(z^2+1\right)^2}$ is analytic everywhere on and inside $\mathcal{C}$ except at $z = -i$. Figure 15.30 shows the contour $|z + \frac{i}{2}| = 1$ oriented positively, with the $\times$'s marking the poles of the function $\dfrac{\sin 2z}{(z^2+1)^2}$. Define $f(z) \triangleq \dfrac{\sin 2z}{(z-i)^2}$. Using Theorem 15.39 with $m = 1$, we calculate

$$\oint_{\mathcal{C}} \frac{\sin 2z}{\left(z^2+1\right)^2}\,dz = \oint_{\mathcal{C}} \frac{\sin 2z}{(z+i)^2(z-i)^2}\,dz = \oint_{\mathcal{C}} \frac{(\sin 2z)/(z-i)^2}{(z+i)^2}\,dz = \frac{2\pi i}{1!}\cdot\frac{d}{dz}\left[\frac{\sin 2z}{(z-i)^2}\right]\Big|_{at\ z=-i}$$

$$= 2\pi i \cdot \left(\frac{2\cos 2z}{(z-i)^2} - \frac{2\sin 2z}{(z-i)^3}\right)\Big|_{at\ z=-i} = 2\pi i \cdot \left(\frac{2\cos(-i2)}{(-i-i)^2} - \frac{2\sin(-i2)}{(-i-i)^3}\right)$$

$$= 2\pi i \cdot \left(\frac{2\cosh 2}{-4} - \frac{-2i\sinh 2}{i8}\right) = i\pi\left(-\cosh 2 + \frac{1}{2}\sinh 2\right). \bigcirc$$

Alternatively, it would take more work to calculate the partial fractions expansion

$$\frac{1}{\left(z^2+1\right)^2} = \frac{\frac{i}{4}}{z-i} + \frac{-\frac{1}{4}}{(z-i)^2} - \frac{\frac{i}{4}}{z+i} + \frac{-\frac{1}{4}}{(z+i)^2},$$

but after that we could use Theorem 15.39 with $m = 0$ and $m = 1$.

15.9.2 Integration of a Laurent Series

If $f(z)$ has a Laurent series (15.53) in Section 15.6, that is,

$$f(z) \triangleq \sum_{j=1}^{\infty} a_{-j}\,(z-z_0)^{-j} + \sum_{j=0}^{\infty} a_j\,(z-z_0)^j,$$

and is an analytic function in an annulus $\mathcal{A}_{r,R}(z_0)$, then we can integrate term by term on a positively oriented circle $\mathcal{C}:\ |z-z_0|=\rho$, where $r<\rho<R$, to get

$$\begin{aligned}\oint_{\mathcal{C}} f(z)\,dz = \cdots + a_{-3}\oint_{\mathcal{C}} \frac{dz}{(z-z_0)^3} + a_{-2}\oint_{\mathcal{C}} \frac{dz}{(z-z_0)^2} \\ + a_{-1}\oint_{\mathcal{C}} \frac{dz}{z-z_0} + a_0\oint_{\mathcal{C}} 1\,dz + a_1\oint_{\mathcal{C}}(z-z_0)\,dz + \cdots = \cdots + 0 + 0 + a_{-1}\cdot 2\pi i \\ + 0 + 0 + \cdots = 2\pi i\, a_{-1},\end{aligned}$$

by using Theorem 15.30 in Section 15.8. This motivates a definition:

Definition 15.29

$\text{Res}[f;\ z_0\,] = a_{-1}$ is called the **residue of** f **at** z_0, assuming f has Laurent series (15.53) in Section 15.6.

Example 15.62

Define $g(z) \triangleq \dfrac{\sin z}{z^4}$. Evaluate (a) $\text{Res}[\,g;\ 0\,]$ and (b) $\displaystyle\oint_{|z|=0.5} g(z)\,dz$.

Method:

(a) Similar to the work in Example 15.39 in Section 15.6, the Maclaurin series of $\sin z$ gives

$$g(z) = \frac{\sin z}{z^4} = z^{-3} - \frac{1}{3!}\,z^{-1} + \sum_{\ell=1}^{\infty} \frac{(-1)^{\ell+1}}{(2\ell+3)!} z^{2\ell-1},$$

so $\text{Res}[\,g;\ 0\,] = a_{-1} = -\frac{1}{3!}$.

(b) Method I: By the earlier calculation that motivated the definition of "residue,"

$$\oint_{|z|=0.5} g(z)\,dz = 2\pi i\cdot a_{-1} = -\frac{\pi i}{3}.$$

Method II: Using Cauchy's integral formula with $m=3$,

$$\oint_{|z|=0.5} g(z)\,dz = \frac{2\pi i}{3!}\cdot\left(\frac{d^3}{dz^3}\big[\sin z\big]\right)\Big|_{at\ z=0} = \frac{2\pi i}{3!}\cdot(-\cos z)\Big|_{at\ z=0} = \frac{2\pi i}{3!}\cdot(-1) = -\frac{\pi i}{3}.\ \bigcirc$$

In general, calculating $f^{(m)}(z)$ may involve a lot of work. Even though this was not true in Example 15.62, Method I was still easier than Method II. In some problems, residues may provide a relatively easy way to evaluate integrals.

Theorem 15.41

Suppose $f(z)$ has an isolated singularity at z_0 and m is a nonnegative integer.

(a) If z_0 is a removable singularity of f, then $\text{Res}[f;\ z_0\,] = 0$.

(b) If z_0 is a pole of order $m+1$ of f, then

$$\text{Res}[f;\ z_0\,] = \frac{1}{m!}\lim_{z\to z_0}\left(\frac{d^m}{dz^m}\Big[\,(z-z_0)^{m+1}f(z)\,\Big]\right). \tag{15.74}$$

(c) If z_0 is a "simple pole" of f, that is, a pole of order one of f, then

$$\text{Res}[f;\ z_0\,] = \lim_{z\to z_0}(z-z_0)f(z)\,. \tag{15.75}$$

Note that (c) is a special case of (b) with $m = 0$.

Example 15.63

Find $\text{Res}[\,z\cot^2 z;\ 0\,]$.

Method: Define $f(z) \triangleq z\cot^2 z = z\cdot\dfrac{\cos^2 z}{\sin^2 z}$. While it may be possible to find a Laurent series of f, it is easier to use Theorem 15.41 to find the residue. But, to do that, we must first find the order of the pole of $f(z)$ at $z = 0$.

First, we can find that $\sin z$ has a zero of order one at $z = 0$ by using the definition of "zero." After explaining that, we can use "preparation" Theorem 15.24 in Section 15.7 to see that

$$\sin z = (z-0)k(z) \tag{15.76}$$

for some function $k(z)$ that is analytic at $z = 0$ and has $k(0) \neq 0$. In fact, (15.76) implies $k(z) = \frac{\sin z}{z}$ for $z \neq 0$, and L'Hôpital's rule, that is, Theorem 15.12 in Section 15.3, calculates that

$$\lim_{z\to 0}\frac{\sin z}{z} = \lim_{z\to 0}\frac{(\sin z)'}{(z)'} = \lim_{z\to 0}\frac{\cos z}{1} = 1,$$

so

$$k(z) \triangleq \left\{\begin{array}{ll}\dfrac{\sin z}{z}, & \text{if } z\neq 0\\[2ex] 1, & \text{if } z = 0\end{array}\right\}.$$

Next, we substitute (15.76) into the formula for $f(z)$ to see that, for $z \neq 0$,

$$f(z) = z\cdot\frac{\cos^2 z}{z^2\,\big(k(z)\big)^2} = \frac{\cos^2 z}{z\,\big(k(z)\big)^2} = \frac{1}{z}\cdot\left(\frac{\cos z}{k(z)}\right)^2 \triangleq \frac{g(z)}{z}\,. \tag{15.77}$$

Note that $g(z) \triangleq \dfrac{\cos^2 z}{\big(k(z)\big)^2}$ is analytic at $z = 0$, because $\cos z$ and $k(z)$ are analytic at $z = 0$ and $k(0) \neq 0$; also, $g(0) = 1$. By Theorem 15.26(b) in Section 15.7 and (15.77), $f(z)$ has a pole of order one at $z = 0$.

So, Theorem 15.41(c) allows us to calculate

$$\text{Res}[\, z\cot^2 z;\ 0\,] = \lim_{z\to 0} z f(z) = \lim_{z\to 0} z\cdot\frac{g(z)}{z} = \lim_{z\to 0} g(z) = g(0) = 1. \bigcirc$$

15.9.3 Cauchy's Residue Theorem

Theorem 15.42

Suppose $\mathcal{C}$ is a simple, closed, and positively oriented contour and $\mathcal{D}$ is a simply connected domain that contains both $\mathcal{C}$ and the interior of $\mathcal{C}$. Suppose $f(z)$ is analytic on $\mathcal{D}$ except at finitely many points or at a sequence of points $z_1, \ldots, z_k, \ldots$. If $f(z)$ is analytic on the interior of $\mathcal{C}$ except at finitely many points $z_1, \ldots, z_n$, then

$$\oint_{\mathcal{C}} f(z)\, dz = 2\pi i \sum_{k=1}^{n} \text{Res}[f;\ z_k\,]. \tag{15.78}$$

Example 15.64

Find $\displaystyle\oint_{|z|=1.5} \frac{1}{z(z+1)(z-2)}\, dz$.

Method: The singularities of $f(z) \triangleq \dfrac{1}{z(z+1)(z-2)}$ are at $z = 0, -1, 2$. Figure 15.31 depicts the singularities as $\times$'s and the contour $\mathcal{C}:\ |z| = 1.5$ with positive orientation. Two of the singularities of $f(z)$ are contained inside $\mathcal{C}$, while $z = 2$ is not. By Theorems 15.41 and 15.42,

$$\begin{aligned}\oint_{\mathcal{C}} f(z)\, dz &= 2\pi i\,(\text{Res}[f;\ 0\,] + \text{Res}[f;\ -1\,]) \\ &= 2\pi i\left(\left(\cancel{z}\cdot\frac{1}{\cancel{z}(z+1)(z-2)}\right)\Big|_{at\ z=0} + \left(\cancel{(z+1)}\cdot\frac{1}{z\cancel{(z+1)}(z-2)}\right)\Big|_{at\ z=-1}\right) \\ &= 2\pi i\left(-\frac{1}{2}+\frac{1}{3}\right) = -\frac{\pi i}{3}. \bigcirc\end{aligned}$$

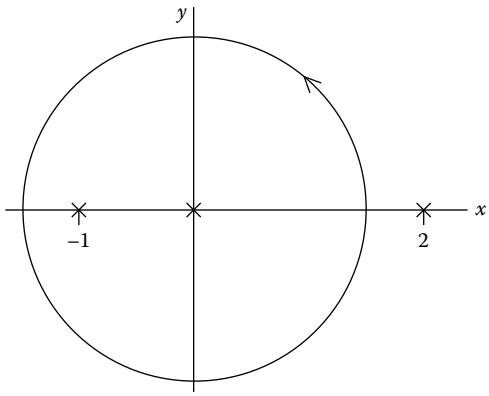


FIGURE 15.31
Example 15.62: $\mathcal{C}:\ |z| = 1.5$ and poles of $f(z)$.

Learn More About It

The book *Fundamentals of Complex Analysis, with Applications to Engineering and Science*, 3rd edition, by E. B. Saff and A. D. Snider, Prentice Hall/Pearson Education, Inc., © 2003, pages 207–209, has an explanation for Theorem 15.40.

15.9.4 Problems

1. Find the isolated singularities of $f(z)=\frac{1}{z(z+1)(z-2)}$ and compute the corresponding residues.
2. Find the isolated singularities of $f(z)=\frac{1}{z^2(z+1)}$ and compute the corresponding residues.
3. Define $f(z)=\frac{1}{z\sin z}$ *only* for z inside the disk $D_2(\frac{\pi}{2})$. Find the isolated singularities of f and compute the corresponding residues.

In problems 4–8, evaluate the integral.

4. $\oint_{|z|=3} \dfrac{z}{(z-i)(z-2i)}\,dz$
5. $\oint_{|z|=\frac{3}{2}} \left(\dfrac{z\cos z}{(z-1)^2}+\dfrac{z\cos z}{z+2}\right)dz$
6. $\oint_{|z|=3} \left(\dfrac{z\cos z}{(z-1)^2}+\dfrac{z\cos z}{z+2}\right)dz$
7. $\oint_{|z|=3} \dfrac{e^z}{z-i\pi}\,dz$
8. $\oint_{|z|=4} \dfrac{e^z}{z-i\pi}\,dz$

In problems 9–12, evaluate the integral using Cauchy's residue theorem.

9. $\oint_{|z-1|=3} \dfrac{1}{z^2-9}\,dz$
10. $\oint_{|z|=3} \dfrac{z}{\sin z}\,dz$
11. $\oint_{|z-\frac{\pi}{2}|=2} \dfrac{z}{\sin z}\,dz$
12. $\oint_{|z|=2} \dfrac{z}{e^z+1}\,dz$
13. Evaluate the integral, assuming ω and ξ are unspecified constants and $R>\omega>0$:

$$\oint_{|z|=R} \frac{ze^{i\xi z}}{z^2+\omega^2}\,dz.$$

14. Evaluate the integral, assuming ω and ξ are unspecified constants and $R > \omega > 0$:

$$\oint_{|z| = R} \frac{ze^{i\xi z}}{z^2 - \omega^2}\, dz.$$

15.10 Real Integrals by Complex Integration Methods

Theorem 15.29 in Section 15.8 gave our first technique for evaluating line integrals in the complex plane by parametrizing the contour and then calculating a definite integral with respect to a real variable.

By "real definite integral" we mean an integral with respect to a real variable. We can use Theorem 15.29 in Section 15.8 in reverse to evaluate some difficult real definite integrals by using Cauchy's residue theorem. This will be a good example of the mathematical "culture" of using the same results in many ways by "looking at it from many angles."

15.10.1 Integration of Periodic Functions

Example 15.65

Evaluate $\displaystyle\int_0^{2\pi} \frac{d\theta}{1 + \cos^2\theta}$.

Method: At first, this seems to have no connection with complex variables. But, integrating over the interval $0 \le \theta \le 2\pi$ suggests a connection with a line integral over a circle.

Parametrize the unit circle $|z| = 1$ by $\mathcal{C}: z = z(\theta) = e^{i\theta}$, $0 \le \theta \le 2\pi$, which has positive orientation. For z on $\mathcal{C}$,

$$\cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2} = \frac{z + z^{-1}}{2} = (2z)^{-1}(z^2 + 1).$$

On $\mathcal{C}$, we have $\dfrac{dz}{d\theta}(\theta) = ie^{i\theta} = iz$; hence, $d\theta = \dfrac{dz}{iz}$. By Theorem 15.29 in Section 15.8,

$$\begin{aligned}\int_0^{2\pi} \frac{1}{1 + \cos^2\theta}\, d\theta &= \oint_{|z|=1} \frac{1}{1 + \left((2z)^{-1}(z^2+1)\right)^2} \frac{dz}{iz} = \oint_{|z|=1} \frac{(2z)^2}{(2z)^2 + \left(z^2+1\right)^2} \frac{dz}{iz} \\ &= -4i \oint_{|z|=1} \frac{z}{4z^2 + \left(z^2+1\right)^2}\, dz = -4i \oint_{|z|=1} \frac{z}{z^4 + 6z^2 + 1}\, dz.\end{aligned}$$

The four singularities are where the denominator is zero, that is, where $0 = z^4 + 6z^2 + 1 = \left(z^2\right)^2 + 6z^2 + 1$, that is, where

$$z^2 = \frac{-6 \pm \sqrt{6^2 - 4}}{2} = -3 \pm 2\sqrt{2},$$

that is,

$$z = \pm i\sqrt{3+2\sqrt{2}}, \quad \pm i\sqrt{3-2\sqrt{2}}.$$

Of these singularities, only $z_3 = i\sqrt{3-2\sqrt{2}}$ and $z_4 = -i\sqrt{3-2\sqrt{2}}$ are inside $\mathcal{C}$. Because

$$f(z) \triangleq \frac{z}{z^4+6z^2+1} = \frac{z}{\left(z^2+3+2\sqrt{2}\right)\left(z-i\sqrt{3-2\sqrt{2}}\right)\left(z+i\sqrt{3-2\sqrt{2}}\right)},$$

both z_3 and z_4 are simple poles of $f(z)$. It is relatively straightforward to use Theorem 15.41(c) in Section 15.9 to calculate that

$$\int_0^{2\pi} \frac{1}{1+\cos^2\theta}\,d\theta = -4i \oint_{|z|=1} f(z)\,dz = -4i \cdot 2\pi i \left(\text{Res}[f(z);\, z_3] + \text{Res}[f(z);\, z_4] \right)$$

$$= \cdots = 8\pi\left(\frac{1}{8\sqrt{2}} + \frac{1}{8\sqrt{2}}\right) = \pi\sqrt{2}. \bigcirc$$

In principle, this technique will work for any real integral of the form

$$\int_0^{2\pi} h(\cos\theta, \sin\theta)\,d\theta,$$

where $h(\cos\theta, \sin\theta)$, for example, $\dfrac{1+\sin\theta+3\cos^2\theta}{2+\sin^2\theta}$, is a rational function of $\cos\theta$ and $\sin\theta$ whose denominator is never zero for $0 \le \theta \le 2\pi$.

15.10.2 Improper Integrals over $(-\infty, \infty)$, $[0, \infty)$, or $(-\infty, 0]$

Example 15.66

Evaluate $\displaystyle\int_0^{\infty} \frac{dx}{\left(x^2+3\right)^2}$.

Method: Define $g(x) = \frac{1}{\left(x^2+3\right)^2}$ and $f(z) = \frac{1}{\left(z^2+3\right)^2}$. We will explain why the improper integral $\int_{-\infty}^{\infty} g(x)\,dx$ exists and equals

$$2\pi i\,\text{Res}[f(z);\, i\sqrt{3}] = \cdots = \frac{\pi}{12\sqrt{3}}.$$

First, the improper integral $\int_{-\infty}^{\infty} g(x)\,dx$ exists because both

$$\lim_{b\to\infty} \int_0^b \frac{dx}{\left(x^2+3\right)^2} \quad \text{and} \quad \lim_{a\to-\infty} \int_a^0 \frac{dx}{\left(x^2+3\right)^2}$$

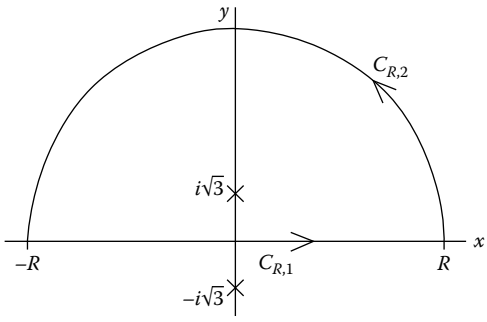


FIGURE 15.32
Complex integration to evaluate a real integral.

exist, by a comparison theorem for definite integrals. Because both of the improper integrals $\int_0^\infty g(x)dx$ and $\int_{-\infty}^0 g(x)dx$ are convergent, we have

$$\int_{-\infty}^{\infty} g(x)\,dx = \lim_{R\to\infty}\int_{-R}^{0} g(x)\,dx + \lim_{R\to\infty}\int_{0}^{R} g(x)\,dx = \lim_{R\to\infty}\int_{-R}^{R} g(x)\,dx.$$

Consider the contour $\mathcal{C}_R = \mathcal{C}_{R,1} + \mathcal{C}_{R,2}$ shown in Figure 15.32, which also indicates by $\times$ the singularities of $f(z)$ at $z = \pm i\sqrt{3}$. We calculate

$$\int_{\mathcal{C}_{R,2}} f(z)\,dz = \int_{\mathcal{C}_{R,2}} \frac{1}{\left(z^2+3\right)^2}\,dz = \int_0^\pi \frac{1}{\left(\left(Re^{i\theta}\right)^2+3\right)^2} iRe^{i\theta}\,d\theta = \int_0^\pi \frac{1}{\left(R^2e^{i2\theta}+3\right)^2} iRe^{i\theta}\,d\theta.$$

Using Lemma 15.1 in the following,

$$\left|\int_{\mathcal{C}_{R,2}} f(z)\,dz\right| = \left|\int_0^\pi \frac{1}{\left(R^2e^{i2\theta}+3\right)^2} iRe^{i\theta}\,d\theta\right| \le \int_0^\pi \left|\frac{1}{\left(R^2e^{i2\theta}+3\right)^2} iRe^{i\theta}\right| d\theta$$

$$\le \int_0^\pi \frac{R}{\left(R^2-3\right)^2}\,d\theta \to 0, \quad \text{as} \quad R\to\infty.$$

Also, $\mathcal{C}_{R,1}: z = x + i0$, $-R \le x \le R$, so

$$\int_{\mathcal{C}_{R,1}} f(z)\,dz = \int_{-R}^{R} g(x)\,dx.$$

Putting things together, we have

$$\lim_{R\to\infty}\int_{\mathcal{C}_{R,1}+\mathcal{C}_{R,2}} f(z)\,dz = \lim_{R\to\infty}\left(\int_{\mathcal{C}_{R,2}} f(z)\,dz + \int_{-R}^{R} g(x)\,dx\right) = 0 + \int_{-\infty}^{\infty} g(x)\,dx.$$

For $R > \sqrt{3}$,

$$\int_{\mathcal{C}_{R,1}+\mathcal{C}_{R,2}} f(z)\,dz = 2\pi i\,\text{Res}[f(z);\ i\sqrt{3}\,].$$

Finally, use the fact that $g(x)$ is an even function to get

$$\int_0^\infty \frac{dx}{\left(x^2+3\right)^2} = \frac{1}{2}\int_{-\infty}^{\infty} g(x)\,dx = \frac{1}{2}\lim_{R\to\infty}\int_{C_{R,1}+C_{R,2}} f(z)\,dz$$

$$= \frac{1}{2}\,2\pi i\,\text{Res}[f(z);\ i\sqrt{3}\,] = \cdots = \frac{\pi\sqrt{3}}{36}.$$

In the earlier calculations, the last step was to use the fact that $f(z) = \frac{1}{\left(z-i\sqrt{3}\right)^2\left(z+i\sqrt{3}\right)^2}$ has a pole of order two at $z = i\sqrt{3}$, along with Theorem 15.41(b) in Section 15.9. ○

Lemma 15.1

For $R > \sqrt{3}$, and all θ in the interval $[0, \pi]$,

$$\left|\frac{1}{\left(R^2e^{i2\theta}+3\right)^2}\,ie^{i\theta}\right| \le \frac{1}{\left(R^2-3\right)^2}. \tag{15.79}$$

Why? The triangle inequality, $|z_1+z_2| \le |z_1| + |z_2|$, implies that $|z_1| \ge |z_1+z_2| - |z_2|$. So, defining $z_1 = R^2e^{i2\theta}+3$ and $z_2 = -3$, we have

$$|R^2e^{i2\theta}+3| = |z_1| \ge |z_1+z_2| - |z_2| = |R^2e^{i2\theta}+3-3| - |-3| = |R^2e^{i2\theta}| - 3$$

$$= |R^2|\,|e^{i2\theta}| - 3 = R^2\cdot 1 - 3 = (R^2-3).$$

It follows that

$$\frac{1}{|R^2e^{i2\theta}+3|} \le \frac{1}{(R^2-3)}$$

and thus that

$$\left|\frac{1}{\left(R^2e^{i2\theta}+3\right)^2}\,ie^{i\theta}\right| = \frac{1}{|R^2e^{i2\theta}+3|^2}\cdot|ie^{i\theta}| = \left(\frac{1}{|R^2e^{i2\theta}+3|}\right)^2\cdot 1 \le \frac{1}{(R^2-3)^2}\,.\ \square$$

The reasoning and techniques we used in Example 15.66 generalize to a theorem.

Theorem 15.43

Suppose $p(x)/q(x)$ is a rational function, *degree* $\big(q(x)\big) \ge 2 +$ *degree* $\big(p(x)\big)$, and $q(x)$ has no real zeros. Then

$$\int_{-\infty}^{\infty}\frac{p(x)}{q(x)}\,dx = 2\pi i\sum_{k=1}^{n}\text{Res}\Big[\frac{p(z)}{q(z)};\ z_k\Big], \tag{15.80}$$

where $z_1, \ldots, z_n$ are the zeros of $q(z)$ in the upper half plane $\mathcal{I}m(z) \ge 0$.

15.10.3 Cauchy Principal Value

As we saw in Section 7.1, an improper integral $\int_{-\infty}^{\infty} g(x)\,dx$ may fail to exist even though $\lim_{R\to\infty}\left(\int_{-R}^{R} g(x)\,dx\right)$ exists. Nevertheless, it can be useful in solving differential equations and other subjects to acknowledge the latter limit's value.

Definition 15.30

The **Cauchy principal value** of $\int_{-\infty}^{\infty} g(x)\,dx$ is defined by

$$P.\,v.\int_{-\infty}^{\infty} g(x)\,dx \triangleq \lim_{R\to\infty}\left(\int_{-R}^{R} g(x)\,dx\right),$$

if the latter exists.

If $|g(x)|$ is integrable on $(-\infty,\infty)$, then the principal value equals $\int_{-\infty}^{\infty} g(x)\,dx$.

Example 15.67

As we saw in Example 7.3 in Section 7.1,

$$P.\,v.\int_{-\infty}^{\infty}\frac{2x}{x^2+1}\,dx = 0$$

even though the improper integral is divergent. ○

Example 15.68

Find $P.\,v.\int_{-\infty}^{\infty}\frac{\cos\omega x}{x^2+1}\,dx$, where ω is a real constant.

Method: Because $\cos\omega x = \mathcal{R}e\left(e^{i\omega x}\right)$, we have

$$P.\,v.\int_{-\infty}^{\infty}\frac{\cos\omega x}{x^2+1}\,dx = \mathcal{R}e\left(P.\,v.\int_{-\infty}^{\infty}\frac{1}{x^2+1}\,e^{i\omega x}\,dx\right).$$

Let $f(z)\triangleq\dfrac{1}{z^2+1}\,e^{i\omega z}$ and use the same contour $\mathcal{C}_R = \mathcal{C}_{R,1}+\mathcal{C}_{R,2}$ found in Example 15.66 and shown in Figure 15.32. At any point on $\mathcal{C}_{R,2}$ given by $z = Re^{i\theta} = x+iy$ that lies in the *upper* half plane, that is, has $y>0$, we have

$$\left|e^{i\omega z}\right| = \left|e^{i\omega(x+iy)}\right| = \left|e^{i\omega x}\right|\ \left|e^{-\omega y}\right| = 1\cdot e^{-\omega y} < 1.$$

So,

$$\left| \int\limits_{\mathcal{C}_{R,2}} \frac{1}{z^2+1}\, e^{i\omega z}\, dz \right| = \left| \int\limits_0^\pi \frac{1}{\left(Re^{i\theta}\right)^2+1}\, e^{i\omega R e^{i\theta}}\, iRe^{i\theta}\, d\theta \right| \le \int\limits_0^\pi \frac{\left|e^{i\omega(x+iy)}\right|}{|R^2 e^{i2\theta}+1|} \cdot \left|iRe^{i\theta}\right| d\theta$$

$$< \int\limits_0^\pi \frac{R}{R^2-1}\, d\theta \to 0, \quad \text{as} \quad R \to \infty.$$

So, similar to the work of Example 15.66, the convergent improper integral is given by

$$\int\limits_{-\infty}^{\infty} \frac{\cos\omega x}{x^2+1}\, dx = P.v. \int\limits_{-\infty}^{\infty} \frac{\cos\omega x}{x^2+1}\, dx \triangleq \lim_{R\to\infty} \left(\int\limits_{-R}^{R} \frac{\cos\omega x}{x^2+1}\, dx \right)$$

$$= \lim_{R\to\infty} \mathcal{R}e \left(\int\limits_{\mathcal{C}_{R,1}} \frac{1}{z^2+1}\, e^{i\omega z}\, dz \right) = \mathcal{R}e \left(2\pi i \operatorname{Res}\Big[\frac{e^{i\omega z}}{z^2+1};\, i \Big] \right)$$

$$= \mathcal{R}e \left(2\pi i \left(\frac{e^{i\omega z}}{z+i}\Big|_{at\ z=i} \right) \right) = \mathcal{R}e \left(2\pi i\, \frac{e^{i\omega\, i}}{i+i} \right) = \pi\, e^{-\omega}. \bigcirc$$

15.10.4 Hilbert Transform

Suppose $f(z) = u(x,y) + iv(x,y)$ is analytic on the upper half plane, $y > 0$, and is continuous on the real axis, $y = 0$. Suppose also that

$$\int\limits_{\mathcal{C}_{R,2}} \left| \frac{f(z)}{z} \right| dz \to 0 \quad \text{as } R \to \infty,$$

where, as previously mentioned, $\mathcal{C}_{R,2} : z = R\,e^{i\theta}$, $0 \le \theta \le \pi$. Using Cauchy's integral formula, we have

$$f(z) = \frac{1}{2\pi i} \int\limits_{\mathcal{C}_R} \frac{f(\zeta)}{\zeta - z}\, d\zeta,$$

for z in the upper half plane, where, as previously mentioned $\mathcal{C}_R = \mathcal{C}_{R,1} + \mathcal{C}_{R,2}$ and $\mathcal{C}_{R,1} : z = x + i0$, $-R \le x \le R$. Further details in a book by Bernard Friedman explain why

$$u(x, 0^+) = P.v. \frac{1}{\pi} \int\limits_{-\infty}^{\infty} \frac{v(\xi)}{\xi - x}\, d\xi, \tag{15.81}$$

and

$$v(x, 0^+) = -P.v. \frac{1}{\pi} \int\limits_{-\infty}^{\infty} \frac{u(\xi)}{\xi - x}\, d\xi, \tag{15.82}$$

for $-\infty < x < \infty$. Equation (15.81) says that $u(x, 0^+)$ is the **Hilbert transform** of its harmonic conjugate $v(x, y)$ and $v(x, 0^+)$ is the Hilbert transform of its harmonic conjugate $-u(x, y)$; the latter uses the result of Problem 15.3.3.21. So, (15.82) really says the same thing as (15.81). Hilbert transforms are useful in signal processing. An example of this is discussed in a book by Saff and Snider.

Learn More About It

Useful discussions of the Hilbert transform are in Section 5.1 of *Lectures on Applications-Oriented Mathematics*, by Bernard Friedman, ed. by Victor Twersky, Holden-Day Inc., © 1969 (reprinted by Wiley Classics Library, © 1991), and in Section 8.5 of *Fundamentals of Complex Analysis, with Applications to Engineering and Science*, 3rd edn., by E. B. Saff and A. D. Snider, Prentice Hall/Pearson Education, Inc., © 2003.

If an integral involves a square root, then one needs to integrate over a contour which bends around the branch cut on a half line. Good references for this include *Operational mathematics*, 3rd edn., by Ruel V. Churchill, McGraw-Hill © 1971, and the book by Saff and Snider cited earlier.

15.10.5 Problems

In problems 1–9, evaluate, that is, find the exact value, of the integral.

1. $\int_0^{2\pi} \dfrac{1}{1+\sin^2\theta}\, d\theta$ $\left[\text{Hint: Recall } \sin\theta = \frac{1}{2i}\left(e^{i\theta} - e^{-i\theta}\right)\right]$
2. $\int_0^{\pi} \dfrac{1}{1+\cos^2\theta}\, d\theta$ [Hint: Use symmetry of $\cos^2\theta$ on the interval $[0, \pi]$ vs. $[\pi, 2\pi]$.]
3. $\int_0^{\pi} \dfrac{1}{1+\sin^2\theta}\, d\theta$ [Hint: Use symmetry of $\sin^2\theta$ on the interval $[0, \pi]$ vs. $[\pi, 2\pi]$.]
4. $\int_{-\infty}^{\infty} \dfrac{1}{\left(x^2+4\right)^2}\, dx$
5. $\int_{-\infty}^{\infty} \dfrac{x^2}{\left(x^2+4\right)^2}\, dx$
6. $\int_0^{2\pi} \dfrac{\sin^2\theta}{1+\cos^2\theta}\, d\theta$
7. $\int_0^{2\pi} \dfrac{1}{2+\cos^2\theta}\, d\theta$
8. *P. v.* $\int_{-\infty}^{\infty} \dfrac{x\sin\omega x}{x^2+1}\, dx$
9. *P. v.* $\int_{-\infty}^{\infty} \dfrac{x^2\cos\omega x}{x^4+1}\, dx$

Key Terms

absolute value: beginning of Section 15.1
analytic: Definition 15.5 in Section 15.3 and Definition 15.6 in Section 15.3
anti-derivative: Definition 15.26 in Section 15.8
arg(z): (15.23) in Section 15.4
Arg(z): (15.24) in Section 15.4
bounded: before Theorem 15.32 in Section 15.8
"branch" of *log(z)*: (15.34) in Section 15.4
Cauchy principal value: Definition 15.30 in Section 15.10
Cauchy-Riemann equations: (15.14) in Section 15.2
complex: beginning of Section 15.1
complex conjugate: beginning of Section 15.1
connected: Definition 15.9 in Section 15.3
continuous: Definition 15.2 in Section 15.2
contour, simple, closed: Definition 15.24 in Section 15.8
converges absolutely: Definition 15.15 in Section 15.6
converges uniformly: Definition 15.17 in Section 15.6
deformation: Definition 15.27 in Section 15.8
DeMoivre's Theorem: after (15.5) in Section 15.1
derivative: Definition 15.4 in Section 15.2
differentiable: Definition 15.4 in Section 15.2
directed line segment: Definition 15.8 in Section 15.3
discrete convolution: end of Section 15.6
diverges: Definition 15.15 in Section 15.6
domain: Definition 15.9 in Section 15.3
entire: Definition 15.7 in Section 15.3
equipotential: Example 15.14 in Section 15.2
essential singularity: Definition 15.22 in Section 15.7
exp(z): (15.26) in Section 15.4
Euler's formula: (15.3) in Section 15.1
extension: Definition 15.3 in Section 15.2
exterior: after Theorem 15.32 in Section 15.8
function of a complex variable: before Example 15.4 in Section 15.1
geometric series: (15.43) in Section 15.6
harmonic: Definition 15.10 in Section 15.3
harmonic conjugate: Definition 15.11 in Section 15.3
Hilbert transform: (15.81) in Section 15.10, (15.82) in Section 15.10
holomorphic: after Definition 15.6 in Section 15.3
image: before Example 15.4 in Section 15.1
imaginary: beginning of Section 15.1
imaginary part: beginning of Section 15.1
incompressible: Example 15.14 in Section 15.2
initial point: Definition 15.23 in Section 15.8
interior: after Theorem 15.32 in Section 15.8
inversion mapping: before Example 15.5 in Section 15.1
irrational: Problem 15.1.4.10

isolated singularity: Definition 15.21 in Section 15.7
Laurent series: Definition 15.19 in Section 15.6
limit of $f(z)$ **exists**: Definition 15.1 in Section 15.2
log(z): (15.28) in Section 15.4
Log(z): (15.29) in Section 15.4
$\boldsymbol{Log_\sigma(z)}$: (15.34) in Section 15.4
mapping: before Example 15.4 in Section 15.1
modulus: beginning of Section 15.1
open: before Theorem 15.9 in Section 15.3
open disk: before Definition 15.1 in Section 15.2
open annulus: Definition 15.18 in Section 15.6
opposite curve: Definition 15.25 in Section 15.8
orthogonal families: before Example 15.14 in Section 15.2
parametrized curve: Definition 15.23 in Section 15.8
piecewise smooth curve: Definition 15.23 in Section 15.8
polar exponential form: (15.4) in Section 15.1
polar form: (15.1) in Section 15.1
pole of order m: Definition 15.22 in Section 15.7
polygonal curve: Definition 15.8 in Section 15.3
polynomial in z of **degree** n: before Theorem 15.3 in Section 15.2
positively oriented: Definition 15.23 in Section 15.8 and Definition 15.24 in Section 15.8
potential flow: Example 15.14 in Section 15.2
power function: (15.35) in Section 15.4
power series about: (15.42) in Section 15.6
principal branch or **principal value** of the power function: after (15.35) in Section 15.4
principal n-th root of unity: before Example 15.4 in Section 15.1
punctured disk: before Definition 15.1 in Section 15.2
punctured plane: Example 15.23 in Section 15.3
radius of convergence: Definition 15.16 in Section 15.6
real: beginning of Section 15.1
real analytic: after (15.49) in Section 15.6
real part: beginning of Section 15.1
removable singularity: Definition 15.22 in Section 15.7
residue of f: Definition 15.29 in Section 15.9
set union: after Example 15.31 in Section 15.5
simple, **closed** curve: Definition 15.23 in Section 15.8
simply connected: Definition 15.28 in Section 15.8
2-D Stokes potential fluid flow: after Example 15.23 in Section 15.3
streamlines: after Example 15.14 in Section 15.2
Taylor Series: Theorem 15.22 in Section 15.6
terminal point: Definition 15.23 in Section 15.8
truncate: before Theorem 15.21 in Section 15.6
zero of order m: Definition 15.20 in Section 15.7

16

Conformal Mapping

16.1 Conformal Mappings and the Laplace Equation

16.1.1 Linear Mappings

A **linear mapping** is a function defined by $f(z) = az + b$, where $a \neq 0$ and a and b are constants that may be complex. Here are special cases:

1. If $f(z) = z + b$, then f is a **translation**.
2. If $f(z) = \alpha z$ and α is real, then f is a **magnification** (or compression or dilation). In the future we may refer to a mapping as a magnification even if $|\alpha| < 1$.
3. If $f(z) = e^{i\varphi}z$, where φ is real, then f is a **rotation**, because $f(re^{i\theta}) = re^{i(\theta+\varphi)}$. Figure 16.1 illustrates this for $\varphi = \frac{\pi}{6}$.
4. If $f(z) = az$, where $a = \alpha e^{i\varphi}$ and α and φ are real, then f is a rotation followed by a magnification.

Finally, we see that a linear mapping $f(z) = az + b$ is a rotation, followed by a magnification, followed by a translation, as illustrated in Figure 16.2.

We observe that a linear mapping "preserves angles and orientation between curves." Let's check this symbolically: Suppose we have two parametrized curves that intersect at a point z_0, that is,

$$\begin{aligned} &\mathcal{C}_1 : z = z_1(t),\ -1 \le t \le 1 \quad \text{and} \quad \mathcal{C}_2 : z = z_2(t),\ -1 \le t \le 1 \\ &\quad \text{with} \quad z_0 = z_1(0) = z_2(0)\,. \end{aligned} \tag{16.1}$$

At z_0 the respective tangent vectors to $\mathcal{C}_1$ and $\mathcal{C}_2$ are $\dot{z}_1(0)$ and $\dot{z}_2(0)$, assuming both are nonzero. This is illustrated in Figure 16.3.

Under the linear mapping $f(z) = az + b$, the images of the parametrized curves $\mathcal{C}_1$ and $\mathcal{C}_2$ are $f(\mathcal{C}_1) \triangleq \{f(z_1(t)) : -1 \le t \le 1\}$ and $f(\mathcal{C}_2) \triangleq \{f(z_2(t)) : -1 \le t \le 1\}$, so at the point $f(z_0)$, the tangent vectors are $a\dot{z}_1(0)$ and $a\dot{z}_2(0)$, respectively. (We assume that $a \neq 0$ as part of the definition of "linear mapping.")

Write $\dot{z}_1(0) = |\dot{z}_1(0)|\, e^{i\theta_1}$ and $\dot{z}_2(0) = |\dot{z}_2(0)|\, e^{i\theta_2}$. Without loss of generality, $-\pi < \theta_1 < \theta_2 \le \theta_1 + \pi$. Then the angle between $\dot{z}_1(0)$ and $\dot{z}_2(0)$ is $\theta_2 - \theta_1$. Writing $a = \alpha e^{i\varphi}$, where α

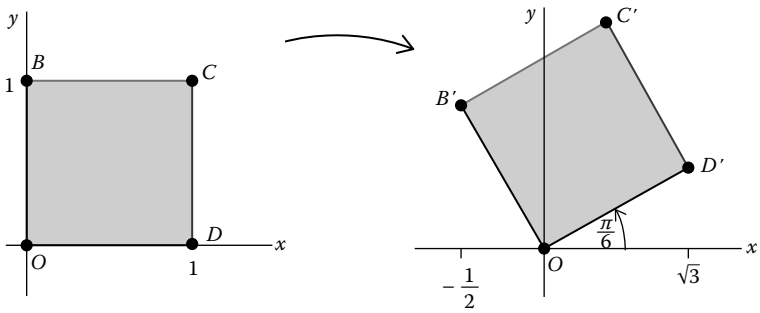


FIGURE 16.1
$f(z) = e^{i\pi/6}\, z$.

and φ are real, the angle between $a\dot{z}_1(0) = \alpha|\dot{z}_1(0)|\, e^{i(\theta_1+\varphi)}$ and $a\dot{z}_2(0) = \alpha|\dot{z}_2(0)|\, e^{i(\theta_2+\varphi)}$ is $(\theta_2 + \varphi) - (\theta_1 + \varphi) = \theta_2 - \theta_1$. This explains why all linear maps preserve angles between curves as well as the orientation $sgn(\theta_2 - \theta_1)$.

We will say that f is "conformal at z_0" if for every two curves $\mathcal{C}_1$ and $\mathcal{C}_2$ that intersect at z_0, the angle from $\mathcal{C}_1$ to $\mathcal{C}_2$ equals the angle from $f(\mathcal{C}_1)$ to $f(\mathcal{C}_2)$ at $f(z_0)$. A conformal map need not be linear.

Theorem 16.1

Suppose f is differentiable at z_0. Then f is conformal at z_0 if, and only if, $f'(z_0) \neq 0$.

Why? Suppose $\mathcal{C}_1$ and $\mathcal{C}_2$ are parametrized as in (16.1). By the chain rule, $\frac{d}{dt}\big[f(z_1(t))\big] = f'(z_1(t))\dot{z}_1(t)$. So, the tangent vectors to $f(\mathcal{C}_1)$ and $f(\mathcal{C}_2)$ at $f(z_0)$ are $f'(z_0)\dot{z}_1(0)$ and $f'(z_0)\dot{z}_2(0)$, respectively, *if* $f'(z_0) \neq 0$; however, if $f'(z_0) = 0$, then there are no tangent vectors to the parametrized curves $f(\mathcal{C}_1)$ and $f(\mathcal{C}_2)$ at $f(z_0)$.

Similar to the explanation why a linear mapping preserves angles, if $f'(z_0) \neq 0$, then the angle between $f(\mathcal{C}_1)$ and $f(\mathcal{C}_2)$ at $f(z_0)$ is $\big(Arg\,\big(f'(z_0)\big) + \theta_2\big) - \big(Arg\,\big(f'(z_0)\big) + \theta_1\big) = \theta_2 - \theta_1$, so f does preserve angles. □

Recall Definition 15.9(b) in Section 15.3.

Definition 16.1

Suppose $f(z)$ is defined on a domain $\mathcal{D}$.

(a) f is **conformal on** $\mathcal{D}$ if f is analytic on $\mathcal{D}$ and $f'(z_0) \neq 0$ at every z_0 in $\mathcal{D}$.
(b) f is **conformal at a boundary point** z_0 if

$$\lim_{z \text{ is in } \mathcal{D} \text{ and } z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} \text{ exists and is } \neq 0.$$

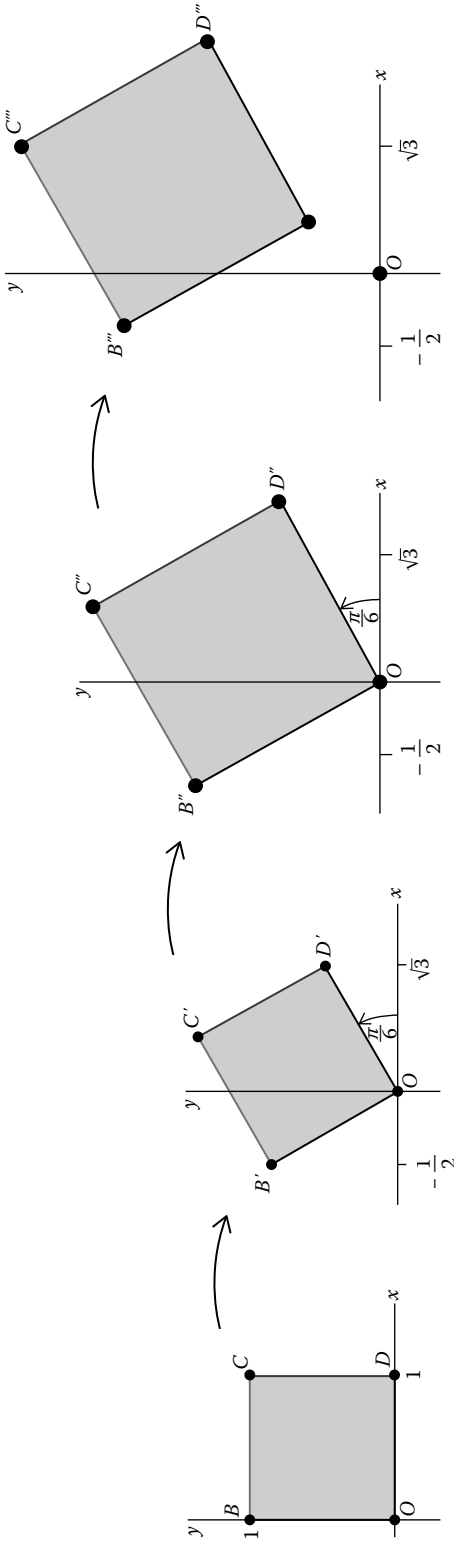


FIGURE 16.2
$f(z) = az + b$.

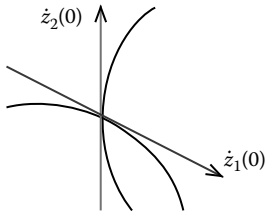


FIGURE 16.3
Tangent vectors $\dot{z}_1(0)$, $\dot{z}_2(0)$.

Example 16.1

$f(z) \triangleq \frac{1+z}{1-z}$ is defined on the domain $\mathcal{D}_1(0) = \{z : |z| < 1\}$. Study conformality of $f(z)$ on its domain and at its boundary points.

Method: $f'(z) = \frac{2}{(1-z)^2}$, so f is conformal on domain $\mathcal{D}_1(0)$ and also at every boundary point of $\mathcal{D}_1(0)$ *except* at $z = 1$. ◯

Theorem 16.2

If g is conformal at z_0 and f is conformal at $g(z_0)$, then $f(g(z))$ is conformal at z_0.

Why? This is a straightforward consequence of the chain rule. □

16.1.2 Harmonic Functions

Recall that a function ϕ is harmonic on a domain $\mathcal{D}$ if $\phi(x, y)$ satisfies Laplace's equation

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = 0$$

on $\mathcal{D}$. Also, recall that if $F(x, y) = \phi(x, y) + i\psi(x, y)$ is analytic on $\mathcal{D}$, then both ϕ and its harmonic conjugate ψ are harmonic on $\mathcal{D}$.

A conformal mapping can transform Laplace's equation in a complicated domain into Laplace's equation in a simpler domain. After solving Laplace's equation in the simpler domain, we can use the inverse of the conformal mapping to get a solution of Laplace's equation in the original domain. This fits into a general mathematical method of reducing a new difficult problem to a simpler problem or even to a problem for which the solution is well known.

To use this method, we need two results:

Theorem 16.3

If $w = f(z)$ is conformal on domain $\mathcal{D}$ in the z-plane then (a) $f(\mathcal{D})$ is a domain in the w- plane, and (b) there exists an inverse map $z = f^{-1}(w)$ that is conformal on $f(\mathcal{D})$.

Under the assumptions of Theorem 16.3, $f(f^{-1}(w)) = w$ on $f(\mathcal{D})$ and $f^{-1}(f(z)) = z$ on $\mathcal{D}$.

Theorem 16.4

If $f(z)$ is conformal on domain $\mathcal{D}$, $w = u + iv = f(z)$, and $\Phi = \Phi(u, v)$ is harmonic on domain $\mathcal{E} = f(\mathcal{D})$, then $\phi(x, y) \triangleq \Phi\left(f(x + iy)\right)$ is harmonic on $\mathcal{D}$.

Why? Let z_0 be any point in $\mathcal{D}$ and denote $w_0 = f(z_0)$. Because Φ is harmonic on $\mathcal{E}$, Theorem 15.14 in Section 15.3 implies that there exists a local harmonic conjugate Ψ such that $g(w) \triangleq \Phi(u, v) + i\Psi(u, v)$ is analytic on some open ball $B_\delta(w_0)$ contained in $\mathcal{E}$. This implies that

$$\Phi\left(f(x + iy)\right) + i\Psi\left(f(x + iy)\right) = g\left(f(x + iy)\right)$$

is analytic on $f^{-1}\left(B_\delta(z_0)\right)$; hence, $\Phi\left(f(x + iy)\right)$ is harmonic on $f^{-1}\left(B_\delta(z_0)\right)$. Since this is true for any z_0 in $\mathcal{D}$, $\phi(x, y)$ is harmonic on $\mathcal{D}$. □

16.1.3 Elementary Functions

We will study the conformality of some familiar functions: exponential, trigonometric, logarithmic, and a new class of functions called "Möbius transformations."

Example 16.2

Why is e^z conformal everywhere?

Method: At all z, there exists $(e^z)' = e^z$, and for all $z = x + iy$ with x, y real, $|(e^z)'| = |e^{x+iy}| = |e^x\, e^{iy}| = |e^x|\, |e^{iy}| = e^x \cdot 1 \neq 0$. ◯

Example 16.3

Where is $\sin z$ conformal?

Method: At all z, there exists $(\sin z)' = \cos z$, so all we have to do is to find the solutions of $\cos z = 0$. For all real x, y, $\cos(x + iy) = \cdots = \cos x \cosh y - i \sin x \sinh y$. Separating the real and imaginary parts of the equation $\cos(x + iy) = 0 + i0$ gives the system of equations

$$\left\{\begin{array}{ll}(1) & 0 = \cos x \cosh y \\ (2) & 0 = \sin x \sinh y\end{array}\right\}.$$

Because $\cosh y > 0$ at all real y, the only solutions of equation (1) have $x = \left(n - \frac{1}{2}\right)\pi$. Substitute that into equation (2) to get

$$0 = \sin\left(\left(n - \frac{1}{2}\right)\pi\right) \sinh y = (-1)^{n+1} \sinh y,$$

so only $y = 0$ solves equation (2) for those values of x.

So, $\sin z$ is conformal everywhere except at the points $z = \left(n - \frac{1}{2}\right)\pi + i0$. ◯

16.1.4 Möbius Transformations

Möbius transformations are defined by

$$w = M(z) \triangleq \frac{az+b}{cz+d}, \tag{16.2}$$

where a, b, c, d are constants that may be complex and $ad - bc \neq 0$. The latter assumption implies at least one of c and d is nonzero. We will see in Section 16.2 that having Möbius transformations in our "library" will help us find a mapping to circles and lines, and in Section 16.3, we will use this to solve Laplace's equation in some complicated domains.

The assumption that $ad - bc \neq 0$ guarantees that $M(z)$ is not a constant function. Why? As we can see in the case when $c \neq 0$,

$$M(z) = \frac{\frac{a}{c}(cz+d) + b - \frac{ad}{c}}{cz+d} = \frac{a}{c} - \frac{ad-bc}{c(cz+d)}. \tag{16.3}$$

In Problem 16.1.5.1, you will discuss the criterion $ad - bc \neq 0$ in the case when $c = 0$ but $d \neq 0$.

We calculate that

$$M'(z) = \frac{ad-bc}{(cz+d)^2}$$

is defined and nonzero everywhere except at $z = -\frac{d}{c}$.

Theorem 16.5

Suppose a Möbius transformation is defined by (16.2):

(a) If $c \neq 0$, then $M(z)$ is conformal except at $z = -\frac{d}{c}$.
(b) If $c = 0$, then $M(z)$ is a linear mapping that is conformal everywhere.

Note that the basic assumption $ad - bc \neq 0$ implies $a \neq 0$ in case (b), that is, if $c = 0$.

The logarithmic transformation $w = Log(z) = \ln|z| + iArg(z)$ is clearly conformal everywhere except on the nonpositive real axis.

Example 16.4

Where is the mapping $w = h(z) \triangleq Log\left(\frac{1+z}{1-z}\right)$ conformal?

Method: Define $\zeta = g(z) \triangleq \frac{1+z}{1-z}$. By Theorem 16.2, $h(z)$ is conformal wherever $g(z)$ is both conformal and not equal to a nonpositive real number. By Theorem 16.5, $g(z)$ is conformal everywhere except at $z = 1$.

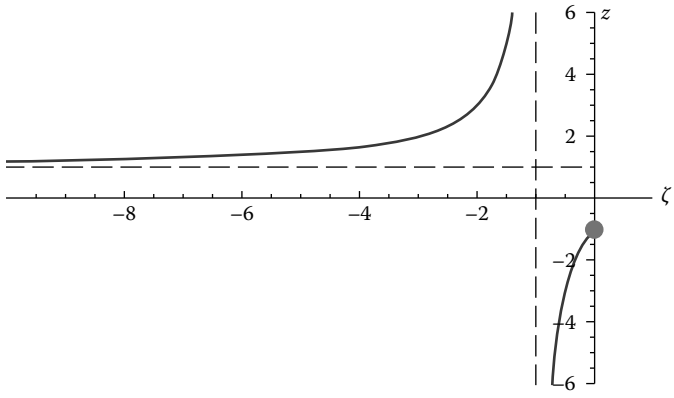


FIGURE 16.4
Example 16.4.

For which z is $\zeta = \frac{1+z}{1-z}$ a nonpositive real number? To answer this, it helps to invert the dependence to get z as a function of ζ:

$$\zeta = \frac{1+z}{1-z} \iff (1-z)\zeta = 1+z \iff \zeta - 1$$
$$= (1+\zeta)z \iff z = k(\zeta) \triangleq \frac{\zeta-1}{\zeta+1} = 1 - \frac{2}{\zeta+1}.$$

The set of all z for which ζ is real and ≤ 0 is the set $k\big((-\infty, 0]\big)$, as shown in Figure 16.4. Basic calculus graphical techniques give that ζ is real and $\leq 0 \iff z \leq -1$ or $z > 1$.

So, the set of z for which ζ is *not* a nonpositive real number is everywhere except the intervals $(-\infty, -1]$ and $(1, \infty)$ on the real axis.

Finally, we conclude that $g(z)$ is conformal on

$$\{z : \mathcal{I}m(z) \neq 0\} \cup \{z : z \text{ is real and } -1 < z < 1\}. \bigcirc$$

Often, using the inverse of a mapping is a valuable technique for understanding the original mapping.

16.1.5 Problems

1. In the case when $c = 0$ and $d \neq 0$, why is the Möbius transformation given in (16.2) not constant if, and only if, $ad - bc \neq 0$?

In problems 2–5, find where the given function is conformal.

2. $Log\left(\frac{2-z}{2+z}\right)$
3. $Log\left(\frac{3+z}{3-z}\right)$
4. $\cos z$

5. $\cosh z$
6. Use the fact that $\sin z = \frac{1}{2i}(e^{iz} - e^{-iz})$ to solve Example 16.3, that is, find where $\sin z$ conformal, by another method.

16.2 Möbius Transformations

Visualizing the effects of a mapping of the complex plane will help us find and use mappings to solve the Laplace equation in Section 16.3. As we saw in Section 15.1, one way to visualize a mapping is to show a region A in the z-plane along with its image, $f(A)$, in the w-plane. In particular, it is useful to know the images of lines and circles under commonly used mappings such as Möbius transformations, given by (16.2) in Section 16.1, that is,

$$w = M(z) = \frac{az + b}{cz + d}.$$

16.2.1 Circles, Lines, and Möbius Transformations

Other than linear mappings, the simplest Möbius transformation is the **inversion mapping** given by

$$w = f(z) = \frac{1}{z}. \tag{16.4}$$

Lemma 16.1

For any $R > 0$ and z_0 in $\mathbb{C}$, define the circle $\mathcal{C} = \{z : |z - z_0| = R\}$.

(a) If 0 is not in $\mathcal{C}$, that is, $|z_0| \neq R$, then the image of $\mathcal{C}$ under the inversion mapping is the circle

$$f(\mathcal{C}) = \left\{ w : |w - w_0| = \frac{R}{\big|\, |z_0|^2 - R^2 \,\big|} \right\} \tag{16.5}$$

whose center is at

$$w_0 = \frac{\overline{z_0}}{|z_0|^2 - R^2}. \tag{16.6}$$

(b) If 0 is in $\mathcal{C}$, that is, $|z_0| = R > 0$, then the image of $\widehat{\mathcal{C}} \triangleq \{z \neq 0 : |z - z_0| = R\}$ under the inversion mapping is

$$\left\{ w = \frac{1}{1 + e^{i\varphi}} \cdot \frac{1}{z_0} : 0 < \varphi \leq 2\pi, \varphi \neq \pi \right\}, \tag{16.7}$$

which is all of the points on a line that passes through $\frac{1}{2z_0}$.

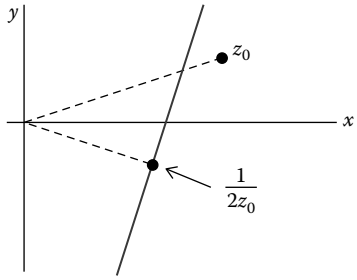


FIGURE 16.5
Lemma 16.1(b).

Why? In Problem 16.2.5.23, you will derive the result in part (a). For part (b), we note that every z in $\widehat{\mathcal{C}}$ can be written in the form

$$z = z_0 + |z_0|\, e^{i\theta} = z_0 + z_0\, e^{-iArg(z_0)}\, e^{i\theta} = z_0 + z_0\, e^{i\varphi}$$

for some $\varphi \neq \pi$ with $0 < \varphi \leq 2\pi$. Under the inversion map, its image is

$$w = \frac{1}{z} = \frac{1}{z_0}\,\frac{1}{1+e^{i\varphi}} = \frac{1}{z_0}\,\frac{1+e^{-i\varphi}}{|1+e^{i\varphi}|^2} = \cdots = \frac{1}{z_0}\left(\frac{1+\cos\varphi}{2(1+\cos\varphi)} - i\,\frac{\sin\varphi}{2(1+\cos\varphi)}\right)$$

$$= \frac{1}{2z_0} - i\left(\frac{\sin\varphi}{1+\cos\varphi}\right)\frac{1}{2z_0}.$$

Figure 16.5 illustrates the image, which is a line whose direction vector is $-\frac{i}{2z_0}$. Multiplication of a complex number by $-i$ rotates it by $\frac{-\pi}{2}$, and in Problem 16.2.5.25, you will explain why $\left\{\frac{\sin\varphi}{1+\cos\varphi} : 0 < \varphi \leq 2\pi \text{ with } \varphi \neq \pi\right\} = (-\infty, \infty)$.

So, the image of $\widehat{\mathcal{C}}$ under the inversion map is a line passing through the point $\frac{1}{2z_0}$. □

Here are special cases: If z_0, the center of the circle that passes through the origin, is real, then under the inversion map the image of the circle is a vertical line in the w-plane; if z_0 is imaginary, then the image is a horizontal line.

Recall the basic assumption $ad - bc \neq 0$. Assuming also $c \neq 0$, the Möbius transformation defined in (16.2) in Section 16.1, that is, $w = M(z) = \frac{az+b}{cz+d}$, can be written as in (16.3) in Section 16.1, that is,

$$M(z) = \frac{a}{c} - \frac{ad - bc}{c(cz + d)}.$$

Briefly we will explain why the image of every circle is either a circle or a line: Suppose $\mathcal{C}$ is any circle *and* $-\frac{d}{c}$ is *not* in $\mathcal{C}$. By (16.3) in Section 16.1, M can be written as composition of linear mappings and the inversion mapping, specifically

$$M(z) = g_3\left(g_2\left(f\left(g_1(z)\right)\right)\right),$$

where
$g_1(z) \triangleq cz + d$
$f(w_1) \triangleq \frac{1}{w_1}$
$g_2(w_2) \triangleq \frac{ad-bc}{c} w_2$
$g_3(w_3) \triangleq \frac{a}{c} - w_3$

Each of the linear maps g_1, g_2, g_3 maps circles to circles and maps lines to lines. In addition, because we assumed $-\frac{d}{c}$ is not in $\mathcal{C}$, $w_1 \triangleq cz + d \neq 0$ is not in the image $g_1(\mathcal{C})$. By Lemma 16.1(a), $f(g_1(\mathcal{C}))$ is a circle. Because g_2 and g_3 are linear mappings, $M(\mathcal{C}) = g_3\left(g_2\left(f\left(g_1(\mathcal{C})\right)\right)\right)$ is a circle.

On the other hand, suppose $\mathcal{C}$ is any circle *and* $-\frac{d}{c}$ *is* in $\mathcal{C}$. Then $w_1 = 0$ is in the image $g_1(\mathcal{C})$, so Lemma 16.1(b) implies $f\left(\widehat{g_1(\mathcal{C})}\right)$ is a line, where

$$\widehat{g_1(\mathcal{C})} \triangleq \left\{w_1 \neq 0 : w_1 = g_1(z) \text{ for some } z \text{ in } \mathcal{C}\right\}.$$

Because g_2 and g_3 are linear mappings, $M(\mathcal{C}) = g_3\left(g_2\left(f\left(\widehat{g_1(\mathcal{C})}\right)\right)\right)$ is a line in the w-plane.

If L is a line in the z-plane, then there are likewise two cases depending on whether 0 is in $g_1(L)$. As before, assume f is the inversion mapping. We define a **punctured line** to be all points on a line except for one excluded point.

Lemma 16.2

Define a line $L = \{z = z_0 + tz_1 : -\infty < t < \infty\}$, where z_0 and z_1 are complex numbers:

(a) If 0 is not in L, then $f(L) \cup \{0\}$ is a circle through the origin.
(b) If 0 is in L, then, defining $\widehat{L} = \{z \neq 0 : z \text{ is in } L\}$, we have $f(\widehat{L})$ as a punctured line.

Why?

$$u + iv = w = \frac{1}{z} \iff x + iy = z = \frac{1}{w} = \frac{1}{u + iv} = \frac{u}{u^2 + v^2} + i\,\frac{-v}{u^2 + v^2}$$

is on the line L, that is, has $x + iy = z = z_0 + tz_1 = x_0 + tx_1 + i(y_0 + ty_1)$ for some t. It follows that

$$\frac{u}{u^2 + v^2} - x_0 = tx_1 \quad \text{and} \quad ty_1 = \frac{-v}{u^2 + v^2} - y_0;$$

hence,

$$\frac{uy_1}{u^2+v^2} - x_0y_1 = tx_1y_1 = \frac{-vx_1}{u^2+v^2} - y_0x_1.$$

So,

$$y_1u + x_1v = \left(x_0y_1 - y_0x_1\right)\left(u^2+v^2\right). \tag{16.8}$$

(a) If 0 is not in L, then the vectors z_0 and z_1 are linearly independent, that is,

$$0 \neq \begin{vmatrix} x_0 & x_1 \\ y_0 & y_1 \end{vmatrix} = x_0y_1 - x_1y_0.$$

In this case, we can divide (16.8) by $\left(x_0y_1 - x_1y_0\right)$ to get

$$u^2+v^2 = \frac{1}{x_0y_1 - x_1y_0}\left(y_1u + x_1v\right),$$

which is a circle whose center is at

$$u_0 + iv_0 = \frac{y_1}{2(x_0y_1 - x_1y_0)} + i\,\frac{x_1}{2(x_0y_1 - x_1y_0)},$$

passes through the origin $(u,v) = (0,0)$, and has radius

$$\frac{\sqrt{x_1^2+y_1^2}}{2\left|x_0y_1 - x_1y_0\right|}.$$

(b) If 0 is in L, then $L = \{tz_1 : -\infty < t < \infty\}$. Define $\widehat{L} = \{tz_1 : t \neq 0\} = \{z \neq 0 : z \text{ is in } L\}$. Then

$$f(\widehat{L}) = \left\{\frac{1}{t}\cdot\frac{1}{z_1} : t \neq 0\right\}$$

is a punctured line. □

16.2.2 Mapping Two Given Circles to Two Concentric Circles

In Section 11.5, we saw how to solve the Laplace equation inside a disk, outside a disk, or inside an annulus. The latter is the region between two concentric circles. Suppose we want to solve the Laplace equation in a region $\mathcal{D}$ that lies between two *non*concentric circles, that is, between two **eccentric circles**.

Here we will see that we can find a Möbius transformation M that maps $\mathcal{D}$ to an annulus in the w-plane. In Section 16.3, we will solve the Laplace equation in the annulus and thus find a solution of the Laplace equation in the original region, $\mathcal{D}$, using Theorem 16.4 in Section 16.1.

Lemma 16.3

If three points w_1, w_2, and w_3 do not lie on a line, that is, are not **collinear**, then there is exactly one circle that contains w_1, w_2, and w_3.

Why? You will explain why in Problem 16.2.5.24.

Theorem 16.6

Suppose real numbers $\delta \neq 0$ and $\gamma > 0$ are given, and denote circles $\mathcal{C}_1 : |z| = 1$ and $\mathcal{C}_2 : |z - \delta| = \gamma$. Suppose $\mathcal{C}_1$ and $\mathcal{C}_2$ have no point in common. Then we can find real numbers α, β such that the Möbius transformation*

$$w = M(z) = \frac{z - \alpha}{z - \beta} \tag{16.9}$$

maps each of the circles $\mathcal{C}_1$ and $\mathcal{C}_2$ to concentric circles centered at 0 in the w-plane, as long as it turns out that

$$z = \beta \text{ is on neither } \mathcal{C}_1 \text{ nor } \mathcal{C}_2 . \tag{16.10}$$

Here,

$$\beta = \frac{1}{\alpha}, \tag{16.11}$$

the real number α satisfies

$$\delta\alpha^2 + (\gamma^2 - \delta^2 - 1)\alpha + \delta = 0, \tag{16.12}$$

and the images of the circles are

$$M(\mathcal{C}_1) : |w| = |\alpha| \quad \text{and} \quad M(\mathcal{C}_2) : |w| = \left| \frac{\delta - \alpha}{\gamma} \right| . \tag{16.13}$$

Why? Note that $\delta \neq 0$ and (16.12) imply $\alpha \neq 0$. Dividing (16.12) by $(-\alpha)$ and substituting in (16.11) implies that

$$(\delta - \alpha)(\delta - \beta) = \gamma^2 . \tag{16.14}$$

In Problem 16.2.5.19, you will explain why the image of $\mathcal{C}_1 : |z| = 1$ is $M(\mathcal{C}_1) : |w| = |\alpha|$.

To explain why the image $M(\mathcal{C}_2)$ is the circle $|w| = \left| \frac{\delta - \alpha}{\gamma} \right|$, first we note that $M(\mathcal{C}_2)$ is a circle by Lemma 16.1 and the assumption that $z = \beta$ is not on $\mathcal{C}_2$. By Lemma 16.3, it will

* This theorem is an interpretation of results in Keener (1988). It appears that the conclusions and derivation of theorem 16.6 give a specific and useful new method.

suffice to show that M maps three distinct (hence, not collinear) points on the circle $\mathcal{C}_2$ to three distinct (hence, not collinear) points on the circle $|w| = \left|\frac{\delta-\alpha}{\gamma}\right|$.

The points $z = \delta + \gamma$ and $\delta \pm i\gamma$ are distinct and lie on the circle $\mathcal{C}_2$. Our calculations as follows will establish that $|M(\delta+\gamma)| = |M(\delta \pm i\gamma)| = \left|\frac{\delta-\alpha}{\gamma}\right|$. First,

$$\begin{aligned}
|M(\delta \pm i\gamma)| &= \left|\frac{\delta \pm i\gamma - \alpha}{\delta \pm i\gamma - \beta}\right| = \left|\frac{(\delta-\alpha) \pm i\gamma}{(\delta-\beta) \pm i\gamma}\right| = \left|\frac{((\delta-\alpha) \pm i\gamma)\gamma}{((\delta-\beta) \pm i\gamma)\gamma}\right| = \left|\frac{(\delta-\alpha)\gamma \pm i\gamma^2}{((\delta-\beta) \pm i\gamma)\gamma}\right| \\
&= \left|\frac{(\delta-\alpha)\gamma \pm i(\delta-\alpha)(\delta-\beta)}{((\delta-\beta) \pm i\gamma)\gamma}\right| = \left|\frac{(\delta-\alpha)(\gamma \pm i(\delta-\beta))}{((\delta-\beta) \pm i\gamma)\gamma}\right| = \frac{|\delta-\alpha|\,\cancel{|\gamma \pm i(\delta-\beta)|}}{|\gamma|\,\cancel{|\delta-\beta) \pm i\gamma|}} \\
&= \frac{|\delta-\alpha|}{|\gamma|},
\end{aligned}$$

because $|\gamma \pm i(\delta-\beta)| = \sqrt{\gamma^2 + (\delta-\beta)^2} = |(\delta-\beta) \pm i\gamma|$.

Also, we calculate

$$\begin{aligned}
|M(\delta+\gamma)| &= \left|\frac{\delta+\gamma-\alpha}{\delta+\gamma-\beta}\right| = \left|\frac{((\delta-\alpha)+\gamma)\gamma}{((\delta-\beta)+\gamma)\gamma}\right| = \left|\frac{(\delta-\alpha)\gamma+\gamma^2}{((\delta-\beta)+\gamma)\gamma}\right| = \left|\frac{(\delta-\alpha)\gamma+(\delta-\alpha)(\delta-\beta)}{((\delta-\beta)+\gamma)\gamma}\right| \\
&= \left|\frac{(\delta-\alpha)(\gamma+(\delta-\beta))}{((\delta-\beta)+\gamma)\gamma}\right| = \frac{|\delta-\alpha|\,\cancel{|\gamma+(\delta-\beta)|}}{|\gamma|\,\cancel{|\delta-\beta)+\gamma|}} = \frac{|\delta-\alpha|}{|\gamma|}.
\end{aligned}$$

All that remains is to explain why the points $M(\delta \pm i\gamma)$, $M(\delta+\gamma)$ are distinct.

In Problem 16.2.5.20 you will explain why $\beta = \frac{1}{\alpha}$, along with Theorem 16.6's hypothesis (16.10), that is, "$z = \beta$ is on neither $\mathcal{C}_1$ nor $\mathcal{C}_2$," imply that the solutions of quadratic equation (16.12) are *not* $\alpha = \pm 1$. In Problem 16.2.5.21, you will derive that

$$M(\delta + i\gamma) = M(\delta - i\gamma) \iff \alpha = \beta \iff \alpha = \pm 1,$$

so hypothesis (16.10) guarantees that $M(\delta+i\gamma) \neq M(\delta-i\gamma)$. In Problem 16.2.5.22, you will derive that

$$M(\delta + i\gamma) = M(\delta + \gamma) \iff (\delta-\alpha)\gamma - \gamma^2 = \mp i\alpha\gamma,$$

which is impossible because δ, γ, and α are real and $\alpha \neq 0$.

Having seen that M maps three noncollinear points on the circle $\mathcal{C}_2$ to three noncollinear points on the circle $|w| = \left|\frac{\delta-\alpha}{\gamma}\right|$ we conclude that M is as desired. ◯

The quadratic equation in (16.12) can be rewritten as

$$\alpha^2 + \frac{\gamma^2 - \delta^2 - 1}{\delta}\,\alpha + 1 = 0, \tag{16.15}$$

so $\alpha_- \cdot \alpha_+ = 1$ and $\alpha_\pm^{-1} = \alpha_\mp$.

Example 16.5

(a) Find a Möbius transformation that maps each of the circles $|z| = 1$ and $|z - 1| = 3$ to a circle whose centers are at 0 in the w-plane. (b) Does your transformation map the region between the circles $|z| = 1$ and $|z - 1| = 3$ to an annulus?

Method: (a) With $\delta = 1$ and $\gamma = 3$, Theorem 16.6 says to let $M(z) = \frac{z-\alpha}{z-\beta}$, where

$$\beta = \frac{1}{\alpha} \quad \text{and} \quad 1 \cdot \alpha^2 + (3^2 - 1^2 - 1)\alpha + 1 = 0.$$

The latter is a quadratic equation whose solutions are

$$\alpha = \alpha_{\pm} \triangleq \frac{-7 \pm 3\sqrt{5}}{2}.$$

As expected, $\alpha_- \cdot \alpha_+ = 1$ so $\alpha_{\pm}^{-1} = \alpha_{\mp}$.

(1) With $\alpha = \alpha_+ = \frac{-7+3\sqrt{5}}{2}$, we get the Möbius transformation

$$M_+(z) \triangleq \frac{z - \frac{-7+3\sqrt{5}}{2}}{z - \frac{-7-3\sqrt{5}}{2}} = \frac{2z + 7 - 3\sqrt{5}}{2z + 7 + 3\sqrt{5}}.$$

By Theorem 16.6's conclusion (16.13), M_+ maps

$$\mathcal{C}_1 : |z| = 1 \quad \mapsto \quad M_+(\mathcal{C}_1) : |w| = \tfrac{7-3\sqrt{5}}{2}, \quad \text{and}$$

$$\mathcal{C}_2 : |z - 1| = 3 \quad \mapsto \quad M_+(\mathcal{C}_2) : |w| = \tfrac{3-\sqrt{5}}{2}.$$

$\mathcal{C}_1$ is inside $\mathcal{C}_2$. Because $3 - \sqrt{5} > 7 - 3\sqrt{5}$, $M_+(\mathcal{C}_1)$ is inside $M_+(\mathcal{C}_2)$. We see that M_+ preserves the interior versus exterior relationship between $\mathcal{C}_1$ and $\mathcal{C}_2$ (Figure 16.6).

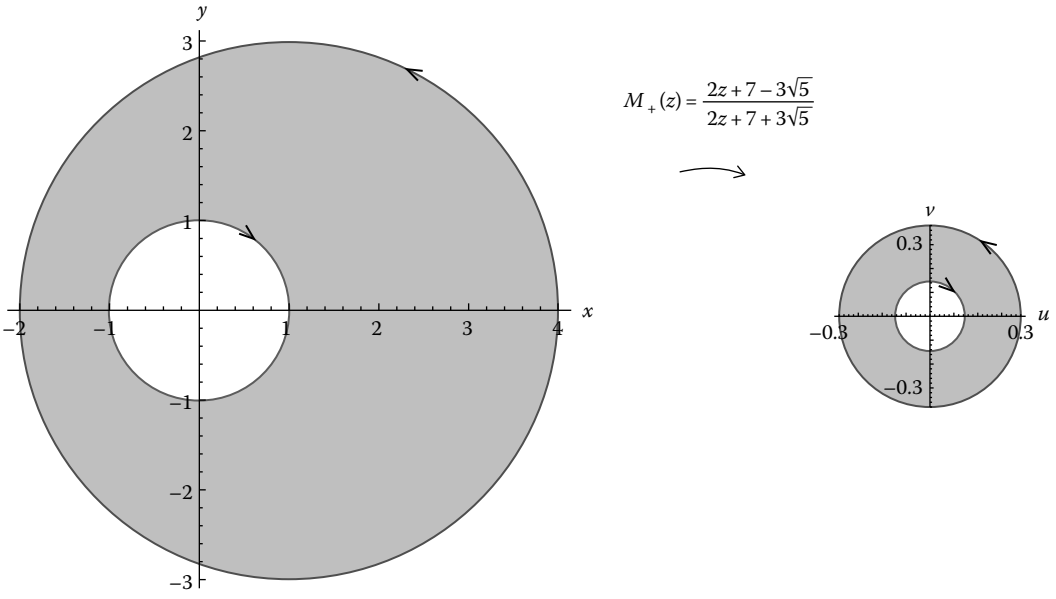


FIGURE 16.6
Example 16.5(a)(1).

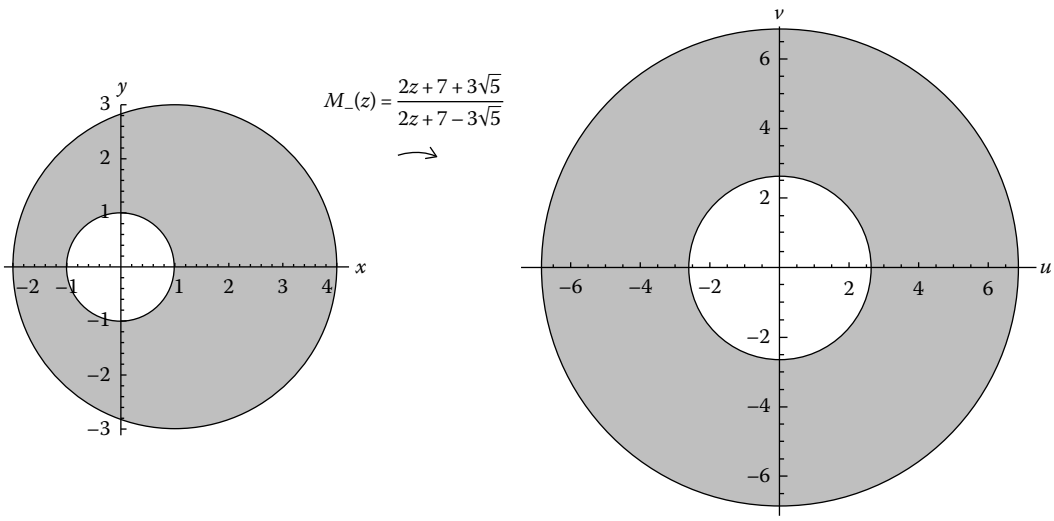


FIGURE 16.7
Example 16.5(a)(2).

(2) With $\alpha = \alpha_- = \frac{-7-3\sqrt{5}}{2}$, we get the Möbius transformation

$$M_-(z) \triangleq \frac{z - \frac{-7-3\sqrt{5}}{2}}{z - \frac{-7+3\sqrt{5}}{2}} = \frac{2z+7+3\sqrt{5}}{2z+7-3\sqrt{5}}.$$

By Theorem 16.6's conclusion (16.13), M_- maps

$$\mathcal{C}_1 : |z| = 1 \quad \mapsto \quad M_-(\mathcal{C}_1) : |w| = \tfrac{7+3\sqrt{5}}{2} \quad \text{and}$$

$$\mathcal{C}_2 : |z-1| = 3 \quad \mapsto \quad M_-(\mathcal{C}_2) : |w| = \tfrac{3+\sqrt{5}}{2}.$$

Because $\mathcal{C}_1$ is inside $\mathcal{C}_2$ and $M_-(\mathcal{C}_2)$ is inside $M_-(\mathcal{C}_1)$, we see that M_- reverses the interior versus exterior relationship between $\mathcal{C}_1$ and $\mathcal{C}_2$ (Figure 16.7).

(b) M_+ has a pole only at $z = -\frac{7+3\sqrt{5}}{2}$, which is not in the connected set $\mathcal{R}$ consisting of the circles $\mathcal{C}_1 : |z| = 1$, $\mathcal{C}_2 : |z-1| = 3$ and all of the points between them. So, M_+, is continuous on $\mathcal{R}$. It follows that the image, $M_+(\mathcal{R})$, is the connected set consisting of $M_+(\mathcal{C}_1)$, $M_+(\mathcal{C}_2)$, and all of the points between them, that is, the annulus $\frac{7-3\sqrt{5}}{2} < r < \frac{3-\sqrt{5}}{2}$.

Similarly, M_- has a pole only at $z = \frac{-7+3\sqrt{5}}{2}$, which is not in the same region $\mathcal{R}$ as defined in the paragraph previously. By the same reasoning process, it follows that the image, $M_+(\mathcal{R})$, is the annulus $\frac{3+\sqrt{5}}{2} < r < \frac{7+3\sqrt{5}}{2}$. ○

What if the given circles are not of the form $|z| = 1$ and $|z-\delta| = \gamma$? The general situation can be handled by Theorem 16.6 after performing some preliminary manipulations.

Example 16.6

Find a Möbius transformation that maps each of the circles $|z| = 5$ and $|z - 2| = 2$ to circles whose centers are at 0 in the w-plane.

Method: Note that 0 being on one of the given circles is not a problem because we will not use the inversion mapping on the given circle.

The change of variables $z = 5\widetilde{z}$ maps the given circles to

$$|5\widetilde{z}| = 5 \quad \text{and} \quad 2 = |z - 2| = |5\widetilde{z} - 2|,$$

that is,

$$|\widetilde{z}| = 1 \quad \text{and} \quad \left|\widetilde{z} - \frac{2}{5}\right| = \frac{2}{5}.$$

With $\delta = \frac{2}{5}$ and $\gamma = \frac{2}{5}$, Theorem 16.6 and (16.15) say to let $M(z) = \frac{z-\alpha}{z-\beta}$, where

$$\beta = \frac{1}{\alpha} \quad \text{and} \quad 0 = \alpha^2 - \frac{5}{2}\alpha + 1 = (\alpha - 2)\left(\alpha - \frac{1}{2}\right),$$

so $\alpha_+ = 2$ and $\alpha_- = \frac{1}{2}$. As expected, $\alpha_- \cdot \alpha_+ = 1$ and $\alpha_\pm^{-1} = \alpha_\mp$.

(1) With $\alpha = \alpha_+ = 2$, we get the Möbius transformation

$$M_+(\widetilde{z}) \triangleq \frac{\widetilde{z} - 2}{\widetilde{z} - \frac{1}{2}} = \frac{2\widetilde{z} - 4}{2\widetilde{z} - 1};$$

substituting back $\widetilde{z} = z/5$, we get the Möbius transformation

$$M_+(z) \triangleq \frac{(2z/5) - 4}{(2z/5) - 1} = \frac{2z - 20}{2z - 5}.$$

By Theorem 16.6's conclusion (16.13), M_+ maps

$$\mathcal{C}_1 : |z| = 5 \quad \mapsto \quad M_+(\mathcal{C}_1) : |w| = 2 \quad \text{and}$$

$$\mathcal{C}_2 : |z - 2| = 2 \quad \mapsto \quad M_+(\mathcal{C}_2) : |w| = 4.$$

Because $\mathcal{C}_2$ is inside $\mathcal{C}_1$ and $M_+(\mathcal{C}_1)$ is inside $M_+(\mathcal{C}_2)$, we see that M_+ reverses the interior versus exterior relationship between $\mathcal{C}_1$ and $\mathcal{C}_2$.

In fact, we note for future reference that if we parametrize $\mathcal{C}_1$ with positive orientation, then $M_+(\mathcal{C}_1)$ is parametrized with negative orientation. So, the interior of the closed parametrized curve $\mathcal{C}_1$ is mapped by M_+ to the exterior of the closed parametrized curve $M_+(\mathcal{C}_1)$.

Similarly, if we parametrize $\mathcal{C}_2$ with negative orientation, then $M_+(\mathcal{C}_2)$ is parametrized with positive orientation. So, the exterior of the closed, parametrized curve $\mathcal{C}_2$ is mapped by M_+ to the interior of the closed curve $M_+(\mathcal{C}_2)$. The orientations are also shown in Figure 16.8.

(2) With $\alpha = \alpha_- = \frac{1}{2}$, we get the Möbius transformation

$$M_-(\widetilde{z}) \triangleq \frac{\widetilde{z} - (1/2)}{\widetilde{z} - 2},$$

that is,

$$M_-(z) = \frac{(z/5) - (1/2)}{(z/5) - 2} = \frac{2z - 5}{2z - 20}.$$

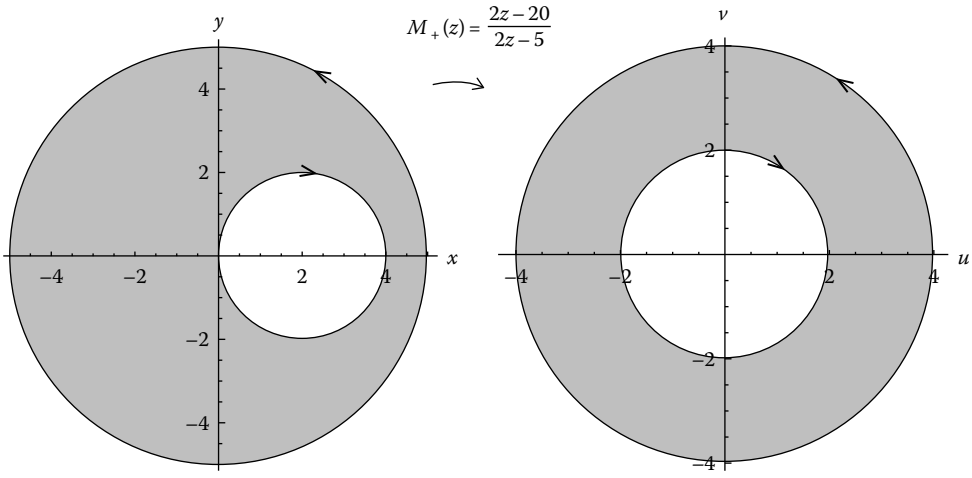


FIGURE 16.8
Example 16.6(1).

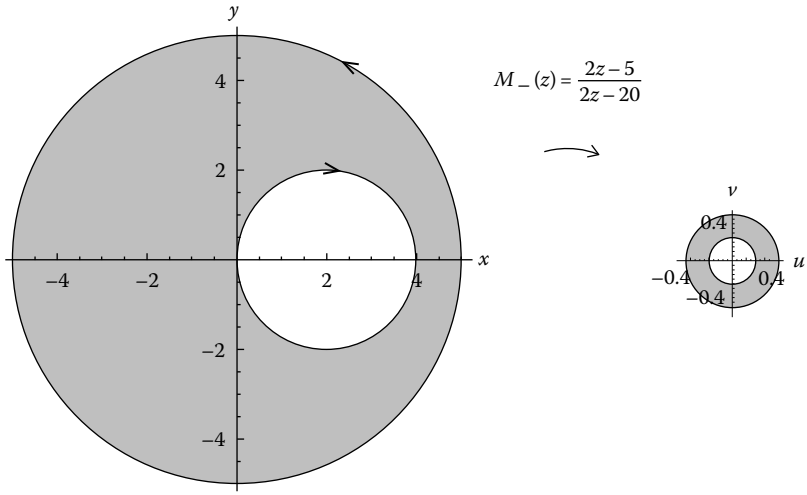


FIGURE 16.9
Example 16.6(2).

By Theorem 16.6's conclusion (16.13), M_- maps

$$\mathcal{C}_1 : |z| = 5 \quad \mapsto \quad M_-(\mathcal{C}_1) : |w| = \tfrac{1}{2} \quad \text{and}$$

$$\mathcal{C}_2 : |z-2| = 2 \quad \mapsto \quad M_-(\mathcal{C}_2) : |w| = \tfrac{1}{4}.$$

Because $\mathcal{C}_2$ is inside $\mathcal{C}_1$ and $M_-(\mathcal{C}_2)$ is inside $M_-(\mathcal{C}_1)$, we see that M_- preserves the interior versus exterior relationship between $\mathcal{C}_1$ and $\mathcal{C}_2$ (Figure 16.9). ○

What if neither of the given circles is centered at the origin? This situation requires an additional preliminary step: translate the center of one of the given circles to the origin by

a linear mapping $z = \widetilde{z} + (constant)$. For example, suppose we wanted to map the circles $|z-3| = 2$ and $|z-1| = 5$ to circles centered at $w = 0$. Let $z = \widetilde{z} + 3$, so the given circles are $|\widetilde{z}| = 2$ and $|\widetilde{z}+2| = 5$. After that, let $\breve{z} = 2\widetilde{z}$, so the given circles are $|\breve{z}| = 1$ and $|\breve{z}+1| = \frac{5}{2}$. After that, we can use the formulas of Theorem 16.6.

What if the given circles are of the forms $|z| = 1$ and $|z-\mu| = \gamma$, where μ is not real? A preliminary rotation of the complex plane takes care of this: Let $\varphi = Arg(\mu)$ and $\widetilde{z} = ze^{-i\varphi}$, so the circle $\gamma = |z-\mu|$ becomes $\gamma = \left|\widetilde{z}e^{i\varphi} - |\mu|e^{i\varphi}\right| = \left|\widetilde{z} - |\mu|\right| \cdot |e^{i\varphi}| = |\widetilde{z} - \delta|$, where $\delta = |\mu|$.

16.2.3 Some Useful Facts about Möbius Transformations

Given four *distinct* points z, z_1, z_2, z_3, their **cross ratio** is defined to be

$$(z, z_1, z_2, z_3) \triangleq \frac{(z-z_1)(z_2-z_3)}{(z-z_3)(z_2-z_1)}.$$

Lemma 16.4

Suppose z, z_1, z_2, z_3 are distinct points and w, w_1, w_2, w_3 are distinct points. Define a mapping $w = g(z)$ implicitly by $(z, z_1, z_2, z_3) = (w, w_1, w_2, w_3)$, that is, by

$$\frac{(z-z_1)(z_2-z_3)}{(z-z_3)(z_2-z_1)} = \frac{(w-w_1)(w_2-w_3)}{(w-w_3)(w_2-w_1)}. \tag{16.16}$$

Then $w = g(z)$ is a Möbius transformation and $w_j = g(z_j)$ for $j = 1, 2, 3$.

Why? The details are tedious but easy: Cross multiplication of (16.16) gives

$$\begin{aligned} &w(w_2-w_3)(z-z_3)(z_2-z_1) - w_1(w_2-w_3)(z-z_3)(z_2-z_1) \\ &\quad = w(w_2-w_1)(z-z_1)(z_2-z_3) - w_3(w_2-w_1)(z-z_1)(z_2-z_3); \end{aligned}$$

hence,

$$w = \frac{w_1(w_2-w_3)(z-z_3)(z_2-z_1) - w_3(w_2-w_1)(z-z_1)(z_2-z_3)}{(w_2-w_3)(z-z_3)(z_2-z_1) - (w_2-w_1)(z-z_1)(z_2-z_3)} \triangleq \frac{az+b}{cz+d} \triangleq g(z). \tag{16.17}$$

In this generality, we will not bother to derive the formulas for the constants a, b, c, d in terms of $z_1, z_2, z_3, w_1, w_2, w_3$. The reader may easily verify from (16.17) that $g(z_j) = w_j$ for $j = 1, 2, 3$.

The reader may also derive the formulas for the constants a, b, c, d and check that $ad - bc \neq 0$ follows from the given information that z_1, z_2, z_3 are distinct and w_1, w_2, w_3 are distinct.

Example 16.7

Find a Möbius transformation that maps the points $-1, 1, i$ to the points $-1, 3, (1 - i2)$, respectively.

Method: Let z_1, z_2, z_3 be $i, 1, -1$ and let w_1, w_2, w_3 be $(1 - i2), 3, -1$; we chose this order because the orders z_1, z_2, z_3 and w_1, w_2, w_3 correspond to travel counterclockwise on the circles $|z| = 1$ and $|w - 1| = 2$. Formula (16.17) gives

$$w = \frac{(1 - i2)\big(3 - (-1)\big)\big(z - (-1)\big)(1 - i) - (-1)\big(3 - (1 - i2)\big)(z - i)\big(1 - (-1)\big)}{\big(3 - (-1)\big)\big(z - (-1)\big)(1 - i) - \big(3 - (1 - i2)\big)(z - i)\big(1 - (-1)\big)}$$

$$= \cdots = \frac{-i2z - i4}{-i2z} = \frac{z + 2}{z} = 1 + \frac{2}{z} \triangleq g(z). \bigcirc$$

Notice that Lemma 16.4 gives us a method for finding a map that takes one arbitrary circle to another arbitrary circle. Theorem 16.6 gives us a method for finding a map that takes two given circles to two circles centered at the origin but whose radii we did not specify.

A resource of information about Möbius and other transformations is the *Dictionary of Conformal Representation*, by H. Kober. For example, in its Section 3.3 on page 8, we find that

$$M(z) \triangleq \frac{z - \alpha}{z - \beta}$$

accomplishes the mapping of circle to circle given by

$$|z - z_0| = r \;\; \mapsto \;\; |w - w_0| = R, \text{ where } w_0 = \frac{z_0 - \alpha - r^2 s}{z_0 - \beta}, \; R = r|s|, \text{ and } s = \frac{-\beta + \alpha}{|z_0 - \beta|^2 - r^2},$$

assuming $z = \beta$ is not on the circle $|z - z_0| = r$.

On the other hand, item (iii) on page 7 of Kober's "dictionary" states that if $z = \beta$ *is* on the circle $|z - z_0| = r$, that is, if $|\beta - z_0| = r$, then M accomplishes the mapping of circle to the line given by

$$|z - z_0| = r, \, z \neq \beta \;\; \mapsto \;\; \mathcal{R}e\left(\overline{\left(\frac{\alpha - \beta}{z_0 - \beta}\right)} \cdot w\right) = -\frac{|\alpha - \beta|^2 + 2\mathcal{R}e\left((z_0 - \alpha)(\alpha - \beta)\right)}{2r^2}.$$

The results of Lemma 16.1(a) are found on page 9 of Kober's dictionary.

16.2.4 Möbius Transformation to or from a Line

Suppose $M(z)$ is a Möbius transformation and L is a line. If the "point at infinity" is on the image $M(L)$, then $M(L)$ cannot be a circle and therefore must be a line. The cross-ratio method can be adapted to this situation. In an informal way, consider

$$(w, w_1, w_2, w_3) = \frac{(w - w_1)(w_2 - w_3)}{(w - w_3)(w_2 - w_1)}.$$

If $w_3 \to \infty$, then $\dfrac{w_2 - w_3}{w - w_3} \to 1$. This motivates defining

$$(w, w_1, w_2, \infty) \triangleq \frac{(w - w_1)}{(w_2 - w_1)}.$$

Example 16.8

Find a Möbius transformation that maps the region $|z| < 1$ to the region $\mathcal{R}e(w) < 0$.

Method: Let $M(z)$ be the desired Möbius transformation. First, let's choose $M(z)$ so that it maps the circle $|z| = 1$ to the line $\mathcal{R}e(w) = 0$; this follows the general principle that if we want to understand an inequality, it helps to first understand the corresponding equality.

We can consider the point ∞ to be on the line $\mathcal{R}e(w) = 0$, so the cross-ratio method defines M implicitly by

$$(z, z_1, z_2, z_3) = (w, w_1, w_2, \infty) = \frac{(w - w_1)}{(w_2 - w_1)},$$

where the circle $|z| = 1$ has points z_1, z_2, z_3 mapped to w_1, w_2, ∞, respectively, on the line $\mathcal{R}e(w) = 0$. The mapping we get depends on the orientations of z_1, z_2, z_3, and w_1, w_2. Figure 16.10 illustrates what are essentially the only two possibilities, although we could choose different magnitudes of w_2.

(a) With $z_1 = 1, z_2 = i, z_3 = -1, w_1 = 0, w_2 = i$, we get

$$-iw = \frac{w - 0}{i - 0} = (w, w_1, w_2, \infty) = (z, z_1, z_2, z_3) = \frac{(z-1)\big(i - (-1)\big)}{\big(z - (-1)\big)(i - 1)} = \cdots = -i\,\frac{z-1}{z+1},$$

that is,

$$w = M_+(z) \triangleq \frac{z-1}{z+1}. \tag{16.18}$$

(b) With $z_1 = 1, z_2 = i, z_3 = -1, w_1 = 0, w_2 = -i$, we get

$$iw = \frac{w - 0}{-i - 0} = (w, w_1, w_2, \infty) = (z, z_1, z_2, z_3) = \frac{(z-1)\big(i - (-1)\big)}{\big(z - (-1)\big)(i - 1)} = \cdots = -i\,\frac{z-1}{z+1},$$

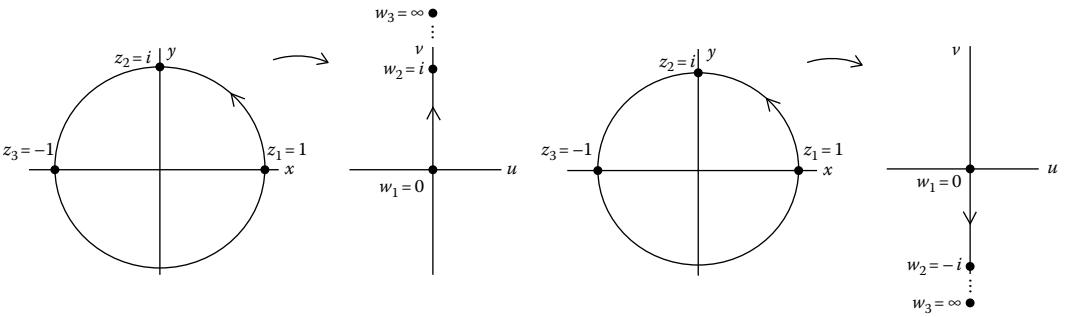


FIGURE 16.10
Example 16.8: Two orientations for $M(z)$.

that is,

$$w = M_-(z) \triangleq -\frac{z-1}{z+1}. \tag{16.19}$$

Each of the Möbius transformations $M_\pm$ maps the circle $|z| = 1$ to the line $\mathcal{R}e(w) = 0$, as shown in Figure 16.10. Which, if either, maps the inside of the circle to the left half plane, $\mathcal{R}e(w) < 0$?

We can test the maps by calculating a "test value":

$$M_\pm(0) = \pm(-1) = \mp 1.$$

Because $M_-(0) = 1$, $M_-(z)$ is not satisfactory.

$M_+(0)$ is in the left half-plane, $\mathcal{R}e(w) < 0$. In fact, for any $z = re^{i\theta}$ with $|z| < 1$,

$$\begin{aligned} M_+(z) &= \frac{re^{i\theta}-1}{re^{i\theta}+1} = \frac{(re^{i\theta}-1)(re^{-i\theta}+1)}{(re^{i\theta}+1)(re^{-i\theta}+1)} = \frac{r^2-1+i2r\sin\theta}{r^2+1+2r\cos\theta} \\ &= \frac{r^2-1}{r^2+1+2r\cos\theta} + i\,\frac{2r\sin\theta}{r^2+1+2r\cos\theta}, \end{aligned}$$

so $r = |z| < 1$ implies

$$\mathcal{R}e\,(M_+(z)) = \frac{r^2-1}{x^2+y^2+1+2x} = \frac{r^2-1}{(x+1)^2+y^2} < 0.$$

So, $M_+(z)$ is a Möbius transformation $M_\pm$ that maps the inside of the circle to the left half-plane, $\mathcal{R}e(w) < 0$. ◯

Any positive multiple of $M_+(z)$, for example, $M(z) = \frac{179(z-1)}{z+1}$, also accomplishes what $M_+(z)$ does in Example 16.8.

Learn More About It

Much useful information about Möbius and other transformations and their applications to solving PDEs is found in *Dictionary of Conformal Representation*, by H. Kober, © 1957, Dover Publication, Inc., New York.

16.2.5 Problems

In each of problems 1–8 find a Möbius transformation that maps both of the given circles to concentric circles centered at the origin.

1. $|z| = 1$ and $|z+1| = \frac{5}{2}$
2. $|z| = 1$ and $|z+1| = 3$
3. $|z| = 3$ and $|z-1| = 1$
4. $|z| = 3$ and $|z+1| = 1$
5. $|z-2| = 1$ and $|z-1| = 3$
6. $|z-1| = 2$ and $|z-2| = 4$

7. $|z| = 1$ and $|z - i| = \frac{5}{2}$
8. $|z| = \frac{1}{2}$ and $|z - i| = 2$

In each of problems 9–12, find a Möbius transformation that maps the given circle in the z-plane to the given circle or line in the w-plane.

9. $|z| = 1 \;\mapsto\; |w - i| = 2$
10. $|z| = 2 \;\mapsto\; |w - i2| = 1$
11. $|z| = 1 \;\mapsto\; \mathcal{I}m(w) = 0$
12. $|z| = 2 \;\mapsto\; \mathcal{I}m(w) = 1$

In each of problems 13 and 14, find a Möbius transformation that maps the given line in the z-plane to the given circle or line in the w-plane.

13. $\mathcal{R}e(z) = 1 \;\mapsto\; |w| = 2$
14. $\mathcal{R}e(z) = 1 \;\mapsto\; \mathcal{I}m(w) = 2$

In each of problems 15–18, find a Möbius transformation that maps the given region in the z-plane to the given region in the w-plane.

15. $|z| < 2 \;\mapsto\; \mathcal{R}e(w) < 0$
16. $|z| < 2 \;\mapsto\; \mathcal{R}e(w) < 1$
17. $|z| > 2 \;\mapsto\; \mathcal{R}e(w) < 1$
18. $|z| > 2 \;\mapsto\; \mathcal{I}m(w) < 1$
19. Explain why the image of $\mathcal{C}_1 : |z| = 1$ is $M(\mathcal{C}_1) : |w| = |\alpha|$ in the derivation of Theorem 16.6, assuming α is real.
20. (a) Explain why $\beta = \frac{1}{\alpha}$, along with Theorem 16.6's hypothesis (16.10), that is, "$z = \beta$ is on neither $\mathcal{C}_1$ nor $\mathcal{C}_2$," imply that the solutions of quadratic equation (16.12) are *not* $\alpha = \pm 1$.
    (b) Explain why $\beta = \frac{1}{\alpha}$, along with $\alpha = 1$ or $\alpha = -1$, would imply $M_{\pm}(z)$ does not map the two circles to circles.
21. (a) In the context of the derivation of Theorem 16.6, derive that

$$M(\delta + i\gamma) = M(\delta - i\gamma) \iff \alpha = \beta.$$

    (b) Derive that $\alpha = \beta \iff \alpha = \pm 1$.
    (c) Conclude that

$$M(\delta + i\gamma) = M(\delta - i\gamma) \iff \alpha = \pm 1.$$

22. Derive that

$$M(\delta + i\gamma) = M(\delta + \gamma) \iff (\delta - \alpha)\gamma - \gamma^2 = \mp i\alpha\gamma$$

    in the context of the derivation of Theorem 16.6.

23. Explain why Lemma 16.1(a) is true, that is, explain why if 0 is not in $\mathcal{C}$, that is, $|z_0| \neq R$, then the image of $\mathcal{C}$ under the inversion mapping is the circle given in (16.5), that is, $f(\mathcal{C}) = \left\{w : |w - w_0| = \frac{R}{|w_0|^2 - R^2}\right\}$, whose center is given in (16.6), that is, is at $w_0 = \frac{z_0}{|z_0|^2 - R^2}$.
24. Explain why Lemma 16.3 is true, that is, explain why if three points w_1, w_2, and w_3 do not lie on a line, then there is exactly one circle that contains w_1, w_2, and w_3.
    (a) First, consider $(w_2 - w_1) = (x_2 - x_1) + i(y_2 - y_1)$ and $(w_3 - w_1) = (x_3 - x_1) + i(y_3 - y_1)$ as vectors in the complex plane. Use linear independence to conclude that

$$0 \neq x_1(y_2 - y_3) - x_2(y_1 - y_3) + x_3(y_1 - y_2).$$

    (b) Discover where the center (h, k) and radius r of the unique circle on which w_1, w_2, and w_3 lie by explaining why

$$(x_1 - h)^2 + (y_1 - k)^2 = r^2 = (x_2 - h)^2 + (y_2 - k)^2$$

    and

$$(x_1 - h)^2 + (y_1 - k)^2 = r^2 = (x_3 - h)^2 + (y_3 - k)^2$$

    and then using those equations to arrive at a linear algebraic system of equations for unknowns h and k.
25. Explain why

$$\left\{\frac{\sin\varphi}{1 + \cos\varphi} : 0 < \varphi \leq 2\pi \text{ with } \varphi \neq \pi\right\} = (-\infty, \infty).$$

16.3 Solving Laplace's Equation Using Conformal Maps

Here we will use conformal maps to solve Laplace's equation in a region in the z-plane by mapping the region into one in the w-plane that is easier to work with. Additionally, even if the original region is not difficult to work with, the solution we get using a conformal map may be in a form much simpler to use than an infinite series, the latter having been produced in Section 11.3 or Section 11.5.

As usual, let $z = x + iy$, where x, y are real. In this section, we will consider conformal maps $\zeta = g(z)$, that is, the image will be in the $\zeta = \xi + i\eta$ plane, where ξ, η are real. (We do *not* write $u + iv = w = g(z)$ because we reserve for later use in fluid flow problems the symbols u, v to stand for the components of fluid velocity in two dimensions, and we reserve w for the "complex velocity," which we will define later in this section.)

Lemma 16.5

Suppose $g(z)$ is a conformal map from a domain $\mathcal{D}$. Then $\mathcal{E} = g(\mathcal{D})$, the image of $\mathcal{D}$ under g, is a domain in the ζ-plane.

Why? g being conformal on $\mathcal{D}$ includes the property that g is analytic and thus continuous on $\mathcal{D}$. Because $\mathcal{D}$ is open, a famous topological result called "the open mapping theorem" implies that $\mathcal{E}$ is an open set in the $\zeta-$plane. The continuity of g and the property that $\mathcal{D}$ is connected imply $\mathcal{E}$ is connected, by another topological theorem. □

Lemma 16.6

Suppose $g(z)$ is a conformal map from a domain $\mathcal{D}$, that is, an open, connected set in the z-plane, to a domain $\mathcal{E} \triangleq f(\mathcal{D})$ in the ζ-plane. *Locally*, that is, in an open ball B_δ in $\mathcal{E}$, $g(z)$ has an inverse function, that is,

$$z = x + iy = g^{-1}(\zeta).$$

Suppose $\Phi = \Phi(\zeta) = \Phi(\xi, \eta)$ solves Laplace's equation in $\mathcal{E}$ and define

$$\phi(x, y) \triangleq \Phi\big(\xi(x, y), \eta(x, y)\big) = \Phi\big(g(\zeta)\big).$$

Then $\phi(x, y)$ satisfies Laplace's equation in $\mathcal{D}$.

Also, suppose a smooth contour Γ is part of the boundary of $\mathcal{E}$.

(a) If Φ satisfies the Dirichlet boundary condition that $\Phi(\xi, \eta) \equiv \Phi_0$ is a constant on Γ, then $\phi(x, y) \equiv \Phi_0$ on $g^{-1}(\Gamma)$.

(b) If Φ satisfies the Neumann boundary condition that $\frac{\partial \Phi}{\partial n_\zeta}(\xi, \eta) \equiv 0$ on Γ, then $\frac{\partial \phi}{\partial n_z}(x, y) \equiv 0$ on $g^{-1}(\Gamma)$.

Why? (a) is obvious; (b) follows from the fact that the conformal map $g^{-1}(\zeta)$ preserves angles, for example, the right angle between the vector $\widehat{\mathbf{n}}_\zeta$, written as a complex number, and $\frac{\partial \Phi}{\partial \xi} + i\,\frac{\partial \Phi}{\partial \eta}$. The latter is referred to as the gradient of Φ, written as a complex number. □

Example 16.9

In the left-hand side of (the earlier) Figure 16.8, the region between two eccentric circles is shaded. Find the electrostatic potential in the region, assuming the potential is one on the outer circle, $|z| = 5$, and zero on the inner circle, $|z - 2| = 2$.

Method: From Example 16.6 in Section 16.2 we know that the Möbius transformation $M = M_+$ given by

$$\zeta = M(z) \triangleq \frac{2z - 20}{2z - 5}$$

maps the circle $|z| = 5$ to the circle $|\zeta| = 2$ and maps the circle $|z-2| = 2$ to the circle $|\zeta| = 4$. Does M map

$$\mathcal{D} \triangleq \{z : |z| < 5 \text{ and } |z-2| > 2\},$$

the region between the curves $|z| = 5$, $|z-2| = 2$, onto

$$\mathcal{E} \triangleq \{\zeta : 2 < |\zeta| < 4\},$$

the region between the curves $|\zeta| = 2$, $|\zeta| = 4$?

Yes. Parametrize $\mathcal{C}_1 : |z| = 5$ with positive orientation and parametrize $\mathcal{C}_2 : |z-2| = 2$ with negative orientation. As discussed in Example 16.6 in Section 16.2, $M(z)$ maps the interior of $\mathcal{C}_1$ to the exterior of $M(\mathcal{C}_1)$ and maps the exterior of $\mathcal{C}_2$ to the interior of $M(\mathcal{C}_2)$, so $M(\mathcal{D}) \subseteq \mathcal{E}$. We can check that the inverse map,

$$z = M^{-1}(\zeta) \triangleq \frac{5\zeta - 20}{2\zeta - 2},$$

maps the exterior of $M(\mathcal{C}_1)$ to the interior of $\mathcal{C}_1$ and maps the interior of $M(\mathcal{C}_2)$ to the interior of $\mathcal{C}_2$, so $M^{-1}(\mathcal{E}) \subseteq \mathcal{D}$. That and $M(\mathcal{D}) \subseteq \mathcal{E}$ explain why $M(\mathcal{D}) = \mathcal{E}$, as we desired.

It is easy to solve Laplace's equation $\Delta\Phi(\xi, \eta) = 0$ in the annulus $\mathcal{E}$ with boundary conditions $\Phi = 0$ on the circle $R = 4$ and $\Phi = 1$ on the circle $R = 2$, where polar coordinates for (ξ, η) are $\xi = R\cos\vartheta, \eta = R\sin\vartheta$. The boundary conditions are constants, so we can look for a solution of Laplace's equation that is independent of ϑ: Laplace's equation that $\Phi = \Phi(R)$ should satisfy is

$$\frac{1}{R}\frac{d}{dR}\left[R\frac{d\Phi}{dR}\right] = 0.$$

This gives

$$\Phi = c_1 + c_2 \ln R,$$

where c_1, c_2 are arbitrary constants. The boundary conditions are

$$1 = \Phi(2) = c_1 + c_2 \ln 2 \quad \text{and} \quad 0 = \Phi(4) = c_1 + c_2 \ln 4,$$

and it's easy to see that the solution is $c_1 = 2, c_2 = -\frac{1}{\ln 2}$. The solution in the ζ-plane is $\Phi = 2 - \frac{1}{\ln 2}\ln R$. Noting that $R = \sqrt{\xi^2 + \eta^2}$, we have $\ln R = \frac{1}{2}\ln(\xi^2 + \eta^2)$, so

$$\Phi = \Phi(\xi, \eta) = 2 - \frac{1}{2\ln 2}\ln(\xi^2 + \eta^2). \tag{16.20}$$

To transform back to the original problem in the $z = x + iy$-plane, we note that

$$\begin{aligned}
\xi + i\eta = \zeta = M(z) &= \frac{2z-20}{2z-5} = \frac{2(x+iy)-20}{2(x+iy)-5} = \frac{(2x-20)+i2y}{(2x-5)+i2y} \\
&= \frac{\big((2x-20)+i2y\big)\big((2x-5)-i2y\big)}{(2x-5)^2+(2y)^2} \\
&= \frac{(2x-20)(2x-5)+4y^2+i2y\big((2x-5)-(2x-20)\big)}{(2x-5)^2+(2y)^2} \\
&= \frac{4x^2+4y^2-50x+100+i30y}{(2x-5)^2+(2y)^2}.
\end{aligned}$$

So,

$$\xi = \frac{4x^2 + 4y^2 - 50x + 100}{(2x-5)^2 + (2y)^2}, \quad \eta = \frac{30y}{(2x-5)^2 + (2y)^2}\,.$$

The solution of the original problem is

$$\phi = \phi(x,y) = 2 - \frac{1}{2\ln 2} \ln\left(\frac{(4x^2 + 4y^2 - 50x + 100)^2 + (30y)^2}{\left((2x-5)^2 + (2y)^2\right)^2} \right). \bigcirc$$

By the way, this could enable us to calculate the electric field **E** between the eccentric circles using

$$\mathbf{E} = -\nabla\phi = -\frac{\partial\phi}{\partial x}\,\hat{\boldsymbol{\imath}} - \frac{\partial\phi}{\partial y}\,\hat{\boldsymbol{\jmath}}.$$

If $\phi = \phi(z)$ has a harmonic conjugate $\psi(z)$, then the function

$$f(z) \triangleq \phi(z) + i\,\psi(z) \tag{16.21}$$

is analytic. The curves $\phi(x,y) = \textit{constant}$ are the equipotentials; the curves $\psi(x,y) = \textit{constant}$ are the **electric field lines**. From (16.20) and

$$\Phi = \Phi(\xi,\eta) = 2 - \frac{1}{\ln 2}\,\ln|\zeta| = 2 - \frac{1}{\ln 2}\,\mathcal{R}e(Log\,\zeta)\,,$$

so it follows that its harmonic conjugate is $\Psi = \Psi(\xi,\eta) = -\frac{1}{\ln 2}\,\mathcal{I}m(Log\,\zeta)$. So the electric field lines are the level sets

$$\textit{constant} = \psi(z) = -\frac{1}{\ln 2}\,Arg(\xi + i\eta) = -\frac{1}{\ln 2}\,Arg(4x^2 + 4y^2 - 50x + 100 + i30y)\,.$$

16.3.1 Boundary Values on a Circle

Suppose $-\pi = \alpha_0 < \alpha_1 < \alpha_2 < \cdots < \alpha_{N-1} < \alpha_N = \pi$. The corresponding points $z_k = Re^{i\alpha_k}$ on the circle $|z| = R$ divide it into N arcs. The problem we want to solve is to find a solution $\phi = \phi(r,\theta)$ of Laplace's equation inside the circle that also has given constant potentials on those arcs, that is, satisfies

$$\phi(R,\theta) = \begin{cases} \phi_1, & -\pi = \alpha_0 \le \theta < \alpha_1 \\ \phi_2, & \alpha_1 \le \theta < \alpha_2 \\ \vdots & \\ \phi_N, & \alpha_{N-1} \le \theta < \alpha_N = \pi \end{cases}. \tag{16.22}$$

Let's map the open disk $|z| < R$ to the upper half-plane, that is, $\mathcal{I}m(\zeta) > 0$, and apply a theorem that gives the solution to the corresponding Dirichlet problem in the ζ-plane. Thus, instead of an infinite series solution to the original problem in the xy-plane as we would get in Section 11.5, we will get a simpler solution formula.

First, we will find a Möbius transformation that maps the open disk $|z| < R$ to the half-plane $\mathcal{I}m(\zeta) > 0$ by using cross ratios. How? We want to both (a) map the three points $z = R, iR, -R$, on the boundary of the disk, to $\zeta = 0, 1, \infty$, respectively, on the boundary of the half plane, and (b) map the interior, that is, points on the "left," of the *positively* oriented circle $|z| = R$ to the *interior*, that is, points on the left, of the positively oriented line $\mathcal{I}m(\zeta) = 0$. The calculations are similar to those we did in Example 16.8 in Section 16.2:

$$\zeta = \frac{\zeta - 0}{1 - 0} = \frac{\zeta - \zeta_1}{\zeta_2 - \zeta_1} = (\zeta, 0, 1, \infty) = (\zeta, \zeta_1, \zeta_2, \infty) = (z, z_1, z_2, z_3) = (z, R, iR, -R)$$

$$= \frac{(z - z_1)(z_2 - z_3)}{(z - z_3)(z_2 - z_1)} = \frac{(z - R)\big(iR - (-R)\big)}{\big(z - (-R)\big)(iR - R)} = \frac{z - R}{z + R} \cdot \frac{iR + R}{iR - R} = \cdots = -i\,\frac{z - R}{z + R}\,.$$

So, we define a mapping by

$$\zeta = M(z) \triangleq -i\,\frac{z - R}{z + R}\,. \tag{16.23}$$

We will need the images of the points $z_k = Re^{i\alpha_k}$. In general,

$$M(Re^{i\theta}) = -i\,\frac{Re^{i\theta} - R}{Re^{i\theta} + R} = -i\,\frac{e^{i\theta} - 1}{e^{i\theta} + 1} = -i \cdot \frac{2i}{2i} \cdot \frac{e^{i\theta} - 1}{e^{i\theta} + 1} \cdot \frac{e^{-i\theta/2}}{e^{-i\theta/2}} \cdot \frac{2}{2}$$

$$= \left(\frac{e^{i\theta/2} - e^{-i\theta/2}}{2i}\right) \Big/ \left(\frac{e^{i\theta/2} + e^{-i\theta/2}}{2}\right) = \frac{\sin(\theta/2)}{\cos(\theta/2)} = \tan\frac{\theta}{2}\,,$$

assuming θ is not an odd multiple of π.

So,

$$\xi_k + i\eta_k = \zeta_k = M(z_k) = \tan\frac{\alpha_k}{2} + i0,\ k = 1, \ldots, N - 1, \tag{16.24}$$

and

$$\zeta_0 = -\infty + i0,\ \zeta_N = \infty + i0,$$

at least as limits.

The function $\Phi\big(\xi,\eta)\big) \triangleq \phi\big(\xi(x,y),\eta(x,y)\big)$, where $\xi + i\eta = M(x+iy)$, should satisfy the Dirichlet problem

$$\left\{\begin{array}{ll} (PDE) & 0 = \dfrac{\partial^2\Phi}{\partial\xi^2} + \dfrac{\partial^2\Phi}{\partial\eta^2}, \text{ for } \eta > 0 \\ (BC) & \Phi(\xi,0) = \left\{\begin{array}{ll} \phi_1, & -\infty < \xi < \xi_1 \\ \phi_2, & \xi_1 \le \xi < \xi_2 \\ \vdots & \\ \phi_{N-1}, & \xi_{N-2} \le \xi < \xi_{N-1} \\ \phi_N, & \xi_{N-1} \le \xi < \infty \end{array}\right\} \end{array}\right\}. \tag{16.25}$$

Fortunately, Matthews and Howell (2006) have solved this problem:

Theorem 16.7

The solution of Dirichlet problem (16.25) is

$$\begin{aligned} \Phi(\xi,\eta) &= \phi_N + \frac{1}{\pi}\sum_{k=1}^{N-1}(\phi_k - \phi_{k+1})Arg(\zeta - \xi_k) \\ &= \phi_N + \frac{1}{\pi}\sum_{k=1}^{N-1}(\phi_k - \phi_{k+1})\cos^{-1}\left(\frac{\eta}{\sqrt{(\xi-\xi_k)^2+\eta^2}}\right). \end{aligned} \tag{16.26}$$

Note that $\eta > 0$ so $Arg(\xi + i\eta) = \cos^{-1}\left(\eta/\sqrt{\xi^2+\eta^2}\right)$.

The original Dirichlet problem on the disk $|z| \le R$ is mapped to the Dirichlet problem (16.25) on the half-plane $\mathcal{I}m(\zeta) \ge 0$ by

$$\begin{aligned} \xi + i\eta = \zeta = M(z) &= -i\frac{z-R}{z+R} = -i\frac{(x-R)+iy}{(x+R)+iy} = -i\frac{\big((x-R)+iy\big)\big((x+R)-iy\big)}{(x+R)^2+y^2} \\ &= \cdots = \frac{2Ry}{(x+R)^2+y^2} + i\frac{R^2-(x^2+y^2)}{(x+R)^2+y^2}. \end{aligned}$$

To use (16.26), it will help to simplify

$$\frac{\eta}{\xi-\xi_k} = \frac{R^2-(x^2+y^2)}{2Ry - \xi_k\big((x+R)^2+y^2\big)} = \frac{R^2-(x^2+y^2)}{2Ry - \big((x+R)^2+y^2\big)\tan(\alpha_k/2)}. \tag{16.27}$$

Using this, the solution to the original Dirichlet problem in the disk $x^2 + y^2 < R^2$ is

$$\phi(x,y) = \phi_N + \frac{1}{\pi}\sum_{k=1}^{N-1}(\phi_k - \phi_{k+1}) \cos^{-1}\left(\frac{R^2 - (x^2+y^2)}{\sqrt{\left(2Ry - \left((x+R)^2+y^2\right)\tan(\alpha_k/2)\right)^2 + \left(R^2-(x^2+y^2)\right)^2}}\right). \tag{16.28}$$

16.3.2 The Joukowsky Map

We will see that one special mapping gives beautiful images of **Zhukovskii airfoils**, wing-like profiles that were important in the history of flight. Those profiles are not only of historical interest but also guides to further interplay between theory and practice.

Zhukovskii, or Joukowsky or Joukowski in older transliterations from the Cyrillic alphabet, used the **Joukowsky map**

$$z = J(\zeta) \triangleq \frac{1}{2}\left(\zeta + \frac{a^2}{\zeta}\right), \tag{16.29}$$

where a is a positive constant. We will see that it produces the images shown in Figure 16.11.

But first, we will see some routine-looking images. Suppose ρ is a positive *constant*. When we map the circle centered at the origin given by

$$\mathcal{C}_\rho : |\zeta| = \rho, \ -\pi < \varphi \le \pi,$$

we get an ellipse, as long as $\rho \neq a$: Writing $\zeta = \rho e^{i\varphi}$, we have

$$x + iy = z = J(\zeta) = \frac{1}{2}\left(\rho\, e^{i\varphi} + \frac{a^2}{\rho}\, e^{-i\varphi}\right),$$

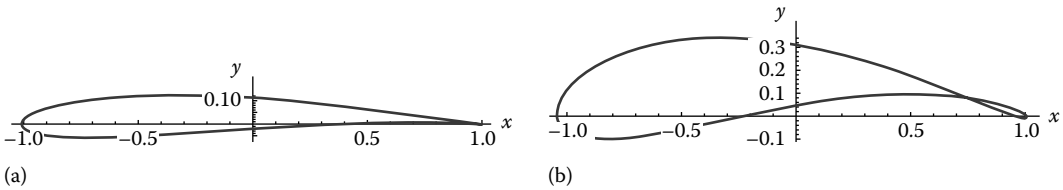


FIGURE 16.11
Images under the Joukowsky map (a) a Zhukovskii airfoil and (b) not a Zhukovskii airfoil.

so

$$\left\{\begin{array}{l} x = \frac{1}{2}\left(\rho\,\cos\varphi + \frac{a^2}{\rho}\,\cos\varphi\right) = \frac{1}{2}\left(\rho + \frac{a^2}{\rho}\right)\cos\varphi \\ \\ y = \frac{1}{2}\left(\rho\,\sin\varphi - \frac{a^2}{\rho}\,\sin\varphi\right) = \frac{1}{2}\left(\rho - \frac{a^2}{\rho}\right)\sin\varphi \end{array}\right\}. \tag{16.30}$$

So, for $0 < \rho \neq a$, (16.30) shows that $J\left(\mathcal{C}_\rho\right) \triangleq \mathcal{E}_\rho$, that is, the ellipse

$$\mathcal{E}_\rho : 1 \equiv \cos^2\varphi + \sin^2\varphi = \frac{x^2}{\frac{1}{4}\left(\rho + \frac{a^2}{\rho}\right)^2} + \frac{y^2}{\frac{1}{4}\left(\rho - \frac{a^2}{\rho}\right)^2}. \tag{16.31}$$

Note that the ellipse has major axis of length $g(\rho) \triangleq \left(\rho + a^2\rho^{-1}\right)$ and minor axis of length $f(\rho) \triangleq \left|\rho - a^2\rho^{-1}\right|$.

For $\rho = a$, the image of $\mathcal{C}_a$ under the Zhukovskii map is a line segment on the *x-axis*. Why? For $\rho = a$, we can think of the image intuitively as being a *degenerate* ellipse whose minor axis has length $f(a) = 0$ and major axis has length $g(a) = 2a$. In fact, because of (16.30), the image $J(\mathcal{C}_a)$ consists of the points

$$(x, y) = \left(\frac{1}{2}\,g(a)\cos\varphi, \frac{1}{2}f(a)\sin\varphi\right) = (a\cos\varphi, 0), \quad -\pi < \varphi \leq \pi,$$

that is, is the line segment $-a \leq x \leq a$, $y = 0$.

We can think of the line segment as a limiting or "degenerate" case of an ellipse. Thinking this way helps because in the study of fluid flow, we can think of one copy of the line segment, specifically the *image* of the half circle $|\zeta| = a$, $0 \leq \varphi \leq \pi$, as being the "upper" boundary of an infinitesimally thin plate and another copy of the line segment as the "lower" boundary of the thin plate because it is the image of the half circle $|\zeta| = a$, $-\pi \leq \varphi \leq 0$.

Supposing $0 < \rho < a$, in Problem 16.3.5.9 you will explain why the points in the interior of the circle $\mathcal{C}_\rho$ are mapped to the exterior of the ellipse $\mathcal{E}_\rho$ and the points in the exterior of $\mathcal{C}_\rho$ are mapped onto the whole z-plane.

We have already seen that there are rich geometrical aspects of the Joukowsky mapping. In addition, we will see that the image of rays $Arg(\zeta) = \varphi_0$ is hyperbolas in the z-plane, as long as φ_0 is not an integer multiple of $\frac{\pi}{2}$, that is, as long as the ray does not lie on the $\xi-$ or η-axis: By (16.30), $x + iy = z = J(\zeta) = J(\rho e^{i\varphi_0})$ implies

$$x = \frac{1}{2}\left(\rho + \frac{a^2}{\rho}\right)\cos\varphi_0 \quad \text{and} \quad y = \frac{1}{2}\left(\rho - \frac{a^2}{\rho}\right)\sin\varphi_0.$$

We will eliminate ρ by using

$$\frac{2x}{\cos\varphi_0} = \rho + \frac{a^2}{\rho} \quad \text{and} \quad \frac{2y}{\sin\varphi_0} = \rho - \frac{a^2}{\rho},$$

from which it follows that

$$\frac{2x}{\cos\varphi_0} + \frac{2y}{\sin\varphi_0} = 2\rho \quad \text{and} \quad \frac{2x}{\cos\varphi_0} - \frac{2y}{\sin\varphi_0} = \frac{2a^2}{\rho}.$$

This gives

$$\frac{x}{\cos\varphi_0} + \frac{y}{\sin\varphi_0} = \rho \quad \text{and} \quad \frac{x}{a^2\cos\varphi_0} - \frac{y}{a^2\sin\varphi_0} = \frac{1}{\rho}.$$

The product of those two identities is

$$1 \equiv \left(\frac{x}{\cos\varphi_0} + \frac{y}{\sin\varphi_0}\right)\left(\frac{x}{a^2\cos\varphi_0} - \frac{y}{a^2\sin\varphi_0}\right),$$

that is, the hyperbola

$$1 = \frac{x^2}{a^2\cos^2\varphi_0} - \frac{y^2}{a^2\sin^2\varphi_0}.$$

16.3.3 Zhukovskii Airfoils

Recall that $J(\zeta)$ maps circles $\mathcal{C}_\rho$ centered at the origin to ellipses, as long as $0 < \rho \neq a$, and maps the circle $\mathcal{C}_a$ to a line segment on the real axis. We will see that the Zhukovskii airfoils are the images of circles that are *not* centered at the origin, specifically given by

$$\mathcal{C}: \ |\zeta - \zeta_0| = |a - \zeta_0|,$$

where $Arg(\zeta_0) = \delta$ and $|\zeta_0| \ll a$. Here is a brief version of the derivation in Panton (2005).

If ζ is a point on the circle $\mathcal{C}$, the law of cosines implies

$$|\zeta|^2 - (2|\zeta_0|\cos\nu)|\zeta| - (a^2 - 2a|\zeta_0|\cos\delta) = 0, \tag{16.32}$$

where $\nu = \delta - \gamma$ and $Arg(\zeta) = \gamma$. You will explain this in Problem 16.3.5.10.

Define a nondimensional parameter by

$$\rho \triangleq \frac{|\zeta|}{a} > 0,$$

and define a "small"* parameter by

$$\varepsilon \triangleq \frac{|\zeta_0|}{a}.$$

* Figure 16.12 does not appear to have $|\zeta_0| \ll a$ in order to make room for labeled quantities.

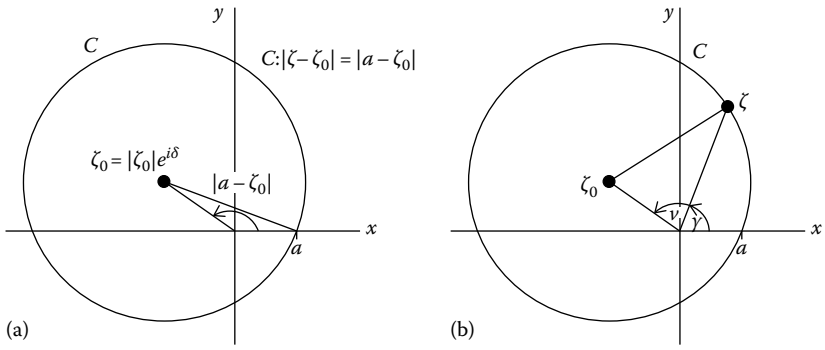


FIGURE 16.12
(a) $\mathcal{C}$ centered at $|\zeta_0|e^{i\delta}$ (b) using the law of cosines.

Using these and then solving the quadratic equation $\rho^2 - (2\varepsilon\cos\nu)\rho - (1 - 2\varepsilon\cos\delta) = 0$ gives

$$\rho \approx 1 + \varepsilon \cdot \begin{Bmatrix} \cos\nu - \cos\delta, & \nu \neq \delta \\ 0, & \nu = \delta \end{Bmatrix}. \tag{16.33}$$

You will explain this in Problem 16.3.5.11.

For $\nu \neq \delta$, thinking of the Joukowsky map leads us to calculate

$$\zeta \approx ae^{i\gamma}\big(1 + \varepsilon(\cos\nu - \cos\delta)\big) \quad \text{and} \quad \frac{a^2}{\zeta} \approx ae^{-i\gamma}\big(1 - \varepsilon(\cos\nu - \cos\delta)\big).$$

Define $K(\zeta) \triangleq \frac{1}{2}\left(\zeta + \frac{1}{\zeta}\right)$. Then the behavior of the Joukowsky map on the circle $\mathcal{C}$ is given by

$$x + iy = z = J(\zeta) = aK\left(\frac{\zeta}{a}\right) = aK\left(\rho e^{i\gamma}\right) = \cdots \approx a\Big(\cos\gamma + i\varepsilon(\cos\nu - \cos\delta)\sin\gamma\Big).$$

So,

$$\frac{x}{a} \approx a\cos\gamma \quad \text{and} \quad \frac{y}{a} \approx \varepsilon(\cos\nu - \cos\delta)\sin\gamma\,. \tag{16.34}$$

Note that

$$\sin\gamma = \pm\sqrt{1 - \left(\frac{x}{a}\right)^2}$$

and

$$\cos\nu - \cos\delta = \cos(\delta - \gamma) - \cos\delta = \cdots = \sin\delta\sin\gamma - \cos\delta(1 - \cos\gamma),$$

and define

$$q(x) \triangleq \varepsilon \sin\delta \left(1 - \left(\frac{x}{a}\right)^2\right) \quad \text{and} \quad t(x) \triangleq \varepsilon |\cos\delta| \left(1 - \left(\frac{x}{a}\right)\right) \sqrt{1 - \left(\frac{x}{a}\right)^2}.$$

Then (16.34) can be rewritten as

$$\frac{y}{a} \approx q(x) \pm t(x)\,. \tag{16.35}$$

The parabola $\frac{y}{a} = q(x)$ is called the **camber line** and the term $t(x)$ is called the **thickness distribution**, in the book by Panton.

The camber line $\frac{y}{a} = q(x)$ reaches its maximum at $x = 0$, and its maximum value is

$$H \approx \varepsilon \sin\delta = \frac{|\zeta_0| \sin\delta}{a}\,,$$

which is called the **camber ratio**.

The thickness distribution $t(x)$ reaches its maximum value

$$T \approx \frac{3\sqrt{3}\varepsilon\ |\cos\delta|}{4} = \frac{3\sqrt{3}|\zeta_0|\ |\cos\delta|}{4a}\,,$$

called the **thickness ratio**, at $x = -\frac{a}{2}$, known as the **quarter-chord point**.

In terms of nondimensionalized quantities $X = \frac{x}{a}, Y = \frac{y}{a}$, the approximate Zhukovskii airfoil is given by

$$Y = H(1 - X^2) \pm \frac{4}{3\sqrt{3}}\, T(1 - X)\sqrt{1 - X^2}\,. \tag{16.36}$$

It can be shown that the left end of the airfoil is the parabolic "nose" and the right end of the airfoil is a cusped "tail."

16.3.4 Lift on Zhukovskii Airfoils

Later in this section, we will study 2D fluid flow around Zhukovskii airfoils. As mentioned earlier in this section, denote by

$$\mathbf{v} = \mathbf{v}(t, \xi, \eta) = u\,\hat{\boldsymbol{\imath}} + v\,\hat{\boldsymbol{\jmath}},$$

the fluid velocity vector field for 2D steady flows in the $\xi\eta$-plane. Potential flow is when there is a scalar potential function Φ such that

$$\mathbf{v} = \nabla\Phi,$$

that is,

$$u(\xi, \eta)\,\hat{\boldsymbol{\imath}} + v(\xi, \eta)\,\hat{\boldsymbol{\jmath}} = \mathbf{v} = \frac{\partial\Phi}{\partial\xi}\,\hat{\boldsymbol{\imath}} + \frac{\partial\Phi}{\partial\eta}\,\hat{\boldsymbol{\jmath}}\,. \tag{16.37}$$

As usual, in the study of complex variables, we denote $\xi + i\eta = \zeta$, so (16.37) becomes

$$\mathbf{v} = \mathbf{v}(\zeta) = u(\zeta) + iv(\zeta) = \frac{\partial \Phi}{\partial \xi} + i\,\frac{\partial \Phi}{\partial \eta}.$$

If $\Phi = \Phi(\zeta)$ has a harmonic conjugate $\Psi(\zeta)$, then the **complex potential** function

$$f(\zeta) \triangleq \Phi(\zeta) + i\,\Psi(\zeta)$$

is analytic, and we define the **complex velocity** $\upsilon = \upsilon(\zeta)$ by

$$\upsilon \triangleq \frac{df}{d\zeta}\,.$$

The Cauchy–Riemann equations tell us that, in fact,

$$\upsilon = \frac{\partial \Phi}{\partial \xi} + i\,\frac{\partial \Psi}{\partial \xi} = \frac{\partial \Phi}{\partial \xi} - i\,\frac{\partial \Phi}{\partial \eta} = \overline{\mathbf{v}(\zeta)},$$

so the "complex velocity" is the complex conjugate of the 2D velocity vector written as a complex number.

By the way, we wrote the Zhukovskii mapping as $z = J(\zeta)$, that is, with independent variable $\zeta = \xi + i\,\eta$ rather than independent variable $w = u + iv$, because we wanted to reserve u and v for the components of fluid velocity in 2D flow.

Let

$$\mathbf{v} = \mathbf{v}(\zeta)$$

be a velocity vector field in the ζ-plane corresponding to complex scalar potential

$$\Phi(\zeta) + i\,\Psi(\zeta) \triangleq v_\infty \left(e^{-i\alpha}\zeta + \frac{a^2 e^{i\alpha}}{\zeta} + \kappa\, Log\left(e^{-i\alpha}\,\zeta\right)\right), \tag{16.38}$$

where v_∞ is a real, positive constant. The velocity vector field derived from the potential in (16.38) both has circulation κ around the circle $\zeta = a$ and corresponds to an angle of attack α. Why is the latter true? Denote $\zeta = \xi + i\,\eta$, so

$$\begin{aligned}
&\Phi(\zeta) + i\,\Psi(\zeta) = \Phi(\xi,\eta) + i\,\Psi(\xi,\eta)\\
&\quad = v_\infty\, \mathcal{R}e\left(\cos(-\alpha)\xi - \sin(-\alpha)\eta + i(\cos(-\alpha)\eta + \sin(-\alpha)\xi) + a^2(\cos\alpha + i\,\sin\alpha)\right)\\
&\qquad \times \frac{(\xi - i\,\eta)}{\xi^2 + \eta^2} + v_\infty\, \mathcal{R}e\left(\kappa\,\left(Log(e^{-i\alpha}) + Log(\zeta)\right)\right);
\end{aligned}$$

hence,

$$\Phi(\xi,\eta) = v_\infty \left(\xi\cos\alpha + \eta\sin\alpha + \frac{a^2\,(\xi\cos\alpha + \eta\sin\alpha)}{\xi^2 + \eta^2} + \kappa\,\ln(\xi^2 + \eta^2)\right).$$

It follows that

$$\begin{aligned} u(\xi,\eta) + iv(\xi,\eta) = \mathbf{v}(\xi,\eta) &= \frac{\partial \Phi}{\partial \xi} + i\,\frac{\partial \Phi}{\partial \eta} \\ &= v_\infty \left(\cos\alpha + i\,\sin\alpha + \frac{a^2\cos\alpha}{\xi^2+\eta^2} - \frac{a^2\cos(\alpha)2\xi^2}{(\xi^2+\eta^2)^2} + i\left(\frac{a^2\sin\alpha}{\xi^2+\eta^2} - \frac{a^2\sin(\alpha)2\eta^2}{(\xi^2+\eta^2)^2}\right)\right) \\ &\quad + v_\infty\,\kappa\,\left(\frac{2\xi}{\xi^2+\eta^2} + i\,\frac{2\eta}{\xi^2+\eta^2}\right). \end{aligned}$$

So, indeed, the velocity "at ∞" is

$$\lim_{\zeta\to\infty} \left(u(\xi,\eta)\hat{\boldsymbol{\imath}} + v(\xi,\eta)\hat{\boldsymbol{\jmath}}\right) = v_\infty\cos\alpha\,\hat{\boldsymbol{\jmath}} + v_\infty\sin\alpha\,\hat{\boldsymbol{\jmath}},$$

which has fluid flow at an angle α with respect to horizontal.

It is known that the lift force on the airfoil is given by

$$F_y = \varrho_0\, v_\infty\,\kappa,$$

where ϱ_0 is the air density. But it seems that any value of κ is allowed, so the lift force is unspecified. It turns out that there is physical reasoning, called the "Kutta condition," that determines the uniquely correct value of κ and thus the lift. As discussed in Panton's book cited earlier, the airfoil has a stagnation point, that is, a place where the complex velocity is zero, at the tail of the airfoil, that is, the point $z = a$. The streamline connected to that point makes an angle $\alpha + \beta$, where α is the **angle of attack** and $\beta \approx 2H$ as $\epsilon = \frac{|z_0|}{a} \to 0^+$. Here, H is the camber ratio, and

$$\sin(\alpha+\beta) = \frac{\kappa}{4\pi\, r_0 v_\infty},$$

where

$$r_0 \approx \left(\frac{1}{4} + \frac{T}{3\sqrt{3}}\right)\cdot 2a.$$

The **chord length** is $\ell = 2a$, if you are reading other sources.

It follows that

$$\kappa = 4\pi\, r_0 v_\infty \sin(\alpha+\beta),$$

and so the lift is

$$F_y = 4\pi\,\varrho_0\, r_0\, v_\infty^2.$$

Further,

$$\beta \approx \epsilon\sin\delta \approx 2H, \text{ as } \epsilon \to 0^+.$$

In terms of the parameters H and T with which we partially characterized Zhukovskii airfoils, the lift is

$$F_y \approx 4\pi a \left(\frac{1}{2}\varrho_0 v_\infty^2\right)\left(1 + \frac{4T}{3\sqrt{3}}\right)\sin(\alpha + 2H), \text{ as } \epsilon \to 0^+.$$

Because air density decreases as altitude increases, the lift force decreases as altitude increases, if the angle of attack and the airspeed stay the same.

Of course, real airplane wings are only approximately Zhukovskii airfoils, in part, because a cusped tail is impossible to machine precisely. Also, fluids don't exactly have 2D potential flow. Nevertheless, Zhukovskii airfoil theory can play a useful role in understanding wings, especially at moderate airspeeds.

Example 16.10

In Section 6.3, we saw a Zhukovskii airfoil that was the image of the circle

$$\mathcal{C}: \zeta \triangleq (-0.07 + \rho_0\cos(t)) + i(0.05 + \rho_0\sin(t)),$$

where $\rho_0 = \sqrt{(1 + 0.07)^2 + (0.05)^2} = |a - \zeta_0|$, under the map (6.30) in Section 6.3, that is,

$$z = x + iy \triangleq \xi \cdot \frac{1 + \xi^2 + \eta^2}{2(\xi^2 + \eta^2)} + i\eta \cdot \frac{-1 + \xi^2 + \eta^2}{2(\xi^2 + \eta^2)}.$$

Find the approximate camber ratio, H, and the thickness ratio, T, and indicate them on a picture of the airfoil. Also, find the lift in terms of the unspecified wind speed at infinity, the air density, and the angle of attack.

Method: From the form of $\mathcal{C}$, we see that its center is at $\zeta_0 = -0.07 + i0.05 = |\zeta_0|\cos\delta + i|\zeta_0|\sin\delta = |\zeta_0|e^{i\delta}$ and $a = 1$. So, the camber ratio is

$$H \approx \frac{|\zeta_0|\sin\delta}{a} = \frac{0.05}{1} = 0.05$$

and the thickness ratio is

$$T \approx \frac{3\sqrt{3}\,|\zeta_0|\,|\cos\delta|}{4a} = \frac{3\sqrt{3}\,|-0.07|}{4 \cdot 1}$$

$$\approx 0.0909326674.$$

The airfoil is shown in Figure 16.13. This is the same airfoil as in Figure 16.11a, except stretched to make room to show H and T. Figure 16.13 was drawn using the approximations in (16.36), while Figure 16.11a was drawn using the exact Joukowsky map, as in (6.30) in Section 6.3. The lack of a cusp at the tail is an artifact of drawing thick curves in order to be visible in the printed copy.

The lift is

$$F_y \approx 4\pi a \left(\frac{1}{2}\varrho_0 v_\infty^2\right)\left(1 + \frac{4T}{3\sqrt{3}}\right)\sin(\alpha + 2H) = 2\pi \cdot 1 \cdot \varrho_0 v_\infty^2(1 + 0.07)\sin(\alpha + 0.1)$$

$$= 2.14\pi\varrho_0 v_\infty^2 \sin(\alpha + 0.1). \bigcirc$$

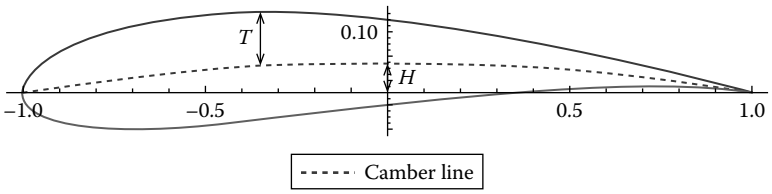


FIGURE 16.13
Zhukovskii airfoil of Example 16.10.

Learn More About It

Complex Variables for Mathematics and Engineering, 5th ed., by John H. Mathews and Russell W. Howell, Jones and Bartlett Publ., Inc., © 2006, has the derivation of Theorem 16.7's solution of a Dirichlet problem.
Also, much additional, useful information about conformal maps, including the Joukowsky map, can be found in Volume I of *Applied and Computational Complex Analysis*, by Peter Henrici, John Wiley & Sons, Inc. © 1986; *Conformal Mapping*, by Zeev Nehari, McGraw-Hill Co., Inc. © 1952; the books by Keener, Panton, Saff, and Snider; and *An Informal Introduction to Theoretical Fluid Mechanics*, by James Lighthill, Oxford University Press © 1986.

16.3.5 Problems

In problems 1–6, solve Laplace's equation for $\phi = \phi(, y)$ in the region between the given circles and satisfying the given boundary conditions.

1. $\phi = 0$ on $|z| = 1$ and $\phi = 1$ on $|z + 1| = \frac{5}{2}$.
2. $\phi = 0$ on $|z| = 1$ and $\phi = 0$ on $|z + 1| = 3$.
3. $\phi = 1$ on $|z| = 3$ and $\phi = -1$ on $|z - 1| = 1$.
4. $\phi = 1$ on $|z - 1| = 2$ and $\frac{\partial \phi}{\partial n} = 0$ on $|z - 2| = 4$.
5. $\phi = 1$ on $|z| = 1$ and $\frac{\partial \phi}{\partial n} = 0$ on $|z - i| = \frac{5}{2}$.
6. $\phi = 0$ on $|z| = \frac{1}{2}$ and $\phi = -1$ on $|z - i| = 2$.

In problems 7 and 8, solve (16.25) with the given boundary conditions.

7. $\phi(R, \theta) = \begin{Bmatrix} 1, & -\pi \le \theta < -\frac{\pi}{3} \\ -1, & -\frac{\pi}{3} \le \theta < \frac{\pi}{3} \\ 1, & \frac{\pi}{3} \le \theta < \pi \end{Bmatrix}$.
8. $\phi(R, \theta) = \begin{Bmatrix} 1, & -\pi \le \theta < -\frac{\pi}{3} \\ 2, & -\frac{\pi}{3} \le \theta < \frac{\pi}{3} \\ 1, & \frac{\pi}{3} \le \theta < \pi \end{Bmatrix}$.
9. Suppose $0 < \rho \neq a$. Recall that (16.31) implies that if ζ is on the circle $\mathcal{C}_\rho$ the Joukowsky map's image point $z = J(\zeta)$ lies on the ellipse $\mathcal{E}_\rho$. Recall that $\mathcal{E}_\rho$

has major axis of length $g(\rho) \triangleq \left(\rho + a^2\rho^{-1}\right)$ and minor axis of length $f(\rho) \triangleq \left|\rho - a^2\rho^{-1}\right|$.

(a) Explain why $g(\rho)$ is decreasing for $0 < \rho < a$ and increasing for $a < \rho$.

(b) Explain why $f(\rho) = -\rho + a^2\rho^{-1}$ for $0 < \rho < a$ and $f(\rho) = \rho - a^2\rho^{-1}$ for $a < \rho$. Explain why $f(\rho)$ is decreasing for $0 < \rho < a$ and increasing for $a < \rho$.

(c) If $a < \rho$, explain why the points in the exterior of $\mathcal{C}_\rho$ are mapped to the exterior of $\mathcal{E}_\rho$.

(d) Explain why the interior of the circle $\mathcal{C}_a$, that is, the disk $\mathcal{D}_a$, is mapped onto the whole z-plane.

(e) If $0 < \rho < a$, explain why the points in the interior of the circle $\mathcal{C}_\rho$ are mapped to the exterior of the ellipse $\mathcal{E}_\rho$.

(f) If $0 < \rho < a$, explain why the exterior of the circle $\mathcal{C}_\rho$ is mapped onto the whole z-plane.

10. Derive (16.32). [Hints: First, consider the triangle whose vertices are the points 0, ζ, and ζ_0. Opposite the central angle whose measure is ν is the side whose length is $|\zeta - \zeta_0|$. Use the law of cosines for this angle. Later, use $|\zeta - \zeta_0| = |a - \zeta_0|$ and $\zeta_0 = |\zeta_0|e^{i\delta}$.]
11. Derive (16.33) starting from (16.32).

In problems 12–14 use (6.30) in Section 6.3, that is, $x(\xi, \eta) \triangleq \xi \cdot \dfrac{1 + \xi^2 + \eta^2}{2(\xi^2 + \eta^2)}$, $y(\xi, \eta) = \eta \cdot \dfrac{-1 + \xi^2 + \eta^2}{2(\xi^2 + \eta^2)}$, to draw the Zhukovskii airfoil with the given center and $a = 1$. Find the camber ratio, H; the thickness ratio, T; and the lift in terms of the unspecified wind speed at infinity, the air density, and the angle of attack.

12. $\zeta_0 = 0.05 - i0.07$
13. $\zeta_0 = -0.07 + i0.07$
14. $\zeta_0 = -0.07 + 0.10$

Key Terms

angle of attack: before Example 16.10 in Section 16.3
camber line: after (16.35) in Section 16.3
camber ratio: after (16.35) in Section 16.3
chord length: before Example 16.10 in Section 16.3
collinear: Lemma 16.3 in Section 16.2
complex velocity: after (16.37) in Section 16.3
conformal: before Theorem 16.1 in Section 16.1; Definition 16.1 in Section 16.1
conformal at a boundary point: Definition 16.1 in Section 16.1
conformal on $\mathcal{D}$: Definition 16.1 in Section 16.1
cross-ratio: before Lemma 16.4 in Section 16.2
eccentric circles: before Lemma 16.3 in Section 16.2
electric field lines: after (16.21) in Section 16.3
inversion mapping: (16.4) in Section 16.2

Joukowsky map: (16.29) in Section 16.3
linear mapping: beginning of Section 16.1
magnification: beginning of Section 16.1
Möbius transformations: (16.2) in Section 16.1
punctured line: before Lemma 16.2 in Section 16.2
quarter chord point: after (16.35) in Section 16.3
rotation: beginning of Section 16.1
thickness distribution: after (16.35) in Section 16.3
thickness ratio: after (16.35) in Section 16.3
translation: beginning of Section 16.1
Zhukovskii airfoils: beginning of Section 16.3.2

References

Keener, J.P., *Principles of Applied Mathematics. Transformation and Approximation*, Section 6.3. Addison-Wesley Publ. Co., New York, 1988.

Mathews, J.H. and Howell, R.W., *Complex Variables for Mathematics and Engineering*, 5th edn. Jones and Bartlett Publ., Inc., Sudbury, MA, 2006.

Panton, R.L., *Incompressible Flow*, 3rd edn., Section 18.13. John Wiley & Sons, Inc., New York, 2005.

17

Integral Transform Methods

17.1 Fourier Transform

Recall Definition 9.3 in Section 9.4, that is, that a function is absolutely integrable on $\mathbb{R}$ if the improper integral $\int_{-\infty}^{\infty} |f(x)|\,dx$ converges. Recall Definition 9.4 in Section 9.4, that is, that if $f(x)$ is absolutely integrable on $\mathbb{R}$, then its **Fourier transform** is defined by the improper integral

$$F(\omega) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x)\, e^{-i\omega x}\, dx. \tag{17.1}$$

Theorem 9.7 (Fourier inversion theorem) in Section 9.4 states that if $f(x)$ is absolutely integrable on $\mathbb{R}$ and is piecewise smooth on every finite interval, then

$$f(x) \doteq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega)\, e^{i\omega x} d\omega, \tag{17.2}$$

in the sense that at every x_0,

$$\frac{f(x_0^+) + f(x_0^-)}{2} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega)\, e^{i\omega x_0} d\omega. \tag{17.3}$$

The Fourier transform and its inversion theorem are very useful for solving some physical problems, as we will see in Section 17.2. A reference for a rigorous proof of the Fourier inversion theorem was given in "Learn More About It" found at the end of Section 9.4.

Just as the symbol $\mathcal{L}$ stands for the Laplace transform operator and $\mathcal{L}^{-1}$ for the inverse Laplace transform operator, we denote the Fourier transform operator and its inverse operator by

$$\mathcal{F}[f(t)] = F(\omega) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t)\, e^{-i\omega t} dt \tag{17.4}$$

and

$$\mathcal{F}^{-1}[F(\omega)] \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega)\, e^{i\omega t} dt. \tag{17.5}$$

As we mentioned in Section 9.4, we caution that there are many slightly different definitions of the Fourier transform and corresponding inverse transform in current use.

In this section, we concentrate on Fourier transforms and inverse transforms for which complex variable integration methods are particularly needed or useful.

Example 17.1

Find the Fourier transform of

$$f(t) = \frac{1}{a^2 + t^2},$$

where the constant $a \neq 0$.

Method: In effect, we will consider four sub-cases corresponding to $a > 0$ versus $a < 0$ and $\omega \geq 0$ versus $\omega \leq 0$.

First, assume $a > 0$ and $\omega \geq 0$. We will calculate the Fourier transform using residues and the "real integral" method of Section 15.10, especially Example 15.68 in Section 15.10. Let $\mathcal{C}_R = \mathcal{C}_{R,1} + \mathcal{C}_{R,2}$, as shown in Figure 17.1a.

Let $z = t + iy$ and extend $f = f(z) = \frac{1}{a^2+z^2}$ to be a function of the complex variable z.

We chose $\mathcal{C}_{R,2}$ to be the bottom half circle because the factor

$$e^{-i\omega z} = e^{-i\omega(t+iy)} = e^{\omega y} e^{-i\omega t} \to 0, \text{ as } y \to -\infty.$$

Noting that $\mathcal{C}_R$ is *negatively* oriented, we calculate

$$\int_{\mathcal{C}_R} \frac{e^{-i\omega z}}{a^2 + z^2}\, dz = -2\pi i \operatorname{Res}\left[\frac{e^{-i\omega z}}{a^2 + z^2};\ -ia\right] = -2\pi i \lim_{z \to -ia} (z + ia) \cdot \frac{e^{-i\omega z}}{z^2 + a^2}$$

$$= -2\pi i \lim_{z \to -ia} \frac{e^{-i\omega z}}{z - ia} = -2\pi i \cdot \frac{e^{-a\omega}}{-i2a} = \frac{\pi e^{-a\omega}}{a}.$$

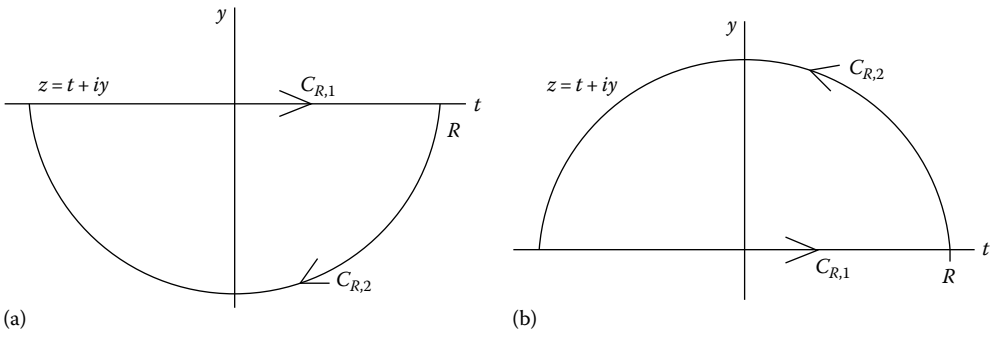


FIGURE 17.1
Large semicircular contours in (a) the lower half plane and (b) the upper half plane.

Continuing with this case of $a > 0$, now we will use the line integral over $\mathcal{C}_R$ to explain why the Fourier transform of $f(t)$ is given by

$$\mathcal{F}\left[\frac{1}{a^2+t^2}\right](\omega) = \frac{1}{\sqrt{2\pi}} \cdot \frac{\pi e^{-a\omega}}{a}, \text{ for } \omega \geq 0.$$

Suppose $R > a$. As $R \to \infty$,

$$\frac{1}{\sqrt{2\pi}} \cdot \frac{\pi e^{-a\omega}}{a} = \frac{1}{\sqrt{2\pi}} \int\limits_{\mathcal{C}_{R,1}} f(z)\, e^{-i\omega z}\, dz = \frac{1}{\sqrt{2\pi}} \int\limits_{-R}^{R} f(t)\, e^{-i\omega t} dt \to \mathcal{F}[f(t)](\omega)$$

because (1) the improper integrals $\int_0^R \frac{1}{a^2+t^2}\, e^{-i\omega t} dt$ and $\int_{-R}^0 \frac{1}{a^2+t^2}\, e^{-i\omega t} dt$ converge and (2) $\int_{\mathcal{C}_{R,2}} \frac{1}{a^2+z^2} e^{-i\omega z}\, dz \to 0$, as $R \to \infty$. The latter follows from the calculating on $\mathcal{C}_{R,2}$: $z = Re^{-i\theta}$, $0 \leq \theta \leq \pi$, that

$$\left|\frac{e^{-i\omega z}}{a^2+z^2}\right| = \left|\frac{e^{-i\omega R e^{-i\theta}}}{a^2+R^2e^{-i2\theta}}\right| = \frac{1}{|a^2+R^2e^{-i2\theta}|} \cdot |e^{-i\omega R\cos\theta}| \cdot |e^{-\omega R\sin\theta}|$$

$$= \frac{1}{|R^2e^{-i2\theta}+a^2|} \cdot 1 \cdot e^{-\omega R\sin\theta} \leq \frac{1}{R^2-a^2} \cdot 1 \cdot 1 = \frac{1}{R^2-a^2}$$

because $0 \leq \theta \leq \pi$ and we assumed $\omega \geq 0$. So,

$$\left|\int\limits_{\mathcal{C}_{R,2}} \frac{e^{-i\omega z}}{a^2+z^2}\, dz\right| \leq \int\limits_0^{\pi} \frac{1}{R^2-a^2}\, d\theta = \frac{\pi}{R^2-a^2} \to 0, \text{ as } R \to \infty.$$

We conclude that if $a > 0$, then, for $\omega \geq 0$,

$$\mathcal{F}[f(t)] = \frac{1}{\sqrt{2\pi}} \cdot \frac{\pi e^{-a\omega}}{a}.$$

Continuing with $a > 0$, if $\omega \leq 0$, then we can use the same method but applied to the *positively* oriented contour $\mathcal{C}_R$ in the upper half plane shown in Figure 17.1b, to get

$$\mathcal{F}[f(t)](\omega) = \sqrt{\frac{\pi}{2}} \cdot \frac{e^{a\omega}}{a}$$

for $\omega \leq 0$. Putting the two cases for ω together, we have that for $a > 0$,

$$\mathcal{F}\left[\frac{1}{a^2+t^2}\right](\omega) = \sqrt{\frac{\pi}{2}} \cdot \frac{e^{-a|\omega|}}{a}. \tag{17.6}$$

On the other hand, if $a < 0$, define $\alpha = |a| = -a > 0$. It follows from (17.6) that

$$\mathcal{F}\left[\frac{1}{a^2+t^2}\right](\omega) = \mathcal{F}\left[\frac{1}{\alpha^2+t^2}\right](\omega) = \sqrt{\frac{\pi}{2}} \cdot \frac{e^{-\alpha|\omega|}}{\alpha} = \sqrt{\frac{\pi}{2}} \cdot \frac{e^{-|a|\ |\omega|}}{-a} = \sqrt{\frac{\pi}{2}} \cdot \frac{e^{-|a|\ |\omega|}}{|a|}.$$

To summarize,

$$\mathcal{F}\left[\frac{1}{a^2+t^2}\right](\omega) = \sqrt{\frac{\pi}{2}} \cdot \frac{e^{-|a|\ |\omega|}}{|a|}. \bigcirc \tag{17.7}$$

Example 17.2

Find the Fourier transform of $f(x) = e^{-x^2/2}$.

Method: The improper integral $\mathcal{F}[e^{-x^2/2}](\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} e^{-i\omega x} dx$ exists because the improper integrals $\int_{-\infty}^{0} e^{-x^2/2} e^{-i\omega x} dx$ and $\int_{0}^{\infty} e^{-x^2/2} e^{-i\omega x} dx$ converge. To find a formula for the Fourier transform, it helps to first complete the square in the exponent

$$-\frac{1}{2}x^2 - i\omega x = -\frac{1}{2}\left(x^2 + i2\omega x\right) = -\frac{1}{2}\left((x+i\omega)^2 - (i\omega)^2\right) = -\frac{1}{2}(x+i\omega)^2 - \frac{\omega^2}{2},$$

so

$$\mathcal{F}[e^{-x^2/2}](\omega) = \frac{e^{-\omega^2/2}}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} e^{-(x+i\omega)^2/2}\, dx. \tag{17.8}$$

At this point, we recall from Problem 7.7.2.8 that

$$\int_{-\infty}^{\infty} e^{-u^2/2}\, du = \sqrt{2\pi}. \tag{17.9}$$

We will use this to evaluate the integral in (17.8). In fact, we will use a contour integral to explain why the integrals in (17.8) and (17.9) are equal, as we would *guess* by making the non-rigorous substitution $u = x + \frac{i\omega}{2}$. Unfortunately, the latter would lead to nonsense such as writing $-\infty + i\omega = -\infty$.

It will turn out that integrating over the large, rectangular contour (Fisher, 1999, pp. 100–101) shown in Figure 17.2 works better than a large semicircular contour in *this* problem. Finding a useful contour is as much of an art as a science, so it helps to be familiar with different examples.

Let $\mathcal{C}_{R,1} : z = -x + i\omega, \ -R \le x \le R$.

We define a function of the complex variable z by $f(z) \triangleq e^{-z^2/2}$, which is analytic everywhere. The Cauchy–Goursat theorem implies that

$$0 = \int_{\mathcal{C}_{R,1}+\cdots+\mathcal{C}_{R,4}} e^{-z^2/2}\, dz = \int_{\mathcal{C}_{R,1}} e^{-z^2/2}\, dz + \cdots + \int_{\mathcal{C}_{R,4}} e^{-z^2/2}\, dz. \tag{17.10}$$

By taking $R \to \infty$, the integrals over $\mathcal{C}_{R,2}$ and $\mathcal{C}_{R,4}$ will go to zero. Why? For $z = \pm R + iy$ on $\mathcal{C}_{R,2}$ or $\mathcal{C}_{R,4}$,

$$|e^{-z^2/2}| = \left|e^{-(\pm R+iy)^2/2}\right| = \left|e^{-(R^2-y^2)/2}\right| \left|e^{\mp iRy}\right| = e^{-(R^2-y^2)/2} \cdot 1.$$

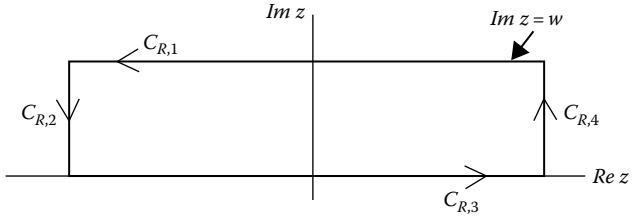


FIGURE 17.2
Large rectangular contour.

It follows that $|e^{-z^2}| \le e^{-(R^2-\omega^2)}$ on $\mathcal{C}_{R,2}$ and on $\mathcal{C}_{R,4}$ because $0 \le y \le \omega$. On $\mathcal{C}_{R,4}$: $z = R + iy$, $0 \le y \le \omega$, we calculate that

$$\left| \int_{\mathcal{C}_{R,4}} e^{-z^2/2}\, dz \right| \le \int_0^{\omega} |e^{-z^2/2}|\, dy \le \int_0^{\omega} e^{-(R^2-\omega^2)/2}\, dy = \omega e^{-(R^2-\omega^2)/2} \to 0,$$

as $R \to \infty$. A similar calculation holds on $\mathcal{C}_{R,2}$.

By (17.10) and the facts that $\lim_{R\to\infty} \int_{-\mathcal{C}_{R,2}} e^{-z^2/2} dz = \lim_{R\to\infty} \int_{-\mathcal{C}_{R,4}} e^{-z^2/2}\, dz = 0$,

$$\lim_{R\to\infty} \int_{-\mathcal{C}_{R,1}} e^{-z^2/2}\, dz = -\lim_{R\to\infty} \int_{\mathcal{C}_{R,1}} e^{-z^2/2}\, dz = \lim_{R\to\infty} \int_{\mathcal{C}_{R,3}} e^{-z^2/2}\, dz = \int_{-\infty}^{\infty} e^{-x^2/2}\, dx = \sqrt{2\pi}.$$

From (17.8),

$$\mathcal{F}[e^{-x^2/2}](\omega) = \frac{e^{-\omega^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x+i\omega)^2/2}\, dx = \frac{e^{-\omega^2/2}}{\sqrt{2\pi}} \cdot \lim_{R\to\infty} \int_{-\mathcal{C}_{R,1}} e^{-z^2/2}\, dz;$$

hence,

$$\mathcal{F}\left[e^{-x^2/2}\right](\omega) = \frac{e^{-\omega^2/2}}{\sqrt{2\pi}} \lim_{R\to\infty} \int_{-R}^{R} e^{-x^2/2}\, dx = \frac{e^{-\omega^2/2}}{\sqrt{2\pi}} \sqrt{2\pi} = e^{-\omega^2/2}. \bigcirc$$

We recall some other results about Fourier transforms and apply them to transforms we have calculated in this section.

Theorem 9.8 in Section 9.4 states that if a is a nonzero constant and $F(\omega) = \mathcal{F}[f(x)](\omega)$ exists, then we have (9.43) in Section 9.4, that is,

$$\mathcal{F}[f(ax)](\omega) = \frac{1}{a} F\left(\frac{\omega}{a}\right) \quad \text{and} \quad \mathcal{F}^{-1}\left[F\left(\frac{\omega}{a}\right)\right](x) = a\, \mathcal{F}^{-1}[F(\omega)](x).$$

Example 17.3

Find $\mathcal{F}\left[e^{-\beta x^2}\right](\omega)$.

Method: Rewrite $e^{-\beta x^2} = f(ax)$, where we choose $f(x) \triangleq e^{-x^2/2}$ because we know its Fourier transform. To find $1/a$, note that $f(ax) = e^{-a^2x^2/2}$, so we need $a^2/2 = \beta$, so $a = \sqrt{2\beta}$. By Theorem 9.8 in Section 9.4 and the conclusion of Example 17.2,

$$\mathcal{F}[e^{-\beta x^2}](\omega) = \frac{1}{a} e^{-(\omega/a)^2/2} = \frac{1}{\sqrt{2\beta}} \cdot e^{-(\omega/\sqrt{2\beta})^2/2} = \frac{1}{\sqrt{2\beta}} \cdot e^{-\omega^2/(4\beta)}. \bigcirc$$

A general result doubles the usefulness of tables of Fourier transforms because we can reverse the roles of Fourier transform and inverse Fourier transform. This was not possible for the Laplace transform.

Theorem 9.9(a) in Section 9.4 states that we have (9.44) in Section 9.4, that is,

$$(a)\ \mathcal{F}^{-1}[F(\omega)](x) = \mathcal{F}[F(\omega)](-x) \quad \text{and} \quad (b)\ \mathcal{F}[f(x)](\omega) = \mathcal{F}^{-1}[f(x)](-\omega),$$

where $\mathcal{F}[f(x)](\omega)$ and $\mathcal{F}[F(\omega)](x)$ are assumed to exist.

Example 17.4

Find $\mathcal{F}\left[\sqrt{\frac{\pi}{2}}\cdot\frac{e^{-|ax|}}{|a|}\right](\omega)$.

Method: Exchanging the roles of x and ω in (17.7), that is, the conclusion of Example 17.1, gives

$$\mathcal{F}\left[\sqrt{\frac{\pi}{2}}\cdot\frac{e^{-|ax|}}{|a|}\right](\omega)=\mathcal{F}^{-1}\left[\sqrt{\frac{\pi}{2}}\cdot\frac{e^{-|ax|}}{|a|}\right](-\omega)=\frac{1}{a^2+(-\omega)^2}=\frac{1}{a^2+\omega^2}.\ \bigcirc$$

17.1.1 Convolution

Recall that we defined the **convolution** on the real line by

$$(f*g)(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}g(x-\xi)f(\xi)\,d\xi.$$

Theorem 9.10 in Section 9.4 (convolution) states that we have (9.45) in Section 9.4, that is,

$$\mathcal{F}[(f*g)(x)]=\mathcal{F}[f(x)]\cdot\mathcal{F}[g(x)]\quad\text{and}\quad(f*g)(x)=\mathcal{F}^{-1}[F(\omega)\,G(\omega)],$$

assuming $F(\omega)\triangleq\mathcal{F}[f(x)](\omega)$ and $G(\omega)\triangleq\mathcal{F}[g(x)](\omega)$ exist.

17.1.2 Problems

In problems 1–4, use the definition to find the Fourier transform of the given function.

1. $f(x)=e^{-|x|}$
2. $f(x)=e^{-a|x|}$, where a is a positive constant
3. $f(x)=\frac{x}{1+x^2}$
4. $f(x)=\frac{x}{a^2+x^2}$
5. Find the inverse Fourier transform of $F(\omega)=e^{-|\omega|}$.

In problems 6 and 7 use the method of Theorem 9.11 in Section 9.4(b) and the results of previous examples to find the Fourier transform or the inverse Fourier transform of the given function.

6. $\mathcal{F}\left[e^{-a|x|}\right]$, where a is a positive constant
7. $\mathcal{F}^{-1}\left[e^{-\beta\omega^2}\right]$
8. Give a plausible explanation why $\mathcal{F}[(f*g)(x)]=\mathcal{F}[f(x)]\cdot\mathcal{F}[g(x)]$, that is, Table 17.1 entry F.12, is true. Your explanation may be a calculation in which you do not pay attention to issues of convergence, that is, it does not have to be mathematically rigorous.
9. If $f(x)$ is an odd, real-valued function on $(-\infty,\infty)$, find a simpler expression for $\mathcal{F}[f(x)]$ in terms of an integral over $0\le x<\infty$ involving $\sin(\omega x)$.

TABLE 17.1

Fourier Transforms: $\beta > 0, k > 0, \tau > 0, a, c, t_0$ Are Nonzero Constants

Formula	$f(t)$	$F(\omega) = \mathcal{F}[f(t)]$ or $F(\omega) = \mathcal{F}[f(x)]$
F.1	$f'(x)$	$i\omega F(\omega)$
F.2	$e^{iat} f(t)$	$F(\omega - a)$
F.3	$f(t) = \begin{cases} 1, & -a < t < a \\ 0, & \lvert t\rvert > a \end{cases}$	$a\cdot\sqrt{\frac{2}{\pi}} \cdot \begin{cases} \frac{\sin(a\omega)}{a\omega}, & \omega \neq 0 \\ 1, & \omega = 0 \end{cases}$
F.4	$\operatorname{sinc}(t) \triangleq \frac{\sin t}{t},\ t \neq 0$	$\sqrt{\frac{\pi}{2}} \cdot \begin{cases} 1, & -1 < \omega < 1 \\ \frac{1}{2}, & \omega = \pm 1 \\ 0, & \lvert\omega\rvert > 1 \end{cases}$
F.5	$\frac{1}{a^2+t^2}$	$\sqrt{\frac{\pi}{2}} \cdot \frac{e^{-\lvert a\omega\rvert}}{\lvert a\rvert}$
F.6	$\frac{e^{-\lvert ax\rvert}}{\lvert a\rvert}$	$\sqrt{\frac{2}{\pi}} \cdot \frac{1}{a^2+\omega^2}$
F.7	$e^{-x^2/2}$	$e^{-\omega^2/2}$
F.8	$f(ax)$	$\frac{1}{a} F\left(\frac{\omega}{a}\right)$
F.9	$e^{-\beta x^2}$	$\frac{1}{\sqrt{2\beta}} \cdot e^{-\omega^2/(4\beta)}$
F.10	$f(t - t_0)$	$e^{-i\omega t_0} F(\omega)$
F.11	$e^{-at}\,\text{Step}(t)$, where $a > 0$	$\frac{1}{\sqrt{2\pi}} \frac{1}{a+i\omega}$
F.12	$(f * g)(t)$	$F(\omega)G(\omega)$
F.13	$f(x)\cos kx$	$\frac{1}{2}\left(F(\omega - k) + F(\omega + k)\right)$
F.14	$f(x)\sin kx$	$\frac{1}{i2}\left(F(\omega - k) - F(\omega + k)\right)$
F.15	$f(x) = \begin{cases} \sin kx, & 0 \le x \le \tau \\ 0, & \text{all other } x \end{cases}$	$\sqrt{\frac{2}{\pi}} \cdot \frac{1}{k^2-\omega^2}\left(e^{-i\omega\tau}\left(-i\omega\sin k\tau - k\cos k\tau\right) + k\right)$

10. Suppose a is a positive constant. Use Table 17.1 entry F.11 and convolution to explain why

$$\mathcal{F}^{-1}\left[\frac{1}{(a+i\omega)^2}\right] = te^{-at}\,\text{Step}(t).$$

11. If $f(x)$ is an even, real-valued function on $(-\infty, \infty)$, explain why

$$\mathcal{F}[f(x)](\omega) = \sqrt{\frac{2}{\pi}} \int_0^\infty f(x)\,\cos(\omega x)\,dx,$$

which is called the **Fourier cosine transform**, denoted by $\mathcal{F}_c[f(x)](\omega)$. [Hints: Explain why $f(x)$ being even and real valued on $(-\infty, \infty)$ implies both (a) $\mathcal{F}_c[f(x)](\omega) = \mathcal{R}e\big(\mathcal{F}[f(x)](\omega)\big)$ and (b) $\mathcal{F}[f(x)](\omega)$ is real.]

17.2 Applications to Partial Differential Equations

In order to use Fourier transforms to solve partial differential equations (PDEs) or ordinary differential equations (ODEs) we need to get a formula for the Fourier transform of a derivative.

Theorem 17.1

Assuming $f(x)$ and $f'(x)$ are absolutely integrable on the real line, and $\lim_{x\to\pm\infty} f(x) = 0$,

$$\mathcal{F}[f'(x)](\omega) = i\omega\mathcal{F}[f(x)](\omega). \tag{17.11}$$

Why? Integration by parts gives

$$\begin{aligned}\mathcal{F}[f'(x)](\omega) &= \frac{1}{\sqrt{2\pi}} \lim_{R\to\infty} \int_{-R}^{R} f'(x)e^{-i\omega x}\,dx \\ &= \frac{1}{\sqrt{2\pi}} \lim_{R\to\infty} \left(\left[f(x)e^{-i\omega x} \right]_{-R}^{R} - \int_{-R}^{R} f(x)(-i\omega e^{-i\omega x})dx \right),\end{aligned}$$

and use of $\lim_{R\to\infty} \left(f(R)e^{-i\omega R} - f(-R)e^{i\omega R} \right) = 0$ explains why

$$= 0 + \frac{i\omega}{\sqrt{2\pi}} \lim_{R\to\infty} \int_{-R}^{R} f(x)e^{-i\omega x}\,dx = i\omega\mathcal{F}[f(x)](\omega). \square$$

The method of separation of variables that we used in Chapter 11 works well for a spatial variable in a finite interval. But when a spatial variable is in an infinite interval, it helps a lot to use integral transforms.

Example 17.5

Solve the heat equation in one space dimension with no sources or sinks. Assume $\lim_{x\to\pm\infty}$ temperature $= 0$ and $\lim_{x\to\pm\infty}$ heat flux $= 0$.

Method: The PDE is

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2},$$

where α is a positive constant and $u = u(x,t)$. We will write the subscript in $\mathcal{F}_x$ to indicate the Fourier transform of the x dependence. As in the use of Laplace transforms, we will denote the transform of a function of x by the corresponding capital letter function of ω:

$$U(\omega,t) \triangleq \mathcal{F}_x[u(x,t)](\omega) \triangleq \int_{-\infty}^{\infty} u(x,t)\,e^{-i\omega x}\,dx.$$

Take the Fourier transform of both sides of the PDE to get

$$\mathcal{F}_x\left[\frac{\partial u}{\partial t}(x,t)\right](\omega) = \alpha\,\mathcal{F}_x\left[\frac{\partial^2 u}{\partial x^2}(x,t)\right](\omega).$$

The right-hand side is simplified by using Theorem 17.1, and we assume that we can interchange the two operations of (a) partial differentiation with respect to t and (b) taking the Fourier transform with respect to x to get

$$\frac{\partial U}{\partial t}(\omega,t) = \alpha\,(i\omega)^2 U(\omega,t);$$

hence,

$$\frac{\partial U}{\partial t} = -\alpha\,\omega^2 U, \tag{17.12}$$

for short.

We can solve (17.12) as if it were the first-order ODE of exponential decay:

$$U(\omega,t) = A(\omega)\,e^{-\alpha\omega^2 t},$$

for an arbitrary function $A(\omega)$.

In terms of the initial heat distribution $u(x,0)$ for $-\infty < x < \infty$, we have

$$A(\omega) = U(\omega,0) = \mathcal{F}_x\left[\frac{\partial u}{\partial t}(x,0)\right](\omega) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} u(x,0)\,e^{-i\omega x}\,dx.$$

This is good progress on finding the solution. But just as for solving ODEs using Laplace transforms, we need to transform back to get the solution as a function of (x,t), not (ω,t):

$$u(x,t) = \mathcal{F}_x^{-1}\Big[U(\omega,t)\Big](x) = \mathcal{F}_x^{-1}\left[e^{-\alpha\omega^2 t}\cdot A(\omega)\right](x).$$

The latter is the inverse Fourier transform of a product, so we can use convolution Theorem 9.12 in Section 9.4 to express the solution as a convolution:

$$u(x,t) = \left(\mathcal{F}_x^{-1}\left[e^{-\alpha\omega^2 t}\right] * \mathcal{F}_x^{-1}\Big[A(\omega)\Big]\right)(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} g(x-\xi,t)u(\xi,0)\,d\xi.$$

Here,

$$g(x,t) \triangleq \mathcal{F}_x^{-1}\left[e^{-\alpha\omega^2 t}\right](x) = \frac{1}{\sqrt{2\alpha t}}\cdot e^{-x^2/(4\alpha t)},$$

by using the result of Example 17.3 in Section 17.1 with $\beta = \alpha t$. So, the solution of the problem is

$$u(x,t) = \frac{1}{2\sqrt{\pi\alpha t}}\int_{-\infty}^{\infty} e^{-(x-\xi)^2/(4\alpha t)}\,u(\xi,0)\,d\xi. \;\bigcirc \tag{17.13}$$

The function

$$G(x,t) \triangleq \frac{1}{\sqrt{2\pi}}\cdot g(x,t) \triangleq \frac{1}{2\sqrt{\pi\alpha t}}\,e^{-x^2/(4\alpha t)}$$

is called the **heat kernel** or **Green's function** or **propagator**, and (17.13) can be rewritten as

$$u(x,t) = \int_{-\infty}^{\infty} G(x-\xi,t)\, u(\xi,0)\, d\xi.$$

The earlier calculations worked well for a PDE on the real line $-\infty < x < \infty$. Other problems take place on a half line and use the next "tool" in our toolbox.

17.2.1 Fourier Cosine and Sine Transforms

Recall from Problem 17.1.2.11 the definition of the **Fourier cosine transform**, which came up naturally when taking the Fourier transform of an even function. Problem 17.1.2.11 asked you to explain why

$$\mathcal{F}\big[f(x)\big](\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x)\, e^{-i\omega x}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x)\, (\cos(\omega x) - i \sin(\omega x))\, dx$$

reduces to the **Fourier cosine transform**, defined by

$$\mathcal{F}_c[f(x)] \triangleq \sqrt{\frac{2}{\pi}} \int_0^{\infty} f(x)\, \cos(\omega x)\, dx, \tag{17.14}$$

if $f(x)$ is even, that is, $f(-x) \equiv f(x)$. If $f(x)$ were to satisfy some differential equation on the half line $0 < x < \infty$ and satisfy $f'(0) = 0$, a requirement automatically met by an even, differentiable function, then the Fourier cosine transform may be a relevant tool.

Similarly, *if* $f(x)$ is odd, then the Fourier transform reduces to the **Fourier sine transform**, defined by

$$\mathcal{F}_s[f(x)] \triangleq \sqrt{\frac{2}{\pi}} \int_0^{\infty} f(x)\, \sin(\omega x)\, dx. \tag{17.15}$$

If $f(x)$ were to satisfy some differential equation on the half line $0 < x < \infty$ and satisfy $f(0) = 0$, a requirement automatically met by an odd, continuous function, then the Fourier sine transform may be a relevant tool.

Another motivation comes from calculating transforms of the second derivative.

Theorem 17.2

Assuming $f(x)$, $f'(x)$, and $f''(x)$ are absolutely integrable on $[0,\infty)$, $\lim_{x\to\infty} f(x) = 0$, and $\lim_{x\to\infty} f'(x) = 0$,

$$\text{(a)} \quad \mathcal{F}_c[f''(x)](\omega) = -\sqrt{\frac{2}{\pi}} \cdot f'(0) - \omega^2 \mathcal{F}_c[f(x)](\omega), \tag{17.16}$$

$$\text{(b)} \quad \mathcal{F}_s[f''(x)](\omega) = \omega \cdot \sqrt{\frac{2}{\pi}} \cdot f(0) - \omega^2 \mathcal{F}_s[f(x)](\omega), \tag{17.17}$$

where $\mathcal{F}_c[f''(x)](\omega)$ or $\mathcal{F}_s[f''(x)](\omega)$ is assumed to exist, respectively.

Why? (a) Using integration by parts twice and ignoring issues of convergence,

$$\begin{aligned}
\mathcal{F}_c[f''(x)](\omega) &= \sqrt{\frac{2}{\pi}} \cdot \lim_{R\to\infty} \int_0^R f''(x)\, \cos(\omega x)\, dx \\
&= \sqrt{\frac{2}{\pi}} \cdot \lim_{R\to\infty} \left(\left[f'(x)\, \cos(\omega x) + \omega f(x) \sin(\omega x) \right]_0^R \quad - \omega^2 \int_0^R f(x)\, \cos(\omega x)\, dx \right) \\
&= \sqrt{\frac{2}{\pi}} \cdot \lim_{R\to\infty} \left(f'(R)\, \cos(\omega R) + \omega f(R) \sin(\omega R) \quad - f'(0) - 0 - \omega^2 \int_0^R f(x)\, \cos(\omega x)\, dx \right) \\
&= \sqrt{\frac{2}{\pi}} \cdot \left(0 + 0 - f'(0) - \omega^2 \int_0^R f(x)\, \cos(\omega x)\, dx \right) = -\sqrt{\frac{2}{\pi}} \cdot f'(0) - \omega^2 \mathcal{F}_c[f(x)](\omega).
\end{aligned}$$

To explain (b) in Problem 17.2.2.10 as follows, you will use arguments similar to the earlier arguments for (a). □

A problem that specifies a nonzero value for $f'(0)$ suggests using the Fourier cosine transform in order to incorporate the value of $f'(0)$. Similarly, a problem that specifies a nonzero value for $f(0)$ suggests using the Fourier sine transform.

In order to use the Fourier cosine or sine transform, it helps to know more of their properties.

Theorem 17.3

(Fourier cosine and sine transform inversion theorems) Assume $f(x)$ is absolutely integrable on $[0, \infty)$ and piecewise smooth on every finite interval contained in $[0, \infty)$.

(a) Assume f is continuous at x and $F_c(\omega) \triangleq \mathcal{F}_c[f(x)](\omega)$ exists. Then

$$f(x) = \mathcal{F}_c^{-1}[F_c(\omega)](\omega) \triangleq \sqrt{\frac{2}{\pi}} \int_0^\infty F_c(\omega)\, \cos(\omega x)\, d\omega. \tag{17.18}$$

(b) Assume f is continuous at x and $F_s(\omega) \triangleq \mathcal{F}_s[f(x)](\omega)$ exists. Then

$$f(x) = \mathcal{F}_s^{-1}[F_s(\omega)](\omega) \triangleq \sqrt{\frac{2}{\pi}} \int_0^\infty F_s(\omega)\, \sin(\omega x)\, d\omega. \tag{17.19}$$

Why? The reader can find references for explanations in the "Learn More About It" section as follows. Note also that, as for the convergence of Fourier series and for the Fourier inversion theorem, at a point where $f(x)$ has a finite jump, we replace $f(x)$ by $\frac{1}{2}\left(f(x^-)+f(x^+)\right)$.

Analogous to Theorem 9.8 in Section 9.4 we have the following:

Theorem 17.4

For $j=s$ or $j=c$, suppose a is a nonzero constant and $F_j(\omega) \triangleq \mathcal{F}_j[f(x)](\omega)$ exists. Then

$$\mathcal{F}_j[f(ax)](\omega) = \frac{1}{a} F_j\left(\frac{\omega}{a}\right) \quad \text{and} \quad \mathcal{F}_j^{-1}\left[F_j\left(\frac{\omega}{a}\right)\right](x) = a\mathcal{F}_j^{-1}[F_j(\omega)](x), \tag{17.20}$$

where $\mathcal{F}_j[f(x)](\omega)$ is assumed to exist.

Theorem 17.5

(Convolutions) Suppose there exist both the Fourier sine transform $F_s(\omega) \triangleq \mathcal{F}_s[f(x)](\omega)$ and the Fourier cosine transforms $G_c(\omega) \triangleq \mathcal{F}_c[g(x)](\omega)$, and $H_c(\omega) \triangleq \mathcal{F}_c[h(x)](\omega)$. Then

$$\mathcal{F}_s^{-1}[G_c(\omega)F_s(\omega)](\omega) = \frac{1}{2}\int_0^\infty \left(g(|x-\xi|) - g(x+\xi)\right) f(\xi)\, d\xi, \tag{17.21}$$

and

$$\mathcal{F}_c^{-1}[G_c(\omega)H_c(\omega)](\omega) = \frac{1}{2}\int_0^\infty \left(g(|x-\xi|) + g(x+\xi)\right) h(\xi)\, d\xi. \tag{17.22}$$

Why? See the "Learn More About It" section as follows.

Example 17.6

Solve the problem

$$\left\{\begin{array}{l} \dfrac{\partial u}{\partial t} = \alpha \dfrac{\partial^2 u}{\partial x^2},\ 0 < x < \infty,\ 0 < t < \infty \\ u(0,t) = \phi(t),\ 0 < t < \infty \\ u(x,0) = f(x),\ 0 < x < \infty \end{array}\right\}, \tag{17.23}$$

along with the boundary conditions that $\lim_{|x|\to\infty} u(x,t) = \lim_{|x|\to\infty} \frac{\partial u}{\partial x}(x,t) = 0$.

This problem corresponds to diffusion from a source at the left end of the half line $0 < x < \infty$.

Method: Because a possibly nonzero $u(0,t)$ is specified, it makes sense to use the Fourier sine transform

$$U(\omega,t) \triangleq \mathcal{F}_{s,x}[u(x,t)](\omega) \triangleq \sqrt{\frac{2}{\pi}} \int_0^\infty u(x,t)\, \sin(\omega x)\, dx.$$

Take the Fourier transform of both sides of the PDE to get

$$\mathcal{F}_{s,x}\Big[\frac{\partial u}{\partial t}(x,t)\Big](\omega) = \alpha \mathcal{F}_{s,x}\left[\frac{\partial^2 u}{\partial x^2}(x,t)\right](\omega).$$

We assume that we can interchange the two operations of (a) partial differentiation with respect to t and (b) taking the Fourier sine transform with respect to x. So, Theorem 17.1 applied to this PDE implies

$$\frac{\partial U}{\partial t}(\omega,t) = \alpha\,\omega \cdot \sqrt{\frac{2}{\pi}} \cdot u(0,t) - \alpha\,\omega^2 U,$$

that is,

$$\frac{\partial U}{\partial t} = \alpha\,\omega \cdot \sqrt{\frac{2}{\pi}} \cdot \phi(t) - \alpha\,\omega^2 U, \tag{17.24}$$

using the assumptions that both $u(x,t) \to 0$ and $\frac{\partial u}{\partial x}(x,t) \to 0$ as $|x| \to \infty$.

We can solve (17.24) as if it were a first-order linear ODE:

$$\frac{\partial U}{\partial t}(\omega,t) + \alpha\,\omega^2 U = \alpha\,\omega \cdot \sqrt{\frac{2}{\pi}} \cdot \phi(t)$$

has integrating factor $\mu(t) = e^{\alpha\omega^2 t}$, so the ODE can be rewritten as

$$\frac{d}{dt}\big[e^{\alpha\omega^2 t} U\big] = \alpha\omega \cdot \sqrt{\frac{2}{\pi}} \cdot e^{\alpha\omega^2 t}\, \phi(t),$$

whose solutions are, implicitly,

$$e^{\alpha\omega^2 t} U(\omega,t) = \alpha\,\omega \cdot \sqrt{\frac{2}{\pi}} \int_0^t e^{\alpha\omega^2 \tau}\, \phi(\tau) + C(\omega).$$

Explicitly,

$$U(\omega,t) = \alpha\,\omega \cdot \sqrt{\frac{2}{\pi}} \int_0^t e^{-\alpha\omega^2 (t-\tau)}\, \phi(\tau) + C(\omega) e^{-\alpha\omega^2 t},$$

where $C(\omega)$ is an arbitrary function of ω.

Substitute in the initial condition to get

$$F_s(\omega) \triangleq \mathcal{F}_{s,x}\big[f(x)\big](\omega) = U(\omega,0) = 0 + C(\omega),$$

so

$$U(\omega,t) = \alpha\,\omega \cdot \sqrt{\frac{2}{\pi}} \int_0^t e^{-\alpha\omega^2 (t-\tau)}\, \phi(\tau) + F_s(\omega) e^{-\alpha\omega^2 t}.$$

To find the solution $u(x,t)$, use the inverse Fourier sine transform

$$u(x,t) = \sqrt{\frac{2}{\pi}}\,\mathcal{F}_{s,x}^{-1}\left[\int_0^t \alpha\,\omega e^{-\alpha\omega^2(t-\tau)}\,\phi(\tau)\,d\tau\right](x) + \mathcal{F}_{s,x}^{-1}\Big[F_s(\omega)e^{-\alpha\omega^2 t}\Big](x). \qquad (17.25)$$

The second term in (17.25) is relatively easy to calculate using the Convolution Theorem result (17.21). To prepare for that, we note that the function $g(x) \triangleq \frac{1}{\sqrt{2\alpha t}}\,e^{-x^2/(4\alpha t)}$ is an even, real-valued function of x, so Problem 17.1.2.11 implies that its Fourier cosine transform equals its Fourier transform. Using the result of Example 17.3 in Section 17.1, the latter can be calculated to be

$$\mathcal{F}\left[\frac{1}{\sqrt{2\alpha t}}\,e^{-x^2/(4\alpha t)}\right] = e^{-\alpha\omega^2 t},$$

so

$$G_c(\omega) = \mathcal{F}_c[g(x)](\omega) = \mathcal{F}\big[g(x)\big] = e^{-\alpha\omega^2 t}.$$

Using (17.21)'s convolution result and our definition that $F_s(\omega) = \mathcal{F}_{s,x}\big[f(x)\big]$, we have

$$\begin{aligned}
\mathcal{F}_{s,x}^{-1}\Big[F_s(\omega)e^{-\alpha\omega^2 t}\Big](x) &= \frac{1}{2}\int_0^\infty \big(g(|x-\xi|) - g(x+\xi)\big)f(\xi)\,d\xi \\
&= \frac{1}{2}\int_0^\infty \left(\frac{1}{\sqrt{2\alpha t}}\,e^{-(x-\xi)^2/(2\alpha t)} - \frac{1}{\sqrt{2\alpha t}}\,e^{-(x+\xi)^2/(2\alpha t)}\right) f(\xi)\,d\xi \\
&= \frac{1}{2\sqrt{2\alpha t}}\int_0^\infty \big(e^{-(x-\xi)^2/(2\alpha t)} - e^{-(x+\xi)^2/(2\alpha t)}\big)f(\xi)\,d\xi.
\end{aligned}$$

To calculate the first term in (17.25), Example 17.7 in the following has

$$\mathcal{F}_{s,x}^{-1}\big[\omega e^{-\alpha\omega^2 t}\big](x) = \frac{x}{(2\alpha t)^{3/2}}\,e^{-x^2/(4\alpha t)},$$

so it follows that

$$\mathcal{F}_{s,x}^{-1}\big[\alpha\,\omega e^{-\alpha\omega^2(t-\tau)}\big](x) = \frac{\alpha x}{\big(2\alpha(t-\tau)\big)^{3/2}}\,e^{-x^2/(4\alpha(t-\tau))}.$$

This enables us to calculate the first term in (17.25):

$$\begin{aligned}
\sqrt{\frac{2}{\pi}}\cdot\mathcal{F}_{s,x}^{-1}\left[\int_0^t \alpha\,\omega e^{-\alpha\omega^2(t-\tau)}\,\phi(\tau)\,d\tau\right](x) &= \sqrt{\frac{2}{\pi}}\cdot\int_0^t \mathcal{F}_{s,x}^{-1}\Big[\alpha\omega e^{-\alpha\omega^2(t-\tau)}\Big](x)\phi(\tau)\,d\tau \\
= \sqrt{\frac{2}{\pi}}\int_0^t \frac{\alpha x}{\big(2\alpha(t-\tau)\big)^{3/2}}\,e^{-x^2/(4\alpha(t-\tau))}\phi(\tau)\,d\tau.
\end{aligned}$$

Put the first and second terms together to conclude that the solution of the problem is

$$\begin{aligned}
u(x,t) = \sqrt{\frac{2}{\pi}}\int_0^t \frac{\alpha x}{\big(2\alpha(t-\tau)\big)^{3/2}}\,e^{-x^2/(4\alpha(t-\tau))}\phi(\tau)\,d\tau \\
+ \frac{1}{2\sqrt{2\alpha t}}\int_0^\infty \big(e^{-(x-\xi)^2/(2\alpha t)} - e^{-(x+\xi)^2/(2\alpha t)}\big)f(\xi)\,d\xi. \bigcirc
\end{aligned}$$

Example 17.7

Explain why

$$\text{(a)} \quad \mathcal{F}_s^{-1}\big[\omega e^{-\alpha\omega^2 t}\big](x) = \frac{x}{(2\alpha t)^{3/2}}\, e^{-x^2/(4\alpha t)} \tag{17.26}$$

and

$$\text{(b)} \quad \mathcal{F}_c^{-1}\big[e^{-\alpha\omega^2 t}\big](x) = \frac{1}{\sqrt{2\alpha t}}\, e^{-x^2/(4\alpha t)}. \tag{17.27}$$

Method: In Problem 17.2.2.11, you will do part (b), which has been discussed earlier.

For part (a), first, we use symmetry to get an integral over the real line in order to later apply a contour integral method in the complex plane:

$$\mathcal{F}_s^{-1}\big[\omega e^{-\alpha\omega^2 t}\big](x) \triangleq \sqrt{\frac{2}{\pi}} \int_0^\infty \omega e^{-\alpha\omega^2 t}\, \sin(\omega x)\, d\omega = \frac{1}{2} \cdot \sqrt{\frac{2}{\pi}} \int_{-\infty}^\infty \omega e^{-\alpha\omega^2 t}\, \sin(\omega x)\, d\omega,$$

because $\omega e^{-\alpha\omega^2 t}\, \sin(\omega x)$ is an even function of the real variable ω.

For convenience, define

$$w(x,t) \triangleq \mathcal{F}_s^{-1}\big[\omega e^{-\alpha\omega^2 t}\big](x),$$

which is what we are calculating. Because $\sin(\omega x) = \frac{1}{2i}\left(e^{i\omega x} - e^{-i\omega x}\right)$,

$$w(x,t) = \frac{1}{2i} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty \omega e^{-\alpha\omega^2 t} \left(e^{i\omega x} - e^{-i\omega x}\right) d\omega = \frac{1}{2i}\big(v_+(x,t) + v_-(x,t)\big), \tag{17.28}$$

where

$$v_\pm(x,t) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty \omega e^{-\alpha\omega^2 t}\, e^{\pm i\omega x} d\omega. \tag{17.29}$$

First, by Theorem 9.8 in Section 9.4, $v_+(x,t)$ is an inverse Fourier transform:

$$v_+(x,t) = \mathcal{F}^{-1}\left[F\left(\frac{\omega}{a}\right)\right] = af(ax) = \frac{1}{\sqrt{2\alpha t}} \cdot f\left(\frac{x}{\sqrt{2\alpha t}}\right), \tag{17.30}$$

where $f(x) = \mathcal{F}^{-1}\big[F(\omega)\big]$, $a = \frac{1}{\sqrt{2\alpha t}}$, and

$$F\left(\frac{\omega}{a}\right) = \omega e^{-\alpha t\, \omega^2} \Longrightarrow F(\omega) = a\omega\, e^{-\alpha t\, a^2\, \omega^2} = \frac{1}{\sqrt{2\alpha t}}\, \omega e^{-\omega^2/2}.$$

So, we need to calculate

$$f(x) = \mathcal{F}^{-1}[F(\omega)] = \mathcal{F}^{-1}\left[\frac{1}{\sqrt{2\alpha t}}\, \omega e^{-\omega^2/2}\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty \frac{1}{\sqrt{2\alpha t}}\, \omega e^{-\omega^2/2}\, e^{i\omega x} d\omega.$$

As usual, we complete the square to get

$$f(x) = e^{-x^2/2} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty \frac{1}{\sqrt{2\alpha t}}\, \omega e^{-(\omega - ix)^2/2}\, d\omega;$$

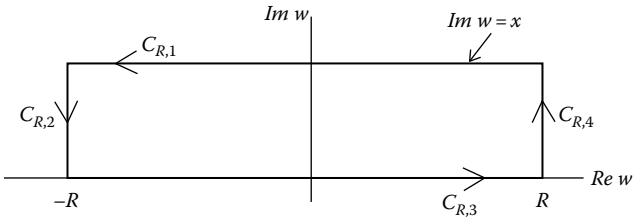


FIGURE 17.3
Large rectangular contour.

hence,

$$f(x) = \frac{1}{\sqrt{2\alpha t}}\, e^{-x^2/2} \cdot \frac{1}{\sqrt{2\pi}} \left(\int_{-\infty}^{\infty} (\omega - ix) e^{-(\omega - ix)^2/2} d\omega + ix \int_{-\infty}^{\infty} e^{-(\omega - ix)^2/2}\, d\omega \right). \tag{17.31}$$

From now on, we consider ω to be a complex variable so that we can use contour integration methods.

For the first integral in (17.31), we use a contour integration as in Example 17.2 in Section 17.1. The large rectangular contour is shown in Figure 17.3. The function $G(\omega) \triangleq (\omega - ix)e^{-(\omega - ix)^2/2}$ is analytic in ω everywhere, so the Cauchy–Goursat theorem and the calculation that

$$\left| \int_{\mathcal{C}_{R,4}} \omega e^{-(\omega - ix)^2}\, d\omega \right| \to 0 \quad \text{and} \quad \left| \int_{\mathcal{C}_{R,2}} \omega e^{-(\omega - ix)^2}\, d\omega \right| \to 0 \quad \text{as } R \to \infty,$$

imply that

$$-\int_{-\infty}^{\infty} (\omega - ix) e^{-(\omega - ix)^2/2}\, d\omega = \lim_{R\to\infty} \int_{-\mathcal{C}_{R,1}} \omega e^{-\omega^2}\, d\omega = \lim_{R\to\infty} \int_{\mathcal{C}_{R,3}} \omega e^{-\omega^2}\, d\omega = \lim_{R\to\infty} 0 = 0,$$

because $\omega e^{-\omega^2}$ is an odd function on the interval $-R < \omega < R$. Similarly, the function $H(\omega) \triangleq e^{-(\omega - ix)^2/2}$ is analytic in ω everywhere, so the Cauchy–Goursat theorem and the calculation that

$$\left| \int_{\mathcal{C}_{R,4}} e^{-(\omega - ix)^2}\, d\omega \right| \to 0 \quad \text{and} \quad \left| \int_{\mathcal{C}_{R,2}} e^{-(\omega - ix)^2}\, d\omega \right| \to 0 \quad \text{as } R \to \infty,$$

imply that

$$\int_{-\infty}^{\infty} e^{-(\omega - ix)^2/2}\, d\omega = \lim_{R\to\infty} \int_{-\mathcal{C}_{R,1}} e^{-\omega^2}\, d\omega = \lim_{R\to\infty} \int_{\mathcal{C}_{R,3}} e^{-\omega^2}\, d\omega = \int_{-\infty}^{\infty} e^{-\omega^2/2}\, d\omega = \sqrt{2\pi},$$

so

$$f(x) = \frac{1}{\sqrt{2\alpha t}} \cdot e^{-x^2/2} \cdot \frac{1}{\sqrt{2\pi}} \cdot (ix) \cdot \sqrt{2\pi} = \frac{1}{\sqrt{2\alpha t}} \cdot (ix) \cdot e^{-x^2/2}.$$

It follows that

$$v_+(x,t) = \frac{1}{\sqrt{2\alpha t}} \cdot f\left(\frac{x}{\sqrt{2\alpha t}}\right) = \frac{1}{\sqrt{2\alpha t}} \cdot \left(\frac{1}{\sqrt{2\alpha t}} \cdot (ix) \cdot e^{-x^2/2}\Big|_{x \mapsto \frac{x}{\sqrt{2\alpha t}}}\right)$$

$$= \frac{1}{\sqrt{2\alpha t}} \cdot \frac{1}{\sqrt{2\alpha t}} \cdot \left(i\,\frac{x}{\sqrt{2\alpha t}}\right) \cdot e^{-x^2/(4\alpha t)} = \frac{ix}{(2\alpha t)^{3/2}} \cdot e^{-x^2/(4\alpha t)}.$$

Similar calculations explain why, again,

$$v_-(x,t) = \frac{ix}{(2\alpha t)^{3/2}} \cdot e^{-x^2/(4\alpha t)}.$$

So, what we have calculated is, by (17.28),

$$\mathcal{F}_s^{-1}\left[\omega e^{-\alpha\omega^2 t}\right](x) = w(x,t) = \frac{1}{2i}\big(v_+(x,t) + v_-(x,t)\big) = \frac{x}{(2\alpha t)^{3/2}} \cdot e^{-x^2/(4\alpha t)}. \bigcirc$$

Theorem 17.6

(Parseval's theorem) Suppose there exist $F_c(\omega) = \mathcal{F}_c[f(x)](\omega)$, $F_s(\omega) = \mathcal{F}_s[f(x)](\omega)$, $G_c(\omega) = \mathcal{F}_c[g(x)](\omega)$, and $G_s(\omega) = \mathcal{F}_c[g(x)](\omega)$. Then

$$\int_0^\infty f(x)\,g(x)\,dx = \int_0^\infty F_c(\omega)\,G_c(\omega)\,d\omega \quad \text{and} \quad \int_0^\infty f(x)\,g(x)\,dx = \int_0^\infty F_s(\omega)\,G_s(\omega)\,d\omega. \tag{17.32}$$

Why? See the "Learn More About It" section as follows.

Learn More About It

Problem 17.2.2.6 is from *Applied Partial Differential Equations*, by Paul DuChateau and David Zachmann, Harper & Row Publishers, Inc., © 1989. Problems 17.2.2.8 and 17.2.2.9 are based on unpublished lecture notes of a course given by Melvin Baron at Columbia University in 1975. A derivation of (17.37), found in Problem 17.2.2.9 as follows, is on pp. 18–20 of *Lectures on Applications-Oriented Mathematics*, by Bernard Friedman, ed., by Victor Twersky, Holden-Day Inc., © 1969 (reprinted by Wiley Classics Library, © 1991).

Derivations of Theorems 17.3, 17.5, and 17.6 are in Section 1.3 of *Fourier Transforms*, by Ian N. Sneddon, Dover Publications, Inc., © 1995.

17.2.2 Problems

Throughout these problems, assume the solution of the PDE satisfies the boundary condition(s) $\lim_{x\to\pm\infty} |u(x,t)| < \infty$, $\lim_{x\to\pm\infty} |u(x,y)| < \infty$, or $\lim_{y\to\pm\infty} |u(x,y)| < \infty$, as appropriate for the spatial domain on which the solution is to be defined.

In problems 1 and 2 solve the problem by using the Fourier sine or cosine transform with respect to x.

1. $$\left\{\begin{array}{c} \frac{\partial u}{\partial t} = \alpha\, \frac{\partial^2 u}{\partial x^2},\ 0 < x < \infty,\ 0 < t < \infty \\ u(0,t) = 0,\ 0 < t < \infty \\ u(x,0) = \begin{cases} 2, & 0 < x < \pi \\ 0, & \pi < x < \infty \end{cases} \end{array}\right\}$$

2. $$\left\{\begin{array}{c} \frac{\partial u}{\partial t} = \alpha\, \frac{\partial^2 u}{\partial x^2},\ 0 < x < \infty,\ 0 < t < \infty \\ \frac{\partial u}{\partial x}(0,t) = 0,\ 0 < t < \infty \\ u(x,0) = \begin{cases} 1, & 0 < x < 3 \\ 0, & 3 < x < \infty \end{cases} \end{array}\right\}$$

3. Solve the heat equation in $\mathbb{R}^2$, that is,

$$\left\{\begin{array}{c} \frac{\partial u}{\partial t} = \alpha \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right), (x,y) \text{ in } \mathbb{R}^2, 0 < t < \infty \\ u(x,y,0) = f(x,y),\ (x,y) \text{ in } \mathbb{R}^2 \end{array}\right\}.$$

   Hint: Take the Fourier transforms with respect to both x and y, that is, use

$$U(\omega, \nu, t) \triangleq \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u(x,y,t) e^{-i\omega x} e^{-i\nu y} dx dy.$$

4. Use the Fourier cosine transform to solve

$$\left\{\begin{array}{c} \frac{\partial u}{\partial t} = \alpha\, \frac{\partial^2 u}{\partial x^2},\ 0 < x < \infty,\ 0 < t < \infty \\ \frac{\partial u}{\partial x}(0,t) = \psi(t),\ 0 < t < \infty \\ u(x,0) = 0,\ 0 < x < \infty \end{array}\right\}.$$

5. Use a Fourier cosine transform and (17.22) to solve the problem in an infinite strip given by

$$\left\{\begin{array}{c} 0 = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2},\ 0 < x < \infty, 0 < y < H \\ \frac{\partial u}{\partial x}(0,y) = 0,\ 0 < y < H \\ \begin{cases} u(x,0) = f(x) \\ u(x,H) = 0 \end{cases}, 0 < x < \infty \end{array}\right\}.$$

   You may assume that $\mathcal{F}_{c,x}^{-1}\left[\frac{\cosh(a\omega)}{\cosh(b\omega)}\right] = \sqrt{2\pi}\frac{\cos\left(\pi a/(2b)\right)\cosh\left(\pi x/(2b)\right)}{b\left(\cosh(\pi x/b) + \cos(\pi a/b)\right)}$, as long as $|\mathcal{R}e(a)| < |\mathcal{R}e(b)|$.

6. Use the Fourier transform to explain why the 1D wave equation

$$\left\{ \begin{array}{c} \frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, -\infty < x < \infty, -\infty < t < \infty \\ \left\{ \begin{array}{c} u(x,0) = f(x) \\ \frac{\partial u}{\partial t}(x,0) = g(x) \end{array} \right\}, -\infty < x < \infty \end{array} \right\}$$

has solutions given by d'Alembert's formula (10.88) in Section 10.5.

7. Explain why (17.27) is true.

Recall from Definition 7.21 in Section 7.7 that the error function erf is given by

$$\text{erf}(\theta) \triangleq \frac{2}{\sqrt{\pi}} \int_0^\theta e^{-y^2}\, dy.$$

Define the **complementary error function**, **erfc**, by

$$\text{erfc}(\theta) \triangleq 1 - \text{erf}(\theta). \tag{17.33}$$

8. Define $I(x) \triangleq \int_0^\infty e^{-a\omega^2} \cos \omega x \, d\omega$, where a is a positive constant. Then

$$I(x) = \frac{1}{2} \cdot \sqrt{\frac{\pi}{a}}\, e^{-x^2/(4a)}. \tag{17.34}$$

Establish (17.34) by the following:

a. Take the ordinary derivative of $I(x)$ with respect to x and move the derivative inside the integral by noting that the improper integral converges absolutely. [Inside the integral, the derivative is the partial derivative with respect to x.] After that, use integration by parts for $\int_0^b \ldots d\omega$ to eventually get

$$\frac{dI}{dx} + \frac{x}{2a} I = 0,$$

and the initial condition

$$I(0) = \int_0^\infty e^{-a\omega^2}\, d\omega = \frac{1}{2} \cdot \sqrt{\frac{\pi}{a}},$$

the latter by using the fact that the error function satisfies $\lim_{\theta \to \infty} erf(\theta) = 1$.

b. Solve part (a)'s ODE and initial condition in so as to evaluate $I(x)$.

9. Let A be a constant. Here you will solve a special case of Example 17.6,

$$\left\{ \begin{array}{c} \frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2},\ 0 < x < \infty,\ 0 < t < \infty \\ u(0,t) = A,\ 0 < t < \infty \\ u(x,0) = 0,\ 0 < x < \infty \end{array} \right\} \tag{17.35}$$

by using a derivation of an inverse Fourier sine transform.

a. Explain why $U(\omega, t)$, the Fourier sine transform of the solution, satisfies an IVP whose solution is

$$U(\omega, t) = \sqrt{\frac{2}{\pi}} \cdot A \cdot \frac{1}{\omega} \left(1 - e^{-\alpha\omega^2 t}\right).$$

b. Break up $u(x, t) = \mathcal{F}_s^{-1}[U(\omega, t)](x)$ as a sum of two inverse Fourier sine transforms. The first involves

$$\frac{2}{\pi} \int_0^\infty \frac{\sin \omega x}{\omega} \, d\omega = 1, \text{ for } x > 0, \tag{17.36}$$

which you may *assume* is true. [Its derivation is done using principal value and residues using an "indented" contour to work around the difficulty at $\omega = 0$.]

c. The second of the inverse Fourier sine transforms involves

$$\frac{2}{\pi} \int_0^\infty \frac{\sin \omega x}{\omega} \, e^{-\alpha\omega^2 t} \, d\omega = \operatorname{erf}\left(\frac{x}{2\sqrt{\alpha t}}\right). \tag{17.37}$$

Establish (17.37) by using the following method: Problem 17.2.2.8 gives the fact that

$$\frac{1}{2} \cdot \sqrt{\frac{\pi}{a}} \, e^{-x^2/(4a)} = I(x) \triangleq \int_0^\infty e^{-a\omega^2} \cos \omega x \, d\omega.$$

Use interchange of the order of integrations to obtain

$$\int_0^x I(u) \, du = \int_0^x \left(\int_0^\infty e^{-a\omega^2} \cos \omega u \, d\omega \right) du = \int_0^\infty \frac{\sin \omega x}{\omega} \, e^{-\alpha\omega^2} \, d\omega. \tag{17.38}$$

After that, use the substitution $a = \alpha t$ and the left-hand side of (17.38) to finish establishing (17.37).

d. Finish up by finding that the solution of the original problem (17.35), is

$$u(x, t) = A \operatorname{erfc}\left(\frac{x}{2\sqrt{\alpha t}}\right).$$

10. Explain why Theorem 17.2(b) is true.
11. Use the result of Example 17.3 in Section 17.1 to explain why (17.27)(b) is true.

17.3 Inverse Laplace Transform

We used the Laplace transform in an operational way in Chapter 4 by using the fact that

$$F(s) = \mathcal{L}[f(t)] \triangleq \int_0^\infty f(t)\, e^{-st}\, dt,\ s > a \iff \quad f(t) = \mathcal{L}^{-1}[F(s)],\ t \geq 0. \tag{17.39}$$

In Example 4.20 in Section 4.4, we solved the IVP $\dot{y} + 2y = 3e^{-4t},\ y(0) = 5$, by calculating that

$$y(t) = \mathcal{L}^{-1}[Y(s)] = \mathcal{L}^{-1}\left[\frac{13}{2}\,\frac{1}{s+2} - \frac{3}{2}\,\frac{1}{s+4}\right] = \frac{13}{2}\,e^{-2t} - \frac{3}{2}\,e^{-4t}$$

because, for example,

$$\mathcal{L}[e^{-2t}] = \frac{1}{s+2},\ s > -2 \iff \mathcal{L}^{-1}\left[\frac{1}{s+2}\right] = e^{-2t}\ t \geq 0.$$

In effect, we found $y(t) = \mathcal{L}^{-1}[Y(s)]$ by using a table of Laplace transforms, such as Table 4.1 in Section 4.5, by writing $F(s)$ as a linear combination of Laplace transforms whose inverse transforms we could look up.

The "table lookup" method works well for many problems, although we may need to manipulate an $F(s)$ into a form that is on the table. Also, a relatively exotic $F(s)$ may require us to consult long, extensive tables of expert knowledge of inverse Laplace transforms.

But we may need to find the inverse Laplace transform of an $F(s)$ that is neither on our table nor can be manipulated into a form on our table. Indeed, finding *new* inverse transforms is how those long, expert level tables were created!

Here is a nonexpert level, hypothetical example of the historical process: After people learned that $\mathcal{L}^{-1}\left[\frac{1}{s+2}\right] = e^{-2t}$, they realized that they would also need to find $\mathcal{L}^{-1}\left[\frac{1}{(s+2)^2}\right]$, for example, in order to solve the IVP $\dot{y} + 2y = e^{-2t},\ y(0) = 0$. In the process of finding the latter, they created a new table entry:

$$\mathcal{L}^{-1}\left[\frac{1}{(s+a)^{k+1}}\right] = \frac{t^k}{k!}\,e^{-at},\ s > -a,$$

for any nonnegative integer k.

We need a method to find inverse Laplace transforms beyond the table lookup method. This will help in more theoretical issues. For example, it would be good to know in general which functions have an inverse Laplace transform.

From now on, we will consider functions $F(s)$ to depend on a complex variable s. This will allow us to use contour integration to find inverse Laplace transforms.

We need some technical definitions first.

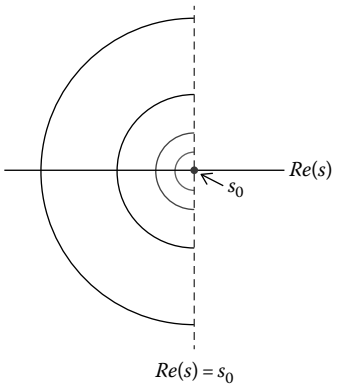


FIGURE 17.4
Left half circles.

Definition 17.1

An **unbounded family of left half circles** (Figure 17.4) is the collection of sets

$$S_{s_0,k} \triangleq \{s : |s - s_0| = R_k \text{ and } \mathcal{R}e(s) \leq s_0\},$$

where s_0 is a fixed real number and the positive radii $R_k \to \infty$ as $k \to \infty$.

A set $\mathcal{S}$ is **countably infinite** if we can list the distinct elements of $\mathcal{S}$ as a sequence $\{s_k\}_{k=1}^{\infty}$.

Definition 17.2

Suppose $S_{s_0,k}$ is an unbounded family of left half circles. We say $F(s) \to 0$ **uniformly on that family** if for every $\varepsilon > 0$, there is a K such that for all $k \geq K$,

$$|F(s)| < \varepsilon \quad \text{for all } s \text{ in } S_{s_0,k}.$$

Definition 17.3

$F(s)$ is of **class L** if (a) $F(s) \to 0$ on some unbounded family of left half circles centered at the real number s_0 and (b) $F(s)$ is analytic, except possibly at finitely many poles or a countably infinite set of poles σ, all of which satisfy $\mathcal{R}e(\sigma) < s_0$.

Recall that by Definition 15.22 in Section 15.7, all poles are isolated singularities.

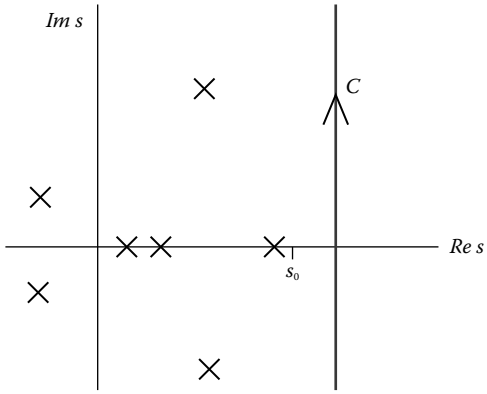


FIGURE 17.5
Bromwich contour.

Definition 17.4

A **Bromwich contour** $\mathcal{C}$ for $F(s)$ is a vertical line contour

$$\mathcal{C}: s = \gamma + iy, \ -\infty < y < \infty,$$

for some real number γ, which lies to the right of all singularities of $F(s)$.

Figure 17.5 illustrates a Bromwich contour for a function, each of whose poles is indicated by an $\times$.

Theorem 17.7

(Laplace inversion theorem) Suppose $F(s)$ is of class **L** using an unbounded family of left half circles centered at the real number s_0. Then there exists a real number $\gamma > s_0$ such that the Bromwich contour $\mathcal{C}: s = \gamma + iy, \ -\infty < y < \infty$, gives

$$\mathcal{L}^{-1}\big[F(s)\big](t) \triangleq \frac{1}{2\pi i} \cdot P.V. \int_{\mathcal{C}} F(s)\, e^{st}\, ds, \ t \geq 0. \tag{17.40}$$

Further,

$$\mathcal{L}^{-1}\big[F(s)\big](t) = \sum_k \text{Res}\big[F(s)\, e^{st};\, s_k\big], \ t \geq 0, \tag{17.41}$$

where $\{s_k\}$ are the poles of $F(s)$.

Definition 17.5

$f(t)$ is of **exponential order** if there are a positive number M and real numbers α, t_0 such that

$$|f(t)| \le Me^{\alpha t}, \quad \text{for all } t \ge t_0.$$

Corollary 17.1

Suppose $f(t)$ is of exponential order, with α as in Definition 17.5, and is piecewise continuous on the interval $[0, \infty)$. Then there exists

$$F(s) \triangleq \mathcal{L}\big[f(t)\big](t) \triangleq \int_0^\infty f(t)\, e^{-st}\, dt, \quad \text{for all } s > \alpha,$$

and $F(s)$ is of class **L**, and

$$\mathcal{L}^{-1}\big[F(s)\big](t) = f(t),\ t \ge 0.$$

It is traditional to write (17.40), the contour integral for the inverse Laplace transform, as

$$\mathcal{L}^{-1}\big[F(s)\big](t) \triangleq \frac{1}{2\pi i} \cdot P.V. \int_{\gamma - i\infty}^{\gamma + i\infty} F(s)\, e^{st}\, ds,\ t \ge 0. \tag{17.42}$$

When we are learning a new method that extends what we know, it is usually a good idea to understand how the old results can be obtained by the new method.

Example 17.8

Evaluate $\mathcal{L}^{-1}\big[\frac{1}{s-3}\big]$ using a Bromwich contour.

Method: The only singularity of $F(s) \triangleq \frac{1}{s-3}$ is at $s = 3$, so $\mathcal{C} : s = 4 + iy,\ -\infty < y < \infty$, is a Bromwich contour. We use the residue theorem and the positively oriented curve $\mathcal{C}_R = \mathcal{C}_{R,1} + \mathcal{C}_{R,2}$ shown in Figure 17.6:

$$\mathcal{C}_{R,1} : z = 4 + iy,\ -R \le y \le R \qquad \text{and} \quad \mathcal{C}_{R,2} : z = 4 + Re^{i\theta},\ \frac{\pi}{2} \le \theta \le \frac{3\pi}{2}.$$

We calculate, for any $R > 0$, that

$$\frac{1}{2\pi i}\left(\int_{\mathcal{C}_{R,1}} F(s)e^{st}\, ds + \int_{\mathcal{C}_{R,2}} F(s)e^{st}\, ds\right) = \frac{1}{2\pi i}\int_{\mathcal{C}_R} F(s)e^{st}\, ds = \frac{1}{2\pi i}\int_{\mathcal{C}_R} \frac{1}{s-3}\, e^{st}\, ds$$

$$= \text{Res}\left[\frac{e^{st}}{s-3}; 3\right] = e^{3t}.$$

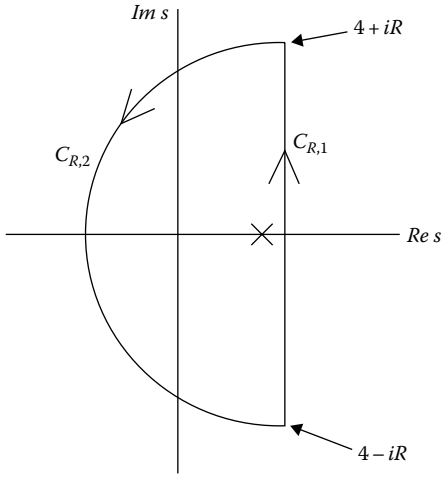


FIGURE 17.6
Example 17.8.

For s on the left half circle $\mathcal{C}_{R,2}: s = 4 + Re^{i\theta}$,

$$|F(s)| = \left|\frac{1}{s-3}\right| = \frac{1}{|1+Re^{i\theta}|} \le \frac{1}{R-1} \to 0, \quad \text{as } R \to \infty,$$

and

$$\mathcal{R}e(s) = \mathcal{R}e\left(4 + Re^{i\theta}\right) \le 4; \text{ hence, } |e^{st}| \le e^{4t}, \quad \text{for } t \ge 0.$$

Choose left half circles with $R = R_k$. Similarly to the argument in Example 17.1 in Section 17.1, we see that $\dfrac{1}{|1+R_k e^{i\theta}|} \le \dfrac{1}{R_k - 1}$; hence,

$$\left|\int_{\mathcal{C}_{R,2}} F(s)\, e^{st}\, ds\right| \le \int_{\mathcal{C}_{R,2}} |F(s)|\, |e^{st}|\, ds \le \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} \frac{1}{|1+R_k e^{i\theta}|}\, e^{4t}\, d\theta \le \pi e^{4t} \cdot \frac{1}{R_k - 1} \to 0,$$

as $k \to \infty$. So,

$$\mathcal{L}^{-1}\left[\frac{1}{s-3}\right](t) = \frac{1}{2\pi i}\ P.V. \int_{4-i\infty}^{4+i\infty} \frac{1}{s-3}\, e^{st}\, ds = \frac{1}{2\pi i} \lim_{R\to\infty} \int_{\mathcal{C}_{R,1}} \frac{1}{s-3}\, e^{st}\, ds$$

$$= \lim_{R\to\infty} \frac{1}{2\pi i} \int_{\mathcal{C}_R} F(s) e^{st}\, ds = \lim_{R\to\infty} \text{Res}\left[\frac{e^{st}}{s-3}; 3\right] = \lim_{R\to\infty} e^{3t} = e^{3t}. \bigcirc$$

Clearly, in a sophomore ODE course, it makes sense to find $\mathcal{L}^{-1}\left[\frac{1}{s-3}\right](t)$ by using the table lookup method rather than the Bromwich contour method. But we will see that the generality of the latter is useful.

Theorem 17.8

Suppose a function of class **L** can be written in the form

$$F(s) = \frac{p(s)}{q(s)}.$$

Suppose $F(s)$ is analytic except at $\{s_k\}$, which is either a finite or a countably infinite set, and that each s_k is both a simple pole of $F(s)$ and satisfies $p(s_k) \neq 0$. Then

$$\mathcal{L}^{-1}\big[F(s)\big](t) = \sum_k \frac{p(s_k)}{q'(s_k)}\, e^{s_k t}. \tag{17.43}$$

Why? The second "preparation" theorem (15.27) in Section 15.7 implies that $q(s) = (s - s_k)g(s)$ where $g(s)$ is analytic at s_k and $q'(s_k) = g(s_k) \neq 0$. When we calculate the residues called for in Theorem 17.7, we get (17.43). □

Formula (17.43) is a generalization of Heaviside's method for partial fractions expansion for a *rational* function $F(s) = p(s)/q(s)$. Note that a rational function $F(s)$ can be of class **L** only if the degree of the polynomial $p(s)$ is strictly less than the degree of the polynomial $q(s)$.

17.3.1 Solving a Wave Equation

Example 17.9

A flat string of length L is tied at its right end at $x = L$. Beginning at time $t = 0$, its left end at $x = 0$ moves up and down with height $\sin t$. Assuming that initially the string is flat and at rest, find the motion of the string for $t \geq 0$. Assume that $c = \sqrt{\varrho/T}$ is constant and $\frac{L}{\pi c}$ is not an integer.

Method: The mathematical problem is

$$\left\{ \begin{array}{c} \frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2},\ 0 < x < L,\ 0 < t < \infty \\ \\ \begin{array}{c} u(0,t) = \sin t,\ 0 < t < \infty \\ u(L,t) = 0, \quad 0 < t < \infty \end{array} \\ \\ u(x,0) = \dfrac{\partial u}{\partial t}(x,0) = 0,\ 0 < x < L \end{array} \right\}. \tag{17.44}$$

Let $\mathcal{L}_t$ denote the Laplace transformation operation with respect to t and define

$$U(x,s) \triangleq \mathcal{L}_t[u(x,t)](s) = \int_0^\infty u(x,t)\, e^{-st}\, dt.$$

By a property of Laplace transforms, the PDE in (17.44) and the initial conditions imply

$$\mathcal{L}_t\Big[\frac{\partial^2 u}{\partial t^2}(x,t)\Big](s) = s^2 U(x,s) - s u(x,0) - \frac{\partial u}{\partial t}(x,0) = s^2 U(x,s) - s \cdot 0 - 0;$$

hence,

$$s^2U(x,s) = c^2\frac{\partial^2 U}{\partial x^2},\ 0 < x < L, \tag{17.45}$$

along with boundary conditions

$$U(0,s) = \mathcal{L}_t[u(0,t)](s) = \mathcal{L}_t[\sin t](s) = \frac{1}{s^2+1}, \quad U(L,s) = \mathcal{L}_t[u(L,t)](s) = \mathcal{L}_t[0](s) = 0.$$

The solutions of the "ODE" (17.45), that is,

$$\frac{\partial^2 U}{\partial x^2} - \frac{s^2}{c^2}U(x,s) = 0,$$

are

$$U(x,s) = c_1(s)\cosh\left(\frac{sx}{c}\right) + c_2(s)\sinh\left(\frac{sx}{c}\right), \tag{17.46}$$

or by using "clairvoyance" as in Example 9.32 in Section 9.6 and Section 11.3,

$$U(x,s) = A(s)\sinh\left(\frac{s}{c}(L-x)\right) + B(s)\sinh\left(\frac{sx}{c}\right), \tag{17.47}$$

for arbitrary functions $A(s), B(s)$. The boundary conditions give

$$0 = U(L,s) = A(s)\cdot 0 + B(s)\sinh\left(\frac{sL}{c}\right) \quad \text{and} \quad \frac{1}{s^2+1} = U(0,s) = A(s)\sinh\left(\frac{sL}{c}\right);$$

hence, $B(s) = 0$. The solution of the entire problem has Laplace transform*

$$U(x,s) = A(s)\sinh\left(\frac{s}{c}(L-x)\right) = \frac{1}{(s^2+1)\sinh\left(\frac{sL}{c}\right)}\cdot\sinh\left(\frac{s}{c}(L-x)\right). \tag{17.48}$$

So, all that remains is to find

$$u(x,t) = \mathcal{L}_t^{-1}\left[\sinh\left(\frac{s}{c}(L-x)\right) \Big/ \left((s^2+1)\sinh\left(\frac{sL}{c}\right)\right)\right](t). \tag{17.49}$$

Recall that for real u, v,

$$\sinh(u+iv) = \sinh u\cos v + i\cosh u\sin v.$$

The poles of $U(x,s)$, as a function of the complex variable s, are where $s^2+1=0$ or $\sinh\left(\frac{sL}{c}\right)=0$, that is, are at

$$s_{0,\pm} = \pm i \quad \text{and} \quad s_{n,\pm} = \pm i\,\frac{n\pi c}{L},\ n = 1,2,\ldots.$$

The poles of $U(x,s)$, as a function of s, are depicted in Figure 17.7. Note that $s=0$ is a removable singularity of $U(x,s)$.

* If we use (17.46), that is, not use clairvoyance as in (17.47), we will still arrive at (17.48) although it would take more work and use of a hyperbolic trigonometric function identity.

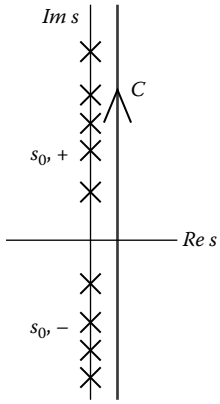


FIGURE 17.7
Example 17.9.

The assumption that "$\frac{L}{\pi c}$ is not an integer" guarantees that all of the poles are simple. In addition, for any small but positive real number ε, for $0 < x < L$, there are left half circles $S_{\varepsilon,k}$

$$\left|\sinh\left(\frac{s}{c}(L-x)\right) \Big/ \sinh\left(\frac{sL}{c}\right)\right|$$

is bounded, so the factor of $\left|\frac{1}{s^2+1}\right|$ in $|U(x,s)|$ shows that $U(x,s)$ is of class **L** in s.

So, we can apply Theorem 17.8 to conclude that

$$u(x,t) = \sum_{j=0}^{\infty} \sum_{\epsilon=\pm 1} \text{Res}\big[U(x,s)e^{st};\, s_{j,\epsilon}\big].$$

Let us calculate those residues: First, for $j=0$ and $\epsilon = \pm 1$, $s_{0,\epsilon} = i\epsilon$, so

$$\begin{aligned}
&\text{Res}\big[U(x,s)e^{st};\, s_{0,\epsilon}\big] \\
&= e^{st}\sinh\left(\frac{s}{c}(L-x)\right) \Big/ \left((s+i\epsilon)\sinh\left(\frac{sL}{c}\right)\right)\Big|_{s=i\epsilon} \\
&= \frac{e^{i\epsilon t}}{i2\epsilon}\cdot \sinh\left(\frac{i\epsilon}{c}(L-x)\right) \Big/ \sinh\left(\frac{i\epsilon L}{c}\right) = \frac{e^{i\epsilon t}}{i2\epsilon}\cdot \sin\left(\frac{\epsilon}{c}(L-x)\right) \Big/ \sin\left(\frac{\epsilon L}{c}\right). \\
&= \frac{e^{\pm it}}{\pm i2}\cdot\left(\pm\sin\left(\frac{L-x}{c}\right)\right)\cdot\left(\pm\sin\left(\frac{L}{c}\right)\right) = \pm\frac{e^{\pm it}}{i2}\cdot\sin\left(\frac{L-x}{c}\right)\cdot\sin\left(\frac{L}{c}\right).
\end{aligned}$$

Add the two residues from the poles $s_{0,\pm}$ to get

$$\frac{e^{it}-e^{-it}}{2i}\cdot\sin\left(\frac{L-x}{c}\right) \Big/ \sin\left(\frac{L}{c}\right),$$

that is,

$$\sin t\cdot\sin\left(\frac{L-x}{c}\right) \Big/ \sin\left(\frac{L}{c}\right).$$

Similarly, for $n \geq 1$ and $\epsilon = \pm 1$, $s_{n,\epsilon} = i\epsilon \cdot \frac{n\pi c}{L}$, so

$$\begin{aligned}
&\operatorname{Res}\left[U(x,s)e^{st};\, s_{n,\epsilon}\right] \\
&= \lim_{s \to i\epsilon n\pi c/L} \left(s - \frac{i\epsilon n\pi c}{L}\right) \cdot e^{st} \\
&\quad \cdot \left(\sinh\left(\frac{s}{c}(L-x)\right) \Big/ \left(1+s^2\right)\right) \cdot \left(1 \Big/ \sinh\left(\frac{sL}{c}\right)\right) \\
&= e^{i\epsilon n\pi ct/L} \cdot \left(\sinh\left(\frac{i\epsilon n\pi (L-x)}{L}\right) \Big/ \left(1 - \left(\frac{n\pi c}{L}\right)^2\right)\right) \\
&\quad \cdot \lim_{s \to i\epsilon n\pi c/L} \left(\left(s - \frac{i\epsilon n\pi c}{L}\right) \Big/ \sinh\left(\frac{sL}{c}\right)\right).
\end{aligned}$$

L'Hôpital's rule gives that this residue equals

$$\begin{aligned}
&= e^{i\epsilon n\pi ct/L} \cdot \left(i \sin\left(\frac{\epsilon n\pi (L-x)}{L}\right) \Big/ \left(1 - \left(\frac{n\pi c}{L}\right)^2\right)\right) \cdot \frac{1}{\frac{L}{c}\cosh(i\epsilon n\pi)} \\
&= e^{i\epsilon n\pi ct/L} \cdot \left(i\epsilon \sin\left(\frac{n\pi (L-x)}{L}\right) \Big/ \left(1 - \left(\frac{n\pi c}{L}\right)^2\right)\right) \cdot \frac{1}{\frac{L}{c}(-1)^n}.
\end{aligned}$$

Noting that $i = \frac{-1}{i}$, *adding* the two residues from the poles $s_{n,\pm}$, gives

$$\frac{2c}{L}(-1)^{n+1} \cdot \frac{e^{in\pi ct/L} - e^{-in\pi ct/L}}{2i} \cdot \sin\left(\frac{n\pi (L-x)}{L}\right) \Big/ \left(1 - \left(\frac{n\pi c}{L}\right)^2\right),$$

that is,

$$\frac{2c}{L}(-1)^{n+1} \cdot \sin\left(\frac{n\pi ct}{L}\right) \cdot \sin\left(\frac{n\pi (L-x)}{L}\right) \Big/ \left(1 - \left(\frac{n\pi c}{L}\right)^2\right).$$

Putting everything together, we get that the solution of the whole problem is

$$u(x,t) = \frac{\sin\left(\frac{L-x}{c}\right)}{\sin\left(\frac{L}{c}\right)} \cdot \sin t \;+\; \frac{2c}{L} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{1 - \left(\frac{n\pi c}{L}\right)^2} \cdot \sin\left(\frac{n\pi (L-x)}{L}\right) \cdot \sin\left(\frac{n\pi ct}{L}\right). \bigcirc$$

Learn More About It

Theorem 17.7, a version of a Laplace inversion theorem, is found on pp. 301–302 of *A First Course in Partial Differential Equations: With Complex Variables and Transform Methods*, by Hans Weinberger, Dover Publications Co. © 1995. Problem 17.3.2.11 is in the exercises of *Transform Methods for Solving Partial Differential Equations*, 2nd edn., by Dean G. Duffy, Chapman & Hall/CRC Press © 2004. [Problem 17.3.2.11 is a simplified version of a problem solved in "Thermal effects on fluid flow and hydraulic fracturing from well bores and cavities in low permeability formations," Y. Wang and E. Papamichos, *Int. J. Numer. Anal. Meth. Geomech.*, **23** (1999) 1819–1834.]

Most of the inverse Laplace transforms given in Table 17.2 involve $\sqrt{s}$, and their derivations need integration over a contour that bends around the branch cut on the half line $\mathcal{R}e(s) \leq 0$. Good references for this include *Operational mathematics*, 3rd edn., by Ruel V, Churchill, McGraw-Hill © 1971; the books by Weinberger and Duffy cited

earlier; and *Fundamentals of Complex Analysis…*, by Saff and Snider cited in Section 15.10.

A source for Table 17.2 is *Handbook of Mathematical Functions*, ed., by Milton Abramowitz and Irene Stegun, Dover Publications © 1964.

17.3.2 Problems

Throughout these problems, assume the solution of the PDE satisfies the boundary condition(s) $\lim_{x\to\pm\infty}|u(x,t)|<\infty$, $\lim_{x\to\pm\infty}|u(x,y)|<\infty$, or $\lim_{y\to\pm\infty}|u(x,y)|<\infty$, as appropriate for the spatial domain.

In problems 1 and 2, find an appropriate Bromwich contour and use it to explain the given inverse Laplace transform.

1. $\mathcal{L}^{-1}\left[\frac{s+1}{(s+1)^2+4}\right]=e^{-t}\cos 2t$
2. $\mathcal{L}^{-1}\left[\frac{1}{s}e^{-as}\right]=\text{step}(t-a)$, assuming a is a positive constant, where step $(t-1)=\begin{cases}0, & t<a\\ 1, & t\ge a\end{cases}$

In problems 3–5, solve the initial value problem for the wave equation on a finite interval with time-dependent boundary condition. Assume $c=\sqrt{\varrho/T}$ is constant. For problem 4, assume $\frac{L}{\pi c}$ is not an integer; for Problems 3 and 5, assume $\frac{2L}{\pi c}$ is not an integer.

TABLE 17.2

Table of Inverse Laplace Transforms, Where a, k Are Nonzero Constants

Formula	$F(s)$	$f(t)=\mathcal{L}^{-1}[F(s)]$
L2.1	$\frac{1}{\sqrt{s}}$	$\frac{1}{\sqrt{\pi t}}$
L2.2	$\frac{1}{\sqrt{s^3}}$	$2\cdot\sqrt{\frac{t}{\pi}}$
L2.3	$\frac{1}{a+\sqrt{s}}$	$\frac{1}{\sqrt{\pi t}}-a\,e^{a^2t}\,\text{erfc}(a\sqrt{t})$
L2.4	$\frac{\sqrt{s}}{s-a^2}$	$\frac{1}{\sqrt{\pi t}}+a\,e^{a^2t}\,\text{erfc}(a\sqrt{t})$
L2.5	$e^{-k\sqrt{s}}$	$\frac{k}{2\sqrt{\pi t^3}}\,e^{-k^2/(4t)}$
L2.6	$\frac{1}{\sqrt{s}}\,e^{-k\sqrt{s}}$	$\frac{1}{\sqrt{\pi t}}\,e^{-k^2/(4t)}$
L2.7	$\frac{1}{s}\,e^{-k\sqrt{s}}$	$\text{erfc}\left(\frac{k}{2\sqrt{t}}\right)$
L2.8	$\frac{1}{s\sqrt{s}}\,e^{-k\sqrt{s}}$	$2\cdot\sqrt{\frac{t}{\pi}}\,e^{-k^2/(4t)}$ - $k\,\text{erfc}\left(\frac{k}{2\sqrt{t}}\right)$
L2.9	$\frac{e^{-k\sqrt{s}}}{a+\sqrt{s}}$	$\frac{1}{\sqrt{\pi t}}\,e^{-k^2/4t}-a\,e^{ak}\,e^{a^2t}\,\text{erfc}\left(\frac{k}{2\sqrt{t}}+a\sqrt{t}\right)$
L2.10	$\frac{1}{\sqrt{1+s^2}}$	$J_0(t)$
L2.11	$\frac{1}{s}\,e^{-s}$	$J_0(2\sqrt{t})$
L2.12	$\frac{1}{s}\arctan\left(\frac{k}{s}\right)$	$Si(kt)\triangleq\int_0^{kt}\frac{\sin\xi}{\xi}\,d\xi$

3. $\left\{\begin{array}{l} \frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2},\ 0 < x < L,\ 0 < t < \infty \\ \left\{\begin{array}{l} u(0,t) = \sin 2t \\ u(L,t) = 0 \end{array}\right\},\ 0 < t < \infty \\ u(x,0) = \frac{\partial u}{\partial t}(x,0) = 0,\ 0 < x < L \end{array}\right\}$

4. $\left\{\begin{array}{l} \frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2},\ 0 < x < L,\ 0 < t < \infty \\ \left\{\begin{array}{l} u(0,t) = 0 \\ u(L,t) = \cos t \end{array}\right\},\ 0 < t < \infty \\ u(x,0) = \frac{\partial u}{\partial t}(x,0) = 0,\ 0 < x < L \end{array}\right\}$

5. $\left\{\begin{array}{l} \frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2},\ 0 < x < L,\ 0 < t < \infty \\ \left\{\begin{array}{l} u(0,t) = 0 \\ \frac{\partial u}{\partial x}(L,t) = \cos t \end{array}\right\},\ 0 < t < \infty \\ u(x,0) = \frac{\partial u}{\partial t}(x,0) = 0,\ 0 < x < L \end{array}\right\}$

6. Generalize Theorem 17.8 for finding $\mathcal{L}^{-1}[F(s)]$, assuming that exactly two of the poles, $s_{0,\pm}$, of the function $F(s) = \frac{p(s)}{q(s)}$ have order $m \geq 2$ and the other poles are simple. As in Theorem 17.8, assume $F(s)$ is of class **L**.

In problems 7 and 8, assume $c = \sqrt{\varrho/T} = 1$ and solve the initial value problem for the wave equation on a half line.

7. $\left\{\begin{array}{l} \frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2},\ 0 < x < \pi,\ 0 < t < \infty \\ \left\{\begin{array}{l} u(0,t) = \sin t \\ u(\pi,t) = 0 \end{array}\right\},\ 0 < t < \infty \\ u(x,0) = \frac{\partial u}{\partial t}(x,0) = 0,\ 0 < x < \pi \end{array}\right\}$

8. $\left\{\begin{array}{l} \frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2},\ 0 < x < \pi,\ 0 < t < \infty \\ \left\{\begin{array}{l} u(0,t) = 0 \\ u(\pi,t) = \cos t \end{array}\right\},\ 0 < t < \infty \\ u(x,0) = \frac{\partial u}{\partial t}(x,0) = 0,\ 0 < x < \pi \end{array}\right\}$

9. Define

$$F(s) = \frac{e^{-x\sqrt{s}}}{s-1}.$$

Use partial fractions [let $q = \sqrt{s}$] to explain why

$$F(s) = \frac{e^{-x\sqrt{s}}}{2}\left(\frac{1}{\sqrt{s}\ -\ 1} - \frac{1}{\sqrt{s}\ +\ 1}\right),$$

and then use entry L2.9 of Table 17.2, the table of inverse Laplace transforms, to explain why this can be evaluated as

$$\mathcal{L}^{-1}[F(s)] = \frac{e^t}{2}\left(e^{-x}\text{erfc}\Big(\frac{x}{2\sqrt{t}} - \sqrt{t}\Big) + e^{x}\text{erfc}\Big(\frac{x}{2\sqrt{t}} + \sqrt{t}\Big)\right).$$

10. Use Laplace transforms to help solve the problem

$$\left\{\begin{array}{l} \frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2}, 0 < x < \infty, 0 < t < \infty \\ \left\{\begin{array}{l} \frac{\partial u}{\partial x}(0,t) = f(t) \\ \lim_{x\to\infty}|u(x,t)| < \infty \end{array}\right\}, 0 < t < \infty \\ u(x,0) = \frac{\partial u}{\partial t}(x,0) = 0, 0 < x < \infty \end{array}\right\}.$$

[Hint: After taking the Laplace transform of the PDE, solve the ODE for functions of s and apply the boundary condition at $x = 0$ and the boundedness as $x \to \infty$.]

11. Use Table 17.2, a table of inverse Laplace transforms, to help solve the system

$$\left\{\begin{array}{l} \frac{\partial u}{\partial t} = \alpha\,\frac{\partial^2 u}{\partial x^2} \\ \\ \frac{\partial v}{\partial t} = \alpha\,\frac{\partial^2 v}{\partial x^2} + \beta\frac{\partial u}{\partial t} \end{array}\right\}, 0 < x < \infty, 0 < t < \infty,$$

along with the boundary conditions

$$\left\{\begin{array}{l} u(0,t) = 1, \frac{\partial v}{\partial x}(0,t) = -1 \\ \\ \lim_{x\to\infty}|u(x,t)| < \infty \\ \lim_{x\to\infty}|v(x,t)| < \infty \end{array}\right\}, 0 < t < \infty.$$

Assume α and β are positive constants. [Hint: Use Laplace transforms to first solve the PDE that only involves $u(x,t)$.]

12. Use Corollary 4.1, a consequence of the second shift theorem for Laplace transforms, to explain why

$$\mathcal{L}^{-1}\left[\frac{1}{s}e^{-cs-k\sqrt{s}}\right] = \text{erfc}\left(\frac{k}{2\sqrt{t-c}}\right)\text{step}(t-c).$$

17.4 Hankel Transforms

In polar coordinates or cylindrical coordinates, the Hankel transform solves differential equations on the interval $0 \le r < \infty$ similarly to the way the Fourier sine or cosine transform solved problems on $0 \le x < \infty$ in Section 17.2. Also, we used Laplace transforms on $0 \le t < \infty$ in Sections 4.4, 4.5, and 17.3. In all cases, (a) differentiations were turned into

algebraic operations, (b) the main difficulty was in finding inverse transforms, and (c) it helped to have tables and complex variable techniques to do that.

Define the **Hankel transform** of functions $f(r)$ by

$$F(k) \triangleq \mathcal{H}_0\left[f(r)\right] \triangleq \int_0^\infty f(r) J_0(kr) r\, dr. \tag{17.50}$$

The Hankel transform inversion is given by

$$\mathcal{H}_0^{-1}\left[F(k)\right] = \int_0^\infty F(k) J_0(rk) k\, dk.$$

The outstanding property of the Hankel transform is that

Theorem 17.9

$$\mathcal{H}_0\left[\frac{1}{r}\frac{d}{dr}\left[r\frac{df}{dr}(r)\right]\right] = -k^2 F(k), \tag{17.51}$$

assuming $f(r)$ satisfies the physical boundary conditions

$$|f(0^+)| < \infty, \text{ where } f(0^+) \triangleq \lim_{r\to 0^+} f(r) \tag{17.52}$$

and

$$\sqrt{r} f(r) \to 0 \text{ and } \sqrt{r} f'(r) \to 0 \text{ as } r \to \infty. \tag{17.53}$$

Why? $J_0(kr)$ satisfies Bessel's differential equation

$$\frac{1}{r}\frac{d}{dr}\left[r\frac{dJ_0}{dr}(kr)\right] + k^2 J_0(kr) = 0 \tag{17.54}$$

and also the boundary conditions (Watson 1922) that

$$\sqrt{r} f(r) \text{ and } \sqrt{r} f'(r) \text{ are bounded as } r \to \infty.$$

We calculate that

$$\mathcal{H}_0\left[\frac{1}{r}\frac{d}{dr}\left[r\frac{df}{dr}(r)\right]\right] = \int_0^\infty \frac{1}{r}\frac{d}{dr}\left[r\frac{df}{dr}(r)\right] J_0(kr)\, r\, dr = \int_0^\infty \frac{d}{dr}\left[r\frac{df}{dr}(r)\right] J_0(kr)\, dr,$$

so integration by parts and the boundary conditions for $f(r)$ and $J_0(kr)$ yield that this improper integral equals

$$= \lim_{a\to 0^+,\, b\to\infty} \left(\left[r\frac{df}{dr}(r)\, J_0(kr) \right]_a^b - \int_a^b r\frac{df}{dr}(r) \cdot \frac{d}{dr}\left[J_0(kr) \right] dr \right)$$

$$= 0 - 0 - \int_0^\infty \frac{df}{dr}(r) \left(r\frac{d}{dr}\left[J_0(kr) \right] \right) dr.$$

A second integration by parts calculates that this equals

$$= - \lim_{a\to 0^+,\, b\to\infty} f(r) \left[\left(r\frac{d}{dr}\left[J_0(kr) \right] \right) \right]_a^b + \int_0^\infty f(r) \left(\frac{d}{dr}\left[r\frac{dJ_0}{dr}(kr) \right] \right) dr,$$

so (17.54) implies that this is

$$= -0 + 0 + \int_0^\infty f(r) \Big(- rk^2 J_0(kr) \Big) dr = -k^2 \int_0^\infty f(r) J_0(kr) r\, dr = -k^2 \mathcal{H}_0\left[f(r) \right] = -k^2 F(k). \quad \square$$

Note that if $f = f(r)$ does not depend on θ, that is, f is circularly symmetric, then the Laplacian operator $\Delta = \nabla^2$ on f is given by $\Delta f(r) = \frac{1}{r}\frac{d}{dr}\left[r\frac{df}{dr}(r) \right]$.

Let's use the theorem to help solve a PDE in polar coordinates.

Example 17.10

A substance* initially has a uniform concentration in the disk $0 \le r < a$ in the xy-plane. For $t > 0$, the substance diffuses in the plane. Find the concentration of the substance for $t > 0$ and (x, y) in $\mathbb{R}^2$.

Method: Let the concentration of the substance be c. The diffusion equation is

$$\frac{\partial c}{\partial t} = D\nabla^2 c,$$

where D is the constant diffusion coefficient, not a differential operator. (The "D" is for "diffusion." If the substance that is diffusing happens to be heat, then $D = \alpha$, in the notation of Chapters 10 and 11.)

The problem seems to be circularly symmetric, so we assume that $c = c(r, t)$. Using the Laplacian in polar coordinates, c should satisfy the PDE

$$\frac{\partial c}{\partial t} = D \cdot \frac{1}{r}\frac{\partial}{\partial r}\left[r\frac{\partial c}{\partial r}(r, t) \right]$$

and the initial condition

$$c(r, 0) = \begin{Bmatrix} c_0, & 0 \le r < a \\ 0, & a \le r < \infty \end{Bmatrix} \tag{17.55}$$

* This is a version of Example 3.5.2 of Duffy (2004).

and the physical boundary conditions

$$|\, c(0^+, t)\,| < \infty \quad \text{and} \quad \lim_{r\to\infty} \sqrt{r}\, c(r,t) = \lim_{r\to\infty} \sqrt{r}\, \frac{\partial c}{\partial r}(r,t) = 0, \text{ for } t > 0.$$

Take the Hankel transform of both sides of the PDE and denote $C(k,t) \triangleq \mathcal{H}_0\,[c(r,t)]$. Using Theorem 17.9, we get

$$\frac{\partial C}{\partial t}(k,t) = -Dk^2 C(k,t)$$

whose solution is

$$C(k,t) = e^{-Dk^2 t} C(k,0).$$

We calculate that

$$C(k,0) = \mathcal{H}_0\,[c(r,0)] = \int_0^\infty c(r,0) J_0(kr) r\, dr = \int_0^a c_0 J_0(kr) r\, dr = c_0 \int_0^a J_0(kr) r\, dr.$$

As it happens (Spiegel 1968), $\int r J_0(r)\, dr = r J_1(r)$, so the change of variables $R = kr$ gives

$$\int r J_0(kr)\, dr = \int \frac{R}{k} J_0(R)\, \frac{dR}{k} = k^{-2} \int R J_0(R)\, dR = k^{-2} R J_1(R) = k^{-1} r J_0(kr).$$

So,

$$C(k,0) = c_0 \cdot \Big[k^{-1} r J_0(kr)\Big]_0^a = c_0 k^{-1} a J_1(ka).$$

The solution of the problem is

$$c(r,t) = \mathcal{H}_0^{-1}\left[a c_0 k^{-1} e^{-Dk^2 t} J_1(ka)\right];$$

hence,

$$c(r,t) = \int_0^\infty a c_0 k^{-1} e^{-Dk^2 t} J_1(ka) J_0(rk) k\, dk = a c_0 \int_0^\infty e^{-Dk^2 t} J_1(ak) J_0(rk) dk. \bigcirc \tag{17.56}$$

In principle, this solution can be computed at any value of (r,t), although the improper integral over the infinite interval $[0,\infty)$ suggests the calculations could be time consuming or inaccurate. In the subject of integral transforms, there are many "tricks of the trade," as we can see from the Dean G. Duffy's book mentioned earlier. Surprisingly, it involves using the general transform $C(k,0)$ rather than its specific evaluation to be $k^{-1} a J_1(ka)$!

The solution of the problem can be written as

$$c(r,t) = \mathcal{H}_0^{-1}\,[C(k,t)\,] = \mathcal{H}_0^{-1}\left[e^{-Dk^2 t}\, C(k,0)\right] = \int_0^\infty e^{-Dk^2 t} \left(\int_0^\infty c(\xi,0) J_0(k\xi) \xi\, d\xi\right) J_0(rk) k\, dk.$$

By changing the order of integration, this becomes*

$$c(r,t) = \int_0^\infty \left(\int_0^\infty e^{-Dk^2 t} J_0(k\xi) J_0(kr) k\, dk\right) c(\xi,0) \xi\, d\xi.$$

* We have corrected the typographical error in Duffy's book that dk should be $k\, dk$.

We can take advantage of the evaluation of "Weber's second exponential integral"*:

$$\int_0^\infty e^{-Dk^2t} J_0(k\xi) J_0(kr) k\, dk = \frac{1}{2Dt} \exp\left(-\frac{r^2+\xi^2}{4Dt}\right) I_0\left(\frac{r\xi}{2Dt}\right), \tag{17.57}$$

where I_0 is the modified Bessel's function of the first kind, of order zero. (We used the function I_0 in Example 11.21) in Section 11.6.

So, the solution of the PDE is

$$c(r,t) = \frac{1}{2Dt} \int_0^\infty \exp\left(-\frac{r^2+\xi^2}{4Dt}\right) I_0\left(\frac{r\xi}{2Dt}\right) c(\xi,0)\xi\, d\xi.$$

Substituting in our specific initial condition given by (17.55), we know the solution of the problem is given by

$$c(r,t) = \frac{c_0}{2Dt} \int_0^a \exp\left(-\frac{r^2+\xi^2}{4Dt}\right) I_0\left(\frac{r\xi}{2Dt}\right) \xi\, d\xi.$$

It is very good that this form of the solution involves an integration over the finite interval $[0,a)$. The only thing that might worry us is the factor of t in many of the denominators. While there appears to be difficulty near $t=0$, in fact for $r \geq 0, \xi > 0$, define $v = t^{-1}$ and calculate

$$\lim_{t\to 0^+} \frac{1}{2Dt} \exp\left(-\frac{r^2+\xi^2}{4Dt}\right) = \lim_{v\to+\infty} \frac{v}{2D} \exp\left(-v \cdot \frac{r^2+\xi^2}{4D}\right) = 0,$$

by using L'Hôpital's rule.

Learn More About It

Valuable resources include *Transform Methods for Solving Partial Differential Equations*, 2nd edn., by Dean G. Duffy, Chapman & Hall/CRC Press © 2004, *Fourier Transforms*, by Ian N. Sneddon, Dover Publications, Inc. © 1995, and *The Theory of Bessel Functions*, by G. N. Watson, Cambridge University Press © 1922 (reprinted 1996).

17.4.1 Problems

Throughout these problems, assume the solution of the PDE satisfies the boundary condition(s) $\lim_{r\to\infty} |u(r,z,t)| < \infty$, $\lim_{z\to\infty} |u(r,z,t)| < \infty$ or $\lim_{r\to\infty} |u(r,\theta,t)| < \infty$, as appropriate for the spatial domain on which the solution is to be defined.

* See p. 273 of Duffy's book, p. 137 of Sneddon's book, or p. 395 of Watson's book.

1. Find conditions on functions $f(r)$ so that $F(k) \triangleq \mathcal{H}_0\left[f(r)\right]$ satisfies

$$k^4 F(k) = \mathcal{H}_0\left[\Delta^2 f(r)\right] = \mathcal{H}_0\left[\frac{1}{r}\frac{d}{dr}\left[r\frac{d}{dr}\left[\frac{1}{r}\frac{d}{dr}\left[r\frac{df}{dr}(r)\right]\right]\right]\right],$$

   where Δ denotes the Laplacian operator and Δ^2 is the **bi-harmonic operator**.
2. You may assume that (Watson 1922, page 389)

$$\int_0^\infty e^{-ak} J_\nu(bk) J_\nu(ck)\, dk = \frac{1}{\pi\sqrt{bc}} \cdot Q_{\nu-\frac{1}{2}}\left(\frac{a^2+b^2+c^2}{2bc}\right), \tag{17.58}$$

   where $Q_\beta(z)$ is called the Legendre function of the second kind, β is not an integer, and a, b, c do not depend on k. Now, for $\nu = 0$, differentiate (17.58) with respect to a to explain why

$$\int_0^\infty e^{-ak} J_0(bk) J_0(ck)\, k\, dk = -\frac{1}{\pi\sqrt{bc}} \cdot \frac{a}{bc}\, Q'_{-\frac{1}{2}}\left(\frac{a^2+b^2+c^2}{2bc}\right). \tag{17.59}$$

3. Let the concentration of a substance be $c = c(r,z)$. Solve the steady-state problem in a half space $z > 0$, that is, $\{(r,z) : 0 \le r < \infty, 0 \le z < \infty\}$, given by the diffusion equation

$$0 = D\nabla^2 c,\ 0 < r < \infty, 0 < z < \infty,$$

   where D is the constant diffusion coefficient and the boundary conditions are

$$|\, c(0^+, z)\,| < \infty, \text{ and } \lim_{r\to\infty} c(r,z) = 0, \text{ for } 0 < z < \infty \text{ and}$$
$$c(r,0) = \begin{cases} c_0, & 0 \le r < a \\ 0, & a \le r < \infty \end{cases}.$$

   Recall that in cylindrical coordinates with circular symmetry, $\nabla^2 c = \frac{1}{r}\frac{\partial}{\partial r}\left[r\frac{\partial c}{tialr}\right] + \frac{\partial^2 c}{\partial z^2}$. [Hint: (17.59) should be useful.]
4. You will explain why the diffusion problem for $0 < r < \infty, 0 < z < \infty, 0 < t < \infty$, given by

$$\left\{\begin{array}{c} \frac{\partial c}{\partial t} = D \cdot \nabla^2 c \\ BCs: \ |\, c(0^+, z, t)\,| < \infty \quad \text{and} \quad \lim_{r\to\infty} c(r,z,t) = 0,\ 0 < z < \infty,\ 0 < t < \infty, \\ \text{and} \quad c(r,0,t) = f(r),\ 0 < r < \infty,\ 0 < t < \infty \\ IC: \ c(r,z,0) = 0,\ 0 \le r < \infty, 0 < z < \infty \end{array}\right\} \tag{17.60}$$

has solution

$$c(r,z,t) = \int_0^\infty J_0(kr)\,\left(1 - e^{-Dk^2t}\right)\sinh(kz)\,F(k)\,k\,dk. \tag{17.61}$$

(a) The boundary condition at $z=0$ suggests using the Fourier sine transform, so define $C(k,\omega,t) = \mathcal{H}_0\left[\mathcal{F}_s\left[c(r,z,t)\right]\right]$. Take both the Hankel and Fourier sine transforms of both sides of the PDE in (17.60) to get

$$\frac{\partial C}{\partial t}(k,\omega t) = -D(\omega^2 + k^2)C - \sqrt{\frac{2}{\pi}}\cdot\omega\cdot F(k),$$

where $F(k) \triangleq \mathcal{H}_0[f(r)]$.

(b) Use the initial condition in (17.60) and solve the ODE for $C(k,\omega,t)$'s dependence on t to get

$$C(k,\omega,t) = \sqrt{\frac{2}{\pi}}\cdot\frac{\omega}{\omega^2+k^2}\cdot F(k)\cdot\left(e^{-D(\omega^2+k^2)t} - 1\right).$$

(c) Use residues to evaluate

$$\mathcal{F}_s^{-1}\left[\frac{\omega}{\omega^2+k^2}\cdot\left(e^{-D(\omega^2+k^2)t} - 1\right)\right] = \sqrt{\frac{2}{\pi}}\int_0^\infty \frac{\omega}{\omega^2+k^2}\cdot\left(e^{-D(\omega^2+k^2)t} - 1\right)\sin\omega z\,dz.$$

Note that the integrand is an even function, so we can replace $\int_0^\infty \ldots d\omega$ by $\frac{1}{2}\int_{-\infty}^\infty \ldots d\omega$.

(d) Finally, take the inverse Hankel transform of $\mathcal{F}_s^{-1}\left[C(k,\omega,t)\right]$ to find that

$$c(r,z,t) = \mathcal{H}_0^{-1}\left[\mathcal{F}_s^{-1}\left[c(r,z,t)\right]\right]$$

satisfies (17.61).

5. In Section 10.4, we mentioned that a model for the vibration of a plate is given by $\varrho\ddot{w} = -D\nabla^4 w + T\nabla^2 w - \varrho g$. Solve the problem

$$\frac{\partial^2 w}{\partial t^2} = -\beta^2\nabla^4 w, 0 < r < \infty,\ 0 < t < \infty,$$

where $w = w(r,t)$ satisfies appropriate boundary and initial conditions and β is a positive constant.

Key Terms

bi-harmonic operator: Problem 17.4.1.1
Bromwich contour: Definition 17.4 in Section 17.3

complementary error function, **erfc**: (17.33) in Section 17.2
convolution on the real line: end of Section 17.1, Problem 17.1.2.8
countably infinite: before Definition 17.3 in Section 17.3
exponential order: Definition 17.5 in Section 17.3
Fourier cosine transform: Problem 17.1.2.11, (17.14) in Section 17.2
Fourier sine transform: (17.15) in Section 17.2
***F(s)* → 0 uniformly on that family**: Definition 17.2 in Section 17.3
Green's function: after (17.13) in Section 17.2
Hankel transform: (17.50) in Section 17.4
heat kernel: after (17.13) in Section 17.2
class L: Definition 17.3 in Section 17.3
propagator: after (17.13) in Section 17.2
unbounded family of left half circles: Definition 17.1 in Section 17.2

References

Duffy, D.G. *Transform Methods for Solving Partial Differential Equations*, 2nd edn. Chapman & Hall/CRC Press, Boca Raton, FL, 2004.

Fisher, S.D. *Complex Variables*, 2nd edn. Dover Publications, New York, 1999, pp. 100–101.

Spiegel, M. R. *Mathematical Handbook of Formulas and Tables*, Schaum's Outlines Series/McGraw-Hill, New York, 1968.

Watson, G.N. *The Theory of Bessel Function*. Cambridge University Press, Cambridge, U.K., 1922, Chapter 7.

18

Nonlinear Ordinary Differential Equations

18.1 Phase Line and Phase Plane

In this chapter, we will learn about nonlinear differential equations. In engineering often our first mathematical model is linear but a more accurate model is nonlinear.

In general, we have an initial value problem (IVP), that is, a system of differential equations and initial condition

$$\left.\begin{matrix}\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t)\\ \mathbf{x}(t_0) = \mathbf{x}_0\end{matrix}\right\}, \tag{18.1}$$

and we will assume that we have existence and uniqueness of a solution to the IVP, meaning that given an initial time t_0 and an initial value $\mathbf{x}(t_0)$, (18.1) has exactly one solution $\mathbf{x}(t)$ for t in some interval $(t_0 - \delta, t_0 + \delta)$ with $\delta > 0$. This is the generalization to systems of the results in Section 3.2. In fact, let us assume that $\mathbf{f}(\mathbf{x}, t)$ is continuously differentiable in $\mathbf{x}$, which is enough to guarantee the existence and uniqueness property. We will say more about this in Section 18.7.

Because of the uniqueness, we can notate that the solution of (18.1) is $\mathbf{x}(t) = \mathbf{x}(t; t_0, \mathbf{x}_0)$, that is, we acknowledge that the solution of IVP (18.1) depends on the initial condition.

In the special case of initial time $t_0 = 0$, we notate $\mathbf{x}(t; t_0, \mathbf{x}_0) = \mathbf{x}(t; \mathbf{x}_0)$; the omission of t_0 will imply that the initial time is $t_0 = 0$. Note that $\mathbf{x}(0; \mathbf{x}_0) = \mathbf{x}_0$.

We saw in Chapters 3 and 4 that if $\mathbf{x}(t) = y(t)$ is one dimensional (1D), that is, $\mathbf{x}$ is in $\mathbb{R}^1$, we can show a solution of (18.1) as a graph in the (t, y)-plane. For example, in Figure 18.1 we show a graph of the solution of $\dot{y} + 10y = -2\sin 2t$, $y(0) = 20/13$, which illustrated the concept of a steady-state oscillation.

Similarly, $\mathbf{x}(t)$ in $\mathbb{R}^2$ can be graphed in $(t, x_1(t), x_2(t))$-space. But suppose the system is **autonomous**, meaning that the right-hand side of the system is $\mathbf{f}(\mathbf{x})$, which does not depend on t, that is, we are studying the IVP

$$\left.\begin{matrix}\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})\\ \mathbf{x}(t_0) = \mathbf{x}_0\end{matrix}\right\}. \tag{18.2}$$

Then, in a sense, all initial times t_0 are the same.

Theorem 18.1

(Autonomous systems) If the system is autonomous, then $\mathbf{x}(t; t_0, \mathbf{x}_0)$, the solution of IVP (18.2), satisfies

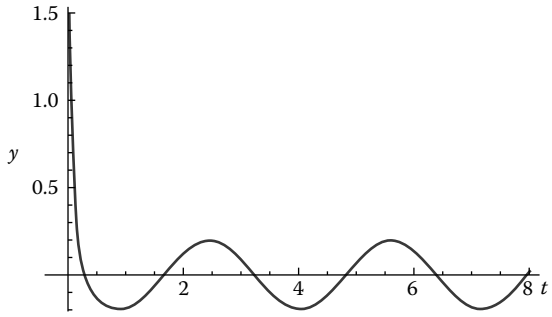


FIGURE 18.1
Steady-state oscillation.

$$\mathbf{x}(t; t_0, \mathbf{x}_0) = \mathbf{x}(t - t_0; \mathbf{x}_0). \tag{18.3}$$

So, for an autonomous system, we can write the solution with any initial time in terms of the solution with initial time $t = 0$ and the same initial value $\mathbf{x}_0$.

Why is Theorem 18.1 correct? By uniqueness of the solution of an IVP, it will suffice to show that $\mathbf{x}(t - t_0; \mathbf{x}_0)$ satisfies (18.2) just as $\mathbf{x}(t; t_0, \mathbf{x}_0)$ does. First, by the chain rule,

$$\frac{d}{dt}\big[\mathbf{x}\big(t - t_0; \mathbf{x}_0\big)\big] = \dot{\mathbf{x}}(t - t_0; \mathbf{x}_0) \cdot \frac{d}{dt}\big[t - t_0\big] = \mathbf{f}\big(\mathbf{x}(t - t_0; \mathbf{x}_0)\big) \cdot 1,$$

so $\mathbf{x}(t - t_0; \mathbf{x}_0)$ satisfies the ordinary differential equation (ODE) in (18.2). Second, as for the initial condition,

$$\mathbf{x}(t - t_0; \mathbf{x}_0)\big|_{t=t_0} = \mathbf{x}(0; \mathbf{x}_0) \triangleq \mathbf{x}_0. \ \square$$

For an autonomous system with $\mathbf{x}$ in $\mathbb{R}^2$, we can display the parametrized curve in $\mathbb{R}^2$ given by

$$\mathcal{C}: \ \mathbf{r} = \mathbf{x}(t; \mathbf{x}_0), \ t \text{ in some interval } \mathcal{I}.$$

Also, we can display the parametrized curves for several initial values $\mathbf{x}_0$. In fact, in Figure 6.14b, we saw several ellipses that are the parametric curve solutions of the undamped harmonic oscillator system

$$\begin{bmatrix} \dot{x} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} v \\ -\frac{k}{m}\, x \end{bmatrix}.$$

When we show several parametrized curve solutions corresponding to different choices of initial values, we get a **phase plane** picture. Note that because of uniqueness of the solution of IVPs, two solutions in the phase plane cannot touch.

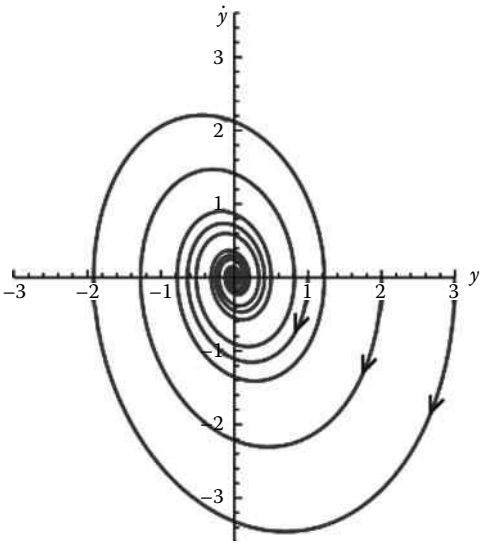


FIGURE 18.2
Phase plane for a damped harmonic oscillator.

Example 18.1

The damped oscillator system $\begin{bmatrix} \dot{y} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} v \\ -2y - 0.4v \end{bmatrix}$ has the phase plane picture shown in Figure 18.2. As we usually do for parametrized curves, the arrow indicates the direction of travel as t increases.

The system in Example 18.1 has the origin as a **stable spiral point**. In general, supposing A is a real 2×2 matrix, the planar, linear, autonomous system

$$\dot{\mathbf{x}} = A\mathbf{x}$$

has a stable spiral point at the origin if the eigenvalues of A are not real but a complex conjugate pair $\lambda_{\pm} = \alpha \pm i\beta$ with $\alpha = \mathcal{R}e(\lambda_{\pm}) < 0$. The origin is an **unstable spiral point** if, instead, $\alpha > 0$.

Example 18.2

The nonlinear system $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} y - x^3 \\ -2x^3 y \end{bmatrix}$ has the phase plane picture shown in Figure 18.3.

Solutions in the phase plane were drawn using *Mathematica*™. For example,

```
s = NDSolve[{x'[t] == y[t] − x[t]^3, y'[t] == −2x[t]^3, x[0] == 1, y[0] == 0},
{x, y}, {t, −20, 20}]
```

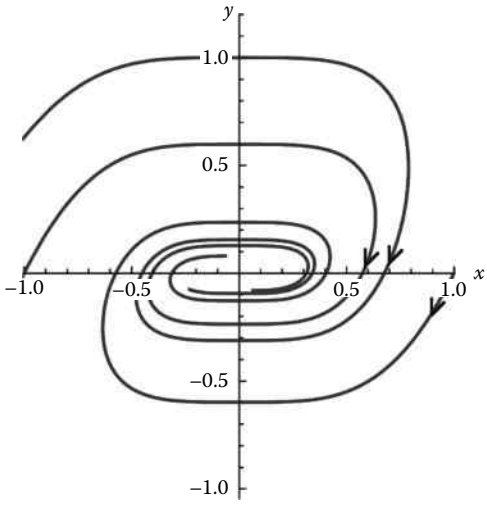


FIGURE 18.3
Phase plane for a nonlinear system.

found an approximate, numerical solution for *one* initial point in the phase plane, and then

```
a = ParametricPlot[Evaluate[{x[t], y[t]}/.s], {t, −20, 20}, PlotStyle− > {Blue, Thick},
PlotRange → {{−1, 1}, {−1.2, 1.2}}, LabelStyle → Directive[FontSize → 16], AxesLabel →
{x, y}, AxesStyle → Thickness[0.00315]]
```

was used to draw the corresponding phase plane curve.

18.1.1 Equilibria

Definition 18.1

System (18.1) has an **equilibrium point** $\bar{\mathbf{x}}$ if

$$\mathbf{x}(t) \equiv \bar{\mathbf{x}}$$

solves (18.1) for t in some open interval.

Other commonly used names for an equilibrium point are **constant solution**, **rest point**, and **critical point**. The plural of "equilibrium point" is **equilibria**.

Theorem 18.2

The autonomous system of ODEs (18.2) has equilibrium point $\bar{\mathbf{x}}$ if, and only if, $\mathbf{f}(\bar{\mathbf{x}}) = \mathbf{0}$.

Why? $\mathbf{x}(t)$ is constant for t in an open interval $\mathcal{I}$ if, and only if, $0 \equiv \dot{\mathbf{x}}(t) = \mathbf{f}\big(\mathbf{x}(t)\big)$ on $\mathcal{I}$. □

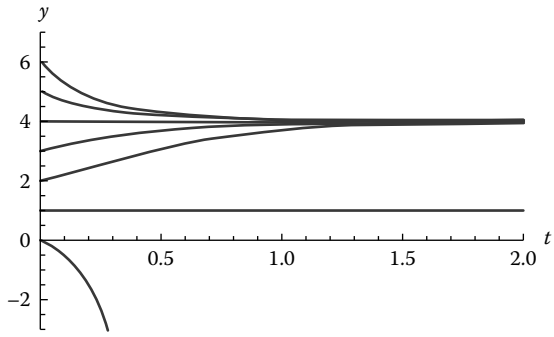


FIGURE 18.4
Example 18.3.

18.1.2 Qualitative Study: The Phase Line

Figure 18.4 shows several solutions of a differential equation of the form

$$\dot{y} = f(y), \tag{18.4}$$

that is, an autonomous ODE in $\mathbb{R}^1$. For this function $f(y)$, ODE (18.4) has constant solutions $y(t) \equiv 1$ and $y(t) \equiv 4$. By Theorem 18.2, $f(1) = f(4) = 0$.

We also see in Figure 18.4 that when $y > 4$, the solution $y(t)$ is a decreasing function of t. Also, when $1 < y < 4$, the solution $y(t)$ is an increasing function of t, and when $y < 1$, the solution $y(t)$ is a decreasing function of t.

Example 18.3

Find a differentiable function $f(y)$ that has among the solutions of $\dot{y} = f(y)$ the functions of t graphed in Figure 18.4.

Method: Because $f(1) = f(4) = 0$, it makes sense for the differentiable function to have the form

$$f(y) = (y-1)(y-4)g(y),$$

where $g(y)$ is differentiable. We know that the product $(y-1)(y-4)$ has

$$Sgn\big((y-1)(y-4)\big) = \begin{Bmatrix} +1, & y < 1 \\ 0, & y = 1 \\ -1, & 1 < y < 4 \\ 0, & y = 4 \\ +1, & y > 4 \end{Bmatrix}.$$

From Figure 18.4, we know where $y(t)$ is increasing or decreasing. This implies we need

$$Sgn\big(f(y)\big) = \begin{Bmatrix} -1, & y < 1 \\ 0, & y = 1 \\ +1, & 1 < y < 4 \\ 0, & y = 4 \\ -1, & y > 4 \end{Bmatrix}, \tag{18.5}$$

so the simplest thing to do is let $g(y) \equiv -1$. So, one example of the desired function $f(y)$ is $f(y) = -(y-1)(y-4)$. ◯

Theorem 18.3

Suppose $f(y)$ is continuously differentiable and $\bar{y}$ is an equilibrium point for (18.4), that is, $\dot{y} = f(y)$. Then

(a) If $y_0 < \bar{y}$, then solution $y(t; y_0) < \bar{y}$ for all $t \geq 0$ for which the solution exists.
(b) If $y_0 > \bar{y}$, then solution $y(t; y_0) > \bar{y}$ for all $t \geq 0$ for which the solution exists.

Why? (a) Uniqueness of solutions and $y_0 \neq \bar{y}$ implies that the two solutions $y(t; y_0)$ and $y(t; y_0) \equiv \bar{y}$ cannot touch in the ty-plane. □

Theorem 18.3 says that the constant solutions act as "barriers" in the ty-plane.
For a first-order scalar ODE, parametrized curve solutions

$$\mathcal{C} : y = y(t),\ \alpha \leq t \leq \beta$$

exist in $\mathbb{R}^1$.

Suppose the ODE is autonomous, that is, is $\dot{y} = f(y)$. If we put all of the parametrized curve solutions in one picture on the real line $\mathbb{R}^1$, we get the **phase line**, the 1D analogue of the phase plane.

Example 18.4

Draw the phase line for the ODE of Example 18.3, that is, $\dot{y} = -(y-1)(y-4)$.

Method: the $Sgn\big(f(y)\big)$ results in (18.5) imply that the phase line is shown in Figure 18.5. ◯

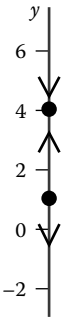


FIGURE 18.5
Example 18.4.

18.1.3 Qualitative Study: The Phase Plane for LCCHS

Suppose A is a 2×2 constant matrix. Consider the LCCHS (linear constant coefficients homogeneous system) of ODEs in $\mathbb{R}^2$, that is,

$$\dot{\mathbf{x}} = A\mathbf{x}, \quad \text{that is,} \quad \begin{Bmatrix} \dot{x} = & ax + by \\ \dot{y} = & cx + dy \end{Bmatrix}, \tag{18.6}$$

for some real constants $a, b, c,$ and d. We know from Chapter 5 that there are several cases for the solutions of (18.6) because there are several cases for A's eigenvalues, λ_1 and λ_2:

(a) λ_1 and λ_2 real and not deficient
(b) λ_1 and λ_2 real, equal, and deficient
(c) λ_1 and λ_2 not real but a complex conjugate pair.

Particular sub-cases have commonly used names and qualitative description of the solution curves in the phase plane.

18.1.4 Saddle Point Case

The **saddle point** case is when $\lambda_2 < 0 < \lambda_1$. Because the eigenvalues are distinct, there is a complete set of basic solutions of the ODE system:

$$\mathbf{x}(t) = c_1 e^{\lambda_1 t}\mathbf{v}_1 + c_2 e^{\lambda_2 t}\mathbf{v}_2\,, \tag{18.7}$$

where c_1, c_2 are arbitrary constants.

A parametrized curve solution with $c_1 \neq 0$ and $c_2 = 0$ will lie on the straight line L_1 through the origin spanned by $\mathbf{v}_1$. Similarly, a parametrized curve solution with $c_1 = 0$ and $c_2 \neq 0$ will lie on the straight line L_2 through the origin spanned by $\mathbf{v}_2$. Following the terminology in the book by Blanchard, Devaney, and Hall mentioned in the following, we refer to the these two types of solutions as **straight line solutions**.

If a solution has both $c_1 \neq 0$ and $c_2 \neq 0$, we will call it a **curved solution**.

Because $\lambda_2 < 0 < \lambda_1$, as $t \to \infty$, in a curved solution, the $c_1 e^{\lambda_1 t}\mathbf{v}_1$ term will dominate. In fact,

$$\mathbf{x}(t) - c_1 e^{\lambda_1 t}\mathbf{v}_1 = c_2 e^{\lambda_2 t}\mathbf{v}_2 \to 0, \text{ as } t \to \infty.$$

So, a curved solution will have L_1 as an asymptote line, as $t \to \infty$.

As $t \to -\infty$, in a curved solution, the $c_2 e^{\lambda_2 t}\mathbf{v}_2$ term will dominate. In fact,

$$\mathbf{x}(t) - c_2 e^{\lambda_2 t}\mathbf{v}_2 = c_1 e^{\lambda_1 t}\mathbf{v}_1 \to 0, \text{ as } t \to -\infty.$$

So, a curved solution will have L_2 as an asymptote line, as $t \to -\infty$.

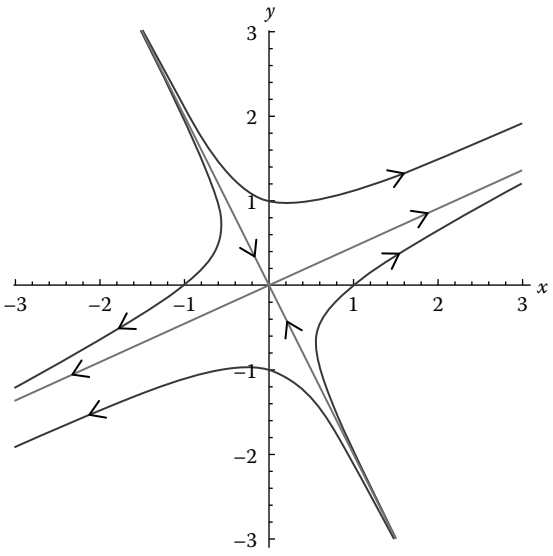


FIGURE 18.6
Example 18.5.

Example 18.5

Analyze and then draw the phase plane for the system $\dot{\mathbf{x}} = \begin{bmatrix} 1 & \sqrt{5} \\ \sqrt{5} & -3 \end{bmatrix} \mathbf{x}$.

Method: It is easy to calculate that the general solution is

$$\mathbf{x}(t) = c_1 e^{2t} \begin{bmatrix} \sqrt{5} \\ 1 \end{bmatrix} + c_2 e^{-4t} \begin{bmatrix} 1 \\ -\sqrt{5} \end{bmatrix},$$

where c_1, c_2 are arbitrary constants. The straight line solutions are on the lines through the vectors $\mathbf{v}_1 = \begin{bmatrix} \sqrt{5} \\ 1 \end{bmatrix}$ and $\mathbf{v}_2 = \begin{bmatrix} 1 \\ -\sqrt{5} \end{bmatrix}$. The curved solutions approach the line through $\mathbf{v}_1$ as $t \to \infty$ and approach the line through $\mathbf{v}_2$ as $t \to -\infty$. The phase plane is sketched in Figure 18.6. ◯

Why do we call the origin a "saddle point" in the case when the two eigenvalues are $\lambda_2 < 0 < \lambda_1$? Because the solutions of the ODE system are level sets of a function that has a saddle point, as we defined it in Section 13.2.

Example 18.6

Analyze and then draw the phase plane for the system $\dot{\mathbf{x}} = \begin{bmatrix} -7 & 3 \\ 2 & -8 \end{bmatrix} \mathbf{x}$.

Method: It is easy to calculate that the general solution is

$$\mathbf{x}(t) = c_1 e^{-5t} \begin{bmatrix} 3/2 \\ 1 \end{bmatrix} + c_2 e^{-10t} \begin{bmatrix} -1 \\ 1 \end{bmatrix},$$

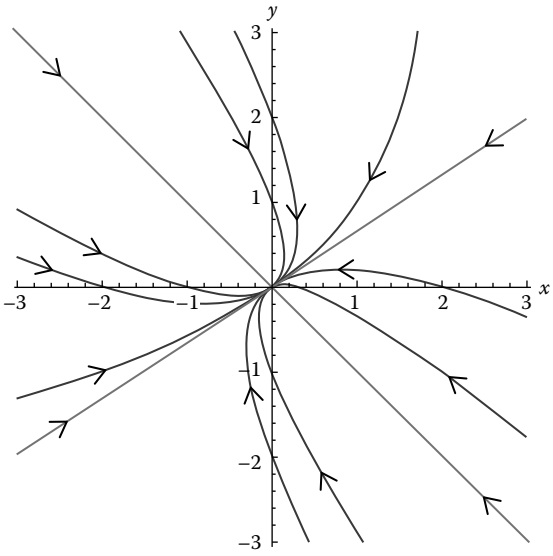


FIGURE 18.7
Example 18.6.

where c_1, c_2 are arbitrary constants. The straight line solutions are on the lines through the vectors $\mathbf{v}_1 = \begin{bmatrix} 3/2 \\ 1 \end{bmatrix}$ and $\mathbf{v}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$. The curved solutions approach the origin parallel to the line through $\mathbf{v}_1$ as $t \to \infty$, because e^{-5t} goes to zero slower than does e^{-10t} as $t \to \infty$.

The curved solutions are roughly parallel to the line through $\mathbf{v}_2$ as $t \to -\infty$, because e^{-10t} is much larger than e^{-5t} as $t \to -\infty$. The phase plane is sketched in Figure 18.7. ◯

Learn More About It

Either *Differential Equations*, 4*th* edn., by Paul Blanchard, Robert L. Devaney, and Glen R. Hall, Brooks/Cole-Thomson, 2011, or *Nonlinear Dynamics and Chaos*, by Steven H. Strogatz, Addison-Wesley, © 1994, is a good introduction to qualitative study of differential equations, including systems.

18.1.5 Problems

For problems 1–5, analyze the system and state what kind of equilibrium point is at the origin.

1. $\dot{\mathbf{x}} = \begin{bmatrix} -1 & \sqrt{3} \\ \sqrt{3} & 1 \end{bmatrix} \mathbf{x}$
2. $\dot{\mathbf{x}} = \begin{bmatrix} 5 & -1 \\ 3 & -1 \end{bmatrix} \mathbf{x}$

3. $\dot{\mathbf{x}} = \begin{bmatrix} 1 & -4 \\ 2 & -3 \end{bmatrix} \mathbf{x}$
4. $\dot{\mathbf{x}} = \begin{bmatrix} -1 & 1 \\ 0 & -2 \end{bmatrix} \mathbf{x}$
5. $\dot{\mathbf{x}} = \begin{bmatrix} 1 & -4 \\ 2 & -3 \end{bmatrix} \mathbf{x}$

For problems 6–8, analyze and then draw the phase plane for the system.

6. $\dot{\mathbf{x}} = \begin{bmatrix} -3 & \sqrt{5} \\ \sqrt{5} & 1 \end{bmatrix} \mathbf{x}$
7. The system of Problem 18.1.5.1
8. The system of Problem 18.1.5.4
9. The IVP $\ddot{x} - x + \frac{3}{2}x^2 = 0$, $x(0) = 1$, $\dot{x}(0) = 0$ has a solution

$$(\star) \quad x(t) = 1 - \left(\frac{1 - e^t}{1 + e^t}\right)^2.$$

   a. Why does $(\star)$ give the unique solution of this IVP?
   b. Use the aforementioned information to find the unique solution of the IVP $\ddot{x} - x + \frac{3}{2}x^2 = 0$, $x(2) = 1$, $\dot{x}(2) = 0$.

18.2 Stability of an Equilibrium Point

In the real world, we sometimes describe a situation as "stable," meaning that it is steady or enduring. For example, it is usually considered a good thing if a nation's economy is stable in the sense of having steady growth. Likewise, when we drive over a bridge, we want it to be stable in the sense that the vibrations of the bridge will be so small that it stands reasonably firm. In these descriptions, we have a qualitative, but not quantitative, concept of stability.

Mathematically there are many precise, quantitative definitions of stability. We have given some for linear homogeneous systems in Chapter 5: Specifically, Definition 5.7 in Section 5.3 gave definitions for constant coefficients ODEs, Definition 5.11 in Section 5.7 gave definitions for constant coefficients difference equations, and Definition 5.12 in Section 5.8 gave definitions for periodic ODEs. It will turn out that these definitions for linear systems are particularizations of definitions for systems in general, that is, systems that may not be linear.

We will study the IVP (18.1) in Section 18.1, that is,

$$\left\{\begin{aligned} \dot{\mathbf{x}} &= \mathbf{f}(\mathbf{x}, t) \\ \mathbf{x}(t_0) &= \mathbf{x}_0 \end{aligned}\right\}.$$

Recall that for a LCCHS $(\star) \quad \dot{\mathbf{x}} = A\mathbf{x}$, we said that the system is asymptotically stable if all its solutions have $\lim_{t\to\infty} \mathbf{x}(t) = \mathbf{0}$, and recall that Theorem 5.11(a) in Section 5.3 said that if all of A's eigenvalues λ satisfy $\mathcal{R}e(\lambda) < 0$, then the system is asymptotically stable.

Because **0** is an equilibrium point for LCCHS ($\star$), our LCCHS definition of asymptotic stability really is a statement about convergence to an equilibrium point. This gives us an idea of how to define stability concepts for the general system of ODEs in (18.1).

Definition 18.2

(a) An equilibrium point $\bar{\mathbf{x}}$ is **stable** if for all $\varepsilon > 0$, no matter how small, and all initial times t_0, there exists some $\delta > 0$, possibly dependent on ε or t_0, such that for all $t \geq t_0$ and initial values $\mathbf{x}_0$ satisfying $||\mathbf{x}_0 - \bar{\mathbf{x}}|| < \delta$,

$$||\mathbf{x}(t; t_0, \mathbf{x}_0) - \bar{\mathbf{x}}|| < \varepsilon. \tag{18.8}$$

(a) An equilibrium point $\bar{\mathbf{x}}$ is **unstable** if it is not stable.

The definition of stability given in Definition 18.2 is sometimes called **stability in the sense of Liapunov**. The latter was a Russian mathematician of the nineteenth century who was one of the founders of the modern approach to dynamical systems.

Figure 18.8 illustrates Definition 18.2: A solution that starts in the ball $B_\delta(\bar{\mathbf{x}}) \triangleq \{\mathbf{x} : ||\mathbf{x}_0 - \bar{\mathbf{x}}|| < \delta\}$ remains in the ball $B_\varepsilon(\bar{\mathbf{x}})$ for all $t \geq t_0$. In the figure, a ball in $\mathbb{R}^n$ appears to be an interval on the **x**-axis. Note in the figure that two solutions appear to intersect at a point $(t_1, \mathbf{y})$, which would violate uniqueness of the solution of IVP (18.1) in Section 18.1, but this misleading appearance is due to displaying n dimensions of **x** in a 1D way. To illustrate the lack of intersection, one of the two solutions is missing part of its graph, around $(t_1, \mathbf{y})$.

Example 18.7

The linear, undamped, harmonic oscillator system is

$$\begin{bmatrix} \dot{y} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} v \\ -\frac{k}{m} y \end{bmatrix}. \tag{18.9}$$

Let $\mathbf{x} = [y \quad v]^T$. We know from Example 6.13 in Section 6.3 that all of the nonzero solutions of (18.9) in the yv-phase plane lie on an ellipse of the form

$$\frac{y^2}{a^2} + \frac{v^2}{(\omega a)^2} = 1, \tag{18.10}$$

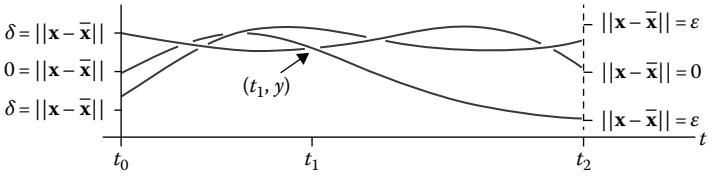


FIGURE 18.8
Stability of $\bar{\mathbf{x}}$.

and are depicted in Figure 6.14. Recall that the constant frequency of vibration is $\omega = \sqrt{k/m}$ and $a = \sqrt{\big(y(0)\big)^2 + \big(v(0)/\omega\big)^2}$ is the amplitude of the undamped harmonic oscillation.

Linear system (18.9) has exactly one equilibrium point, namely, $\mathbf{x} = (y, v) = (0, 0)$. By Definition 18.2, $\mathbf{0}$ is stable: For any $\varepsilon > 0$, no matter how small, let $\delta = \varepsilon/(\sqrt{2} \cdot \max\{1, \omega\})$ to see that if $\big\|\big(y(0), v(0)\big)\big\| < \delta$, then every point on the solution $\big(y(t), v(t)\big)$, which lies on the ellipse (18.10), has

$$||\big(y(t), v(t)\big)|| \leq \sqrt{2} \cdot \max\{1, \omega\} \cdot ||\big(y(0), v(0)\big)|| < \varepsilon.$$

The generalization of Example 18.7 was given in Theorem 5.11 (a) and (b) in Section 5.3; LCCHS $\dot{\mathbf{x}} = A\mathbf{x}$ is either asymptotically stable or neutrally stable, that is, all its solutions are bounded on $[0, \infty)$, if all of A's eigenvalues λ satisfy $\mathcal{R}e(\lambda) \leq 0$ and no deficient eigenvalue λ has real part equal to zero. *For an LCCHS*, being either asymptotically stable or neutrally stable is the same concept as stability in the sense of Liapunov.

It turns out that the concept of asymptotic stability is more sophisticated for nonlinear systems than for LCCHS and thus needs to be broken down into two concepts.

Definition 18.3

An equilibrium point $\bar{\mathbf{x}}$ is an **attractor** if for all initial times t_0, there exists some $\delta > 0$, possibly dependent on t_0, such that for all initial values $\mathbf{x}_0$ satisfying $||\mathbf{x}_0 - \bar{\mathbf{x}}|| < \delta$,

$$\lim_{t \to \infty} ||\mathbf{x}(t; t_0, \mathbf{x}_0) - \bar{\mathbf{x}}|| = 0. \tag{18.11}$$

While condition (18.8) says that solutions stay near the equilibrium point, condition (18.11) says that solutions get arbitrarily closer and closer to the equilibrium point.

For an LCCHS $\dot{\mathbf{x}} = A\mathbf{x}$, $\mathbf{0}$ is the only equilibrium point, and it is an attractor exactly when all of the eigenvalues of the constant matrix A have negative real part.

Definition 18.4

An equilibrium point $\bar{\mathbf{x}}$ is **asymptotically stable** if it is both stable and an attractor.

Again, for an LCCHS $\dot{\mathbf{x}} = A\mathbf{x}$, $\mathbf{0}$ is the only equilibrium point, and it is an attractor exactly when all of the eigenvalues of the constant matrix A have negative real part. So, the Definition 18.4 of asymptotic stability for a general, possibly nonlinear system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$, agrees with Definition 18.4's asymptotic stability when the system happens to be an LCCHS.

Unfortunately, for a nonlinear system, if $n \geq 2$, then an equilibrium point in $\mathbb{R}^n$ can be both unstable and an attractor. Suppose $\mathbf{0}$ is an equilibrium point and $\mathbf{x}(t)$ are the solutions. Intuitively, in two or more dimensions, we can have both (1) $\mathbf{x}(t) \to \mathbf{0}$ as $t \to \infty$, hence $\mathbf{0}$ is an attractor, and (2) $\mathbf{x}(t)$ goes far away from $\mathbf{0}$ on a finite time interval. If we have a sequence of solutions $\mathbf{x}^{(k)}(t)$ whose initial conditions $\mathbf{x}^{(k)}(0) \to \mathbf{0}$ as $k \to \infty$ yet there is a sequence of times t_k at which $||\mathbf{x}^{(k)}(t_k)|| \geq 1$, then $\mathbf{0}$ is unstable. This would correspond to

the colloquial saying, "In the long run, everything will be fine, but in the short run, things will be ruined." A specific example of this is mentioned in "Learn More About It."

18.2.1 Stability from Linearization

Recall that a system of ODEs is "autonomous" if $\mathbf{f}(\mathbf{x}, t)$ does not depend on t. We say a system of ODEs is **periodic** if the dependence of $\mathbf{f}(\mathbf{x}, t)$ on t is periodic, that is, there is a $\mathrm{T} > 0$ such that $\mathbf{f}(\mathbf{x}, t + \mathrm{T}) \equiv \mathbf{f}(\mathbf{x}, t)$.

We know a lot about the stability of the equilibrium point $\mathbf{0}$ of an LCCHS $\dot{\mathbf{x}} = A\mathbf{x}$, or a periodic linear homogeneous system $\dot{\mathbf{x}} = A(t)\mathbf{x}$, by using Theorem 5.11 in Section 5.3 or 5.25 in Section 5.8, respectively. Now we will see how to use that knowledge to get conclusions about the stability of an equilibrium point of a nonlinear system of ODEs $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t)$.

Suppose an autonomous system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ has an equilibrium point $\bar{\mathbf{x}}$, that is, $\mathbf{f}(\bar{\mathbf{x}}) = \mathbf{0}$. Using linear approximation, we have that

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(\bar{\mathbf{x}}) + A(\mathbf{x} - \bar{\mathbf{x}}) + \mathbf{g}(\mathbf{x} - \bar{\mathbf{x}})$$

for $\mathbf{x}$ near $\bar{\mathbf{x}}$, where

$$A \triangleq \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\bar{\mathbf{x}})\right]. \tag{18.12}$$

Definition 18.5

A quantity $\boldsymbol{\varepsilon}$ is $o(h^p)$ if there exists $\lim_{h \to 0^+} \dfrac{||\boldsymbol{\varepsilon}||}{h^p} = 0$.

Note that

$$\mathbf{g}(\mathbf{x} - \bar{\mathbf{x}}) \triangleq \mathbf{f}(\mathbf{x}) - \mathbf{f}(\bar{\mathbf{x}}) - A(\mathbf{x} - \bar{\mathbf{x}}) = o(||(\mathbf{x} - \bar{\mathbf{x}}||), \text{ as } \mathbf{x} \to \bar{\mathbf{x}}.$$

Why? Because we assumed that $\mathbf{f}(\mathbf{x})$ is continuously differentiable.

But, $\bar{\mathbf{x}}$ is an equilibrium point, so $\mathbf{f}(\bar{\mathbf{x}}) = \mathbf{0}$. The system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ is actually

$$\dot{\mathbf{y}} = A\mathbf{y} + \mathbf{g}(\mathbf{y}), \tag{18.13}$$

where $\mathbf{y} \triangleq \mathbf{x} - \bar{\mathbf{x}}$ gives the displacement from equilibrium.

We call the system

$$\dot{\mathbf{y}} = A\mathbf{y} \tag{18.14}$$

the **linearization** of $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ near the equilibrium point $\bar{\mathbf{x}}$.

Using (18.13), all textbooks of advanced ODEs establish the following result. We will explain why it is reasonable in Section 18.3.

Theorem 18.4

(Stability from linearization)

(a) If $\mathbf{y} = \mathbf{0}$ is asymptotically stable for the linearization (18.14), that is, all eigenvalues of the constant matrix A have negative real part, then $\bar{\mathbf{x}}$ is asymptotically stable for $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$.

(b) If the linearization (18.14) has at least one eigenvalue whose real part is positive, then $\bar{\mathbf{x}}$ is unstable for $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$.

Remarks

It is important to understand what Theorem 18.4 says *and* what it does not say. If the linearization has an eigenvalue having zero real part, then neither Theorem 18.4(a) nor Theorem 18.4(b) gives us a conclusion about the original, nonlinear system. The next two examples show that the situation is complicated.

Example 18.8

In Problem 18.4.5.12, you will explain why the nonlinear system $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} y - x^3 \\ -2x^3 \end{bmatrix}$ has $\mathbf{0}$ as a stable equilibrium point. This is true even though the linearization about $\mathbf{0}$ is unstable. The linearization is $\dot{\mathbf{x}} = A\mathbf{x}$, where

$$A = \begin{bmatrix} \frac{\partial}{\partial x}\left[y - x^3\right] & \frac{\partial}{\partial y}\left[y - x^3\right] \\ \frac{\partial}{\partial x}\left[-2x^3\right] & \frac{\partial}{\partial y}\left[-2x^3\right] \end{bmatrix}\Bigg|_{(x,y)=(0,0)} = \begin{bmatrix} -3x^2 & 1 \\ -6x^2 & 0 \end{bmatrix}\Bigg|_{(x,y)=(0,0)} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Because A has a deficient zero eigenvalue, the linearization is unstable. ○

Lemma 18.1

If a system has the form

$$\dot{\mathbf{x}} = A\mathbf{x} + \mathbf{g}(\mathbf{x} - \bar{\mathbf{x}}) \tag{18.15}$$

and all of the terms in $\mathbf{g}(\mathbf{x} - \bar{\mathbf{x}})$ are nonlinear in the sense of being $o(||\mathbf{x} - \bar{\mathbf{x}}||)$, then the linearization of (18.15) about $\mathbf{x} = \bar{\mathbf{x}}$ is

$$\dot{\mathbf{x}} = A\mathbf{x}.$$

In particular, if $\mathbf{g} \equiv \mathbf{0}$, then the linearization is $\dot{\mathbf{x}} = A\mathbf{x}$.

Example 18.9

In Example 18.11, we will explain why the nonlinear system

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \beta(x^2 + y^2) \begin{bmatrix} x \\ y \end{bmatrix} \tag{18.16}$$

has **0** as an asymptotically stable equilibrium point if the constant $\beta < 0$ but has **0** as an unstable equilibrium point if the constant $\beta > 0$. This is true even though the linearization about **0** is stable: The linearization is $\dot{\mathbf{x}} = A\mathbf{x}$, where

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

by Lemma 18.1. Because A has eigenvalues $\pm i$, the linearization is stable. ◯

Examples 18.8 and 18.9 illustrate that if the linearization about an equilibrium point $\bar{\mathbf{x}}$ has an eigenvalue whose real part is zero, then the original nonlinear system's equilibrium point may be asymptotically stable or unstable. It is also possible that the original nonlinear system's equilibrium point may be stable but not asymptotically stable.

Example 18.10

The undamped nonlinear pendulum of length ℓ can be modeled* by the system

$$\begin{bmatrix} \dot{\theta} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} \frac{1}{\ell} v \\ -g \sin\theta \end{bmatrix}.$$

The inverted position is the equilibrium point $(\theta, v) = (\pi, 0)$. Use linearization to explain why the inverted pendulum is unstable.

Method: The linearization is $\begin{bmatrix} \dot{\theta} \\ \dot{v} \end{bmatrix} = A \begin{bmatrix} \theta \\ v \end{bmatrix}$, where we calculate

$$A = \begin{bmatrix} \frac{\partial}{\partial\theta}\left[\frac{1}{\ell} v\right] & \frac{\partial}{\partial v}\left[\frac{1}{\ell} v\right] \\ \frac{\partial}{\partial\theta}\left[-g \sin\theta\right] & \frac{\partial}{\partial v}\left[-g \sin\theta\right] \end{bmatrix}\Bigg|_{(\theta,v)=(\pi,0)}$$

$$= \begin{bmatrix} 0 & \frac{1}{\ell} \\ -g\cos\theta & 0 \end{bmatrix}\Bigg|_{(\theta,v)=(\pi,0)} = \begin{bmatrix} 0 & \frac{1}{\ell} \\ g & 0 \end{bmatrix}.$$

Because the eigenvalues of A are $\pm\sqrt{g/\ell}$, the linearization has an eigenvalue with positive real part. By Theorem 18.4, we conclude that the inverted pendulum is unstable. ◯

By the way, $\omega = \sqrt{g/\ell}$ is the frequency of the linear ODE for an undamped pendulum obtained by approximating $\sin\theta \approx \theta$ near the equilibrium $(\theta, v) = (0, 0)$.

One elementary topic in control theory is how to apply force to stabilize the inverted pendulum, for example, how to balance a long stick on your finger.

* See also Problem 7.2.5.21 for the calculation of the potential energy and see the discussion of calculus of variations for the pendulum in Section 14.2.

18.2.2 Using $r(t)$

If $\mathbf{x}(t)$ is in $\mathbb{R}^n$, define

$$r(t) \triangleq ||\mathbf{x}(t)|| = \sqrt{\mathbf{x}(t)^T\mathbf{x}(t)}.$$

Some nice examples establish stability properties by using $r(t)$.

Example 18.11

Suppose β is a constant. Study the stability properties of equilibria of the system

$$\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix} = \begin{bmatrix} y+\beta x\left(x^2+y^2\right)\\ -x+\beta y\left(x^2+y^2\right)\end{bmatrix}. \tag{18.17}$$

Method: When $\beta = 0$, the only equilibrium point is $(x, y) = (0, 0)$, because $0 = \dot{x} = y + 0 \cdot x\left(x^2+y^2\right) = y$ and $0 = \dot{y} = -x + 0 \cdot y\left(x^2+y^2\right) = -x$ imply $(x, y) = (0, 0)$.

If $\beta = 0$, (18.17) is the LCCHS

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1\\ -1 & 0\end{bmatrix}\mathbf{x},$$

whose eigenvalues $\pm i$. When $\beta = 0$, the equilibrium point $(0, 0)$ is stable but not asymptotically stable, by Theorem 5.11 in Section 5.3.

Define $r = \sqrt{x^2+y^2} \geq 0$.

If $\beta \neq 0$, then an equilibrium point (x, y) must satisfy $0 = \dot{x} = y + \beta x\left(x^2+y^2\right)$ and $0 = \dot{y} = -x + \beta y\left(x^2+y^2\right)$; hence, $y = -\beta x\left(x^2+y^2\right) = -\beta x r^2$ and $x = \beta y\left(x^2+y^2\right) = \beta y r^2$. It follows that necessarily

$$r^2 = x^2 + y^2 = \left(\beta r^2\right)^2 + \left(-\beta r^2\right)^2 = 2\beta^2 r^4;$$

hence,

$$0 = r^2(1 - 2\beta r^2). \tag{18.18}$$

So, there is an equilibrium point with $r = 0$, that is, at $(x, y) = (0, 0)$, no matter what is the sign of β.

In addition, (18.8) has another solution $r = \frac{1}{\sqrt{2\beta}}$, *if* $\beta > 0$. In this case, $x = \beta y \cdot r^2 = \beta y \cdot \frac{1}{2\beta} = \frac{y}{2}$ and $y = -\beta x \cdot r^2 = -\frac{x}{2}$; hence $2x = y = -\frac{1}{2}x$, hence $\frac{3}{2}x = 0$, hence $x = 0$, and hence $y = 2x = 0$, which gives the contradiction $0 < \frac{1}{2\beta} = r^2 = x^2 + y^2 = 0$. So, $(x, y) = (0, 0)$ is the only equilibrium point, *even if* $\beta > 0$.

Define

$$V(t) \triangleq \left(r(t)\right)^2 = \left(x(t)\right)^2 + \left(y(t)\right)^2.$$

Fix a value of $\beta \neq 0$. We calculate

$$\begin{aligned}\dot{V}(t) &= 2x(t)\dot{x}(t) + 2y(t)\dot{y}(t) = 2x(t)\big(\cancel{y(t)} + \beta x(t)V(t)\big) + 2y(t)\big(-\cancel{x(t)} + \beta y(t)V(t)\big)\\ &= 2\beta V(t)\left(\left(x(t)\right)^2 + \left(y(t)\right)\right) = 2\beta\big(V(t)\big)^2.\end{aligned}$$

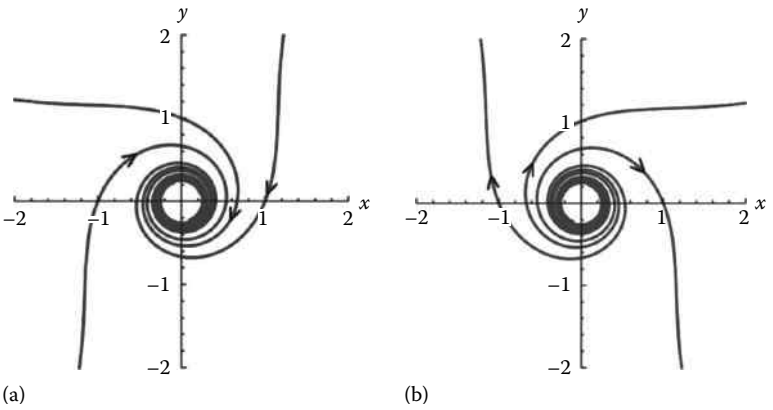


FIGURE 18.9
Phase planes for nonlinear systems. (a) $\beta = -0.4$ and (b) $\beta = 0.4$.

This separable differential equation has solution

$$V(t) = \frac{1}{V_0 - 2\beta t},$$

where $V_0 \triangleq V(0)$.

If $\beta < 0$, then $\dot{V}(t) = 2\beta\big(V(t)\big)^2 < 0$ for all $(x(t), y(t)) \neq (0,0)$. It follows that $(0,0)$ is stable. In addition, $\big(x(t)\big)^2 + \big(y(t)\big)^2 = V(t) \to 0$ as $t \to \infty$; hence, $(0,0)$ is an attractor. It follows that $(0,0)$ is asymptotically stable if $\beta < 0$.

If $\beta > 0$, then

$$\big(x(t)\big)^2 + \big(y(t)\big)^2 = V(t) = \frac{1}{V_0 - 2\beta t} \to \infty, \text{ as } t \to \frac{V_0}{2\beta}^+.$$

It follows that $(0,0)$ is unstable if $\beta > 0$. ◯

Figure 18.9 shows the phase plane for (a) $\beta = -0.4$ and (b) $\beta = 0.4$.

Learn More About It

While the idea of stabilizing an inverted pendulum is an elementary topic, it involves some sophisticated mathematics of control theory. See Sections 3.3 and 9.1 of *Mathematics for Dynamic Modeling*, by Edward Beltrami, Academic Press, Inc., © 1987.

An equilibrium point that is both unstable and an attractor is given in Example 5.7.6 on pages 130–131 of *Ordinary Differential Equations*, 2*nd* ed., by Garrett Birkhoff and Gian-Carlo Rota, Blaisdell Publishing Co., © 1969.

18.2.3 Problems

In problems 1–10, for the given system, (a) find all of the equilibrium points and (b) study their stability using linearization, if possible.

1. $\begin{Bmatrix} \dot{x}_1 = & x_1 & + & 4x_2 \\ \dot{x}_2 = & -x_1 & - & 3x_2 \end{Bmatrix}$

2. $\left\{\begin{array}{l}\dot{x}_1 = x_1 + 3x_2 \\ \dot{x}_2 = -x_1 - x_2 \\ \dot{x}_3 = -x_3 + 1\end{array}\right\}$

3. $\left\{\begin{array}{l}\dot{x}_1 = -x_1\left(1 - x_1^2 - x_2^2\right) \\ \dot{x}_2 = -x_2\left(1 - x_1^2 - x_2^2\right)\end{array}\right\}$

4. $\left\{\begin{array}{l}\dot{x} = -5x + 6 + y^2 \\ \dot{y} = x - y\end{array}\right\}$

5. $\left\{\begin{array}{l}\dot{x} = 8x - y^2 \\ \dot{y} = -y + x^2\end{array}\right\}$

6. $\left\{\begin{array}{l}\dot{x} = -x - x^2 + xy \\ \dot{y} = -y + xy - y^2\end{array}\right\}$
   [Hint: Factor each of the equilibrium equations.]

7. $\left\{\begin{array}{l}\dot{x} = 2\sin x + y \\ \dot{y} = \sin x - 3y\end{array}\right\}$
   [Hint: Solve each of the equilibrium equations for y in terms of x and then solve an equation of the form $f(x) = y = g(x)$ for x.]

8. $\left\{\begin{array}{l}\dot{x} = \frac{1}{1+2y} - 0.2 \\ \\ \dot{y} = 0.3x - 0.6\end{array}\right\}$
   [This is a model for a biochemical reaction.]

9. $\left\{\begin{array}{l}\dot{x} = -0.1x + 0.02xy \\ \dot{y} = 0.2y - 0.4xy - 0.05y^2\end{array}\right\}$

10. $\left\{\begin{array}{l}\dot{x} = x + y^2 \\ \dot{y} = x + y\end{array}\right\}$

11. The motion of a pendulum in a resisting medium can be modeled by $\ddot{\theta} + 2b\dot{\theta} + \omega^2\theta = 0$.
    (a) Write the second-order ODE as a system of two first-order ODEs.
    (b) Using part (a), find all of the equilibrium points and study their stability using linearization.

12. Suppose $\mathbf{f}(\mathbf{x})$ is continuously differentiable at $\mathbf{x} = \mathbf{c}$ for some constant vector $\mathbf{c}$. If ($\star$) $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ has a solution $\mathbf{x}(t)$ satisfying $\lim_{t\to\infty} \mathbf{x}(t) = \mathbf{c}$, then explain why $\mathbf{c}$ must be an equilibrium point for ($\star$).

13. Suppose β is a constant. Study the stability of the origin $(x_1, x_2, x_3) = (0, 0, 0)$ the system

$$\begin{bmatrix}\dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3\end{bmatrix} = \begin{bmatrix}x_2 - x_3 + \beta x_1\left(x_1^2 + x_2^2 + x_3^2\right) \\ x_3 - x_1 + \beta x_2\left(x_1^2 + x_2^2 + x_3^2\right) \\ x_1 - x_2 + \beta x_3\left(x_1^2 + x_2^2 + x_3^2\right)\end{bmatrix}.$$

18.3 Variation of Parameters Using Linearization

Recall from (5.44) in Section 5.4 that the solution of

$$\dot{\mathbf{y}} = A(t)\mathbf{y}(t) + \mathbf{q}(t) \tag{18.19}$$

is given by

$$\mathbf{y}(t) = e^{(t-t_0)A}\mathbf{y}(t_0) + \int_{t_0}^{t} e^{(t-s)A}\mathbf{q}(s)\, ds. \tag{18.20}$$

Now consider an **almost linear** system

$$\dot{\mathbf{y}} = A\mathbf{y}(t) + \mathbf{g}(\mathbf{y}, t), \tag{18.21}$$

where $||\mathbf{g}(\mathbf{y}, t)||$ is $o(||\,\mathbf{y}\,||)$ as $||\,\mathbf{y}\,|| \to 0$.

Equation (18.21) is shorthand for

$$\dot{\mathbf{y}}(t) = A\mathbf{y}(t) + \mathbf{g}(\mathbf{y}(t), t). \tag{18.22}$$

If we define $\mathbf{q}(t) \triangleq \mathbf{g}(\mathbf{y}(t), t)$, we see that the solution of (18.21) is given *implicitly* by

$$\mathbf{y}(t) = e^{(t-t_0)A}\mathbf{y}(t_0) + \int_{t_0}^{t} e^{(t-s)A}\mathbf{g}(\mathbf{y}(s), s)\, ds. \tag{18.23}$$

From (18.23), we will see why Theorem 18.4 in Section 18.2, that is, "Stability from linearization," makes sense in a non-rigorous way.

In the first case, suppose that all eigenvalues of the constant matrix A have negative real part. Then the matrix exponential $e^{tA} \to O$, the zero matrix, as $t \to \infty$. So, we have no difficulty understanding why $e^{(t-t_0)A}\mathbf{y}(t_0)$, the first term on the right-hand side of (18.23), goes to $\mathbf{0}$ as $t \to \infty$, no matter what initial value, $\mathbf{y}(t_0)$, is chosen.

As to why the second term on the right-hand side of (18.23) also goes to $\mathbf{0}$ as $t \to \infty$, the reasoning is a little more complicated because the only thing we know about $\mathbf{g}(\mathbf{y}(s), s)$ is that the dependence of $\mathbf{g}(\mathbf{y}, t)$ on $\mathbf{y}$ is $o(||\,\mathbf{y}\,||)$ as $||\,\mathbf{y}\,|| \to \mathbf{0}$. We refer the reader to "Learn More About It" for references that use more advanced mathematics to clear up the complications.

18.3.1 Saddle Point Theorem

Suppose $\bar{\mathbf{x}}$ is an equilibrium point for an autonomous system of ODEs

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) \tag{18.24}$$

in $\mathbb{R}^n$ and the constant linearization matrix at $\bar{\mathbf{x}}$ is given in (18.12) in Section 18.2, that is,

$$A \triangleq \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\bar{\mathbf{x}})\right].$$

Suppose A has k eigenvalues whose real parts are positive and $n-k$ eigenvalues whose real parts are negative.

The integer k is to satisfy $0 \leq k \leq n$. If $k=0$, then $\bar{\mathbf{x}}$ is asymptotically stable, by Theorem 18.4(a) in Section 18.2; if $k \geq 1$, then $\bar{\mathbf{x}}$ is unstable, by Theorem 18.4(b) in Section 18.2.

Note that we are assuming here that no eigenvalue of A is zero or is imaginary.

So far, we have just been discussing stability of $\bar{\mathbf{x}}$, but we will see that much more is known about nonlinear systems, even just using the linearization.

In the special case of planar *linear* system $\dot{\mathbf{x}} = A\mathbf{x}$, that is, $n=2$, $\bar{\mathbf{x}} = \mathbf{0}$ in $\mathbb{R}^2$ is

- A stable node if $k=0$
- A saddle point if $k=1$
- An unstable node if $k=2$

(18.25)

Recall that for a linear system $\dot{\mathbf{x}} = A\mathbf{x}$, the linearization about the equilibrium point is also $\dot{\mathbf{x}} = A\mathbf{x}$.

Example 18.12

Consider the planar system

$$\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix} = \begin{bmatrix}x-y^2\\ -y\end{bmatrix}. \tag{18.26}$$

Find all of the solutions and show the phase plane.

Method: Generally we cannot find the solutions of a nonlinear system of ODEs, but we can for this example and for some similar examples because one of the ODEs in the system involves only one of the unknowns.

The second ODE in (18.26) is $\dot{y} = -y$, whose solutions are $y(t) = y_0 e^{-t}$, where $y_0 = y(0)$. Substitute into the first ODE in (18.26) to get

$$\dot{x} = x - y_0^2 e^{-2t}.$$

This linear ODE can be solved easily by the method of undetermined coefficients or the method of integrating factor:

$$x(t) = \left(x_0 - \frac{1}{3}y_0^2\right)e^t + \frac{1}{3}y_0^2 e^{-2t}.$$

So, the solutions are

$$\left\{\begin{aligned} x(t) &= \left(x_0 - \tfrac{1}{3}y_0^2\right)e^t + \tfrac{1}{3}y_0^2 e^{-2t}\\ y(t) &= y_0 e^{-t}\end{aligned}\right\}.$$

The solutions behave like

$$x(t) = \left(x_0 - \frac{1}{3}y_0^2\right)e^t, \quad y(t) = 0, \text{ as } t \to \infty,$$

that is, solutions are asymptotic to the line $y=0$.

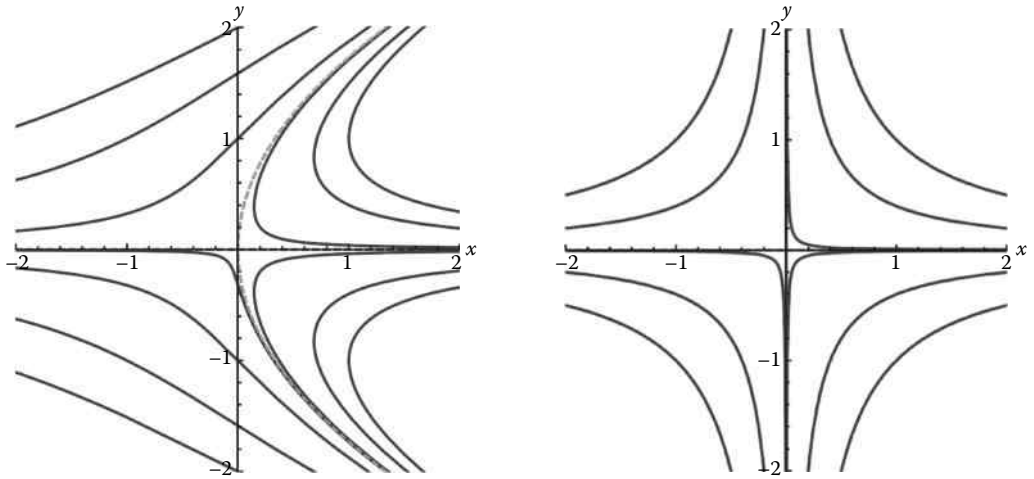


FIGURE 18.10
Example 18.12. (a) $\dot{x} = x - y^2, \dot{y} = -y$ (b) $\dot{x} = x, \dot{y} = -y$.

The solutions behave like

$$x(t) = \frac{1}{3} y_0^2 e^{-2t}, \quad y(t) = y_0 e^{-t}, \text{ as } t \to -\infty,$$

that is, solutions are asymptotic to the parabola $x = \frac{1}{3}y^2$.

Figure 18.10a shows the phase plane for system (18.26): Solutions are drawn using *Mathematica*. The parametrized curves

$$x = x(t), \quad y(t) \equiv 0$$

and

$$x(t) = \frac{1}{3} y(t)^2, \quad y = y(t)$$

are also solutions and are drawn in dashed curves. The solutions lying on the parabola $x = \frac{1}{3}y^2$ are obtained by taking initial conditions $x_0 = \frac{1}{3} y_0^2$. ◯

For Example 18.12, the corresponding linearization $\mathbf{x} = A\mathbf{x}$ is $\dot{x} = x$, $\dot{y} = -y$, and its phase plane is drawn in Figure 18.10b. The phase plane for the nonlinear system looks like a distorted, "sheared to the right" version of the phase plane for the linearization. We will see that this relationship between the nonlinear systems and its linearization can be made precise and turned into a theorem, whether in $\mathbb{R}^2$ or more generally in $\mathbb{R}^n$, for $n \geq 2$.

It turns out that the dashed solution curves for the nonlinear system are directly related to the dashed solution curves of the linearization. The linear system $\dot{x} = x$, $\dot{y} = -y$ has a saddle point at the origin, and the lines $y = 0$ and $x = 0$ have straight line solutions corresponding to the eigenvectors $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$, respectively. The dashed curves in Figure 18.10a, that is, the line $y = 0$ and the parabola $x = \frac{1}{3} y^2$, are called the "unstable manifold" and "stable manifold" and are images of the lines, that is, vector subspaces, $y = 0$ and $x = 0$, respectively.

Consider the case $1 < k < n$, so some of the eigenvalues of the linearization, $\lambda_1, \ldots, \lambda_k$, have positive real part and some, $\lambda_{k+1}, \ldots, \lambda_n$, have negative real part. We will assume* that A has corresponding linearly independent sets of eigenvectors $\{\mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(k)}\}$ and $\{\mathbf{e}^{(k+1)}, \ldots, \mathbf{e}^{(n)}\}$.

There is a projection P_+ onto $\mathcal{V}_+ \triangleq Span\{\mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(k)}\}$, that is, an $n \times n$ matrix satisfying

$$\left(P_+\right)^2 = P_+ \quad \text{and} \quad A\mathbf{p}_+ \text{ is in } \mathcal{V}_+ \text{ for all } \mathbf{p}_+ \text{ in } \mathcal{V}_+.$$

The latter property is referred to as "$\mathcal{V}_+$ is invariant under A." The $_+$ in the notation refers to the assumption that the real parts of $\lambda_1, \ldots, \lambda_k$ are positive.

Correspondingly, the matrix $P_- \triangleq I - P_+$ is a projection onto $\mathcal{V}_- \triangleq Span\{\mathbf{e}^{(k+1)}, \ldots, \mathbf{e}^{(n)}\}$ and $\mathcal{V}_-$ is invariant under A.

Definition 18.6

A set $\mathcal{S}$ is **invariant** for (18.24) if for every initial value $\mathbf{x}_0$ in S, the solution $\mathbf{x}(t; \mathbf{x}_0)$ is in $\mathcal{S}$ for all t.

Recall that the omission of t_0 from the solution notation $\mathbf{x}(t; \mathbf{x}_0)$ implies that the initial time is $t_0 = 0$.

Definition 18.7

A set $\mathcal{S}$ is a **twice continuously differentiable surface** given by a map Φ at a point $\bar{\mathbf{x}}$ **tangent to** a vector subspace $\mathcal{V}$ of $\mathbb{R}^n$ if there is a positive constant ε and a continuously differentiable map $\mathbf{\Phi}$ from $\mathcal{V}$ to $\mathbb{R}^n$ such that

- $\mathcal{S} = \{\mathbf{\Phi}(\bar{\mathbf{x}} + \mathbf{p}) : \mathbf{p} \text{ is in } \mathcal{V} \text{ and } ||\mathbf{p}|| < \varepsilon\}$.
- $\mathbf{\Phi}(\bar{\mathbf{x}} + \mathbf{p}) = \bar{\mathbf{x}} + \mathbf{p} + o(||\mathbf{p}||)$ as $||\mathbf{p}|| \to \mathbf{0}$.

We refer to $\mathcal{S}$ as a k-**dimensional surface** if the dimension of the vector subspace $\mathcal{V}$ is k.

A 1D surface is a parametrized curve, and a 2D surface is a parametrized surface.

Example 18.13

Explain why the dashed curves of Example 18.12 are twice continuously differentiable 1D surfaces.

Method: The dashed line $y = 0$ is the set

$$\mathcal{U} = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix} + c_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} : -\infty < c_1 < \infty \right\},$$

* If instead, there is a deficient eigenvalue, then we would need to use a generalized eigenvector(s), for example, a $\mathbf{w}$ with $(A - \lambda I)\,\mathbf{w} = \mathbf{v}$, assuming $(A - \lambda I)\,\mathbf{v} = \mathbf{0}$.

that is, uses the map

$$\mathbf{\Phi}_+(\mathbf{0}+\mathbf{p}_+) \triangleq \mathbf{0}+\mathbf{p}_+, \text{ for } \mathbf{p}_+ \text{ in } Span\left\{\begin{bmatrix}1\\0\end{bmatrix}\right\}.$$

The parabola $x = \frac{1}{3}y^2$ is the set

$$\mathcal{S} = \left\{\begin{bmatrix}0\\0\end{bmatrix} + c_1\begin{bmatrix}0\\1\end{bmatrix} + \frac{1}{3}c_1^2\begin{bmatrix}1\\0\end{bmatrix} : -\infty < c_1 < \infty\right\},$$

that is, uses the map

$$\mathbf{\Phi}_-(\mathbf{0}+\mathbf{p}_-) \triangleq \mathbf{0}+\mathbf{p}_- + \frac{1}{3}||\mathbf{p}_-||^2\begin{bmatrix}1\\0\end{bmatrix}, \text{ for } \mathbf{p}_- \text{ in } Span\left\{\begin{bmatrix}0\\1\end{bmatrix}\right\}. \bigcirc$$

Definition 18.8

A set $\mathcal{S}$ is a **stable manifold** for (18.24) if there are positive constants ε, r_-, K_- and map $\mathbf{\Phi}_-$ such that

- $\mathcal{S}$ is invariant for (18.24).
- $\mathcal{S}$ is a twice continuously differentiable surface givenbymap $\mathbf{\Phi}_-$ at $\bar{\mathbf{x}}$ tangent to $\mathcal{V}_-$.
- $||\mathbf{x}(t;\mathbf{\Phi}(\bar{\mathbf{x}}+\mathbf{p}_-)|| \le K_- e^{-r_- t}$ for all $\mathbf{p}_-$ in $\mathcal{V}_-$ with$||\mathbf{p}_-|| < \varepsilon,$ for all $t \ge 0$.

Note that $K_- e^{-r_- t} \to 0$ as $t \to \infty$.

Definition 18.9

A set $\mathcal{U}$ is an **unstable manifold** for (18.24) at $\bar{\mathbf{x}}$ if there are positive constants ε, r_+, K_+ such that

- $\mathcal{U}$ is invariant for (18.24).
- $\mathcal{U}$ is a twice continuously differentiable surface given by map $\mathbf{\Phi}_+$ at $\bar{\mathbf{x}}$ tangent to $\mathcal{V}_+$.
- $||\mathbf{x}(t;\mathbf{\Phi}(\bar{\mathbf{x}}+\mathbf{p}_+)|| \le K_+ e^{r_+ t}$ for all $\mathbf{p}_+$ in $\mathcal{V}_+$ with $||\mathbf{p}_+|| < \varepsilon,$ for all $t \le 0$.

Note that $K_+ e^{r_+ t} \to 0$ as $t \to -\infty$.

Theorem 18.5

(Saddle point theorem) Suppose $\mathbf{f}(\mathbf{x})$ is twice continuously differentiable for $\mathbf{x}$ in some open ball $\{\mathbf{x}:||\mathbf{x}-\bar{\mathbf{x}}|| < \varepsilon\}$ and the linearization matrix A at $\bar{\mathbf{x}}$ has k eigenvalues whose real part is positive and $n-k$ eigenvalues whose real part is negative. Then the autonomous system (18.24) has a k-dimensional unstable manifold and an $(n-k)$-dimensional stable manifold at $\bar{\mathbf{x}}$.

The "Learn More About It" section will refer you to where you can find a formal proof of this fundamental theorem of nonlinear ODEs.

18.3.2 Periodic Solutions

Suppose that in

$$\dot{\mathbf{y}} = A(t)\mathbf{y} + \mathbf{q}(t), \tag{18.27}$$

the matrix $A(t)$ and the vector $\mathbf{q}(t)$ are both periodic with period T. Note that if $A(t)$ is constant, then it is periodic with any period.

Theorem 5.26 in Section 5.8 (noncritical systems) established that if the corresponding linear homogeneous system of ODEs

$$\dot{\mathbf{y}} = A(t)\mathbf{y} \tag{18.28}$$

has no T-periodic solution, then the nonhomogeneous system (18.27) is guaranteed to have a T-periodic solution. In fact, that periodic solution is given by

$$\mathbf{y}(t) = X(t)\left(\mathbf{y}_0 + \int_0^t (X(s))^{-1}\,\mathbf{q}(s)ds\right),$$

where $X(t)$ is the principal fundamental matrix of (18.28), that is, $\dot{X}(t) = A(t)X(t)$, $X(0) = I$, and

$$\mathbf{y}_0 = (I - X(\mathsf{T}))^{-1}\,X(\mathsf{T})\int_0^{\mathsf{T}} (X(s))^{-1}\,\mathbf{q}(s)ds.$$

In particular, if $A(t) \equiv A$ is constant and has no eigenvalue that is an integer multiple of $i2\pi/\mathsf{T}$ and $\mathbf{q}(t)$ is periodic with period T, then the nonhomogeneous system (18.27) is guaranteed to have a T-periodic solution given by

$$\mathbf{y}(t) = \left(e^{-\mathsf{T}A} - I\right)^{-1}\int_0^{\mathsf{T}} e^{(t-s)A}\mathbf{q}(s)ds + \int_0^t e^{(t-s)A}\mathbf{q}(s)ds. \tag{18.29}$$

Using more advanced methods of analysis, specifically fixed point theory, one can establish the existence of a periodic solution of a nonlinear system of ODEs.

Theorem 18.6

(Weak perturbation of a noncritical system) Suppose $\mathbf{g}(t, x)$ satisfies

- $\mathbf{g}(t, \mathbf{0}) = \mathbf{0}$,
- $\mathbf{g}(t + \mathsf{T}, \mathbf{x}) \equiv \mathbf{g}(t, \mathbf{x})$, and
- $\mathbf{g}(t, \mathbf{x})$ is continuously differentiable in $\mathbf{x}$ at $\mathbf{x} = \mathbf{0}$,

and the linear homogeneous system of ODEs

$$\dot{\mathbf{y}} = A(t)\mathbf{y} \tag{18.30}$$

has no T-periodic solution. Then the nonhomogeneous system

$$\dot{\mathbf{x}} = A(t)\mathbf{x}(t) + \mathbf{g}(t, x) \tag{18.31}$$

is guaranteed to have a T-periodic solution. In fact, that periodic solution satisfies

$$\mathbf{x}(t) = X(t)\left(\mathbf{x}_0 + \int_0^t (X(s))^{-1}\,\mathbf{g}\big(s, x(s)\big)ds\right)$$

for some appropriate initial value $\mathbf{x}_0$, where $X(t)$ is the principal fundamental matrix of the noncritical linear system (18.30).

Example 18.14

Study $(\star)$ $\dot{\mathbf{x}} = \begin{bmatrix} -1 & 3 \\ 1 & 1 \end{bmatrix}\mathbf{x} + \begin{bmatrix} (\cos 2t)x_1^2 \\ 0 \end{bmatrix}$ for the existence of a π-periodic solution.

Method: Using the method of Section 5.2, we construct the principal fundamental matrix for the corresponding linear homogeneous system of ODEs $(\star\star)$ $\dot{\mathbf{x}} = \begin{bmatrix} -1 & 3 \\ 1 & 1 \end{bmatrix}\mathbf{x}$: Using eigenvalues and eigenvectors, we find the general solution of $(\star\star)$:

$$\mathbf{x} = c_1 e^{2t}\begin{bmatrix} 1 \\ 1 \end{bmatrix} + c_2 e^{-2t}\begin{bmatrix} -3 \\ 1 \end{bmatrix},$$

which gives a fundamental matrix

$$Z(t) = \begin{bmatrix} e^{2t} & -3e^{-2t} \\ e^{2t} & e^{-2t} \end{bmatrix}$$

and thus principal fundamental matrix

$$X(t) = Z(t)\big(Z(0)\big)^{-1} = \begin{bmatrix} e^{2t} & -3e^{-2t} \\ e^{2t} & e^{-2t} \end{bmatrix}\left(\frac{1}{4}\begin{bmatrix} 1 & 3 \\ -1 & 1 \end{bmatrix}\right) = \frac{1}{4}\begin{bmatrix} e^{2t} + 3e^{-2t} & 3e^{2t} - 3e^{-2t} \\ e^{2t} - e^{-2t} & 3e^{2t} + e^{-2t} \end{bmatrix}.$$

Because $(\star\star)$ has no π-periodic solution, Theorem 18.6 guarantees that the original, nonlinear system $(\star)$ has a π-periodic solution $\mathbf{x}(t)$ that satisfies

$$\mathbf{x}(t) = \frac{1}{4}\begin{bmatrix} e^{2t} + 3e^{-2t} & 3e^{2t} - 3e^{-2t} \\ e^{2t} - e^{-2t} & 3e^{2t} + e^{-2t} \end{bmatrix}\cdot$$

$$\cdot\left(\mathbf{x}_0 + \int_0^t \left(\frac{1}{4}\begin{bmatrix} e^{2s} + 3e^{-2s} & 3e^{2s} - 3e^{-2s} \\ e^{2s} - e^{-2s} & 3e^{2s} + e^{-2s} \end{bmatrix}\right)^{-1}\begin{bmatrix} (\cos 2s)x_1^2(s) \\ 0 \end{bmatrix} ds\right),$$

that is,

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \frac{1}{4}\begin{bmatrix} e^{2t} + 3e^{-2t} & 3e^{2t} - 3e^{-2t} \\ e^{2t} - e^{-2t} & 3e^{2t} + e^{-2t} \end{bmatrix}\cdot$$

$$\cdot\left(\mathbf{x}_0 + \int_0^t \left(\frac{1}{4}\begin{bmatrix} 3e^{2s} + e^{-2s} & -3e^{2s} + 3e^{-2s} \\ -e^{2s} + e^{-2s} & e^{2s} + 3e^{-2s} \end{bmatrix}\right)\begin{bmatrix} (\cos 2s)x_1^2(s) \\ 0 \end{bmatrix} ds\right),$$

that is,

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \frac{1}{4}\begin{bmatrix} e^{2t}+3e^{-2t} & 3e^{2t}-3e^{-2t} \\ e^{2t}-e^{-2t} & 3e^{2t}+e^{-2t} \end{bmatrix} \left(\mathbf{x}_0 + \frac{1}{4}\int_0^t \begin{bmatrix} 3e^{2s}+e^{-2s}(\cos 2s)x_1^2(s) \\ -e^{2s}+e^{-2s}(\cos 2s)x_1^2(s) \end{bmatrix} ds \right). \tag{18.32}$$

This is a system of nonlinear Volterra integral equations that can be stated abstractly as $\mathbf{x} = N(\mathbf{x})$. The most straight forward way to find the solution function $\mathbf{x}(t)$ is to take a first guess for the solution as $\mathbf{x}_1(t) \equiv \mathbf{v}$ for some constant vector $\mathbf{v}$ and then calculate the sequence of vector-valued functions $\mathbf{x}_2(t) \triangleq N(\mathbf{x}_1(t)), \mathbf{x}_3(t) \triangleq N(\mathbf{x}_2(t)), \ldots$. Hopefully, the sequence of vector-valued functions will converge to a solution $\mathbf{x}_\infty(t)$ that would be the π-periodic solution of the original system ($\star$). ◯

Learn More About It

Example 18.12 is essentially Example 9.28 in *Dynamics and Bifurcations*, by Jack K. Hale and Hüseyin Koçak, Springer-Verlag, © 1991.

Formal statements and explanations of the saddle point theorem 18.5 are found in Theorem II.6.1 of *Ordinary Differential Equations*, by Jack K. Hale, Robert E. Krieger Publishing Company, © 1980. Also in that book, existence of invariant manifolds for periodic systems is established in its Theorem IV.3.1, and its Theorem IV.2.1 has an explanation of Theorem 18.6.

18.3.3 Problems

In problems 1 and 2, for the planar system, find all of the solutions and show the phase plane.

1. $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} x \\ -x^2 - y \end{bmatrix}$
2. $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} x - y^4 \\ -y \end{bmatrix}$

In problems 3 and 4, study the system for the existence of a 2π-periodic solution.

3. $\dot{\mathbf{x}} = \begin{bmatrix} 1 & 1 \\ 3 & -1 \end{bmatrix}\mathbf{x} + \begin{bmatrix} x_1^3 \\ \sin t \end{bmatrix}$
4. $\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix}\mathbf{x} + \begin{bmatrix} \cos t \\ x_1^2 \end{bmatrix}$

18.4 Liapunov Functions

Liapunov functions can be used to establish stability of an equilibrium point of a system of nonlinear ODEs. The basic idea is that a Liapunov function can measure something that "corresponds" to the distance from an equilibrium point. We may calculate that as time increases that "distance" is nonincreasing or decreasing, in which case we can conclude

that the equilibrium is stable or asymptotically stable. Or we may calculate that as time increases, that "distance" is increasing, in which case we may be able to conclude that the equilibrium is unstable.

For linear constant coefficients homogeneous systems, Theorem 5.11 in Section 5.3 may give definitive conclusions about stability, at least assuming we can find all of the exact eigenvalues. And we learned in Section 18.2 how to use linearization to get results about stability of an equilibrium point of a nonlinear system of ODEs; however, in the neutrally stable case, linearization fails to give useful conclusions. Liapunov functions give us another tool for analyzing nonlinear systems even if linearization is not conclusive.

We motivate the concepts with two linear examples.

Example 18.15

Let $\mathbf{x} = [y \quad v]^T$ solve the system of differential equations

$$\begin{bmatrix} \dot{y} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} v \\ -\frac{k}{m} y \end{bmatrix}. \tag{18.33}$$

Physically, this is the undamped, linear problem of vibration of a mass m on a massless spring whose spring constant is k, so it makes sense to define the total energy of the system by

$$E = E(y, v) \triangleq \frac{1}{2}\left(ky^2 + mv^2\right),$$

that is, the sum of the potential energy in the spring and the kinetic energy of the mass at the end of the spring. Discuss the behavior of the energy of solutions for $t \geq 0$.

Method: Let $E(t) \triangleq E\big(y(t), v(t)\big)$ where $\mathbf{x}(t) = [y(t) \quad v(t)]^T$ solves (18.33). The chain rule gives

$$\dot{E}(t) = ky(t)\dot{y}(t) + mv(t)\dot{v}(t).$$

But $\mathbf{x}(t)$ solves the system of differential equations, so

$$\dot{E}(t) = ky(t)v(t) + mv(t)\left(-\frac{k}{m} y(t)\right) \equiv 0.$$

So, the energy is constant in time. ◯

Physically, this makes perfect sense: Because there is no damping, we have conservation of the total energy of the system.

In a sense, $E = E(y, v)$ measures the square of the distance between a point (y, v) in the phase plane and the equilibrium point at $(0, 0)$. The conclusion of Example 18.15 says that the solution of the system of differential equations (18.33) stays on the same level set of E as where it begins at $t = 0$. This is illustrated in Figure 18.11a for the specific example of $k/m = 2$.

Example 18.15 also illustrates the subject of Hamiltonian mechanics because ODE system (18.33) can be rewritten as

$$\left\{ \begin{array}{l} \dot{q} = \dfrac{\partial H}{\partial p} \\ \dot{p} = -\dfrac{\partial H}{\partial q} \end{array} \right\},$$

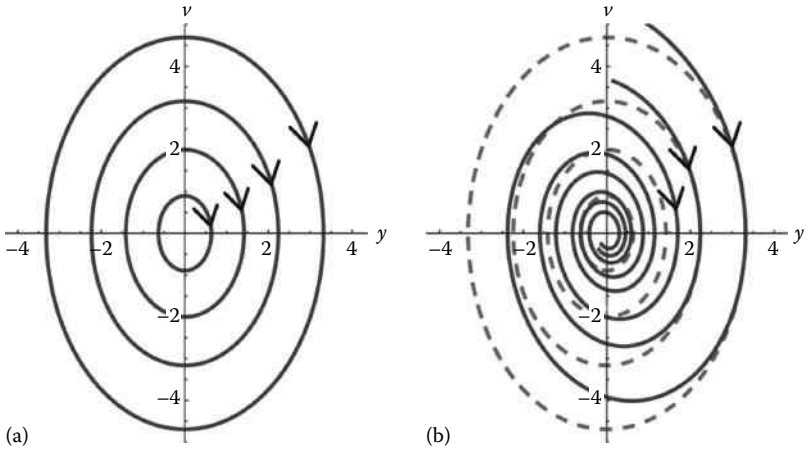


FIGURE 18.11
(a) Example 18.15 and (b) Example 18.16.

where the momentum is $p = mv$, the position is $q = y$, and the Hamiltonian function is $H(p,q) \triangleq \frac{1}{2}kq^2 + \frac{1}{2m}p^2$. This will be generalized in Problem 18.4.5.20, with Problem 18.4.5.21 as an example.

The first thing to know about Hamiltonian systems is that the Hamiltonian is an invariant, that is, $\dot{H} = 0$. As such, they model physical systems without energy dissipation. If we can show that H is "positive definite," as defined later in Definition 18.10, then the Liapunov function theory as follows in Theorem 18.7(a) will establish stability.

Example 18.16

Repeat the instructions of Example 18.15 for the system of differential equations

$$\begin{bmatrix} \dot{y} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} v \\ -\frac{k}{m}y - \frac{b}{m}v \end{bmatrix}. \tag{18.34}$$

Physically, this is the linear problem of vibration of a mass m on a massless spring whose spring constant is k and is experiencing a linear damping force whose coefficient is b. As in Example 18.15 define the total energy of the system by

$$E = E(y,v) \triangleq \frac{1}{2}\left(ky^2 + mv^2\right),$$

that is, the sum of the potential energy in the spring and the kinetic energy of the mass at the end of the spring. Discuss the behavior of the energy of solutions for $t \geq 0$.

Method: Let $E(t) \triangleq E\big(y(t), v(t)\big)$ where $\mathbf{x}(t) = [y(t) \quad v(t)]^T$ solves (18.34). The chain rule gives

$$\dot{E}(t) = ky(t)\dot{y}(t) + mv(t)\dot{v}(t) = ky(t)v(t) + mv(t)\left(-\frac{k}{m}y(t) - \frac{b}{m}v(t)\right) = -bv(t)^2 \leq 0.$$

So, the energy is nonincreasing in time. ◯

Physically, this makes perfect sense: Because there is damping, we do not expect the total energy of the system to be conserved.

The conclusion of Example 18.16 says that the solution of the system of differential equations (18.34) stays on the same level set of E or moves to a lower energy level as time goes by. This is illustrated in Figure 18.11b for the specific example of $k/m = 2$ and $b/m = 0.3$.

18.4.1 Definite Functions

Recall from Definition 2.19 in Section 2.6 that a real, square matrix B is positive definite if $\mathbf{x}^T B\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$. The concept of definiteness can be generalized to great use in the study of stability of differential equations.

Throughout this section, we will assume that $V = V(\mathbf{x})$ is a real-valued function of $\mathbf{x}$ in $\mathbb{R}^n$ that satisfies

$$(1) \qquad V(\mathbf{0}) = 0$$

and

$$(2) \quad \nabla V = \nabla V(\mathbf{x}) = \left[\frac{\partial V}{\partial x_1} \quad \dots \quad \frac{\partial V}{\partial x_n}\right]^T \text{ defined and continuous on some open set } \mathcal{S}. \tag{18.35}$$

Recall that an open ball about $\bar{\mathbf{x}}$ is defined by $B \triangleq B_R(\bar{\mathbf{x}}) \triangleq \{\mathbf{x} : ||\mathbf{x} - \bar{\mathbf{x}}|| < R\}$, where $||\mathbf{x}|| = \left(x_1^2 + \cdots + x_n^2\right)^{1/2}$ is the usual Euclidean norm. A set $\mathcal{S}$ is **open** if for every $\bar{\mathbf{x}}$ in S, there is some $R > 0$, possibly dependent on $\bar{\mathbf{x}}$, such that $B_R(\bar{\mathbf{x}})$ is contained in S.

Define also the **punctured ball** $\widehat{B} = \widehat{B}_R(\bar{\mathbf{x}}) \triangleq \{\mathbf{x} : 0 < ||\mathbf{x} - \bar{\mathbf{x}}|| < R\}$.

Definition 18.10

Suppose $V(\mathbf{x})$ satisfies basic requirements (1) and (2). We say $V(\mathbf{x})$ is

(a) **Positive definite** if $V(\mathbf{x}) > 0$ for all $\mathbf{x}$ in the punctured ball $\widehat{B}_R(\mathbf{0})$
(b) **Positive semi-definite** if $V(\mathbf{x}) \geq 0$ for all $\mathbf{x}$ in the ball $B_R(\mathbf{0})$
(c) **Negative semi-definite** if $V(\mathbf{x}) \leq 0$ for all $\mathbf{x}$ in $B_R(\mathbf{0})$
(d) **Negative definite** $V(\mathbf{x}) < 0$ for all $\mathbf{x}$ in $\widehat{B}_R(\mathbf{0})$

For future reference and to connect the general concept to what we have already seen, note that if B is a real $n \times n$ matrix then

$$V(\mathbf{x}) \triangleq \mathbf{x}^T B\mathbf{x} \tag{18.36}$$

is an example of a function that satisfies our basic requirements (1) and (2). A function $V(\mathbf{x})$ given by (18.36) is called a **quadratic form**.

Usually the easiest kind of positive definite function to work with is $V(\mathbf{x}) \triangleq \mathbf{x}^T B\mathbf{x}$, where B is a positive definite function. And the simplest such choice is $V(\mathbf{x}) \triangleq \mathbf{x}^T\mathbf{x} = x_1^2 + \cdots + x_n^2$, that is, the square of the distance from the equilibrium point $\mathbf{0}$.

Example 18.17

(a) $E(y,v) \triangleq \frac{1}{2}\left(ky^2 + mv^2\right)$ is positive definite on any punctured ball $\widehat{B}_R(\mathbf{0})$.

(b) $\dot{E}(y,v) \triangleq -bv^2$ is negative semi-definite on any ball $B_R(\mathbf{0})$.

(c) $V(x_1,x_2) \triangleq x_1^2 + (x_1 - x_2)^2$ is positive definite on any punctured ball $\widehat{B}_R(\mathbf{0})$.

Method: Only (c) needs explanation as to why it is positive definite rather than merely positive semi-definite: If $V(x_1,x_2)$ is not positive at some point (x_1,x_2), then $0 = x_1^2 + (x_1 - x_2)^2$ would imply that both $0 = x_1$ and $0 = x_1 - x_2$; hence, $x_2 = x_1 = 0$ also. So, $V(\mathbf{x}) > 0$ at every $\mathbf{x} \neq \mathbf{0}$, that is, V is positive definite on any punctured ball $\widehat{B}_R(\mathbf{0})$. ○

If $V(\mathbf{x})$ is positive definite on some ball punctured $\widehat{B}_R(\mathbf{0})$, then the level sets may not be as nice as the ellipses we saw in Figure 18.11, but we can still use the values of $V(\mathbf{x})$ to get enough information about the closeness of $\mathbf{x}$ to $\mathbf{0}$ so as to be useful for studying stability of an equilibrium point of a system of differential equations.

Suppose we have an *autonomous* system of differential equations (18.24) in Section 18.3, that is,

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}),$$

and

(18.24) in Section 18.3 has existence and uniqueness of solutions of IVPs and equilibrium point $\mathbf{0}$ (18.37)

and suppose $V = V(\mathbf{x})$ satisfies the basic requirements (1) and (2). By the chain rule,

$$\frac{d}{dt}\left[\, V\big(\mathbf{x}(t)\big)\,\right] = \frac{\partial V}{\partial x_1}\big(\mathbf{x}(t)\big) \cdot \frac{dx_1}{dt}(t) + \cdots + \frac{\partial V}{\partial x_n}\big(\mathbf{x}(t)\big) \cdot \frac{dx_n}{dt}(t) = \nabla V\big(\mathbf{x}(t)\big) \bullet \dot{\mathbf{x}}(t). \tag{18.38}$$

This suggests defining a new function on $\mathbb{R}^n$ by

$$\dot{V}_{\mathbf{f}}(\mathbf{x}) \triangleq \mathbf{f}\big(\mathbf{x}\big) \bullet \nabla V(\mathbf{x}). \tag{18.39}$$

So, (18.38) states that

$$\frac{d}{dt}\left[\, V\big(\mathbf{x}(t)\big)\,\right] = \dot{V}_{\mathbf{f}}\big(\mathbf{x}(t)\big). \tag{18.40}$$

Example 18.18

Find $\dot{V}_{\mathbf{f}}(\mathbf{x})$ if $V(x,y) \triangleq 2x^2 + y^2$ and $\mathbf{x}(t) \triangleq [x(t) \quad y(t)]^T$ solves the system of differential equations

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -x + y^2 \\ -2y - 2xy \end{bmatrix}. \tag{18.41}$$

Method: $\dot{V}_{\mathbf{f}}(\mathbf{x}) = \begin{bmatrix} -x + y^2 \\ -2y - 2xy \end{bmatrix} \bullet \begin{bmatrix} 4x \\ 2y \end{bmatrix} = (-x + y^2) \cdot 4x + (-2y - 2xy) \cdot 2y = -4x^2 - 4y^2$. ○

Theorem 18.7

(Liapunov theory) Suppose $V(\mathbf{x})$ is positive definite on a punctured ball $\widehat{B}_R(\mathbf{0})$.

(a) If, in addition, $\dot{V}_{\mathbf{f}}$ is negative semi-definite on the corresponding ball $B_R(\mathbf{0})$, then $\mathbf{0}$ is stable for (18.24) in Section 18.3.
(b) If, in addition, $\dot{V}_{\mathbf{f}}$ is negative definite on $\widehat{B}_R(\mathbf{0})$, then $\mathbf{0}$ is asymptotically stable for (18.24) in Section 18.3.

Intuitively, if a function V satisfies the properties that it (a) has minimum value of $V(\bar{\mathbf{x}}) = 0$ at an equilibrium point $\bar{\mathbf{x}}$, (b) takes on only positive values near the equilibrium point, and (c) has $V\big(\bar{\mathbf{x}}(t)\big)$ decreasing for all $t > 0$, then we expect that solutions tend to $\bar{\mathbf{x}}$ as $t \to \infty$.

The complete mathematical explanation for Theorem 18.7 requires some care because $V\big(\mathbf{x}(t)\big)$ does not directly measure distance from the origin. We know $V(\mathbf{x})$ is continuous on $B_R(\mathbf{0})$ because $V(\mathbf{x})$ satisfies (2). What we *can* use is the fact that $V(\mathbf{x})$ being positive definite on $\widehat{B}_R(\mathbf{0})$ implies that for any ε with $0 < \varepsilon < R$,

$$m_\varepsilon \triangleq \min\{V(\mathbf{x}) : \varepsilon \leq ||\,\mathbf{x}\,|| \leq R\}$$

is positive, because Theorem 13.9 in Section 13.2 implies that $V(\mathbf{x})$ attains its global minimum on the annulus $\{\mathbf{x} : \varepsilon \leq ||\,\mathbf{x}\,|| \leq R\}$. Also, for any δ with $0 < \delta < \varepsilon$,

$$M_\delta \triangleq \max\{V(\mathbf{x}) : 0 \leq ||\,\mathbf{x}\,|| \leq \delta\} < \infty,$$

again by the continuity of $V(\mathbf{x})$.

Example 18.19

Study the stability of $\mathbf{0}$ for the system of ODEs (18.41).

Method: The results of Example 18.18 say that $V(x,y) \triangleq 2x^2 + y^2$ has $\dot{V}_{\mathbf{f}}(\mathbf{x}) = -4x^2 - 4y^2$. It is obvious that V is positive definite on any punctured ball $\widehat{B}_R(\mathbf{0})$ and $\dot{V}_{\mathbf{f}}$ is negative definite on $\widehat{B}_R(\mathbf{0})$, so $\mathbf{0}$ is asymptotically stable for the system of ODEs (18.41). ◯

Example 18.20

Study the stability of $\mathbf{0}$ for the system of ODEs (18.34).

Method: Using the physically motivated energy function $E = E(y,v) \triangleq \frac{1}{2}\left(ky^2 + mv^2\right)$ has $\dot{E}_{\mathbf{f}} = -bv^2$. Assuming $k, m, b > 0$ on physical grounds, it is obvious that E is positive definite on any punctured ball $\widehat{B}_R(\mathbf{0})$ and $\dot{E}_{\mathbf{f}}$ is negative semi-definite on $\widehat{B}_R(\mathbf{0})$, so $\mathbf{0}$ is stable for the system of ODEs (18.34). ◯

The results of Example 18.20 are unsatisfying because we *know* that $(y,v) = (0,0)$ is asymptotically stable for the linear, damped harmonic oscillator system (18.34), because the LCCHS has both eigenvalues having negative real part. (See Theorem 5.11 in Section 5.3). So, Theorem 18.7 gives us a correct but fairly useless conclusion about system (18.34).

We will see in Section 18.5 that there are better results for the linear, damped harmonic oscillator system (18.34) using the "LaSalle invariance principle."

Definition 18.11

Suppose $V(\mathbf{x})$ satisfies basic requirement (2) of (18.35), that is, differentiability, on some open set $\mathcal{G}$ and $V(\mathbf{x})$ has a global minimum value on $\mathcal{G}$. Suppose an autonomous system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ has existence and uniqueness of solutions for initial values in $\mathcal{G}$. If $\dot{V}_{\mathbf{f}}$ is negative semi-definite on $\mathcal{G}$, then we say $V(\mathbf{x})$ is a **Liapunov function** on $\mathcal{G}$ for the system of ODEs $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$.

Example 18.21

Use the function $V(x, y) = \frac{9}{2}x^2 + \frac{1}{2}y^2$ to get better conclusions about stability of the origin than we can get using linearization for the system

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} y - 2x(x^2 + y^2) \\ -9x - 2y(x^2 + y^2) \end{bmatrix}.$$

Method: $V(x, y) = \frac{9}{2}x^2 + \frac{1}{2}y^2$ is positive definite on any punctured ball $\widehat{B}_R(\mathbf{0})$. We calculate

$$\dot{V}_{\mathbf{f}}(x, y) = 9x\dot{x} + y\dot{y} = 9x\big(y - 2x(x^2 + y^2)\big) + y\big(-9x - 2y(x^2 + y^2)\big)$$

$$= -2\big(9x^2 + y^2\big)\big(x^2 + y^2\big).$$

Because $\dot{V}_{\mathbf{f}}(x, y)$ is negative definite on $\widehat{B}_R(\mathbf{0})$, the origin is asymptotically stable.

On the other hand, the linearization about the equilibrium point $(0, 0)$ is $\dot{\mathbf{x}} = A\mathbf{x}$, where

$$A = \begin{bmatrix} \frac{\partial}{\partial x}\big[\, y - 2x(x^2 + y^2)\,\big] & \frac{\partial}{\partial y}\big[\, y - 2x(x^2 + y^2)\,\big] \\ \frac{\partial}{\partial x}\big[\, -9x - 2y(x^2 + y^2)\,\big] & \frac{\partial}{\partial y}\big[\, -9x - 2y(x^2 + y^2)\,\big] \end{bmatrix}\Bigg|_{(x,y)=(0,0)}$$

$$= \begin{bmatrix} -2(x^2 + y^2) - 4x^2 & 1 - 4xy \\ -9 - 4xy & -2(x^2 + y^2) - 4y^2 \end{bmatrix}\Bigg|_{(x,y)=(0,0)} = \begin{bmatrix} 0 & 1 \\ -9 & 0 \end{bmatrix}.$$

Because the eigenvalues of A are $\pm i3$, the linearization is neutrally stable, and thus, Theorem 18.4 in Section 18.2 does not give a conclusion about the stability of the original, nonlinear system of ODEs. ◯

18.4.2 Liapunov Functions and Quadratic Forms

Suppose $V(\mathbf{x})$ is a quadratic form, that is, $V(\mathbf{x}) = \mathbf{x}^T B\mathbf{x}$, and the system of ODEs has the form

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) \triangleq A\mathbf{x} + \mathbf{g}(\mathbf{x}), \tag{18.42}$$

where A is a real, constant matrix. In Problem 6.7.6.33, you explained why

$$\nabla V(\mathbf{x}) = \nabla \mathbf{x}^T B\mathbf{x} = \big(B + B^T\big)\mathbf{x}. \tag{18.43}$$

We calculate

$$\dot{V}_{\mathbf{f}}(\mathbf{x}) \triangleq \mathbf{f}(\mathbf{x}) \bullet \nabla V(\mathbf{x}) = \big(A\mathbf{x} + \mathbf{g}(\mathbf{x})\big) \bullet \big(B + B^T\big)\mathbf{x} = \mathbf{x}^T A^T \big(B + B^T\big)\mathbf{x} + \big(\mathbf{g}(\mathbf{x})\big)^T \big(B + B^T\big)\mathbf{x}. \tag{18.44}$$

It is common in this subject to rewrite

$$\mathbf{x}^T A^T \big(B + B^T\big)\mathbf{x} = \mathbf{x}^T A^T B\mathbf{x} + \big(A\mathbf{x}\big)^T \big(B^T\mathbf{x}\big) = \mathbf{x}^T A^T B\mathbf{x} + \big(B^T\mathbf{x}\big)^T \big(A\mathbf{x}\big) = \mathbf{x}^T A^T B\mathbf{x} + \mathbf{x}^T BA\mathbf{x},$$

so we can rewrite (18.44) as

$$\dot{V}_{\mathbf{f}}(\mathbf{x}) = \mathbf{x}^T \big(A^T B + BA\big)\mathbf{x} + \big(\mathbf{g}(\mathbf{x})\big)^T \big(B + B^T\big)\mathbf{x} = \mathbf{x}^T C\mathbf{x} + \big(\mathbf{g}(\mathbf{x})\big)^T \big(B + B^T\big)\mathbf{x}, \tag{18.45}$$

where

$$C \triangleq A^T B + BA. \tag{18.46}$$

Now, suppose that $\mathbf{g}(\mathbf{x})$ is $o(||\,\mathbf{x}\,||)$ as $||\,\mathbf{x}\,|| \to 0$, that is, that

$$\lim_{||\,\mathbf{x}\,|| \to 0} \frac{||\mathbf{g}(\mathbf{x})||}{||\,\mathbf{x}\,||} = 0. \tag{18.47}$$

If, in addition, C is negative definite, then for $||\,\mathbf{x}\,||$ sufficiently small, $||\big(\mathbf{g}(\mathbf{x})\big)^T \big(B + B^T\big)\mathbf{x}|| \to 0$ faster than does $\mathbf{x}^T C\mathbf{x}$, as $||\,\mathbf{x}\,|| \to 0$. It follows that if the ball $B_R(\mathbf{0})$ is sufficiently small, that is, if R is sufficiently small, then $\dot{V}_{\mathbf{f}}(\mathbf{x})$ will be negative definite on $\widehat{B}_R(\mathbf{0})$. This explains our next result.

Theorem 18.8

Suppose $\mathbf{g}(\mathbf{x})$ is $o(||\,\mathbf{x}\,||)$ as $||\,\mathbf{x}\,|| \to 0$ and there exists a positive definite constant matrix B for which $A^T B + BA$ is negative definite. Then $\mathbf{0}$ is asymptotically stable for the system of ODEs (18.42).

By the way, the matrix equation (18.46), where A and C are given and B is to be solved for, is called a **Liapunov equation**.

There is a "converse" of Theorem 18.8:

Theorem 18.9

Suppose $\mathbf{0}$ is asymptotically stable for the LCCHS $\dot{\mathbf{x}} = A\mathbf{x}$. Then there exists a real, constant matrix B for which the quadratic form $V(\mathbf{x}) \triangleq \mathbf{x}^T B\mathbf{x}$ satisfies the hypotheses of Theorem 18.7(a), that is, B is positive definite and $\dot{V}_{\mathbf{f}}(\mathbf{x})$ is negative definite.

Why? Let

$$B \triangleq \int_0^\infty e^{tA^T} e^{tA}\, dt. \tag{18.48}$$

In Problem 18.4.5.14, you will check that for B given in (18.48), $V(\mathbf{x}) \triangleq \mathbf{x}^T B\mathbf{x}$ satisfies the hypotheses of Theorem 18.7(a). □

The matrix B is reminiscent of the matrix M defined in (5.74) in Section 5.6 that we used in studying a control theory problem. Actually, in the history of systems of differential equations, the solution B of the Liapunov equation (18.46) preceded the matrix M used in control theory.

18.4.3 Instability

Definition 18.12

Suppose $V(\mathbf{x})$ satisfies our basic requirements (1), that is, $V(\mathbf{0}) = 0$, and (2), that is, continuous differentiability on some open set $\mathcal{S}$. We say a function $V(\mathbf{x})$ is **indefinite** on $\mathcal{S}$ if, in addition, for *every* ball B_R (contained in S), there are $\mathbf{x}_\pm$ such that $V(\mathbf{x}_+) > 0 > V(\mathbf{x}_-)$.

Theorem 18.10

(a) Suppose $V(\mathbf{x})$ is indefinite on $\mathcal{S}$ and either $\dot{V}_\mathbf{f}(\mathbf{x})$ is positive definite or is negative definite on some punctured ball $\widehat{B}_R(\mathbf{0}) \subseteq \mathcal{S}$. Then $\mathbf{0}$ is unstable for the system of ODEs (18.42).

(b) Suppose both $V(\mathbf{x})$ and $\dot{V}_\mathbf{f}(\mathbf{x})$ are positive definite on some punctured ball. Then $\mathbf{0}$ is unstable for the system of ODEs (18.42).

Example 18.22

The undamped nonlinear pendulum of length ℓ can be modeled* by the system

$$\begin{bmatrix} \dot{\theta} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{\ell}\, v \\ -g \sin\theta \end{bmatrix}.$$

near the inverted position, that is, near the equilibrium point $(\theta, v) = (\pi, 0)$. Use a Liapunov function to establish instability of that equilibrium point.

Method: On physical grounds, the energy in the system is $E(\theta, v) = mg\ell(1 - \cos\theta) + \frac{1}{2}\, mv^2$. It is easy to calculate that $\dot{E}_\mathbf{f}(\theta, v) \equiv 0$, that is, there is conservation of energy in this frictionless model. So, we cannot use this function $E(\theta, v)$ in Instability Theorem 18.10.

Because the equilibrium point is not $\mathbf{0} = (0, 0)$, it may help to make a change of variables: Let $\xi = \theta - \pi$, and note that $\sin\theta = \sin(\xi + \pi) = -\sin\xi$, so the system becomes

* See Problem 7.2.5.21 for the calculation of the potential energy.

$$\begin{bmatrix}\dot{\xi}\\ \dot{v}\end{bmatrix} = \begin{bmatrix}\dfrac{1}{\ell}\,v\\ g\sin\xi\end{bmatrix},$$

and we will study the stability of the equilibrium point $(\xi, v) = (0, 0)$.

Noting that $\cos\theta = \cos(\xi + \pi) = -\cos\xi$, the energy in the system is $\widetilde{V}(\xi, v) \triangleq mg\ell(1 + \cos\xi) + \frac{1}{2}\,mv^2$.

The fact that $\widetilde{V}(0, 0) = 2mg\ell \neq 0$, that is, $(\tilde{1})$ is not true, is not a major issue: Instead of working with $\widetilde{V}$, we can subtract from it $2mg\ell$.

But in order to use Instability Theorem 18.10(a) we would need a function V with either $\dot{V}_{\mathbf{f}}$ being positive definite or being negative definite. We will see that the artful addition of an indefinite term to $E(\theta, v)$ can also make $\dot{V}_{\mathbf{f}}(\theta, v)$ be definite. Define

$$V(\xi, v) \triangleq \widetilde{V}(\xi, v) - 2mg\ell + mg\ell v\sin\xi = mg\ell(-1 + \cos\xi) + \frac{1}{2}\,mv^2 + mg\ell v\sin\xi.$$

Then, using $\dot{\xi} = v/\ell$ and $\dot{v} = g\sin\xi$,

$$\begin{aligned}\dot{V}_{\mathbf{f}}(\xi, v) &= -mg\ell\sin\xi(\dot{\xi}) + mv\dot{v} + mg\ell\dot{v}\sin\xi + mg\ell v\cos\xi(\dot{\xi})\\ &= -\cancel{mg(\sin\xi)v} + \cancel{mvg\sin\xi} + mg^2\ell\sin^2\xi + mgv^2\cos\xi,\end{aligned}$$

that is,

$$\dot{V}_{\mathbf{f}}(\xi, v) = mg\big(g\ell\sin^2\xi + v^2\cos\xi\big);$$

hence, $\dot{V}_{\mathbf{f}}(\xi, v) > 0$ for all (ξ, v) in the punctured ball $\widehat{B} = \{(\xi, v) : 0 < \sqrt{\xi^2 + v^2} < \frac{\pi}{2}\}$. So, $\dot{V}_{\mathbf{f}}(\xi, v)$ is positive definite on $\widehat{B}$.

Second, why is $V(\xi, v)$ indefinite with respect to $(\xi, v) = (0, 0)$? Taylor approximations give $-1 + \cos\xi \approx -\frac{1}{2}\,\xi^2$ and $\sin\xi \approx \xi$, so

$$V(\xi, v) = mg\ell(-1 + \cos\xi) + \frac{1}{2}\,mv^2 + mg\ell v\sin\xi \approx -\frac{1}{2}\,mg\ell\xi^2 + \frac{1}{2}\,mv^2 + mg\ell v\xi \triangleq W(\xi, v).$$

These approximations get better and better as $(\xi, v) \to (0, 0)$.

By completing the square, we have

$$\begin{aligned}W(\xi, v) &= -\frac{1}{2}\,mg\ell\xi^2 + mg\ell v\xi + \frac{1}{2}\,mv^2 = -\frac{1}{2}\,mg\ell\Big(\xi^2 - 2v\xi\Big) + \frac{1}{2}\,mv^2\\ &= -\frac{1}{2}\,mg\ell(\xi - v)^2 + \frac{1}{2}\,m(g\ell + 1)v^2.\end{aligned}$$

On the ray $\xi = v$, $W(\xi, v) = \frac{1}{2}\,m(g\ell + 1)v^2$, so we conclude that V takes on some positive values no matter how small is the ball $\widehat{B}_\delta(\bar{\mathbf{0}})$, that is, no matter how small $\delta > 0$ is chosen.

On the other hand, by completing the square another way, we have

$$\begin{aligned}W(\xi, v) &= \frac{1}{2}\,mv^2 + mg\ell v\xi - \frac{1}{2}\,mg\ell\xi^2 = \frac{1}{2}\,m\Big(v^2 + 2\ell v\xi\Big) - \frac{1}{2}\,mg\ell\xi^2\\ &= \frac{1}{2}\,m(v + g\ell\xi)^2 - \frac{1}{2}\,mg\ell(g\ell + 1)\xi^2.\end{aligned}$$

On the ray $v = -g\ell\xi$, $W(\xi, v) = -\frac{1}{2}\,mg\ell(g\ell + 1)\xi^2$, so we conclude that V takes on some negative values on $\widehat{B}_\delta(\bar{\mathbf{0}})$ no matter how small $\delta > 0$ is chosen.

So, we conclude that $V(\xi, v)$ is indefinite with respect to $(\xi, v) = (0, 0)$; hence, the inverted undamped nonlinear pendulum is unstable. ○

Certainly the details of showing instability of the inverted undamped nonlinear pendulum in Example 18.22 were more difficult than in Example 18.10 in Section 18.2. But it is good to see a physical example of the application of instability theorem, for Liapunov functions.

Note that to use instability theorem, it would have worked as well to let $V(\theta, v) \triangleq E(\theta, v) - 2mg\ell - mg\ell v \sin\theta$.

18.4.4 Stability of Another Equilibrium Point

When studying the stability of an equilibrium point $\bar{\mathbf{x}}$, we could make the change of variables $\mathbf{y} = \mathbf{x} - \bar{\mathbf{x}}$, that is, $\mathbf{x} = \mathbf{y} + \bar{\mathbf{x}}$. Then the stability property(s) of $\bar{\mathbf{x}}$ for $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ is the same as the stability property(s) of $\mathbf{y} = \mathbf{0}$ for $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y} + \bar{\mathbf{x}})$.

Alternatively, we can define definiteness properties for functions on a ball or on an open set around a point $\bar{\mathbf{x}}$, and then the Liapunov theorems apply in this generalization.

We will always assume that

$$(\tilde{1}) \qquad V(\bar{\mathbf{x}}) = 0$$

and

$$(\tilde{2}) \quad \nabla V = \nabla V(\mathbf{x}) = \left[\frac{\partial V}{\partial x_1} \quad \cdots \quad \frac{\partial V}{\partial x_n}\right]^T \text{ defined and continuous on some open set } \mathcal{S}.$$

Definition 18.13

Suppose $V(\mathbf{x})$ satisfies basic requirements $(\tilde{1})$ and $(\tilde{2})$. We say $V(\mathbf{x})$ is

(a) **Positive definite with respect to** $\bar{\mathbf{x}}$ if $V(\mathbf{x}) > 0$ for all $\mathbf{x}$ in a punctured ball $\widehat{B}_R(\bar{\mathbf{x}})$
(b) **Positive semi-definite with respect to** $\bar{\mathbf{x}}$ if $V(\mathbf{x}) \geq 0$ for all $\mathbf{x}$ in $B_R(\bar{\mathbf{x}})$
(c) **Negative semi-definite with respect to** $\bar{\mathbf{x}}$ if $V(\mathbf{x}) \leq 0$ for all $\mathbf{x}$ in $B_R(\bar{\mathbf{x}})$
(d) **Negative definite with respect to** $\bar{\mathbf{x}}$ if $V(\mathbf{x}) < 0$ for all $\mathbf{x}$ in $\widehat{B}_R(\bar{\mathbf{x}})$
(e) **Indefinite with respect to** $\bar{\mathbf{x}}$ if for *every* ball $B_R(\bar{\mathbf{x}})$ (contained in S), there are $\mathbf{x}_\pm$ such that $V(\mathbf{x}_+) > 0 > V(\mathbf{x}_-)$

It is also possible to define these concepts using an open set $\mathcal{S}$ containing $\bar{\mathbf{x}}$ instead of a punctured ball, for example, $V(\mathbf{x})$ is **positive definite with respect to** $\bar{\mathbf{x}}$ if $V(\mathbf{x}) > 0$ for all $\mathbf{x} \neq \bar{\mathbf{x}}$ in $\mathcal{S}$.

Here is one such result.

Theorem 18.11

Suppose $V(\mathbf{x})$ is positive definite with respect to $\bar{\mathbf{x}}$ and $\dot{V}_{\mathbf{f}}(\mathbf{x})$ is negative definite with respect to $\bar{\mathbf{x}}$. Then $\bar{\mathbf{x}}$ is asymptotically stable for $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$.

18.4.5 Problems

1. Examine each of the functions to see if it is indefinite, or positive definite or semi-definite, or negative definite or semi-definite, or none of the above, with respect to the given point. Use the definitions to explain your conclusions
   (a) $V(x,y) = x^2 + 3y^2$ with respect to $(0,0)$
   (b) $V(x,y) = \sin^2 x + y^2$ with respect to $(\pi, 0)$
   (c) $V(x,y) = x^3 + (y-1)^6$ with respect to $(0,1)$

In problems 2–6, use $V(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{x}$ to study the stability of the origin for the given system.

2. $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -x - \frac{1}{3}x^3 - x\cos y \\ -y - y^3 \end{bmatrix}$.
3. $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -y - x\sin^2 x \\ x - y\sin^2 x \end{bmatrix}$.
4. $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} x - y^2 \\ y + xy \end{bmatrix}$.
5. $\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} -x_1 - 2x_2^2 \\ -x_2 + 2x_1x_2 \\ -3x_3 + x_4 \\ -x_3 - 2x_4 \end{bmatrix}$.
6. $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} xp(x,y) \\ -yq(x,y) \end{bmatrix}$,
   where we assume that $p(x,y)$ is negative definite and $q(x,y)$ is positive definite on some ball centered at the origin $(x,y) = (0,0)$.

In problems 7 and 8, use a function of the form $V(x,y) = Ax^2 + By^4$, where A, B are constants, to study the stability of the origin for the given system.

7. $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -x^3 - 3xy^4 \\ x^2y - 2y^3 - y^5 \end{bmatrix}$.
8. $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -x^3 - 3xy^4 \\ 2x^2y - 5y^3 \end{bmatrix}$.
9. Use a function of the form $V(x,y) = Ax^2 + By^2$, where A, B are constants, to study the stability of the origin for the system $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} x^3y^2 - 3x^3 \\ -5x^4y - y \end{bmatrix}$.
10. Use a function of the form $V(x,y) = Ax^2 + Bx^4 + Cy^2$, where A, B, C are constants, to study the stability of the origin for the system $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -x^3 + xy^2 \\ -y - 3x^2y - 5x^4y \end{bmatrix}$.
11. Use the function $V(x,y) = \frac{9}{2}y^2 + \frac{1}{2}v^2$ to get better conclusions about stability of the origin than we can get using linearization for the system
    $\begin{bmatrix} \dot{y} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} v + y(y^2 + v^2) \\ -9y + v(y^2 + v^2) \end{bmatrix}$.
12. Use the function $V(x,y) = x^4 + y^2$ to help conclude that the system

$$\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix} = \begin{bmatrix}y - x^3\\ -2x^3\end{bmatrix}$$

has **0** as a stable equilibrium point even though the linearization about **0** is unstable.

13. Use the function $V(x, y) = x^2 - y^2$ to help conclude that **0** is an unstable equilibrium point for the system
$$\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix} = \begin{bmatrix}-x + y^3\\ x^2 + y\end{bmatrix}.$$
14. Check that for B given in (18.48), $V(\mathbf{x}) \triangleq \mathbf{x}^T B\mathbf{x}$ satisfies the hypotheses of Theorem 18.7(a).

In problems 15–17, use Theorem 18.9 to find a Liapunov function for

15. The system of Problem 5.3.6.11, that is,
$$\dot{\mathbf{x}} = \begin{bmatrix}0 & 1\\ -1 & -2\end{bmatrix}\mathbf{x}.$$
16. The system of Example 5.13 in Section 5.2, that is,
$$\dot{\mathbf{x}} = \begin{bmatrix}0 & 1\\ -6 & -5\end{bmatrix}\mathbf{x}.$$
17. $\dot{\mathbf{x}} = \begin{bmatrix}-3 & 1\\ 1 & -3\end{bmatrix}\mathbf{x}.$
18. (Small project) Study the stability of the origin for the system ($\star$) $\begin{bmatrix}\dot{x}\\ \dot{v}\end{bmatrix} = \begin{bmatrix}v\\ -f(x)\end{bmatrix}$, using the function $H(x, v) \triangleq \frac{1}{2}v^2 + F(x)$, where $F(x) \triangleq \int_0^x f(\xi)\, d\xi$.
    a. Give a condition(s) on the function $f(x)$ that implies that the origin is stable for ($\star$).
    b. Give a condition(s) on the function $f(x)$ that implies that the origin is unstable for ($\star$).

    Note that for a scalar problem $\ddot{x} + f(x) = 0$, $F(x)$ has a natural interpretation as the potential energy and $H(x, v)$ has a natural interpretation as the sum of the kinetic and potential energies.
19. Assume α is an unspecified constant. Use the function $V(x, y) = x^2 + y^2$ to study the stability of the origin for the system of ODEs
$\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix} = \begin{bmatrix}3x + \alpha x(x^2 + y^2)\\ -3y + \alpha y(x^2 + y^2)\end{bmatrix}$. [Hint: Consider separately the cases $\alpha < 0$, $\alpha = 0$, and $\alpha > 0$.]
20. For the "Hamiltonian system" in $\mathbb{R}^{2n}$
$$\left\{\begin{array}{l}\dot{\mathbf{q}} = \frac{\partial H}{\partial \mathbf{p}}\\ \dot{\mathbf{p}} = -\frac{\partial H}{\partial \mathbf{q}}\end{array}\right\},$$
suppose $H(\mathbf{q}, \mathbf{p})$ is positive definite on the punctured ball $\widehat{B}_\delta(\mathbf{0}, \mathbf{0})$. Explain why $(\mathbf{p}, \mathbf{q}) = (\mathbf{0}, \mathbf{0})$ is stable by showing that $H(\mathbf{p}, \mathbf{q})$ is a Liapunov function.
21. Consider the two mass and three horizontal springs system of Example 5.5 in Section 5.1. Rewrite it in the form of a Hamiltonian system, as in Problem 18.4.5.20,

using momentum coordinates $p_1 = m_1\dot{x}_1$ and $p_2 = m_2\dot{x}_2$. Use the result of Problem 18.4.5.20 to discuss the stability of $(\mathbf{q}, \mathbf{p}) = (\mathbf{0}, \mathbf{0})$ in $\mathbb{R}^4$.

22. Suppose $V(0,0) = 0$, $V(x,y)$ is positive definite on the punctured ball $\widehat{B}_\delta(0,0)$ for some $\delta > 0$, and $\nabla V(0,0) = \mathbf{0}$. For the **gradient system** of ODEs

$$\left\{\begin{array}{l} \dot{x} = -\frac{\partial V}{\partial x} \\ \dot{y} = -\frac{\partial V}{\partial y} \end{array}\right\},$$

(a) Explain why $(0,0)$ is stable, and (b) discuss whether $(0,0)$ can fail to be asymptotically stable.

18.5 Short Take: LaSalle Invariance Principle

Recall that our results for Example 18.20 in Section 18.4 were unsatisfying: We knew that Liapunov theory did not give as good a conclusion as we could get from using eigenvalues for that linear system of ODEs.

Suppose we have an *autonomous* system of differential equations (18.24) in Section 18.3, that is, $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$.

Definition 18.14

A set $\mathcal{S}$ is **positively invariant** for (18.24) in Section 18.3 if for every initial value $\mathbf{x}_0$ in S, then the solution $\mathbf{x}(t; \mathbf{x}_0)$ is in $\mathcal{S}$ forward in time, that is, for all $t \geq 0$.

Recall that the omission of t_0 from the solution notation $\mathbf{x}(t; \mathbf{x}_0)$ implies that the initial time is $t_0 = 0$. Recall from Definition 2.32 in Section 2.10 that a set $\mathcal{S}$ is **closed** if every sequence $\{\mathbf{x}_k\}_{k=1}^{\infty}$ in $\mathcal{S}$ that is convergent has its limit in $\mathcal{S}$.

Definition 18.15

The **closure** of a set $\mathcal{G}$, denoted by $\overline{\mathcal{G}}$, consists of all points in $\mathcal{G}$ *and* all limits of sequences in $\mathcal{G}$, that is,

$$\overline{\mathcal{G}} = \{\mathbf{x} : \mathbf{x} = \lim_{k\to\infty} \mathbf{g}_k \text{ for some sequence } \{\mathbf{g}_k\}_{k=0}^{\infty} \text{ in } \mathcal{G}\}.$$

Note that every $\mathbf{g}$ in $\mathcal{G}$ is the limit of the constant sequence $\{\mathbf{g}\}_{k=0}^{\infty}$ in $\mathcal{G}$.

Theorem 18.12

A set is closed if, and only if, it equals its closure.

Definition 18.16

(a) A set $\mathcal{F}$ is a closed, positively invariant subset of $\mathcal{S}$ if

$$\mathcal{F} \subseteq \mathcal{S}, \quad \mathcal{F} \text{ is closed}, \quad \text{and} \quad \mathcal{F} \text{ is positively invariant.}$$

(b) $\mathcal{E}$ is a **maximal, closed, positively invariant** subset of $\mathcal{S}$ if $\mathcal{E}$ is a closed and positively invariant subset of $\mathcal{S}$ and

$$\mathcal{F} \subseteq \mathcal{E} \text{ whenever } \mathcal{F} \text{ is a closed, positively invariant subset of } \mathcal{S}.$$

Definition 18.17

Given a point $\mathbf{x}$ and a closed set $\mathcal{S}$, $dist(\mathbf{x}, \mathcal{S}) \triangleq \min\{||\mathbf{s} - \mathbf{x}|| : \mathbf{s} \text{ is in } \mathcal{S}\}$.

Theorem 18.13

(LaSalle invariance principle) Suppose $V(\mathbf{x})$ is a Liapunov function for an autonomous system (18.24) in Section 18.3, that is, on some open set $\mathcal{G}$ in $\mathbb{R}^n$, $\dot{V}_{\mathbf{f}}$ is negative semi-definite and $V(\mathbf{x})$ has a global minimum value. Define

$$\mathcal{S} \triangleq \{\, \mathbf{x} \text{ in } \overline{\mathcal{G}} : \ \dot{V}_{\mathbf{f}}(\mathbf{x}) = 0 \,\},$$

and suppose $\mathcal{E}$ is a maximal, closed, positively invariant subset of $\mathcal{S}$. If $\mathbf{x}(t)$ is a bounded solution of ODE system (18.24) in Section 18.3 and $\mathbf{x}(t)$ remains in $\mathcal{G}$ for $t \geq 0$, then

$$\mathbf{x}(t) \to \mathcal{E} \text{ as } t \to \infty. \tag{18.49}$$

One typical application of the LaSalle invariance principle is to show that $\mathcal{E}$ consists of a *finite* number of equilibrium points of the autonomous ODE system. In this case, (18.49) implies that every bounded solution of the autonomous system has to converge to a specific equilibrium point, which may be a very desirable property in a physical system.

Example 18.23

Study the stability of $\mathbf{0}$ for the damped oscillator problem in Example 18.20 in Section 18.4.

Method: Assume $k, m, b > 0$ on physical grounds and let $\mathcal{G} = \mathbb{R}^2$. The function $E = E(y, v) \triangleq \frac{1}{2}\left(ky^2 + mv^2\right)$ has $\dot{E}_{\mathbf{f}} = -bv^2$, so $\mathcal{S} = \{(y, v) : \ v = 0\}$. The only initial condition $\mathbf{x}_0$ in $\mathcal{S}$ that has $\mathbf{x}(t; \mathbf{x}_0)$ remaining in $\mathcal{S}$ for all $t > 0$ is $\mathbf{x}_0 = (y_0, 0) = (0, 0)$, that is, the only motion that has zero velocity, $v(t)$, for all $t \geq 0$ is the solution that starts at $(0, 0)$. So, the *only* closed, positively invariant subset of $\mathcal{S}$ is the set $\{(0, 0)\}$. So, the maximal, closed, positively invariant set in $\mathcal{S}$ is $\mathcal{E} = \{(0, 0)\}$.

So, every solution $\mathbf{x}(t) \to \mathbf{0}$ as $t \to \infty$. This says that $\mathbf{0}$ is a global attractor. We already knew that $\mathbf{0}$ is stable, so $\mathbf{0}$ is asymptotically stable. ◯

18.5.1 Stability of a Set

Definition 18.18

(a) A set $\mathcal{S}$ is **stable** if for all $\varepsilon > 0$, no matter how small, and all initial times t_0, there exists some $\delta > 0$, possibly dependent on ε or t_0, such that for all $t \geq t_0$ and initial values $\mathbf{x}_0$ satisfying $dist(\mathbf{x}_0, \mathcal{S}) < \delta$,

$$\text{dist}(\mathbf{x}(t; t_0, \mathbf{x}_0), \mathcal{S}) < \varepsilon. \tag{18.50}$$

(a) A set $\mathcal{S}$ is **unstable** if it is not stable.

Definition 18.19

A *set* $\mathcal{S}$ is an **attractor** if for all initial times t_0, there exists some $\delta > 0$, possibly dependent on t_0, such that for all initial values $\mathbf{x}_0$ satisfying $dist(\mathbf{x}_0, \mathcal{S}) < \delta$,

$$\lim_{t \to \infty} \text{dist}(\mathbf{x}(t; t_0, \mathbf{x}_0), \mathcal{S}) = 0. \tag{18.51}$$

Note that the points $\mathbf{s}(t)$ in $\mathcal{S}$ that minimize the distance to $\mathbf{x}(t; t_0, \mathbf{x}_0)$, that is, "shadow" $\mathbf{x}(t; t_0, \mathbf{x}_0)$, may, in principle, not behave in any organized way.

Definition 18.20

A set $\mathcal{S}$ is **asymptotically stable** if it is both stable and an attractor.

18.5.2 Problems

In problems 1–3, use the LaSalle invariance principle and the given function $V(x, y)$ to study the stability of the origin for the given system.

1. $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} y \\ -x - y^3 \end{bmatrix}$; $V(x, y) = \frac{1}{2}(x^2 + y^2)$
2. $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} y - x^3 \\ -2x^3 \end{bmatrix}$; $V(x, y) = x^4 + y^2$
3. $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -y - x\sin^2 x \\ x - y\sin^2 x \end{bmatrix}$; $V(x, y) = \frac{1}{2}(x^2 + y^2)$
4. The second-order nonlinear ODE $\ddot{y} + g(y)\dot{y} + y = 0$ is equivalent to the nonlinear system $\begin{bmatrix} \dot{y} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} v \\ -y - g(y)v \end{bmatrix}$. Assume that $g(y)$, an unspecified function on $\mathbb{R}^1$, is positive definite. Use the LaSalle invariance principle and the function $V(y, v) = \frac{1}{2}(y^2 + v^2)$ to study the stability of the equilibrium point $(y, v) = (0, 0)$.
5. What does the LaSalle invariance principle say about the stability of the set consisting of all of the equilibrium points for the system of Example 5.8 in Section 5.2?

6. Study the stability of the equilibria $(\theta, v) =$ (a) $(0, 0)$ and (b) $(\pi, 0)$ for the damped nonlinear pendulum
$$\begin{bmatrix} \dot{\theta} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} v \\ -k \sin\theta - bv \end{bmatrix},$$
where k is an unspecified positive constant.

18.6 Limit Cycles

Suppose a planar *autonomous* ODE system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ has a periodic solution $\mathbf{x}(t)$ with period T. It gives a simple, closed, parametrized curve $\mathcal{C}: \mathbf{x} = \mathbf{x}(t)$, $0 \le t \le \mathrm{T}$, in the phase plane.

As in Chapters 6 and 15, we may abuse the definitions by referring to $\mathcal{C}$ as both (1) a parametrized curve, $\mathcal{C}: \mathbf{x} = \mathbf{x}(t)$, $0 \le t \le \mathrm{T}$, and (2) the set $\{\mathbf{x}(t) : 0 \le t \le \mathrm{T}\}$, that is, the set consisting of all of the points $\mathbf{x}(t)$, $0 \le t \le \mathrm{T}$.

But, by Theorem 18.1 in Section 18.1, any **time translate** $\mathbf{x}(t - t_0)$ will also give a periodic solution of the same planar *autonomous* ODE system and will give the same set of points as does $\mathcal{C}$ in the phase plane.

Definition 18.21

For a planar autonomous system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ on a region $\mathcal{D}$, a periodic solution $\mathbf{x}(t)$ is **unique** in $\mathcal{D}$, **up to time translation**, if the only periodic solutions of $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ that are in $\mathcal{D}$ are the time translates of $\mathbf{x}(t)$.

Definition 18.22

(a) A planar autonomous ODE system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ has a **limit cycle** $\mathcal{C}$ if it is not an equilibrium point, it is the only periodic solution inside some open annulus, up to time translation, and for some initial value in that open annulus the solution $\to$ (the set $\mathcal{C}$) as $t \to \infty$ or as $t \to -\infty$.

(b) A limit cycle is **stable** if inside that open annulus all solutions $\to$ (the set $\mathcal{C}$) as $t \to \infty$.

(c) A limit cycle is **unstable** if inside that open annulus all solutions $\to$ (the set $\mathcal{C}$) as $t \to -\infty$.

Note: Usually, a solution $\mathbf{x}(t)$ does not get closer and closer to a specific point on a stable limit cycle $\mathcal{C}$ but instead get closer and closer to the whole curve, as illustrated in Figure 18.12. Many people call $\mathcal{C}$ the **orbit** of the periodic solution, and thus, what we call stable they may call **orbitally stable**.

In Section 18.2, we learned how to use

$$r(t) \triangleq ||\mathbf{x}(t)|| = \sqrt{\mathbf{x}(t)^T \mathbf{x}(t)}$$

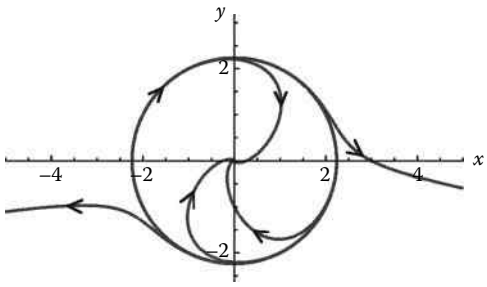


FIGURE 18.12
Example 18.24.

to study stability of the origin for some nonlinear systems of ODEs. Here we will use similar calculations, along with the phase line for 1D ODEs discussed in Section 18.1, to study limit cycles and their stability.

Example 18.24

Study limit cycles of the form $r(t) \equiv constant$ and their stability for the system

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -5x + 2y + x\left(x^2 + y^2\right) \\ -2x - 5y + y\left(x^2 + y^2\right) \end{bmatrix}. \tag{18.52}$$

Method: Because $r^2 = x^2 + y^2$, we calculate

$$r\dot{r} = x\dot{x} + y\dot{y} = x\left(-5x + 2y + x\left(x^2 + y^2\right)\right) + y\left(-2x - 5y + y\left(x^2 + y^2\right)\right) = -5r^2 + r^4,$$

so

$$\dot{r} = -5r + r^3. \tag{18.53}$$

The latter is an autonomous ODE in $\mathbb{R}^1$, so we can use the phase line to study its equilibria and their stability. Only $r = ||\,\mathbf{x}\,|| \geq 0$ is relevant to our problem (18.53). The phase line is shown in Figure 18.13.

The equilibria of (18.52), $r = 0$ and $r = \sqrt{5}$, give periodic solutions, $x(t)^2 + y(t)^2 \equiv 0$ and $x(t)^2 + y(t)^2 \equiv 5$, of (18.52). We see that $x^2 + y^2 \equiv 5$ gives an unstable limit cycle. Note that the origin is an equilibrium point and is thus not a limit cycle.

Note also that (18.53) has no periodic solution other than the equilibrium at the origin and on the circle $r = \sqrt{5}$. Why not? Because if the initial value of a solution has

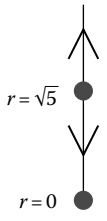


FIGURE 18.13
Example 18.24.

$0 < r(0) \neq 5$, then the phase line in Figure 18.13 shows that at no time $T > 0$ can we have $r(T) = r(0)$, which implies that at no time $T > 0$ can we have $\mathbf{x}(T) = \mathbf{x}(0)$. So, no solution other than on the circle $r = \sqrt{5}$ or at the origin can return to where it started after T units of time, and thus, the only possible limit cycle is on the circle $r = \sqrt{5}$. ○

The phase plane for (18.52) was shown in Figure 18.12.

18.6.1 Periodic Linearization

If the system of ODEs $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t)$ is "periodic," that is, there is a $T > 0$ such that $\mathbf{f}(\mathbf{x}, t + T) \equiv \mathbf{f}(\mathbf{x}, t)$, then the linearization about an equilibrium point $\bar{\mathbf{x}}$ is also periodic. Why? Linear approximation gives

$$\mathbf{f}(\mathbf{x}, t) = \mathbf{f}(\bar{\mathbf{x}}, t) + A(t)(\mathbf{x} - \bar{\mathbf{x}}) + \mathbf{g}(\mathbf{x} - \bar{\mathbf{x}}, t)$$

for $\mathbf{x}$ near $\bar{\mathbf{x}}$, where

$$A(t) \triangleq \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\bar{\mathbf{x}}, t)\right]. \tag{18.54}$$

The **linearization** of $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t)$ near the equilibrium point $\bar{\mathbf{x}}$ is

$$\dot{\mathbf{y}} = A(t)\mathbf{y}. \tag{18.55}$$

Using Theorem 5.25 in Section 5.8, we have

Theorem 18.14

(Stability from linearization)

a. If $\mathbf{y} = \mathbf{0}$ is asymptotically stable for the periodic linearization (18.55), that is, all characteristic multipliers of (18.55) have modulus $|\mu| < 1$, then $\bar{\mathbf{x}}$ is asymptotically stable for $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t)$.
b. If the periodic linearization (18.55) has at least one characteristic multiplier whose modulus is greater than one, then $\bar{\mathbf{x}}$ is unstable for $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t)$.

18.6.2 Linearization about a Periodic Solution

Suppose the system of ODEs $(\star)$ $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t)$ is either autonomous, that is, $\mathbf{f}(\mathbf{x}, t) = \mathbf{f}(\mathbf{x})$, or periodic with period $T > 0$. Suppose that $(\star)$ has a solution $\bar{\mathbf{x}}(t)$ that is also periodic* with period T. We can study the stability of the periodic solution by linearization about it.

Linear approximation gives

$$\mathbf{f}(\mathbf{x}, t) = \mathbf{f}(\bar{\mathbf{x}}(t), t) + A(t)\big(\mathbf{x} - \bar{\mathbf{x}}(t)\big) + \mathbf{g}(\mathbf{x} - \bar{\mathbf{x}}(t), t)$$

* If the system is periodic, we assume it is periodic with the same period as the solution $\mathbf{x}(t)$.

for $\mathbf{x}$ near $\bar{\mathbf{x}}(t)$, where

$$A(t) \triangleq \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\bar{\mathbf{x}}(t), t)\right] \tag{18.56}$$

is a square matrix. The **linearization** of $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t)$ near the periodic solution $\bar{\mathbf{x}}(t)$ is

$$\dot{\mathbf{y}} = A(t)\mathbf{y}. \tag{18.57}$$

In fact, every equilibrium point $\bar{\mathbf{x}}$ is periodic with any period we want, so linearization about an equilibrium point is a particular case of linearization about a periodic solution.

Theorem 18.15

For an autonomous system $\dot{\mathbf{x}} = f(\mathbf{x})$, the periodic linearization (18.57) about a nonconstant periodic solution always has at least one characteristic multiplier equal to one.

Why? Given $\bar{\mathbf{x}}(t)$, a non constant periodic solution of $\dot{\mathbf{x}} = f(\mathbf{x})$, define $\mathbf{y}^{(1)}(t) \triangleq \dot{\bar{\mathbf{x}}}(t)$. Using the chain rule, take the time derivative of $\dot{\bar{\mathbf{x}}} = \mathbf{f}(\mathbf{x}, t)$ to get

$$\frac{d}{dt}[\mathbf{y}^{(1)}(t)] = \ddot{\bar{\mathbf{x}}} = \frac{d}{dt}[\,\mathbf{f}\big(\bar{\mathbf{x}}(t)\big)\,] = \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\big(\bar{\mathbf{x}}(t)\big)\right]\dot{\bar{\mathbf{x}}}(t) = A(t)\mathbf{y}^{(1)}(t). \tag{18.58}$$

We assumed $\bar{\mathbf{x}}(t)$ is not constant, so $\mathbf{y}^{(1)}(t)$ is not identically zero. Letting the first column of a fundamental matrix be $\mathbf{y}^{(1)}(t)$, it follows that there is at least one characteristic multiplier equal to one. □

18.6.3 Levinson–Smith Theorem

For a special class of second-order autonomous ODEs, there is a theorem that established the existence of a stable limit cycle. The **Lienard equation** was motivated by nonlinear ODEs of electronics and has the form

$$\ddot{x} + f(x)\dot{x} + g(x) = 0. \tag{18.59}$$

Define $F(x) = \int_0^x f(\xi)d\xi$ and $G(x) = \int_0^x g(\xi)d\xi$.

Basic assumptions about $f(x)$ and its definite integral $F(x)$ are that

$$\left\{\begin{array}{l} \bullet\ f(x) \text{ is even and continuous on } -\infty < x < \infty, \\ \bullet\ \text{There exists } a > 0 \text{ such that } F(x) < 0 \text{ for } 0 < x < a \text{ and } F(x) > 0 \text{ for } a < x, \\ \bullet\ F(x) \to \infty \text{ as } x \to \infty. \end{array}\right\}. \tag{18.60}$$

A pictorial example of such a function $f(x)$ is given in Figure 18.14a.

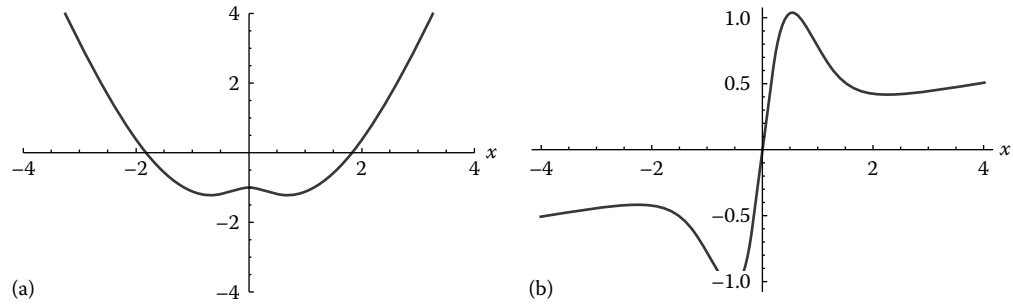


FIGURE 18.14
Pictorial examples of (a) $f(x)$ and (b) $g(x)$.

Basic assumptions about $g(x)$ and its definite integral $G(x)$ are that

$$\left\{\begin{array}{l} \bullet\ g(x) \text{ is odd and continuously differentiable on } -\infty < x < \infty, \\ \bullet\ xg(x) > 0 \text{ for } x > 0 \\ \bullet\ G(x) \to \infty \text{ as } |x| \to \infty. \end{array}\right\}. \tag{18.61}$$

A pictorial example of such a function $g(x)$ is given in Figure 18.14b.

In Problem 18.6.5.6, you will explain why the famous **Van der Pol equation**,

$$\ddot{x} - \eta\left(1 - x^2\right)\dot{x} + x = 0, \tag{18.62}$$

is an example of such an ODE if $\eta > 0$.

Theorem 18.16

(Levinson–Smith) Assume that $f(x)$ satisfies hypotheses (18.60) and $g(x)$ satisfies hypotheses (18.61). Then the second-order nonlinear ODE (18.59) has a stable limit cycle, and it is unique in $\mathbb{R}^2$, up to translation in time.

A thorough and correct explanation for Theorem 18.16 can be found in many books on advanced ODEs, including the book by Miller and Michel mentioned in "Learn More About It."

But we can give an intuitive argument why such a theorem is plausible. Consider the oscillator ODE

$$\ddot{y} + p\dot{y} + qy = 0.$$

If the damping coefficient is positive, all solutions have $y(t) \to 0$ as $t \to \infty$. If there were such a thing as "negative damping," then all solutions would have $|y(t)| \to \infty$ as $t \to \infty$, except for the equilibrium solution $y(t) \equiv 0$.

Consider a solution of the nonlinear ODE (18.59) that starts near $(x, \dot{x}) = (0, 0)$. Because $p \triangleq f(x) < 0$ for small $|x|$, it's as if there is negative damping and the solution in the $(x, \dot{x})$-phase plane grows away from the origin as t increases.

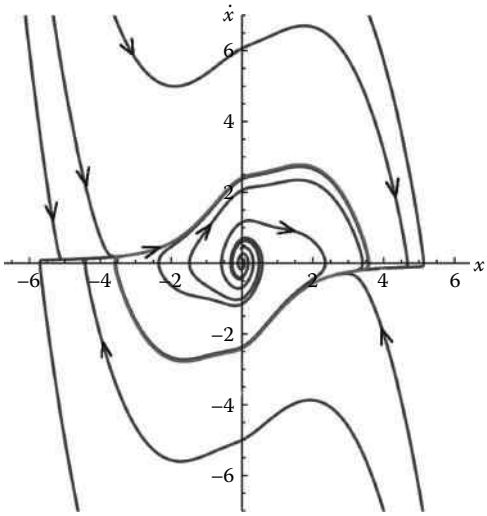


FIGURE 18.15
Levinson–Smith theorem: stable limit cycle.

On the other hand, consider a solution of the nonlinear ODE (18.59) that starts far away from $(x, \dot{x}) = (0, 0)$. Because $p \triangleq f(x) > 0$ for large $|x|$, it's as if there is positive damping and the solution in the $(x, \dot{x})$-phase plane moves toward the origin as t increases.

Also, because of the assumptions on $q \triangleq g(x)$, it is as if there is a positive quasi-frequency of vibration; hence, solutions tend to spiral around the origin.

Put all of this intuition together, and it becomes more reasonable to believe that there is a closed curve solution in the phase plane, somewhere between "near the origin" and "far from the origin." That closed curve would have $(x(\mathrm{T}), \dot{x}(\mathrm{T})) = (x(0), \dot{x}(0))$ for some T, which would be the period of the periodic solution.

Figure 18.15 shows the phase plane for

$$\ddot{x} + \left(x^2\left(0.5 - 2.5e^{-1.4|x|}\right) - 1\right)\dot{x} + \frac{0.2x^3 + x\left(2.5e^{-0.8x^2} + 1\right)}{2x^2 + 1} = 0. \tag{18.63}$$

The limit cycle passes through a point $(x, \dot{x}) \approx (-2.4, 0)$. There is no physical significance to this particular example, but it demonstrates the conclusion of the Levinson–Smith theorem for a complicated-looking ODE. The graphs of $f(x)$ and $g(x)$ in Figure 18.14 produced the stable limit cycle shown in Figure 18.15.

18.6.4 Hopf Bifurcation

A great source of periodic solutions of nonlinear ODEs is the "Poincaré–Andronov–Hopf bifurcation."

The word **bifurcation** refers to a situation where as a parameter, say λ, is varied, the nature of solutions changes. For example, when you reach a bifurcation point or "fork in the road," you have a choice as to which branch of the road to take. Here's an example.

Example 18.25

Suppose λ is a parameter. Study the phase line for the scalar ODE $\dot{y} = y^2 - 2y + \lambda$. For which value(s) of λ does the nature of solutions change as λ varies, that is, where is the bifurcation point(s) of the parameter λ?

Method: This autonomous ODE can be rewritten as $\dot{y} = f(y)$, where

$$f(y) = (y^2 - 2y) + \lambda = (y^2 - 2y + 1) - 1 + \lambda = (y-1)^2 + (\lambda - 1).$$

The equilibrium solutions are those values of y for which $f(y) = 0$, that is,

1. $y = 1 \pm \sqrt{1-\lambda}$, if $\lambda < 1$.
2. $y = 0$, if $\lambda = 1$.
3. D.N.E., if $\lambda > 1$.

The phase lines for the three cases are shown in Figure 18.16. For $\lambda < 1$, there is one stable equilibrium and one unstable equilibrium. The equilibria collapse into a single unstable equilibrium at $\lambda = 1$, and then the equilibrium disappears for $\lambda > 1$. So, there is a bifurcation point at $\lambda = 1$. ◯

A **Poincaré–Andronov–Hopf bifurcation**, or **Hopf bifurcation** for short, is when a *small* periodic solution appears or disappears as λ is varied. The basic situation is that at $\lambda = 0$, the linearization of the ODE system about **0** has an imaginary pair of eigenvalues; hence, we have a critical case, unlike the situation in Theorem 18.6 in Section 18.3. How that imaginary pair of eigenvalues may behave as λ varies from 0 is illustrated in Figure 18.17.

To be specific, we have

Theorem 18.17

Assume that δ is a fixed positive number and an $n \times n$ matrix $A = A(\lambda)$ depends on a single parameter, λ, in the interval $-\delta < \lambda < \delta$. Assume that

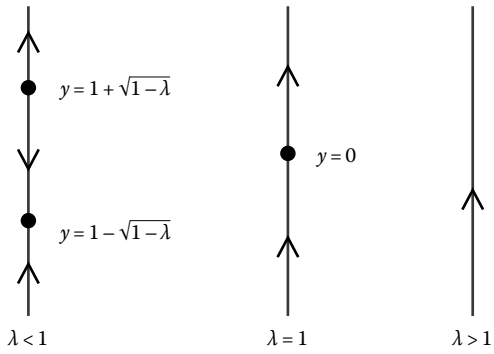


FIGURE 18.16
Example 18.25's phase lines.

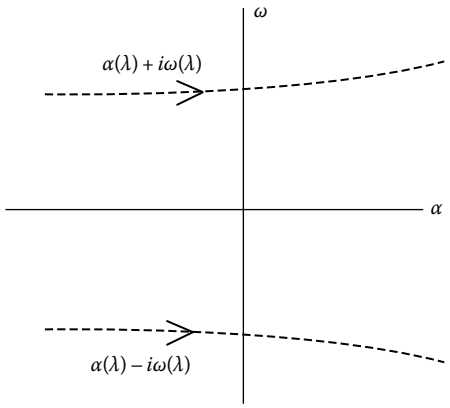


FIGURE 18.17
Eigenvalues $\alpha(\lambda) \pm i\omega(\lambda)$.

- $(n-2)$ eigenvalues of $A(\lambda)$ have negative real part for $-\delta < \lambda < \delta$.
- A has a pair of eigenvalues $\alpha(\lambda) \pm i\omega(\lambda)$ that crosses the imaginary axis with nonzero speed as λ passes through 0, meaning $\alpha(0) = 0, \omega(0) = \omega_0$, and $\frac{d\alpha}{d\lambda}(0) \neq 0$.
- $\mathbf{g}(\mathbf{x}, \lambda)$ is three times continuously differentiable at $(\mathbf{x}, \lambda) = (\mathbf{0}, 0)$.
- $\mathbf{g}(\mathbf{0}, \lambda) \equiv \mathbf{0}$ and $\frac{\partial \mathbf{g}}{\partial \mathbf{x}}(\mathbf{0}, \lambda) \equiv O$, a zero matrix.

(18.64)

Then there exists a parametrized curve $\big(\mathbf{x}_a(t), \lambda(a)\big), 0 \leq a \leq a_0$, of nontrivial periodic solutions of the system of ODEs

$$\dot{\mathbf{x}} = A(\lambda)\mathbf{x} + \mathbf{g}(\mathbf{x}, \lambda). \tag{18.65}$$

Further, there is an invertible transformation P_a of $\mathbb{R}^n$ so that

$$P_a \mathbf{x}_a(t) \approx [a\cos\omega_0 t \quad a\sin\omega_0 t \quad 0 \quad \ldots \quad 0]^T, \text{ as } a \to 0^+.$$

The "a" is for "approximate amplitude."

Theorem 18.18

Assume everything is as in Theorem 18.17. Then, **generically**, $\lambda(a)$ has $\lambda(a) \approx \textit{constant} \cdot a^2$, and for each fixed value of $\lambda(a)$, the periodic solution $\mathbf{x}_a(t)$ is unique for $\dot{\mathbf{x}} = A\big(\lambda(a)\big)\mathbf{x} + \mathbf{g}(\mathbf{x}, \lambda(a))$ in an open ball around $\mathbf{0}$, up to time translation.

Theorem 18.18 says that generically, which means "usually" is a sense that we will not make precise, the situation is like that of a planar ODE system that in polar coordinates is given by the ODE

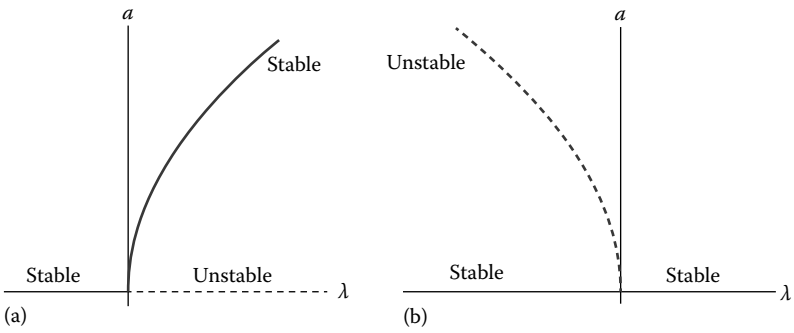


FIGURE 18.18
(a) $\dot{r} = r \cdot (\lambda - r^2)$, (b) $\dot{r} = r \cdot (\lambda + r^2)$.

$$\dot{r} = r \cdot (\lambda - r^2) \tag{18.66}$$

or by the ODE

$$\dot{r} = r \cdot (\lambda + r^2). \tag{18.67}$$

For (18.66), for each positive value of λ near zero, there is a stable periodic solution whose amplitude is $\approx \sqrt{\lambda}$. The **bifurcation diagram** depicts the change in the nature of periodic solutions as the parameter λ changes. The bifurcation diagram for (18.66) is shown in Figure 18.18a. Note that the equilibrium point **0** loses stability as λ increases past zero; the equilibrium point **0** is a periodic solution, too.

For (18.67), the bifurcation diagram in Figure 18.18b shows that for each *negative* value of λ, at least near zero, there is an unstable periodic solution whose amplitude is $\approx \sqrt{-\lambda}$.

Theorem 18.19

(Bendixson–Dulac theorem) Define $\mathbf{x} = [x \;\; y]^T$ and assume $\mathbf{f}(\mathbf{x}) = [f(\mathbf{x}) \;\; g(\mathbf{x})]^T$ is continuously differentiable in an open set $\mathcal{D}$. If there is a continuously differentiable function $\mu(\mathbf{x})$ such that

$$\frac{\partial}{\partial x}\Big[\,\mu(\mathbf{x})f(\mathbf{x})\,\Big] + \frac{\partial}{\partial y}\Big[\,\mu(\mathbf{x})g(\mathbf{x})\,\Big]$$

takes on only one sign in $\mathcal{D}$, then the autonomous ODE system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ has no nonconstant periodic solution in $\mathcal{D}$.

Why? You will explain this in Problem 18.6.5.10.

Learn More About It

Theorem 18.16 is proven in Theorem 7.3.1 of *Ordinary Differential Equations*, by Richard K. Miller and Anthony N. Michel, Dover Publications, 2007, © 1982. Theorems 18.17 and 18.18 are Theorems 11.15 and 11.18, respectively, of *Dynamics and Bifurcations*, by Jack K. Hale and Hüseyin Koçak, Springer-Verlag, © 1991.
A student who wants to explain the Levinson–Smith theorem should work on Problem 16.5, pp. 402–403, of *Theory of Ordinary Differential Equations*, by Earl A. Coddington and Norman Levinson, McGraw-Hill, © 1955.

18.6.5 Problems

In problems 1–4, study limit cycles of the form $r(t) \equiv$ *constant* and their the stability for the system. Also, give a rough sketch of the limit cycle(s) and their stability implicit in the arrow of time on nearby solution curves.

1. $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} y + x(1 - \sqrt{x^2+y^2})(2 - \sqrt{x^2+y^2}) \\ -x + y(1 - \sqrt{x^2+y^2})(2 - \sqrt{x^2+y^2}) \end{bmatrix}$
2. $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} 3x + 2y - x\big(x^2+y^2\big) \\ -2x + 3y - y\big(x^2+y^2\big) \end{bmatrix}$
3. $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -x + y + 2\,x\,e^{-\sqrt{x^2+y^2}} \\ -x - y + 2\,y\,e^{-\sqrt{x^2+y^2}} \end{bmatrix}$
4. $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} 3x + 2y - x\sqrt{x^2+y^2} \\ -2x + 3y - y\sqrt{x^2+y^2} \end{bmatrix}$
5. Suppose $f(x,y)$ is twice continuously differentiable in some open set $\mathcal{O}$ and the **gradient system** of ODEs given by

$$(\star) \qquad \left\{ \begin{aligned} \dot{x} &= \tfrac{\partial f}{\partial x} \\ \dot{y} &= \tfrac{\partial f}{\partial y} \end{aligned} \right\}$$

has a solution $\mathbf{x}(t) = [x(t) \quad y(t)]^T$ that is periodic with period T; hence, in particular $\mathbf{x}(\mathsf{T}) = \mathbf{x}(0)$. Explain why $\mathbf{x}(t)$ is in fact constant and thus is an equilibrium point for $(\star)$. [Hint: Define $h(t) \triangleq f(x(t), y(t))$, explain why $\dot{h}(t) > 0$ at all t unless $\boldsymbol{\nabla} f\big(x(t), y(t)\big) \equiv \mathbf{0}$, and then use Rolle's theorem to get a contradiction if $\mathbf{x}(t)$ were not constant.]

6. Explain why the Van der Pol equation (18.62) satisfies the hypotheses of the Levinson–Smith Theorem 18.16 if $\eta > 0$.
7. Suppose $z(t)$ satisfies a differential equation of the form $(\star)$ $\ddot{z} + F(\dot{z}) + z = 0$, where $F(x)$ is odd and continuously differentiable on $-\infty < x < \infty$.
   (a) Define $x(t) \triangleq \dot{z}(t)$ and explain why $x(t)$ satisfies the ODE $(\star\star)$ $\ddot{x} + f(x)\dot{x} + x = 0$, where $f(x) \triangleq F'(x)$.
   (b) Suppose in addition that (1) there is a constant $a > 0$ such that $F(x) < 0$ for $0 < x < a$ and $F(x) > 0$ for $a < x < \infty$ and (2) $F(x) \to \infty$ as $x \to \infty$. Explain why ODE $(\star\star)$ has a stable limit cycle.
   (c) Use part (b) to conclude $(\star)$ has at most one limit cycle.
8. Explain why $\ddot{x} + \left(x^4 - 1\right)\dot{x} + x^3 = 0$ has a stable limit cycle.
9. Explain why $\ddot{x} + \left(x^2 - 1\right)\dot{x} + 0.01x + \sin x = 0$ has a stable limit cycle.
10. Explain why Theorem 18.19 is true: If there were a nonconstant periodic solution $\mathcal{C} : \mathbf{x} = \mathbf{x}(t), 0 \le t \le \mathsf{T}$, with period T, in $\mathcal{D}$ then define $\overline{\mathcal{D}}$ to be the interior of the simple, closed, continuously differentiable parametrized curve $\mathcal{C}$. Note that

$$\mu(\mathbf{x}(\mathsf{T}))\mathbf{f}(\mathbf{x}(\mathsf{T})) = \mu(\mathbf{x}(0))\mathbf{f}(\mathbf{x}(0))$$

    and use the divergence theorem for the vector field $\mu(\mathbf{x})\mathbf{f}(\mathbf{x})$ in $\overline{\mathcal{D}}$ to get a contradiction.
11. Apply Theorem 18.19 with $\mu(x, y) = e^{-x}$ to explain why the system

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -x^3 \\ -x^3 y - y^3 \end{bmatrix}$$

    has no periodic solution.
12. Suppose λ is a parameter. Study the phase line for the scalar ODE $\dot{y} = \sin y - \lambda$. For which value(s) of λ does the nature of solutions change as λ varies, that is, where is the bifurcation point(s) of the parameter λ?
13. Consider the system in Example 18.11 in Section 18.2. Study the limit cycles as the parameter β varies. For which value(s) of β does the nature of solutions change as β varies, that is, where is the bifurcation point(s) of the parameter β?
14. Consider the system in Example 18.16 in Section 18.4. Study the limit cycles as the damping coefficient b varies. For which value(s) of b does the nature of solutions change as b varies, that is, where is the bifurcation points of the parameter b?

18.7 Existence, Uniqueness, and Continuous Dependence

Here we will study basic issues concerning solutions of systems of (possibly) nonlinear ODEs. First results are generalizations of the results for scalar ODEs that we saw in Section 3.2. Throughout we will be studying the system of ODEs

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(t, \mathbf{x}) \tag{18.68}$$

and also the IVP

$$\left\{ \begin{array}{l} \dfrac{d\mathbf{x}}{dt} = \mathbf{f}(t, \mathbf{x}) \\ \mathbf{x}(t_0) = \mathbf{x}_0 \end{array} \right\}. \tag{18.69}$$

Definition 18.23

A **solution** of an ODE system (18.68) is a function $\mathbf{x}(t)$ that is defined and differentiable on an open interval I and satisfies the system of equations (18.68) on I.

Open intervals can be of the form $(a, b) = \{t : a < t < b\}$, where $a < b$, $(-\infty, \infty) = \{t : -\infty < t < \infty\}$, $(-\infty, b) = \{t : -\infty < t < b\}$, or $(a, \infty) = \{t : a < t < \infty\}$.

A solution has to satisfy the system of ODEs on an open interval. A practical way to find the interval is to substitute the supposed solution into the system of ODEs and during, or after, that, choose the open interval I. But we usually do not have a formula for the solution, so we may need to appeal to a theorem to guarantee that a solution exists.

Theorem 18.20

(Peano's existence theorem) The IVP (18.69) has at least one solution as long as $\mathbf{f}(t, \mathbf{x})$ is continuous in $(t, \mathbf{x})$ in a closed "rectangle" given by

$$\mathcal{R}_{\alpha,\beta}(t_0, \mathbf{x}_0) \triangleq \{(t, \mathbf{x}) : \ t_0 - \alpha \leq t \leq t_0 + \alpha, \ ||\mathbf{x} - \mathbf{x}_0|| \leq \beta\}, \tag{18.70}$$

in $\mathbb{R}^{n+1}$, for some positive scalars α, β.

As far as it goes, this is a good result. However, it doesn't say how many solutions there are for the IVP. In science and engineering, we prefer that a mathematical model of a physical system should give only one prediction about future behavior. Also, the theorem does not tell us for how long a time interval the solution exists. If it turns out that the solution only exists for 10^{-40} s in the future, that would probably not be very useful for making predictions in a physical problem.

In Section 3.2, we got existence and uniqueness results using conditions of differentiability of the right-hand side of the ODE. It turns out that we can relax the conditions somewhat using another kind of condition.

Definition 18.24

A function $\mathbf{f}(t, \mathbf{x})$ satisfies a **local Lipschitz condition** in $\mathbf{x}$ for $(t, \mathbf{x})$ in an open set $\mathcal{D}$ in $\mathbb{R}^{n+1}$ if there is a positive **local Lipschitz constant** K such that

$$||\mathbf{f}(t,\mathbf{x}_1) - \mathbf{f}(t,\mathbf{x}_2)|| \leq K||\mathbf{x}_1 - \mathbf{x}_2|| \tag{18.71}$$

for all $(t,\mathbf{x}_1)$, $(t,\mathbf{x}_2)$ in $\mathcal{D}$.

Theorem 18.21

If $\dfrac{\partial \mathbf{f}}{\partial \mathbf{x}}(t,\mathbf{x})$ is continuous on a closed "rectangle" $\mathcal{R}_{\alpha,\beta}(t_0,\mathbf{x}_0)$, for some positive scalars α,β, then $\mathbf{f}(t,\mathbf{x})$ satisfies a local Lipschitz condition in $\mathbf{x}$ for $(t,\mathbf{x})$ in any open set $\mathcal{D}$ contained in $\mathcal{R}_{\alpha,\beta}(t_0,\mathbf{x}_0)$.

Theorem 18.22

(Picard's theorem with interval of existence) Suppose that

- $\mathbf{f}(t,\mathbf{x})$ is continuous in $(t,\mathbf{x})$ and
- $\mathbf{f}(t,\mathbf{x})$ satisfies a local Lipschitz condition, with local Lipschitz constant K, in $\mathbf{x}$ for $(t,\mathbf{x})$ in a closed *rectangle* $\mathcal{R}_{\alpha,\beta}(t_0,\mathbf{x}_0)$, (18.72)

for some positive scalars α,β, where t_0 and $\mathbf{x}_0$ are the same as in the initial condition $\mathbf{x}(t_0)=\mathbf{x}_0$. Suppose that there is positive constant M such that for all $(t,\mathbf{x})$ in $\mathcal{R}_{\alpha,\beta}(t_0,\mathbf{x}_0)$, we have

$$||\mathbf{f}(t,\mathbf{x})|| \leq M.$$

If we choose $\bar{\alpha}$ and $\bar{\beta}$ sufficiently small that

$$0 < \bar{\alpha} \leq \alpha, 0 < \bar{\beta} \leq \beta, M\bar{\alpha} \leq \bar{\beta}, \text{ and } K\bar{\alpha} < 1,$$

then IVP (18.69) has exactly one solution on the time interval $I_{\bar{\alpha}} \triangleq [t_0 - \bar{\alpha}, t_0 + \bar{\alpha}]$ and the points $\big(t,\mathbf{x}(t)\big)$ remain in the closed "rectangle" $\mathcal{R}_{\bar{\alpha},\bar{\beta}}$ for all t in the interval $I_{\bar{\alpha}}$.

Theorem 18.22 is illustrated in Figure 3.4.

The condition that $M\bar{\alpha} \leq \bar{\beta}$ has an interesting physical interpretation: $||\dot{\mathbf{x}}|| = ||\mathbf{f}\big(t,\mathbf{x}(t)\big)||$ is the speed of an object whose position, $\mathbf{x}(t)$, satisfies ODE (18.68), so M is an upper bound on that speed. So, $M\bar{\alpha} \leq \bar{\beta}$ says, by distance = speed × time, that the object can't go further away from $\mathbf{x}(t_0)$ than a distance of $\bar{\beta}$, that is, $||\mathbf{x}(t)-\mathbf{x}_0|| \leq \bar{\beta}$, that is, the graph of $\mathbf{x}(t)$ versus t stays inside the "rectangle" $\mathcal{R}_{\bar{\alpha},\bar{\beta}}$.

While we will not give a complete explanation for Theorem 18.22, we can mention "Picard iterates," a "building block" of an explanation. First, note that IVP (18.69) is *equivalent* to the integral equation

$$\mathbf{x}(t) = \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{f}\big(s,\mathbf{x}(s)\big)\,ds,\ t_0 - \bar{\alpha} \leq t \leq t_0 + \bar{\alpha}. \tag{18.73}$$

By equivalent we mean that a continuously differentiable function $\mathbf{x}(t)$ satisfies IVP (18.69) if, and only if, $\mathbf{x}(t)$ satisfies integral equation (18.73).

Let $\mathbf{x}_1(t)$ be any function that is defined and continuously differentiable on the interval $I_{\tilde{\alpha}}$ that satisfies $||\mathbf{x}_1(t) - \mathbf{x}_0|| \leq \bar{\beta}$ for all t in $I_{\tilde{\alpha}}$, where $\bar{\alpha}$ and $\bar{\beta}$ are as in Theorem 18.22. We call $\mathbf{x}_1(t)$ the first Picard iterate. The second Picard iterate is defined to be

$$\mathbf{x}_2(t) \triangleq \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{f}\big(s, \mathbf{x}_1(s)\big)\, ds,\ t_0 - \bar{\alpha} \leq t \leq t_0 + \bar{\alpha},$$

and succeeding **Picard iterates** are defined recursively by

$$\mathbf{x}_{k+1}(t) \triangleq \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{f}\big(s, \mathbf{x}_k(s)\big)\, ds,\ t_0 - \bar{\alpha} \leq t \leq t_0 + \bar{\alpha}.$$

An explanation of Theorem 18.22 consists of showing that the sequence of iterates $\{\mathbf{x}_k(t)\}_{k=1}^{\infty}$ (a) converges to a function $\mathbf{x}_\infty(t)$ that is defined and continuously differentiable on $I_{\tilde{\alpha}}$ and (b) $\mathbf{x}_\infty(t)$ satisfies integral equation (18.73). The most efficient way to establish those things is by using a "contraction mapping theorem."

Before proceeding we need to know a property about norms and definite integrals of vector-valued functions: If $\mathbf{g}(t)$ is continuous for t in $[\,a, b\,]$, then

$$\left|\left| \int_a^b \mathbf{g}(t)dt \right|\right| \leq \int_a^b ||\mathbf{g}(t)||\, dt.$$

Convergence of the sequence of Picard iterates is the same concept as the convergence of the sequence of partial sums of a Fourier series in Section 9.1

While we will not give those explanations, we can use integral equation (18.73) to give a brief explanation for the uniqueness conclusion of Theorem 18.22, that is, that (18.73) can have no more than one solution.

For, if *distinct* functions $\mathbf{x}(t)$ and $\mathbf{z}(t)$ both satisfy (18.73), then

$$\begin{aligned} ||\mathbf{x}(t) - \mathbf{z}(t)|| &= \left|\left| \left(\cancel{\mathbf{x}_0} + \int_{t_0}^{t} \mathbf{f}\big(s, \mathbf{x}(s)\big)\, ds \right) - \left(\cancel{\mathbf{x}_0} + \int_{t_0}^{t} \mathbf{f}\big(s, \mathbf{z}(s)\big)\, ds \right) \right|\right| \\ &\leq \left| \int_{t_0}^{t} ||\mathbf{f}\big(s, \mathbf{x}(s)\big) - \mathbf{f}\big(s, \mathbf{z}(s)\big)||\, ds \right| \leq \left| \int_{t_0}^{t} K||\mathbf{x}(s) - \mathbf{z}(s)||\, ds \right|. \end{aligned} \tag{18.74}$$

If we define the positive constant

$$\Delta \triangleq \max_{t_0 - \bar{\alpha} \leq t \leq t_0 + \bar{\alpha}} ||\mathbf{x}(t) - \mathbf{z}(t)||\,,$$

then (18.74) implies that for all t in $I_{\bar{\alpha}}$,

$$\Delta \leq \left| \int_{t_0}^{t} K\Delta \, ds \right| = K\Delta|t - t_0| < K\Delta\bar{\alpha} = (K\bar{\alpha})\Delta. \tag{18.75}$$

But, a hypotheses of Theorem 18.22 is that $K\bar{\alpha} < 1$, so (18.75) implies that

$$\Delta < 1 \cdot \Delta,$$

which is absurd. So, the original assumption that *distinct* functions $\mathbf{x}(t)$ and $\mathbf{z}(t)$ both satisfy (18.73) was incorrect. This establishes the uniqueness of the solution of the integral equation (18.73) and thus of the IVP (18.69). □

Example 18.26

For the scalar IVP $\dot{y} = y$, $y(0) = 1$, use $y_1(t) \equiv 1$ as the initial guess for the solution and find the Picard iterates $y_2(t)$, $y_3(t)$, and $y_4(t)$. Take a guess about the formula for the Picard iterate $y_k(t)$ and use that to find the Picard iterate $y_{k+1}(t)$. Find $y_\infty(t) \triangleq \lim_{k\to\infty} y_k(t)$ and compare it to $y(t)$, the exact solution of the IVP.

Method: $y_1(t) \equiv 1$, $y_2(t) = y(0) + \int_0^t y_1(s)ds = 1 + \int_0^t 1ds = 1 + t$,

$$y_3(t) = y(0) + \int_0^t y_2(s)ds = 1 + \int_0^t (1 + s)ds = 1 + \Big[s + \frac{1}{2} s^2 \Big]_0^t$$
$$= 1 + t + \frac{1}{2} t^2,$$

$$y_4(t) = y(0) + \int_0^t y_3(s)ds = 1 + \int_0^t \Big(1 + s + \frac{1}{2} s^2\Big)ds = 1 + \Big[s + \frac{1}{2!} s^2 + \frac{1}{3!} s^3 \Big]_0^t$$
$$= 1 + t + \frac{1}{2!} t^2 + \frac{1}{3!} t^3.$$

We guess

$$y_k(t) = 1 + t + \frac{1}{2!} t^2 + \frac{1}{3!} t^3 + \cdots + \frac{1}{(k-1)!} t^{k-1},$$

and then

$$y_{k+1}(t) = y(0) + \int_0^t y_k(s)ds = 1 + \int_0^t \Big(1 + s + \frac{1}{2} s^2 + \cdots + \frac{1}{(k-1)!} s^{k-1}\Big)ds$$
$$= 1 + \Big[s + \frac{1}{2!} s^2 + \cdots + \frac{1}{k!} s^k \Big]_0^t = 1 + t + \frac{1}{2!} t^2 + \cdots + \frac{1}{k!} t^k.$$

This suggests

$$y_\infty(t) \triangleq \lim_{k\to\infty} y_k(t) = 1 + t + \frac{1}{2!} t^2 + \cdots + \frac{1}{k!} t^k + \cdots,$$

which equals $y(t) = e^t$, the exact solution of the IVP. ◯

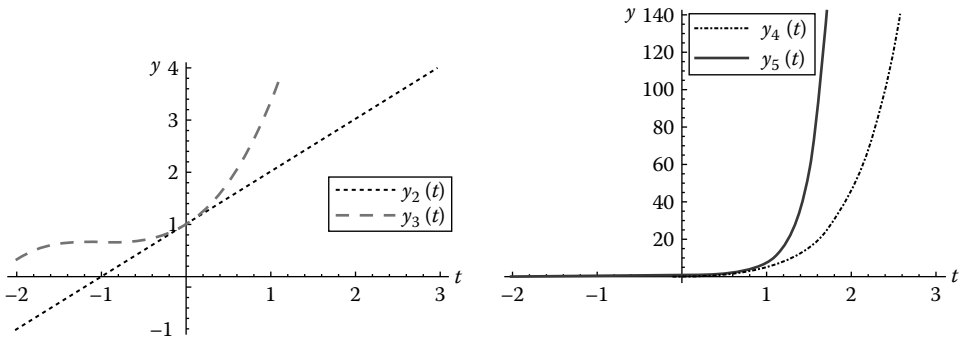


FIGURE 18.19
Picard iterates for $\dot{y} = y,\ y(0) = 1$.

Figure 18.19 shows graphs of the iterates $y_2(t)$, $y_3(t)$, $y_4(t)$, and $y_5(t)$. The latter three give pretty good approximations of the exact solution on a finite time interval. But $y_4(t)$ and $y_5(t)$ give very bad approximations for $t > 1$! While $y_\infty(t) = \lim_{k\to\infty} y_k(t)$ is perfect, the iterates may be troublesome "far" from the initial time.

18.7.1 Continuous Dependence

Recall that an equilibrium point for a system of ODEs is stable if for all nearby initial values, the solution stays near the equilibrium for *all* future time. A related concept deals with how solutions depend, for a *short* amount of future time, on the initial time, initial value, and parameters in the system of ODEs. This concept is called "continuous dependence."

As a first step in this direction, consider how solutions depend on the initial value: Consider IVP

$$\left\{ \begin{array}{l} \dfrac{d\mathbf{x}}{dt} = \mathbf{f}(t, \mathbf{x}) \\ \mathbf{x}(t_1) = \mathbf{x}_1 \end{array} \right\}, \tag{18.76}$$

where we are allowed to vary the initial time, t_1, and the initial value, $\mathbf{x}_1$.

Theorem 18.23

(Continuous dependence on initial condition) Suppose that $\mathbf{f}(t, \mathbf{x})$ satisfies hypotheses (18.72) of Picard's Theorem 18.22 and let the unique solution of IVP (18.76) be

$$\mathbf{x}(t; t_1, \mathbf{x}_1).$$

Then $\mathbf{x}(t; t_1, \mathbf{x}_1)$ is continuous in $(t, t_1, \mathbf{x}_1)$. Let $\bar{\alpha}$ be as in the conclusion of Picard's Theorem 18.22. It follows that for all $\varepsilon > 0$, no matter how small, and all final time $T < \bar{\alpha}$, there exists some $\delta > 0$ such that

$$|t_1 - t_0| + ||\mathbf{x}_1 - \mathbf{x}_0|| < \delta \quad \text{implies} \quad ||\mathbf{x}(t; t_1, \mathbf{x}_1) - \mathbf{x}(t; t_0, \mathbf{x}_0)|| < \varepsilon \text{ for } 0 \le t \le T. \tag{18.77}$$

Notice that this result, (18.77), does not imply stability, which is a property that is to hold for all $t \geq t_0$. Indeed, result (18.77) holds whether or not $\mathbf{x}_0$ is an equilibrium point. Nevertheless, continuous dependence on the initial value $\mathbf{x}_1$ is useful in theoretical stability proofs, including those using Liapunov functions in Section 18.4. In fact, continuous dependence plays a technical role in many fundamental ODE methods, for example, the Poincaré map in Section 18.8.

Numerical methods also use continuous dependence. Suppose a numerical method for an initial value problem for $\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x})$ produces an approximation $\mathbf{x}_k$ for $\mathbf{x}(t_k)$. In order to be explicit in this discussion, suppose we are using Euler's method. For the next time step, there are four types of errors that can make $\mathbf{x}_{k+1}$ differ from $\mathbf{x}(t_{k+1})$:

1. The error in approximating $\mathbf{x}(t_{k+1}) \approx \mathbf{x}_{k+1} + \triangle t\, \mathbf{f}\big(t_k, \mathbf{x}(t_k)\big)$ because $\triangle t$ is not arbitrarily small
2. The error in approximating $\mathbf{f}\big(t_k, \mathbf{x}(t_k)\big) \approx \mathbf{f}(t_k, \mathbf{x}_k)$ because $\mathbf{x}(t_k) \approx \mathbf{x}_k$
3. Round off error during arithmetic operations, and
4. Approximation errors in the machine's hard wired function evaluations

The first kind of error can be influenced by the choice of numerical method, for example, using a Runge-Kutta formula versus Euler's formula. The second kind of error is kept under theoretical control using continuous dependence, which says that the error in the previous step will propagate into the next step but hopefully will not have a drastic effect. Likewise, continuous dependence helps control the effect of the third and fourth kinds of error.

Also, continuous or differentiable dependence on initial conditions is also useful in the shooting method briefly discussed in Section 8.8. For example, suppose we want to solve an ODE-BVP

$$\left\{ \begin{array}{c} y'' = f(x, y, y') \\ y(a) = \alpha,\ y(b) = \beta \end{array} \right\}.$$

Let $y(x; z)$ be the solution of the IVP

$$\left\{ \begin{array}{c} y'' = f(x, y, y') \\ y(a) = \alpha,\ y'(a) = z \end{array} \right\},$$

where we have continuous, or better yet, differentiable, dependence of $y(x; z)$ on z. Then we can use a numerical method for root finding to solve $y(b; z) = \beta$ for $z = z^\star$. The solution of the original ODE-BVP is then $y(x; z^\star)$. In practice, yet another numerical method may be needed to give an approximation to $y(x; z)$.

Theorem 18.24

(**Differentiable dependence** on initial condition) Suppose that $\mathbf{f}(t, \mathbf{x})$ is continuously differentiable on some ball $B_\delta = \{(t, \mathbf{x}) : |t - t_1| + ||\mathbf{x} - \mathbf{x}_1|| < \delta\}$ and let the unique solution of IVP (18.76) be

$$\mathbf{x}(t; t_1, \mathbf{x}_1).$$

Then $\mathbf{x}(t; t_1, \mathbf{x}_1)$ is differentiable in $(t, t_1, \mathbf{x}_1)$, and the matrix

$$Y(t) = \left[\frac{\partial \mathbf{x}}{\partial \mathbf{x}_1}(t; t_1, \mathbf{x}_1) \right]$$

satisfies the linear variational equation

$$\dot{Y} = \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\big(t, x(t; t_1, \mathbf{x}_1)\big) \right] Y \tag{18.78}$$

for t in some interval $(t - \widetilde{\delta}, t + \widetilde{\delta})$ for some $\widetilde{\delta} > 0$ satisfying $0 < \widetilde{\delta} < \delta$.

We also have results in which we can vary the right-hand sides of the system of ODEs. It is good that we have such results because in the real world we cannot assume that we can perfectly measure the initial value $\mathbf{x}_1$, the initial time t_0, or parameters in the system of ODEs.

Suppose the system of ODEs is $\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}, \boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ are parameters in the system.

Theorem 18.25

(Continuous dependence on everything) Suppose that for $(t, \mathbf{x})$ in a closed rectangle $\mathcal{R}_{\alpha,\beta}(t_0, \mathbf{x}_0)$ and all $\boldsymbol{\lambda}$ with $||\boldsymbol{\lambda} - \boldsymbol{\lambda}_0|| < \alpha$,

- $\mathbf{f}(t, \mathbf{x}, \boldsymbol{\lambda})$ is continuous in $(t, \mathbf{x}, \boldsymbol{\lambda})$, and
- $\mathbf{f}(t, \mathbf{x}, \boldsymbol{\lambda})$ satisfies a local Lipschitz condition in $\mathbf{x}$, with local Lipschitz constant K,

(18.79)

for some $\alpha, \beta > 0$, and let the unique solution of IVP (18.76) be

$$\mathbf{x}(t; t_1, \mathbf{x}_1, \boldsymbol{\lambda}).$$

Then $\mathbf{x}(t; t_1, \mathbf{x}_1, \boldsymbol{\lambda})$ is continuous in $(t, t_1, \mathbf{x}_1, \boldsymbol{\lambda})$. Let $\bar{\alpha}$ be as in the conclusion of Picard's Theorem 18.22. It follows that for all $\varepsilon > 0$, no matter how small, and all final time $T < \bar{\alpha}$, there exists some $\delta > 0$ such that $||\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_0|| + |t_1 - t_0| + ||\mathbf{x}_1 - \mathbf{x}_0|| < \delta$ implies

$$||\mathbf{x}(t; t_1, \mathbf{x}_1, \boldsymbol{\lambda}_1) - \mathbf{x}(t; t_0, \mathbf{x}_0, \boldsymbol{\lambda}_0)|| < \varepsilon \text{ for } |t - t_1| \leq T. \tag{18.80}$$

An informal definition of "chaos" is that there is "sensitive dependence on initial conditions," in which no matter how close are two initial values, the *long-term* future behavior appears to depend randomly on the initial value. This phenomenon will be discussed further in Section 18.8. A related concept is that of a "*strange* or *chaotic* attractor," a set that is an attractor according to Definition 18.19 in Section 18.5 but has a Smale horseshoe map structure as will be described in Section 18.8.

Note that continuous dependence only applies locally in time, that is, for $|t - t_1| \leq T < \bar{\alpha}$, whereas sensitive dependence on initial conditions refers to behavior as $t \to \infty$.

Continuous dependence on the right-hand sides, $\mathbf{f}(t, \mathbf{x}, \boldsymbol{\lambda})$, is part of studying the "structural stability" of systems of ODEs. It is a curious intuitive result that systems with chaos tend to be structurally stable in the sense that small perturbations of chaotic systems tend to be chaotic and preserve the property of having a strange attractor.

Learn More About It

There is a "uniform contraction mapping theorem" that not only gets the existence and uniqueness conclusions of Picard's theorem but also the continuous dependence conclusions of Theorems 18.23 through 18.25. Textbooks for graduate level study of ODEs, such as *Ordinary Differential Equations*, by Jack K. Hale, Robert E. Krieger Publishing Co., © 1980, Section I.3, have such theorems and their justifications.

Problem 18.7.2.7 is a version of Problem 2.3.1 in *Differential Equations and Dynamical Systems*, 3*rd* ed., by Lawrence Perko, Springer-Verlag, Inc., © 2001.

Ordinary Differential Equations, by Garrett Birkhoff and Gian-Carlo Rota, Blaisdell Publ. Co., © 1969, has a nice result, Theorem 6.4.3, that compares solutions of two systems, $\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x})$ and $\dot{\mathbf{x}} = \mathbf{g}(t, \mathbf{x})$, without assuming the latter satisfies a local Lipschitz condition.

18.7.2 Problems

1. For the scalar IVP $\dot{y} = y,\ y(2) = 1$, use $y_1(t) \equiv 1$ as the initial guess for the solution and find the Picard iterates $y_2(t)$, $y_3(t)$, and $y_4(t)$. Take a guess about the formula for the Picard iterate $y_k(t)$ and use that to find the Picard iterate $y_{k+1}(t)$. Find $y_\infty(t) \triangleq \lim_{k\to\infty} y_k(t)$ and compare it to $y(t)$, the exact solution of the IVP. [Hint: It helps to leave things factored, for example, $\int_2^t (s-2)ds = \left[\ \frac{1}{2}\ (s-2)^2\ \right]_2^t = \frac{1}{2}\ (t-2)^2$ rather than write it as $\int_2^t (s-2)ds = \left[\ \frac{1}{2}\ s^2 - 2s\ \right]_2^t = \frac{1}{2}\ (t^2 - 2^2) - 2(t-2)$.]
2. For the scalar IVP $\dot{y} = y^2,\ y(2) = 1$, use $y_1(t) \equiv 1$ as the initial guess for the solution and find the Picard iterates $y_2(t)$, $y_3(t)$, $y_4(t)$, and $y_5(t)$. Take a guess about $y_\infty(t) \triangleq \lim_{k\to\infty} y_k(t)$ and compare it to $y(t)$, the exact solution of the IVP.
   [Hint: You may want to use *Mathematica* to find some of the iterates, for example, using commands y[1, t_] : = 1, y[2, t] = y[1, t] + $\int_0^t$ (y[1, s])^2 ds, . . . , y[5, t] = y[1, t] + $\int_0^t$ (y[4, s])^2 ds.]
3. For the IVP system $\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}\mathbf{x},\ \mathbf{x}(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, use $\mathbf{x}_1(t) \equiv \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ as the initial guess for the solution and find the Picard iterates $\mathbf{x}_2(t)$, $\mathbf{x}_3(t)$, $\mathbf{x}_4(t)$, and $\mathbf{x}_5(t)$. Take a guess about the formula for the Picard iterate $\mathbf{x}_k(t)$ and use that to find $\mathbf{x}_\infty(t) \triangleq \lim_{k\to\infty} \mathbf{x}_k(t)$ and compare it to $\mathbf{x}(t)$, the exact solution of the IVP.
4. For the IVP system $\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ -4 & 0 \end{bmatrix}\mathbf{x},\ \mathbf{x}(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, use $\mathbf{x}_1(t) \equiv \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ as the initial guess for the solution and find the Picard iterates $\mathbf{x}_2(t)$, $\mathbf{x}_3(t)$, $\mathbf{x}_4(t)$, and $\mathbf{x}_5(t)$. Take a guess about the formula for the Picard iterate $\mathbf{x}_k(t)$ and use that to find $\mathbf{x}_\infty(t) \triangleq \lim_{k\to\infty} \mathbf{x}_k(t)$ and compare it to $\mathbf{x}(t)$, the exact solution of the IVP.

5. (a) Explain why every continuously differentiable function $\mathbf{x}(t)$ that satisfy IVP (18.69) must also satisfy integral equation (18.73).
   (b) Explain why every continuously differentiable function that $\mathbf{x}(t)$ satisfies integral equation (18.73) must also satisfy IVP (18.69).
6. In this problem, you will use Picard's theorem in the special case of linear systems in $\mathbf{R}^n$, that is, $(\star)\quad \dot{\mathbf{x}} = A(t)\mathbf{x}$. Throughout, assume that $A(t)$ is continuous on an interval $\mathcal{I} = \{t : a \le t \le b\}$. Also, you may assume that (1) the norm of a matrix B satisfies $||B\mathbf{x}|| \le ||B||\,||\mathbf{x}||$ for all $\mathbf{x}$ for which the matrix multiplication is defined and (2) if the matrix $A(t)$ is continuous in t on $\mathcal{I}$, then there exists

$$K = \max_{a \le t \le b} ||A(t)|| < \infty.$$

   (a) Define $\mathbf{f}(t, \mathbf{x}) = A(t)\mathbf{x}$. By linearity, $A(t)\mathbf{x}_1 - A(t)\mathbf{x}_2 = A(t)(\mathbf{x}_1 - \mathbf{x}_2)$. Use this and the preceding assumptions to explain why $\mathbf{f}(t, \mathbf{x})$ satisfies conditions (18.72).
   (b) Explain why the other hypotheses of Picard's theorem are true for $(\star)$ and thus conclude that for every t_0 satisfying $a < t_0 < b$ and every $\mathbf{x}_0$ in $\mathbf{R}^n$, there exists an $\bar{\alpha}$ such that a solution of IVP $\dot{\mathbf{x}} = A(t)\mathbf{x},\ \mathbf{x}(t_0) = \mathbf{x}_0$ exists on the time interval $t_0 - \bar{\alpha} < t < t_0 + \bar{\alpha}$.
   (c) Because the solution exists on $t_0 - \bar{\alpha} < t < t_0 + \bar{\alpha}$, explain why we can use the IVP

$$\dot{\mathbf{x}} = A(t)\mathbf{x},\ \mathbf{x}\left(t_0 + \frac{3\bar{\alpha}}{4}\right) \text{ given},$$

   to continue the solution to exist on the whole time interval

$$t_0 - \bar{\alpha} < t < t_0 + \frac{5\bar{\alpha}}{4},$$

   and then continue the solution to exist on

$$t_0 - \bar{\alpha} < t < t_0 + \frac{7\bar{\alpha}}{4},$$

   etc. Continuing, we get the solution existing on all of the time interval $t_0 - \bar{\alpha} < t < b$. Working backward in time, in a similar way, we get the solution existing on all of the time interval $a < t < b$.
7. Suppose A is a constant matrix and define $\mathbf{x}(t; t_1, \mathbf{x}_1)$ to be the unique solution of the IVP $\dot{\mathbf{x}} = A\mathbf{x},\ \mathbf{x}(t_1) = \mathbf{x}_1$. Express $\mathbf{x}(t; t_1, \mathbf{x}_1)$ in terms of e^{tA} and then calculate the matrix $Y(t) \triangleq \dfrac{\partial \mathbf{x}}{\partial \mathbf{x}_1}(t; t_1, \mathbf{x}_1)$. Compare your conclusion with the results of Theorem 18.24.
8. Suppose $A(t)$ is continuous and periodic with period T. Define $\mathbf{x}(t; t_1, \mathbf{x}_1)$ to be the unique solution of the IVP $\dot{\mathbf{x}} = A\mathbf{x},\ \mathbf{x}(t_1) = \mathbf{x}_1$. Express $\mathbf{x}(t; t_1, \mathbf{x}_1)$ in terms of the

Floquet representation and then calculate the matrix $Y(t) \triangleq \dfrac{\partial \mathbf{x}}{\partial \mathbf{x}_1}(t; t_1, \mathbf{x}_1)$. Compare your conclusion with the results of Theorem 18.24.

9. Assume λ is a positive constant and $\mathbf{x}_1$ is a fixed but unspecified nonzero vector. For the ODE-IVP $\mathbf{x} = \begin{bmatrix} 0 & 1 \\ -\lambda & 1 \end{bmatrix} \mathbf{x},\ \mathbf{x}(0) = \mathbf{x}_1$,
   (a) Find the solution $\mathbf{x}(t; \lambda)$.
   (b) Calculate

$$\mathbf{z}(t; \lambda) \triangleq \frac{\partial \mathbf{x}}{\partial \lambda}(t; \lambda).$$

   (c) Find an ODE-IVP that is satisfied by the function $\mathbf{z}(t)$ you found in part (b).

10. For the IVP $\dfrac{dy}{dt} = 1 + y^2,\ y(0) = y_1$,
   (a) Use separation of variables to find the solution, $y(t; y_1)$, assuming $y_1 \neq 0$.
   (b) find the IVP satisfied by

$$\mathbf{z}(t) \triangleq \frac{\partial y}{\partial y_1}(t; y_1).$$

11. For the scalar IVP

$$\left\{ \begin{array}{c} \dfrac{dy}{dt} = 1 + y^2 \\ y(0) = 0 \end{array} \right\},$$

   (a) Explain why the method of separation of variables produces a solution $y(t)$ that exists only on a finite time interval of the form $-\delta < t < \delta$.
   (b) Explain why Picard's theorem guarantees the existence and uniqueness of a solution on an interval of the form $-\bar{\alpha} \leq t \leq \bar{\alpha}$, and
   (c) Compare the *largest* value of $\bar{\alpha}$ that Picard's theorem produces with the value of δ. Conclude that Picard's theorem theoretical guarantee of a time interval of existence is not as good as the actual time interval of existence!

18.8 Short Take: Horseshoe Map and Chaos

Scientist and mathematicians discovered that nonlinear ordinary differential or difference equations can have solutions that behaved in seemingly unpredictable ways. One such phenomenon is "sensitive dependence on initial conditions," in which no matter how close are two initial values, the *long-term* future behavior appears to depend randomly on the initial value. This is called "chaos."

This was poetically expressed as the "butterfly effect," the idea that the tiny wind from the flapping wings of a butterfly in Mexico could have a big effect on the rain falling

in Mongolia. Indeed, one of the earliest works connecting the mathematics of sensitive dependence to the physical world was Lorenz's system of three ODEs model (Lorenz, 1963) of the atmosphere.

This is true even though the differential or difference equations themselves were "deterministic," that is, did not include random variables as terms or factors. But even random variables, for example, "Brownian motion," have structure and characteristics, and people did find structure in deterministic chaos. These structures include "strange attractors."

Moreover, scientists and engineers learned how to use chaos in mathematical models, for example, in neurophysiology and communication theory.

One well-developed theory that makes the aforementioned concepts precise is the phenomenon of **period doubling**: A scalar or system of differential or difference equations can have periodic solutions whose periods depend on a parameter in the system model, and those periods could successively double and double as the parameter changes. We have seen in Section 5.8 that a system could have a specific initial value that gives a periodic solution. A periodic solution's period tells us when the state returns to its initial value. It makes sense that an initial value near that specific value may wander further away the longer amount of time it has to "wander." An analogy is that a pet animal left alone can be harder to find the longer he is left unattended. So, having an ever increasing period could lead to increasingly sensitive dependence on initial conditions.

What makes the "period doubling route to chaos" even more appealing is that as the period doubles and doubles endlessly, the number of periodic solutions can increase endlessly. So, an initial condition could be near many different periodic solutions as it wanders. When the state of chaos is reached, there can be a "dense set" of periodic solutions in the sense that every initial condition is arbitrarily close to some periodic solution.

One successful model that makes all of the aforementioned text precise is the **Smale horseshoe map**. Consider an **alphabet** consisting of two letters, $\{A, B\}$ and a bi-infinite sequence $\{x_k\}_{k=-\infty}^{\infty}$ with each x_k being either A or B. The set of all such sequences is infinite. In fact, if we look at their truncations $\{x_k\}_{k=-n}^{n}$, then we have $N_n = 2^{2n}$ of them: As the size $2n$ of the truncations increases endlessly, then N_n, the number of possible truncated sequences, increases endlessly. In fact,

$$\lim_{n\to\infty} \frac{\ln N_n}{2n} = \ln 2$$

is called the **entropy**. Because the entropy is positive, a random choice of a bi-infinite sequence using this alphabet could be anything. This is related to the saying that a "give an infinite number of monkeys an infinite amount of time with word processing software, and they will reproduce the works of Shakespeare" mixed in with meaningless streams of characters such as "vpeiwq jidfkjlasddfkpreoptyiofgnksdlpfbdlkd. . ." I struck my keyboard at random. Mixed in were three words, "as," "opt," and "of" as well as a sort of word "pre(-)op."

We have seen in Section 6.6 examples of linear mappings of the plane $\mathbb{R}^2$. Nonlinear mappings

$$\mathbf{x} \mapsto \mathbf{y} = \mathbf{F}(\mathbf{x})$$

can be much more complicated. The horseshoe map bends a region $\mathcal{D}$ into a horseshoe shape, as depicted in Figure 18.20.

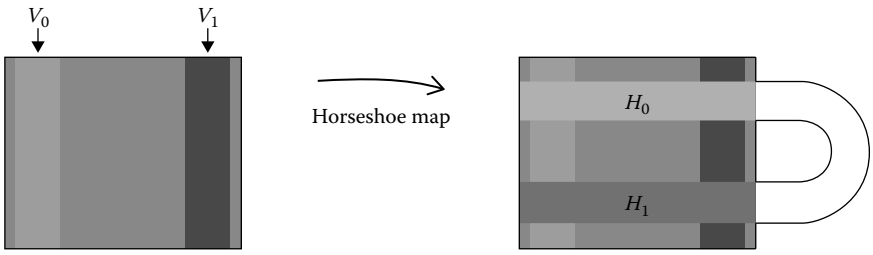


FIGURE 18.20
Horseshoe map.

A map defines a system of nonlinear difference equations

$$\mathbf{x}_{k+1} = \mathbf{F}(\mathbf{x}_k)$$

whose solutions are sequences

$$\{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots\}, \quad \text{that is,} \quad \left\{\mathbf{x}_0, \mathbf{F}(\mathbf{x}_0), \mathbf{F}\big(\mathbf{F}(\mathbf{x}_0)\big), \ldots\right\}.$$

For difference equations, there are concepts of stable and unstable manifolds, $\mathcal{S}$ and $\mathcal{U}$, respectively, that is, curves passing through $\bar{\mathbf{x}}$ along which solutions $\{\mathbf{x}_k\}_{k=1}^{\infty}$ either tend toward or away from $\bar{\mathbf{x}}$ as $k \to \infty$. The stable and unstable manifolds are invariant forward in time, k: For example, if your initial value $\mathbf{x}_0$ starts on $\mathcal{S}$, then the iterates $\mathbf{x}_1, \mathbf{x}_2, \ldots$ are on $\mathcal{S}$, that is, $\mathbf{F}(\mathcal{S}) \subseteq \mathcal{S}$.

The horseshoe map on $\mathcal{D}$ takes two vertical strips V_0, V_1 into two horizontal strips H_0, H_1, respectively, for example, $\mathbf{F}(V_0) \subseteq H_0$. The action of the horseshoe map on $\mathcal{D}$ squeezes space in one direction tangent to the stable manifold and elongates space in the other direction tangent to the unstable manifold. We can imagine that this has the effect that points that are initially close can move much further away as time, k, increases. This gives sensitive dependence on initial conditions. The details of how this happens require some lengthy explanations, but we will only mention the conclusions that in part the map behaves like it was acting on the space of bi-infinite sequences of letters chosen from $\{A, B\}$. This gives both a dense set of periodic solutions and the kind of randomness that follows from having positive entropy.

ODEs

So, what does a horseshoe map have to do with ODEs? The answer is that the solutions of a system of ODEs can define a nonlinear map.

The first way that we can get a map of the plane is if we have an autonomous system of three ODEs, that is, our state space is $\mathbb{R}^3$. Suppose that we have a periodic solution curve $\mathcal{C} : \mathbf{x}(t), -\infty < t < \infty$ that passes through a point $\mathbf{x}(0)$ and we have a plane Π that is perpendicular to $\mathcal{C}$ at $\mathbf{x}(0)$, that is, the tangent vector $\dot{\mathbf{x}}(0)$ is normal to Π. Solutions that start on Π and are near to $\mathbf{x}(0)$ will return to Π in roughly the same amount of time as it takes for $\mathbf{x}(t)$ to return to $\mathbf{x}(0)$, because of continuous dependence of solutions on the initial value. So, a point $\mathbf{x}$ on Π, near to $\mathbf{x}(0)$, will have an image $\mathbf{F}(\mathbf{x})$ on Π. $\mathbf{F}$ is called the induced **Poincaré map**. If the Poincaré map behaves like the horseshoe map on a region on Π, then the system of ODEs in $\mathbb{R}^3$ will exhibit chaos.

People have been able to show that under special circumstances a "homoclinic orbit" will induce a horseshoe map and thus chaos. A **homoclinic orbit** is a solution curve $\mathcal{C}$: $\mathbf{x} = \mathbf{x}(t), -\infty < t < \infty$ that is not periodic but instead has

$$\lim_{t\to\infty} \mathbf{x}(t) = \bar{\mathbf{x}} = \lim_{t\to-\infty} \mathbf{x}(t).$$

Because of uniqueness of solutions, the homoclinic orbit never actually touches the equilibrium point $\bar{\mathbf{x}}$. One special circumstance we want is that the homoclinic orbit should be on both the unstable and stable manifolds of $\bar{\mathbf{x}}$. Because the state space is $\mathbb{R}^3$, the stable manifold or the unstable manifold will have dimension two, and the other will have dimension one. The other special circumstance we want is that the tangent spaces to those two manifolds should intersect **transversally**, meaning that the tangent vectors to the stable and unstable manifolds should, together, span the whole space.

The second way that we can get a map of the plane is if we have a periodic, nonautonomous system of two ODEs that is a small perturbation of an autonomous planar ODE. Because of periodicity, solutions do depend on the initial time. Suppose we have a homoclinic or heteroclinic orbit for the corresponding autonomous planar ODE. Then it is possible to show that there may be a dense set of periodic solutions and sensitive dependence on initial conditions, that is, chaos.

Learn More About It

A useful book of illustrations for dynamics is *Dynamics: The Geometry of Behavior*, particularly *Part Three: Global Behavior*, by Ralph H. Abraham and Christopher D. Shaw, Aerial Press, Inc.

A nice exposition of the Smale horseshoe map is in Section 3.5 of *An Introduction to Dynamical Systems*, by D. K. Arrowsmith and C. M. Place, Cambridge University Press, © 1990. Other good references for period doubling and chaos and their mathematical foundations are Section 11.3 of *Methods of Bifurcation Theory*, by Shui-Nee Chow and Jack K. Hale, Springer-Verlag, © 1982; *Nonlinear Dynamics and Chaos*, by Steven H. Strogatz, Addison-Wesley, © 1994; and Section 12.3 of *Nonlinear Dynamics and Chaos*, by J. M. T. Thompson and H. B. Stewart, John Wiley & Sons, © 1986.

A famous result of chaos in nonlinear difference equations is in "Period three implies chaos," by Tien-Yien Li and James A. Yorke, *Am. Math. Mon.* **82** (1975), 985–992.

18.9 Short Take: Delay Equations

An example of a delay equation is

$$\frac{d}{dt}\big[\, x(t) \,\big] = ax(t) + bx(t-r), \tag{18.81}$$

where r is a positive constant and a, b are constant coefficients. In order to know the value of the derivative $\dot{x}(t)$ for times in the interval $0 \le t \le r$, we need to know the values of

$ax(t) + bx(t - r)$ for $0 \le t \le r$; hence, we need to know the values of

$$x(\theta) \text{ for } -r \le \theta \le 0.$$

So, in order to solve the delay equation (18.81) over as large a time interval as possible, we need to know not just the initial value $x(0)$ but also the **history** of $x(t)$ for $-r \le t \le 0$.

A useful notation is

$$x_t(\theta) \triangleq x(t + \theta) \text{ for } -\mathrm{r} \le \theta \le 0. \tag{18.82}$$

We see that the initial history is given by

$$x_0(\theta) \triangleq x(\theta) \text{ for } -\mathrm{r} \le \theta \le 0. \tag{18.83}$$

We can think of a delay equation as describing the time evolution of $x_t(\theta)$ as a function of t, and the state space in which this is taking place is the vector space $\mathcal{C}$ consisting of all functions $\phi(\theta)$ that are defined and continuous in θ for $-r \le \theta \le 0$, that is,

$$\mathcal{C} \triangleq C[-r, 0].$$

$\mathcal{C}$ is a complete normed vector space when given the norm

$$||\phi||_\infty \triangleq \max_{-r \le \theta \le 0} |\phi(\theta)|.$$

We can study the existence and uniqueness issues of delay equations using Picard iterates just as we did for ODEs in Section 18.7. For example, the scalar delay equation (18.81), along with the initial history $x_0(\theta)$, is equivalent to the integral equation

$$x(t) = x(0) + \int_0^t \big(ax(s) + bx(s - r)\big)ds. \tag{18.84}$$

So, at a time $t > 0$, the history $x_t = x_t(\theta)$ is given by

$$x_t(\theta) = x(0) + \int_0^{t+\theta} \big(ax(s) + bx(s - r)\big)ds, \quad -r \le \theta \le 0.$$

Theorem 18.26

(Existence and uniqueness) For the IVP

$$\left\{ \begin{array}{c} \dfrac{d}{dt}\big[\, x(t) \,\big] = f(t, x(t), x_t) \\ x_{t_0}(\theta) \triangleq \phi_0(\theta), \quad -r \le \theta \le 0 \end{array} \right\}, \tag{18.85}$$

suppose that for all (t,x) in a closed rectangle $\mathcal{R}_{\alpha,\beta}\big(t_0,\phi_0(0)\big)$ and all φ with $||\varphi-\phi_0||_\infty<\beta$,

- $f(t,x,\phi)$ is continuous in (t,x,ϕ) and
- $f(t,x,\phi)$ satisfies local Lipschitz condition

$$|f(t,x,\phi)-f(t,y,\psi)|\le K\left(|x-y|+||\phi-\psi||_\infty\right), \tag{18.86}$$

for some $\alpha,\beta>0$. Then for some $\bar\alpha>0$, IVP (18.85) has a unique solution $x_t=x_t(\theta)$, $-r\le\theta\le 0$ defined for $t_0<t<t_0+\bar\alpha$.

You may have noticed that Theorem 18.81 concludes with the existence (and uniqueness) of a solution only forward in time, that is, for $t_0<t<t_0+\bar\alpha$. It is an unfortunate reality that we cannot expect to have the existence and uniqueness of a solution backward in time. Reference to an example of this will be in "Learn More About It" at the end of the section.

If the initial history $x_0=x_0(\theta)\triangleq\phi(\theta)$ is in $\mathcal{C}$, that is, is continuous on the interval $-r\le\theta\le 0$, it makes sense that there should be a unique solution of (18.81) that exists as a continuously differentiable function for $0<t<\infty$. In addition, as long as the initial history is differentiable from the left at $\theta=0$ and satisfies the "compatibility condition"

$$\lim_{\triangle\theta\to 0^-}\frac{\phi(\triangle\,\theta)-\phi(0)}{\triangle\,\theta}\triangleq\dot\phi(0^-)=a\phi(0)+b\phi(-r),$$

then the solution $x(t)$ will be differentiable at $t=0^+$.

18.9.1 Characteristic Equation

For the delay equation (18.81), try to find solution(s) in the form $x(t)=e^{\lambda t}$ where λ is a constant. Substituting this into the delay equation, we get

$$\lambda e^{\lambda t}=\dot x(t)=ax(t)+bx(t-r)=ae^{\lambda t}+be^{\lambda(t-r)},$$

and then multiply through by $e^{-\lambda t}$ to get

$$\lambda=a+be^{-\lambda r}. \tag{18.87}$$

Unfortunately, this is not an algebraic equation in λ. In fact, there can be infinitely many distinct roots $\lambda_1,\lambda_2,\ldots$, so delay equation (18.81) can have general solution

$$x(t)=\sum_{n=1}^{\infty}c_ne^{\lambda_n t},$$

where $c_1,c_2,\ldots$ are arbitrary constants. Indeed, in order to satisfy the initial history condition, we need to choose $c_1,c_2,\ldots$ so that

$$\phi(\theta)=x_0(\theta)=\sum_{n=1}^{\infty}c_ne^{\lambda_n\theta},\quad -r\le\theta\le 0.$$

We have seen in the study of Fourier series that, in principle, we need an infinite family of functions in order to expand a function in a series of functions.

In fact, delay equations are like PDEs in this respect. Recall that when solving the heat equation, we need to specify the initial heat distribution over space; for a delay equation, we need to specify the initial history over time.

18.9.2 Euler's Method

The simplest numerical method for ODEs is Euler's method. We can use it also for delay equations: Partition the history time interval $-r \le t \le 0$ into subintervals of equal length $\triangle t = \dfrac{r}{N}$ and define

$$x_i = x(t_i), \quad \text{for} \quad i = -N, -N+1, \ldots, 0, 1, 2, \ldots .$$

Note that the initial history function $\phi(\theta) = x_0(\theta) = x(\theta)$ gives the values

$$x_i = \phi(i \triangle t), \quad \text{for} \quad i = -N, -N+1, \ldots, 0.$$

For $i = 1, 2, \ldots$ we use the delay equation to approximate $\dot{x}(t_i)$. For example, for the delay equation (18.81), Euler's method would give

$$x_{i+1} = x_i + \triangle t \cdot \left(a x_i + b x_{i-N}\right), \quad \text{for} \quad i = 0, 1, 2, \ldots .$$

Any other finite differences numerical method, such as a Runge–Kutta method, can be implemented in a similar way.

Learn More About It

Introduction to Functional Differential Equations, by Jack K. Hale and Sjoerd M. Verduyn Lunel, Springer-Verlag, © 1993, is a good reference for the study of all kinds of delay equations. Property 4.2 on pp. 68–69 gives an example where uniqueness backward in time fails to be true even though there is uniqueness forward in time, thus illustrating one of the many ways that delay equations can be much different from ODEs.

There are also models that involve both partial derivative and delays, as in *Theory and Applications of Partial Functional Differential Equations*, by Jianhong Wu, Springer-Verlag, © 1996.

18.9.3 Problems

In problems 1–3, find the characteristic equation of the delay equation or system of equations.

1. $\frac{d}{dt}\left[\, x(t) \,\right] = ax(t) + bx\left(t - \frac{r}{2}\right) + cx(t - r)$
2. $\frac{d^2}{dt^2}\left[\, x(t) \,\right] = ax(t) + bx(t - r)$
3. $\frac{d}{dt}\left[\, \mathbf{x}(t) \,\right] = A\mathbf{x}(t) + B\mathbf{x}(t - r)$, where $\mathbf{x}$ is in $\mathbb{R}^n$ and A and B are $n \times n$ matrices.

Key Terms

almost linear: (18.21) in Section 18.3
alphabet: Section 18.8
asymptotically stable: Definition 18.4 in Section 18.2
attractor: Definition 18.3 in Section 18.2
autonomous: before (18.2) in Section 18.1
bifurcation: before Figure 18.16
bifurcation diagram: Figure 18.18(a)
closed: after Definition 18.14 in Section 18.5
closure: Definition 18.15 in Section 18.5
constant solution: Definition 18.1 in Section 18.1
continuous dependence: Theorem 18.23 in Section 18.7
differentiable dependence: Theorem 18.24 in Section 18.7
critical point: Definition 18.1 in Section 18.1
curved solution: after (18.7) in Section 18.1
entropy: Section 18.8
equilibrium point, **equilibria**: Definition 18.1 in Section 18.1
generically: Theorem 18.18 in Section 18.6
gradient system: Problems 18.4.5.22, 18.6.5.5
history: after (18.81) in Section 18.9
homoclinic orbit: Section 18.8
Hopf bifurcation: before Figure 18.17
indefinite function: Definition 18.12 in Section 18.4, Definition 18.13 in Section 18.4
invariant set: Definition 18.6 in Section 18.3
k-**dimensional surface**: Definition 18.7 in Section 18.3
Liapunov equation: (18.46) in Section 18.4
Liapunov function: Definition 18.11 in Section 18.4
Lienard equation: (18.59) in Section 18.6
limit cycle: Definition 18.22 in Section 18.6
linear variational equation: Theorem 18.24 in Section 18.7
Lipschitz condition: Definition 18.24 in Section 18.7
Lipschitz constant: Definition 18.24 in Section 18.7
linearization: (18.14) in Section 18.2, (18.55) in Section 18.6, (18.57) in Section 18.6
maximal, closed, positively invariant: Definition 18.18 in Section 18.5
negative definite: Definition 18.10 in Section 18.4, Definition 18.13 in Section 18.4
negative semi-definite: Definition 18.10 in Section 18.4, Definition 18.13 in Section 18.4
open set: before Definition 18.10 in Section 18.4
orbit: after Definition 18.22 in Section 18.6
orbitally stable: after Definition 18.22 in Section 18.6
period doubling: beginning of Section 18.8
periodic: before Definition 18.5 in Section 18.2
phase line: before Example 18.4 in Section 18.1
phase plane: before Example 18.1 in Section 18.1
Picard iterates: after (18.73) in Section 18.7
Poincaré–Andronov–Hopf bifurcation: before Figure 18.17
Poincaré map: Section 18.8
positive definite: Definition 18.10 in Section 18.4, Definition 18.13 in Section 18.4
positive semi-definite: Definition 18.10 in Section 18.4, Definition 18.13 in Section 18.4

positively invariant: Definition 18.14 in Section 18.5
punctured ball: before Definition 18.10 in Section 18.4
quadratic form: (18.36) in Section 18.4
rest point: Definition 18.1 in Section 18.1
saddle point: before (18.7) in Section 18.1
Smale horseshoe map: Section 18.8
solution: Definition 18.23 in Section 18.7
stability in the sense of Liapunov: Definition 18.2 in Section 18.2
stable: Definition 18.2 in Section 18.2, Definition 18.22 in Section 18.6
stable manifold: Definition 18.8 in Section 18.3
stable spiral point: Example 18.1 in Section 18.1
straight line solutions: after (18.7) in Section 18.1
time translate: before Definition 18.21 in Section 18.6
transversally: Section 18.8
twice continuously differentiable surface: Definition 18.7 in Section 18.3
unique up to time translation: Definition 18.21 in Section 18.6
unstable: Definition 18.2 in Section 18.2, Definition 18.22 in Section 18.6
unstable manifold: Definition 18.9 in Section 18.3
unstable spiral point: after Example 18.1 in Section 18.1
Van der Pol equation: (18.62) in Section 18.6

Mathematica Commands

s = NDSolve[{x′[t] == y[t] − x[t]^3, y′[t] == −2x[t]^3, x[0] == 1, y[0] == 0}, {x, y}, {t, −20, 20}]: after Example 18.2
a = ParametricPlot[Evaluate[{x[t], y[t]}/.s], {t, −20, 20}, PlotStyle− > {Blue, Thick}, PlotRange → {{−1, 1}, {−1.2, 1.2}}, LabelStyle → Directive[FontSize → 16], AxesLabel → {x, y}, AxesStyle → Thickness[0.00315]]: after Example 18.2
y[1, t_] := 1, y[2, t] = y[1, t] + $\int_0^t$ (y[1, s])^{2}ds, . . . , y[5, t] = y[1, t] + $\int_0^t$ (y[4, s])^{2}ds: Problem 18.7.2.3

Reference

Lorenz, E.N. Deterministic nonperiodic flow. *J. Atmos. Sci.* 20, 130–141, 1963.

Appendix A: Partial Fractions

A.1 Partial Fractions Expansions

A partial fractions expansion is a way to decompose a rational function $p(x)/q(x)$ into simpler rational functions. A rational function is **proper** if the degree of the numerator is strictly less than the degree of the denominator. For example,

$$\frac{p(x)}{q(x)} = \frac{-x^4 + 5x^2 - 17x + 11}{x(x+3)^2\left(x^2 + x + 1\right)}$$

is a proper rational function because the degree of the denominator is five and the degree of the numerator is four.

A.1.1 Real Form

First, a **real linear factor** is a first-degree polynomial $(ax + b)$, where a and b are real and $a \neq 0$. A **quadratic irreducible over the reals**, or **irreducible quadratic** for short, is a second-degree polynomial $(ax^2 + bx + c)$ that has no real roots and for which $a \neq 0$. For example, $x^2 + 2x + 5$ is an irreducible quadratic.

Given a proper rational function $p(x)/q(x)$, in principle, we can factor the denominator, $q(x)$, as a product of linear factors and irreducible quadratic factors. For example,

$$q(x) = x^5 + 7x^4 + 16x^3 + 15x^2 + 9x = \cdots = x(x+3)^2\left(x^2 + x + 1\right).$$

If the denominator $q(x)$'s factors include $(ax + b)^m$, the mth power of a linear factor, *but* not $(ax+b)^{m+1}$, then the *form* of the partial fractions expansion of a proper rational function $p(x)/q(x)$ will include the terms

$$\frac{A_1}{(ax+b)} + \frac{A_2}{(ax+b)^2} + \cdots + \frac{A_m}{(ax+b)^m}$$

and no other terms having $(ax + b)$ in a denominator. If $m = 1$, then the terms reduce to the single term $A_1/(ax + b)$.

If an irreducible quadratic $(ax^2 + bx + c)$ is a factor of $q(x)$, then the *form* of the partial fractions expansion of a proper rational function $p(x)/q(x)$ will include the term

$$\frac{C_1 x + E_1}{(ax^2 + bx + c)}.$$

If the mth power of an irreducible quadratic $(ax^2 + bx + c)$ is a factor of $q(x)$ *but* $(ax^2 + bx + c)^{m+1}$ is not a factor of $q(x)$, then the *form* of the partial fractions expansion of a proper rational function $p(x)/q(x)$ will include the terms

$$\frac{C_1x + E_1}{(ax^2 + bx + c)} + \cdots + \frac{C_mx + E_m}{(ax^2 + bx + c)^m}$$

and no other terms having $(ax^2 + bx + c)$ in a denominator.

Example A.1

The form of the partial fractions expansion of $\dfrac{-x^4 + 5x^2 - 17x + 11}{x(x+3)^2\left(x^2 + x + 1\right)}$ is

$$\frac{-x^4 + 5x^2 - 17x + 11}{x(x+3)^2\left(x^2 + x + 1\right)} = \frac{A}{x} + \frac{B_1}{(x+3)} + \frac{B_2}{(x+3)^2} + \frac{Cx + E}{\left(x^2 + x + 1\right)}, \tag{A.1}$$

where A, B_1, B_2, C, E are constants.

Example A.2

Find the partial fractions expansion of $\dfrac{x^2 - 17x + 11}{x^2(x+1)^3}$.

Method: The partial fractions expansion is

$$\frac{x^2 - 17x + 11}{x^2(x+1)^3} = \frac{A_1}{x} + \frac{A_2}{x^2} + \frac{B_1}{(x+1)} + \frac{B_2}{(x+1)^2} + \frac{B_3}{(x+1)^3}, \tag{A.2}$$

where A_1, A_2, B_1, B_2, B_3 are constants to be found. (Note that $x^2 = (1 \cdot x + 0)^2$ is a power of a linear factor.) How can we find the constants' values? The first step is to multiply through (A.2) by the denominator of the left-hand side (LHS), that is, by $q(x) = x^2(x+1)^3$. This gives

$$x^2 - 17x + 11 = A_1x(x+1)^3 + A_2(x+1)^3 + B_1x^2(x+1)^2 + B_2x^2(x+1) + B_3x^2. \tag{A.3}$$

The easiest constants to find are A_2 and B_3, by substituting in the real zeros of $q(x)$: At $x = 0$, (A.3) is

$$0^2 - 17 \cdot 0 + 11 = A \cdot 0 + A_2 \cdot (0+1)^3 + B_1 \cdot 0 + B_2 \cdot 0 + B_3 \cdot 0 = A_2,$$

so $A_2 = 11$. At $x = -1$, (A.3) is

$$29 = (-1)^2 - 17 \cdot (-1) + 11 = A_1 \cdot 0 + A_2 \cdot 0 + B_1 \cdot 0 + B_2 \cdot 0 + B_3 \cdot (-1)^2 = B_3;$$

hence, $B_3 = 29$.

To find the remaining constants A_1, B_1, B_2, there are many methods. One method is to separate the terms in (A.3) by powers of x. In principle, this gives a system of five linear equations in unknowns A_1, A_2, B_1, B_2, B_3, but we can substitute the values $A_2 = 11$ and $B_3 = 29$ that we already found by substituting in $x = 0$ and $x = -1$. That would reduce the work to solving a system of three linear equations in the three unknowns A_1, B_1, B_2.

A method we prefer is to substitute the values $A_2 = 11$ and $B_3 = 29$ into (A.3) and then keep on the right-hand side (RHS) only the terms involving the three remaining unknowns A_1, B_1, B_2:

$$x^2 - 17x + 11 = A_1x(x+1)^3 + 11(x+1)^3 + B_1x^2(x+1)^2 + B_2x^2(x+1) + 29x^2,$$

which implies

$$x^2 - 17x + 11 - 11(x^3 + 3x^2 + 3x + 1) - 29x^2$$

$$= x^2 - 17x + 11 - 11(x+1)^3 - 29x^2 = A_1x(x+1)^3 + B_1x^2(x+1)^2 + B_2x^2(x+1),$$

that is,

$$-11x^3 - 61x^2 - 50x = A_1x(x+1)^3 + B_1x^2(x+1)^2 + B_2x^2(x+1).$$

Notice that on the RHS, we can factor out $x(x+1)$. In order to have the LHS = RHS, it follows that we must be able to factor out $x(x+3)$ from the LHS. This gives

$$\frac{1}{x(x+1)} \cdot \left(-11x^3 - 61x^2 - 50x\right) = A_1(x+1)^2 + B_1x(x+1) + B_2x,$$

which can be rewritten as

$$-(11x+50) = -\frac{11x^2 + 61x + 50x}{x+1} = A_1(x+1)^2 + B_1x(x+1) + B_2x\,. \tag{A.4}$$

To solve

$$-11x - 50 = A_1(x+1)^2 + B_1x(x+1) + B_2x \tag{A.5}$$

for A_1, B_1, B_2, substitute in $x = 0$ to get $-50 = A_1$, and substitute $x = -1$ into (A.5) to get $-39 = -B_2$; hence, $B_2 = 39$. Substitute the values of A_1 and B_2 into (A.5) to get

$$-11x - 50 = -50(x+1)^2 + B_1x(x+1) + 39x,$$

so

$$-11x - 50 + 50(x^2 + 2x + 1) - 39x = B_1x(x+1);$$

hence,

$$50 = \frac{50x^2 + 50x}{x(x+1)} = B_1.$$

So, the partial fractions expansion is

$$\frac{x^2 - 17x + 11}{x^2(x+1)^3} = \frac{-50}{x} + \frac{11}{x^2} + \frac{50}{(x+1)} + \frac{39}{(x+1)^2} + \frac{29}{(x+1)^3} \cdot \bigcirc$$

Note that if we had made an error before arriving at (A.5), we might have heard an "alarm bell ringing." For example, if instead of (A.5) we had

$$-11x - 50 + \frac{\neq 0}{x(x+1)} = A_1(x+1)^2 + B_1x(x+1) + B_2x,$$

then we would have reached a contradiction, because the polynomial on the RHS cannot equal a non-polynomial on the LHS. So, if the division that arrives at (A.5) has a nonzero remainder, then an alarm bell has rung.

When an alarm bell such as that rings, it tells us that an error has happened somewhere; but, unfortunately, it does not say what went wrong.

Example A.3

Find the partial fractions expansion of $(x^2 - 3x + 5)/((x+1)(x^2+x+1))$.

Method: The partial fractions expansion is

$$\frac{x^2 - 3x + 5}{(x+1)(x^2+x+1)} = \frac{A}{(x+1)} + \frac{Cx + E}{(x^2+x+1)}, \tag{A.6}$$

where A, C, E are constants to be found. How? The first step is to multiply through (A.6) by the denominator of the LHS, that is, by $q(x) = (x+1)(x^2+x+1)$. This gives

$$x^2 - 3x + 5 = A(x^2+x+1) + (Cx+E)(x+1). \tag{A.7}$$

Substitute $x = -1$ into (A.7) to get $9 = A$, and then substitute that into (A.7) to get

$$x^2 - 3x + 5 = 9(x^2+x+1) + (Cx+E)(x+1). \tag{A.8}$$

After that, we could leave on the RHS of (A.8) only what involves C and E:

$$-8x^2 - 12x - 4 = x^2 - 3x + 5 - 9(x^2+x+1) = (Cx+E)(x+1);$$

hence,

$$-8x - 4 = \frac{-8x^2 - 12x - 4}{x+1} = (Cx+E).$$

Because there is no remainder after the division, an alarm bell did not ring.

So, the partial fractions expansion is $(x^2 - 3x + 5)/((x+1)(x^2+x+1)) = 9A/(x+1) + (-8x-4)/(x^2+x+1)$. ◯

An alternative method of solving (A.7), that is, $x^2 - 3x + 5 = A(x^2+x+1) + (Cx+E)(x+1)$, is to sort by powers of x:

$$1 \cdot x^2 - 3x + 5 = A(x^2+x+1) + (Cx+E)(x+1) = Ax^2 + Ax + A + Cx^2 + Cx + Ex + E$$
$$= (A+C)x^2 + (A+C+E)x + (A+E).$$

The x^2 terms matched between the LHS and the RHS require that $1 = A + C$; matching the x^1 terms requires that $-3 = A + C + E$. Matching the x^0 terms requires that $5 = A + E$. So, we would need to solve the system of three equations in three unknowns given in compact form by

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} A \\ C \\ E \end{bmatrix} = \begin{bmatrix} 1 \\ -3 \\ 5 \end{bmatrix}.$$

So,

$$\begin{bmatrix} A \\ C \\ E \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ -3 \\ 5 \end{bmatrix} = \begin{bmatrix} 9 \\ -8 \\ -4 \end{bmatrix}.$$

This gives the same conclusion as in Example A.3. Sorting by powers of x can be a useful method. Its only disadvantage is that there is no alarm bell that can ring. If we use partial fractions as one step in the middle of a long problem, it is a good idea to check that the conclusion of the partial fractions expansion is correct.

A.1.2 Complex Form

If $p(z)/q(z)$ is a proper rational function, then in principle* in practice, we can factor $g(z)$ into a product of linear factors $(z - a)$. So, we may choose not to worry about quadratic factors that are irreducible over the reals and instead use only linear factors, possibly involving complex numbers.

Example A.4

Find the complex partial fractions expansion of $(4z - i7)/(z^3 + 2z^2 + 5z)$.

Method: $(4z - i7)/(z^3 + 2z^2 + 5z)$

$$= \frac{4z - i7}{z(z + 1 - i2)(z + 1 + i2)} = \frac{A}{z} + \frac{B}{z + 1 - i2} + \frac{C}{z + 1 + i2}$$

$$4z - i7 = A(z + 1 - i2)(z + 1 + i2) + Bz(z + 1 + i2) + Cz(z + 1 - i2).$$

Substituting in $z = 0$, $z = -1 + i2$, and $z = -1 - i2$ successively yields

$$A = \frac{-i7}{5}, \quad B = \frac{-4 + i}{(-1 + i2)i4} = \cdots = \frac{7 - i6}{20}, \quad \text{and}$$

$$C = \frac{-4 - i15}{(-1 - i2)(-i4)} = \cdots = \frac{-7 + i34}{20}.$$

So, we have the partial fractions expansion

$$\frac{4z - i7}{z^3 + 2z^2 + 5z} = \frac{1}{20}\left(\frac{-i28}{z} + \frac{7 - i6}{z + 1 - i2} + \frac{-7 + i34}{z + 1 + i2}\right). \bigcirc$$

* But, as we've seen in Chapter 8, solving a nonlinear equation may be difficult, even if the equation is of the form *polynomial* $= 0$, if the polynomial has a high degree.

Appendix B: Laplace Transforms Definitions and Derivations

B.1 Laplace Transforms

Definition B.1

$\mathcal{L}[f(t)\,](s) \triangleq \int_0^\infty f(t)e^{-st}\,dt$, for those values of s for which the improper integral converges.

Definition B.2

A function $f(t)$ is of **exponential order α** if there is a positive constant M such that $|f(t)| \leq Me^{\alpha t}$ for all $t \geq 0$. Here, α is a real number.

Theorem B.1

If $f(t)$ is of exponential order α, then the improper integral $\mathcal{L}[f(t)](s)$ converges for real numbers $s > \alpha$.

B.1.1 Properties and Their Derivations

Note that implicitly, each identity is asserted to be true only when both sides of the identity exist.

1. $\mathcal{L}[f(t) + g(t)](s) = \mathcal{L}[f(t)\,](s) + \mathcal{L}[\,g(t)](s)$.

   Derivation: For those s for which improper integrals $\int_0^\infty f(t)e^{-st}\,dt$ and $\int_0^\infty g(t)e^{-st}\,dt$ both exist, there exists

$$\begin{aligned}\mathcal{L}[f(t)\,](s) + \mathcal{L}[\,g(t)](s) &= \lim_{b\to\infty}\int_0^b f(t)e^{-st}\,dt + \lim_{b\to\infty}\int_0^b g(t)e^{-st}\,dt \\ &= \lim_{b\to\infty}\left(\int_0^b f(t)e^{-st}\,dt + \int_0^b g(t)e^{-st}\,dt\right) = \lim_{b\to\infty}\int_0^b \big(f(t)+g(t)\big)e^{-st}\,dt \\ &\triangleq \int_0^\infty \big(f(t)+g(t)\big)e^{-st}\,dt \triangleq \mathcal{L}[f(t)+g(t)\,](s)\,.\ \square\end{aligned}$$

2. $\mathcal{L}[\, cf(t)\,](s) = c\mathcal{L}[f(t)\,](s)$, any constant c.

   Derivation: For those s for which the improper integral $\int_0^\infty f(t)e^{-st}\,dt$ exists, there exists

$$c\mathcal{L}[f(t)\,](s) = c\cdot\lim_{b\to\infty}\int_0^b f(t)e^{-st}\,dt = \lim_{b\to\infty}\int_0^b cf(t)e^{-st}\,dt \triangleq \int_0^\infty cf(t)e^{-st}\,dt$$

$$\triangleq \mathcal{L}[\, cf(t)](s)\,. \;\square$$

3. $\mathcal{L}[\, e^{at}\,](s) = \frac{1}{s-a}$.

   Derivation: For $s > a$, there exists

$$\mathcal{L}[\, e^{at}\,](s) \triangleq \lim_{b\to\infty}\int_0^b e^{at}e^{-st}\,dt = \lim_{b\to\infty}\int_0^b e^{-(s-a)t}\,dt = \lim_{b\to\infty}\Big[\,\frac{-1}{s-a}\,e^{-(s-a)t}\,\Big]_0^b$$

$$= \lim_{b\to\infty}\frac{-1}{s-a}\left(e^{-(s-a)b}-1\right) = \frac{-1}{s-a}\,(0-1) = \frac{1}{s-a}\,. \;\square$$

4. $\mathcal{L}[\,\dot{y}\,](s) = s\mathcal{L}[\, y(t)\,](s) - y(0)$.

   Derivation: Integration by parts gives

$$\mathcal{L}[\,\dot{y}\,](s) \triangleq \lim_{b\to\infty}\int_0^b \dot{y}(t)e^{-st}\,dt = \lim_{b\to\infty}\left(\Big[\, y(t)\,e^{-st}\,\Big]_0^b - \int_0^b y(t)\big(-se^{-st}\big)\,dt\right)$$

$$= \lim_{b\to\infty}\left(y(b)\,e^{-sb} - y(0) + s\int_0^b y(t)e^{-st}\,dt\right) = 0 - y(0) + s\int_0^\infty y(t)e^{-st}\,dt$$

$$= s\mathcal{L}[\, y(t)\,](s) - y(0)\,. \;\square$$

5. $\mathcal{L}[\, t^n\,](s) = n!/s^n,\ n = 0, 1, 2, \ldots$.

   Derivation: For $n = 0$,

$$\mathcal{L}[\, t^0\,](s) = \lim_{b\to\infty}\int_0^b 1\cdot e^{-st}\,dt = \lim_{b\to\infty}\Big[\,\frac{e^{-st}}{-s}\,\Big]_0^b = \lim_{b\to\infty}\left(\frac{e^{-sb}}{-s} + \frac{1}{s}\right) = \frac{1}{s}\,.$$

For integer $n \geq 1$, define $a_n \triangleq \mathcal{L}[\, t^n \,](s)$. Integration by parts gives

$$a_n \triangleq \lim_{b\to\infty} \int_0^b t^n e^{-st}\, dt = \lim_{b\to\infty} \left[\, t^n \cdot \frac{e^{-st}}{-s} \,\right]_0^b - \int_0^b n\, t^{n-1} \cdot \frac{e^{-st}}{-s}\, dt$$

$$= \lim_{b\to\infty} \left[\, t^n \cdot \frac{e^{-st}}{-s} \,\right]_0^b + \frac{n}{s} \cdot \mathcal{L}[\, t^{n-1}](s)$$

$$= \lim_{b\to\infty} \left(-\frac{b^n}{3e^{sb}} - 0 \right) + \frac{n}{s} \cdot a_{n-1} = 0 + \frac{n}{s} \cdot a_{n-1},$$

using L'Hôpital's rule. Induction explains why $a_n = n!/s^n$. □

6. $\mathcal{L}[\, \cos \omega t \,](s) = \dfrac{s}{s^2 + \omega^2}$

Derivation: Using (3.9) in Section 3.1, that is, $\int e^{at} \cos bt dt = \dfrac{1}{a^2 + b^2}\, e^{at}(b \sin bt + a \cos bt) + c$,

$$\mathcal{L}[\, \cos \omega t \,](s) \triangleq \lim_{b\to\infty} \int_0^b \cos(\omega t) e^{-st}\, dt = \lim_{b\to\infty} \left[\frac{e^{-st}}{(-s)^2 + \omega^2} \left(\omega \sin \omega t - s \cos \omega t\right) \right]_0^b$$

$$= \frac{1}{s^2 + \omega^2} \lim_{b\to\infty} \left(e^{-sb}(\omega \sin \omega b - s \cos \omega b) + s \right)$$

$$= \frac{1}{s^2 + \omega^2}\, (0 + s) = \frac{s}{s^2 + \omega^2}\, . \, \square$$

7. $\mathcal{L}[\, \sin \omega t \,](s) = \dfrac{\omega}{s^2 + \omega^2}$.

Derivation: Using (3.10) in Section 3.1, that is, $\int e^{at} \sin bt dt = \dfrac{1}{a^2 + b^2}\, e^{at}(a \sin bt - b \cos bt) + c$,

$$\mathcal{L}[\, \sin \omega t \,](s) \triangleq \lim_{b\to\infty} \int_0^b \sin(\omega t) e^{-st}\, dt = \lim_{b\to\infty} \left[\frac{e^{-st}}{(-s)^2 + \omega^2} \left(-s\, \sin \omega t - \omega \cos \omega t\right) \right]_0^b$$

$$= \frac{1}{s^2 + \omega^2} \lim_{b\to\infty} \left(e^{-sb}(-s \sin \omega b - \omega \cos \omega b) + \omega \right)$$

$$= \frac{1}{s^2 + \omega^2}\, (0 + \omega) = \frac{\omega}{s^2 + \omega^2}\, . \, \square$$

8. $\mathcal{L}[\, e^{at}f(t)](s) = F(s-a)$, where $F(s) = \mathcal{L}[f(t)\,](s)$.

   Derivation:

$$\mathcal{L}[\, e^{at}f(t)\,](s) \triangleq \lim_{b\to\infty}\int_0^b e^{at}f(t)e^{-st}\,dt = \lim_{b\to\infty}\int_0^b f(t)e^{-(s-a)t}\,dt$$

$$= \left(\lim_{b\to\infty}\int_0^b f(t)e^{-st}\,dt\right)\Big|_{s\mapsto(s-a)} = \mathcal{L}[\, e^{at}f(t)\,]\Big|_{s\mapsto(s-a)} = F(s-a),$$

   where $F(s) = \mathcal{L}[f(t)](s)$. □

9. $\mathcal{L}[\,\ddot{y}\,](s) = s^2\mathcal{L}[\,y](s) - sy(0) - \dot{y}(0)$.

   Derivation: Using property (4),

$$\mathcal{L}[\,\ddot{y}](s) = \mathcal{L}\Big[\,\frac{d}{dt}\big[\,\dot{y}\,\big]\,\Big](s) = s\mathcal{L}[\,\dot{y}(t)](s) - \dot{y}(0) = s\big(s\mathcal{L}[\,y(t)](s) - y(0)\big) - \dot{y}(0)$$

$$= s^2\mathcal{L}[\,y](s) - sy(0) - \dot{y}(0).\ \square$$

10. $\mathcal{L}[\,\mathrm{step}(t-c)](s) = \dfrac{e^{-cs}}{s}$.

    Derivation: For $s > 0$,

$$\mathcal{L}[\,\mathrm{step}(t-c)](s) \triangleq \lim_{b\to\infty}\int_0^b \mathrm{step}(t-c)e^{-st}\,dt = \lim_{b\to\infty}\left(\int_0^c 0\cdot e^{-st}\,dt + \int_c^b 1\cdot e^{-st}\,dt\right)$$

$$= \lim_{b\to\infty}\left(0 + \Big[\,\frac{e^{-st}}{-s}\Big]_c^b\right) = \lim_{b\to\infty}\left(\frac{e^{-sc}}{-s} - \frac{1}{-s}\right) = \frac{e^{-cs}}{s}\,.\ \square$$

11. $\mathcal{L}[\,g(t)\,\mathrm{step}(t-c)](s) = e^{-cs}\cdot\mathcal{L}[\,g(t+c)](s)$.

    Derivation:

$$\mathcal{L}[\,g(t)\,\mathrm{step}(t-c)](s) \triangleq \lim_{b\to\infty}\int_0^b g(t)\mathrm{step}(t-c)e^{-st}\,dt$$

$$= \lim_{b\to\infty}\left(\int_0^c 0\cdot g(t)\cdot e^{-st}\,dt + \int_c^b 1\cdot g(t)e^{-st}\,dt\right).$$

Using the change of variables $\tau = t - c$,

$$\mathcal{L}[\,g(t)\text{step}(t-c)](s) = \lim_{b\to\infty} \int_0^{b-c} 1 \cdot g(\tau + c) e^{-s(\tau+c)}\, d\tau = \lim_{b\to\infty} \int_0^{b-c} g(\tau + c) e^{-s\tau}\, e^{-cs}\, d\tau$$

$$= e^{-cs} \lim_{(b-c)\to\infty} \int_0^{b-c} g(\tau + c) e^{-s\tau}\, d\tau = e^{-cs} \lim_{(b-c)\to\infty} \int_0^{b-c} g(t + c) e^{-st}\, dt$$

$$\triangleq e^{-cs}\mathcal{L}[\,g(t+c)](s)\,. \;\square$$

12. $\mathcal{L}[(f * g)(t)](s) = \mathcal{L}[f(t)](s) \cdot \mathcal{L}[\,g(t)](s)$.

Derivation:

$$\mathcal{L}[(f * g)(t)](s) \triangleq \lim_{b\to\infty} \int_0^b \left(\int_0^t f(t-u) g(u) du \right) e^{-st}\, dt.$$

The triangle in the (t, u)-plane described by $0 \le t \le b, 0 \le u \le t$ can be described as $0 \le u \le b, u \le t \le b$, so changing the order of integration gives

$$\mathcal{L}[(f * g)(t)](s) = \lim_{b\to\infty} \int_0^b \left(\int_u^b f(t-u) g(u)\, dt \right) e^{-st}\, du$$

$$= \lim_{b\to\infty} \int_0^b \left(\int_u^b f(t-u)\, e^{-s(t-u)}\, e^{-su}\, dt \right) g(u)\, du.$$

The change of variables $\tau = t - u$ in the inner integral implies

$$\mathcal{L}[(f * g)(t)](s) = \lim_{b\to\infty} \int_0^b \left(\int_0^{b-u} f(\tau) e^{-s\tau}\, d\tau \right) g(u) e^{-su}\, du$$

$$= \lim_{b\to\infty} \int_0^b \big(\approx F(s)\big) g(u) e^{-su}\, du,$$

where the $\approx$ acknowledges that the inner integral is approaching $F(s) \triangleq \mathcal{L}[f(t)](s)$ as $b \to \infty$. So,

$$\mathcal{L}[(f * g)(t)](s) = \lim_{b\to\infty} \big(\approx F(s)\big) \int_0^b g(u) e^{-su}\, du = F(s)G(s). \;\square$$

13. $\mathcal{L}[\,t\sin\omega t](s) = 2\omega s/(s^2+\omega^2)^2$.

Derivation: Using integration by parts with $u = t$ and $dv = e^{-st}\sin\omega t\,dt$, and (3.10) in Section 3.1, that is,

$$\int e^{at}\sin bt dt = \frac{1}{a^2+b^2}\,e^{at}\big(a\sin bt - b\cos bt\big) + c,$$

$$\begin{aligned}
\mathcal{L}[\,t\sin\omega t](s) &\triangleq \lim_{b\to\infty}\int_0^b t\sin\omega t\,e^{-st}\,dt \\
&= \lim_{b\to\infty}\left(\left[\,t\cdot\frac{e^{-st}}{(-s)^2+\omega^2}\,(-s\sin\omega t - \omega\cos\omega t)\,\right]_0^b\right. \\
&\quad \left. -\int_0^b \frac{e^{-st}}{(-s)^2+\omega^2}\,(-s\sin\omega t - \omega\cos\omega t)\;dt\right) \\
&= \lim_{b\to\infty}\left(\frac{be^{-sb}}{s^2+\omega^2}\,\big(-s\sin\omega b - \omega\cos\omega b\big) - 0\right) \\
&\quad - \frac{1}{s^2+\omega^2}\,(-s\mathcal{L}[\,\sin\omega t](s) - \omega\mathcal{L}[\,\cos\omega t](s)) \\
&= 0 - \frac{1}{s^2+\omega^2}\left(-s\cdot\frac{\omega}{s^2+\omega^2} - \omega\cdot\frac{s}{s^2+\omega^2}\right) = \frac{2\omega s}{(s^2+\omega^2)^2}\,.\ \square
\end{aligned}$$

14. $\mathcal{L}[\,t\cos\omega t](s) = (s^2-\omega^2)/(s^2+\omega^2)^2$.

Derivation: Using integration by parts with $u = t$ and $dv = e^{-st}\cos\omega t\,dt$, and (3.9) in Section 3.1, that is,

$$\int e^{at}\cos bt dt = \frac{1}{a^2+b^2}\,e^{at}\,\big(b\sin bt + a\cos bt\big) + c,$$

$$\begin{aligned}
\mathcal{L}[\,t\cos\omega t](s) &\triangleq \lim_{b\to\infty}\int_0^b t\cos\omega t\,e^{-st}\,dt \\
&= \lim_{b\to\infty}\left(\left[\,t\cdot\frac{e^{-st}}{(-s)^2+\omega^2}\,(\omega\sin\omega t - s\cos\omega t)\,\right]_0^b\right. \\
&\quad \left. -\int_0^b \frac{e^{-st}}{(-s)^2+\omega^2}\,(\omega\sin\omega t - s\cos\omega t)\;dt\right) \\
&= \lim_{b\to\infty}\left(\frac{be^{-sb}}{s^2+\omega^2}\,\big(\omega\sin\omega b - s\cos\omega b\big) - 0\right)
\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{s^2+\omega^2}\big(\omega\mathcal{L}[\sin\omega t](s) - s\mathcal{L}[\cos\omega t](s)\big) \\
&= 0 - \frac{1}{s^2+\omega^2}\left(\omega\cdot\frac{\omega}{s^2+\omega^2} - s\cdot\frac{s}{s^2+\omega^2}\right) = \frac{s^2-\omega^2}{(s^2+\omega^2)^2}\,. \square
\end{aligned}
$$

15. $\mathcal{L}[\,\delta(t-c)](s) = e^{-cs}$.
Derivation: Using "approximate delta functions," we calculate

$$
\begin{aligned}
\mathcal{L}[\,\delta(t-c)](s) &= \lim_{n\to\infty}\mathcal{L}[\,\delta_n(t)](s) \\
&= \lim_{n\to\infty}\mathcal{L}\left[n\cdot\left(\operatorname{step}\Big(t-\big(c-\frac{1}{2n}\big)\Big) - \operatorname{step}\Big(t-\big(c+\frac{1}{2n}\big)\Big)\right)\right](s) \\
&= \lim_{n\to\infty} n\cdot\frac{1}{s}\cdot\left(e^{-\left(c-\frac{1}{2n}\right)s} - e^{-\left(c+\frac{1}{2n}\right)s}\right).
\end{aligned}
$$

Substitute $\varepsilon = 1/2n$ and use L'Hôpital's rule to get

$$
\begin{aligned}
\mathcal{L}[\,\delta(t-c)](s) &= \frac{1}{s}\cdot\lim_{\varepsilon\to 0^+}\frac{e^{-\left(c-\varepsilon\right)s} - e^{-\left(c+\varepsilon\right)s}}{2\varepsilon} = \frac{1}{s}\cdot\lim_{\varepsilon\to 0^+}\frac{se^{-\left(c-\varepsilon\right)s} + se^{-\left(c+\varepsilon\right)s}}{2} \\
&= \frac{1}{s}\cdot\frac{se^{-cs}+se^{-cs}}{2} = e^{-cs}\,. \square
\end{aligned}
$$

16. $\mathcal{L}[f(t)](s) = \dfrac{1}{1-e^{-s\mathsf{T}}}\displaystyle\int_0^{\mathsf{T}} f(t)e^{-st}\,dt$, for $s>0$, if $f(t)$ is periodic with period T.
Derivation:

$$
\begin{aligned}
\mathcal{L}[f(t)](s) &\triangleq \lim_{b\to\infty}\int_0^b f(t)e^{-st}\,dt = \lim_{N\to\infty}\int_0^{(N+1)\mathsf{T}} f(t)e^{-st}\,dt \\
&= \lim_{N\to\infty}\sum_{n=0}^{N}\int_{n\mathsf{T}}^{(n+1)\mathsf{T}} f(t)e^{-st}\,dt.
\end{aligned}
$$

Using periodicity and the substitution $\tau = t - n\mathsf{T}$, we have

$$
\begin{aligned}
\int_{n\mathsf{T}}^{(n+1)\mathsf{T}} f(t)e^{-st}\,dt &= \int_0^{\mathsf{T}} f(\tau+n\mathsf{T})e^{-s(\tau+n\mathsf{T})}\,d\tau = \int_0^{\mathsf{T}} f(\tau)e^{-s\tau}\,e^{-ns\mathsf{T}}\,d\tau \\
&= e^{-ns\mathsf{T}}\int_0^{\mathsf{T}} f(\tau)e^{-s\tau}\,d\tau.
\end{aligned}
$$

So, noting that $|e^{-s\mathsf{T}}| < 1$ for $s > 0$, using the sum of a geometric series, we calculate

$$\mathcal{L}[f(t)](s) = \lim_{N\to\infty} \sum_{n=0}^{N} \left(e^{-ns\mathsf{T}} \int_0^{\mathsf{T}} f(\tau) e^{-s\tau}\, d\tau \right) = \left(\int_0^{\mathsf{T}} f(\tau) e^{-s\tau}\, d\tau \right) \lim_{N\to\infty} \sum_{n=0}^{N} \left(e^{-s\mathsf{T}} \right)^n$$

$$= \left(\int_0^{\mathsf{T}} f(\tau) e^{-s\tau}\, d\tau \right) \sum_{n=0}^{\infty} \left(e^{-s\mathsf{T}} \right)^n = \left(\int_0^{\mathsf{T}} f(\tau) e^{-s\tau}\, d\tau \right) \cdot \frac{1}{1 - e^{-s\mathsf{T}}} \,. \; \square$$

When s is a complex variable, all of the aforementioned results are also true if $s > \alpha$ is replaced by $\mathcal{R}e(s) > \alpha$.

Appendix C: Series Solutions of ODEs

C.1 Power Series Solutions of ODEs

We will start with a simple example that demonstrates the basic methods.

Example C.1

Find a power series solution of the ODE $\dot{y} = 2y$.

Method: Substitute a series solution in the form

$$y(t) = \sum_{n=0}^{\infty} c_n t^n \tag{C.1}$$

into the ODE to get

$$\dot{y} = \sum_{n=0}^{\infty} n\, c_n t^{n-1} = \sum_{n=0}^{\infty} 2c_n t^n = 2y\,. \tag{C.2}$$

But, on the left-hand side, we really have

$$\dot{y} = \frac{d}{dt}\left[\, c_0 t^0 + c_1 t^1 + c_2 t^2 + c_3 t^3 + \cdots \right] = 0 + c_1 t^0 + 2c_2 t^1 + 3c_3 t^2 + \cdots\,,$$

which can be rewritten as

$$\dot{y} = \sum_{n=0}^{\infty} (n+1) c_{n+1} t^n.$$

So, to solve our ODE means to solve

$$\sum_{n=0}^{\infty} (n+1) c_{n+1} t^n = \sum_{n=0}^{\infty} 2c_n t^n,$$

that is,

$$\sum_{n=0}^{\infty} \big((n+1)c_{n+1} - 2c_n\big)\, t^n = 0,$$

for the coefficients $c_0, c_1, c_2, \ldots$. This implies the **recursion relation**

$$(n+1)c_{n+1} - 2c_n = 0,$$

that is,

$$c_{n+1} = \frac{2}{n+1}\, c_n,$$

which is actually a first-order difference equation. Unfortunately, we cannot use a characteristic equation to help solve it. But fortunately, it is simple enough that we can guess the solution: Because

$$c_1 = \frac{2}{1}\, c_0 = 2c_0, \quad c_2 = \frac{2}{2}\, c_1 = \frac{2^2}{2}\, c_0, \quad \text{and} \quad c_3 = \frac{2}{3}\, c_2 = \frac{2^3}{3 \cdot 2}\, c_0, \ldots,$$

we guess that

$$c_n = \frac{2^n}{n!}\, c_0,\ n = 1, 2, \ldots$$

We can easily check that it satisfies the recursion relation. Thus, the solutions of the ODE are

$$y(t) = \sum_{n=0}^{\infty} c_n t^n = c_0 t^0 + c_1 t^1 + c_2 t^2 + c_3 t^3 + \cdots = c_0 \cdot \left(1 + \frac{2^1}{1!}\, t^1 + \frac{2^2}{2!}\, t^2 + \frac{2^3}{3!}\, t^3 + \cdots\right)$$

$$= c_0 \left(1 + \frac{2t}{1!} + \frac{(2t)^2}{2!} + \frac{(2t)^3}{3!} + \cdots\right) = c_0 \cdot e^{2t},$$

where c_0 is an arbitrary constant. Of course, we knew this already, but it is good to see that our new method gives a result that agrees with our old method for solving $\dot{y} = 2y$. ◯

Example C.2

Find a power series solution of the ODE $\ddot{y} + 4y = 0$.

Method: Substitute a series solution in the form

$$y(t) = \sum_{n=0}^{\infty} c_n t^n \tag{C.3}$$

into the ODE to get

$$\sum_{n=0}^{\infty} n(n-1)\, c_n t^{n-2} + \sum_{n=0}^{\infty} 4c_n t^n = 0\,. \tag{C.4}$$

But the first term is

$$\ddot{y} = \frac{d^2}{dt^2}\left[c_0 t^0 + c_1 t^1 + c_2 t^2 + c_3 t^3 + \cdots\right] = 0 + 0 + 2 \cdot 1 c_2 t^0 + 3 \cdot 2 c_3 t^1 + \cdots$$

$$= \sum_{n=2}^{\infty} n \cdot (n-1)\, c_n t^{n-2}\,,$$

which can be rewritten as

$$\ddot{y} = \sum_{n=0}^{\infty} (n+2)(n+1) c_{n+2} t^n.$$

So, to solve our ODE means to solve

$$\sum_{n=0}^{\infty} (n+2)(n+1)c_{n+2}t^n + \sum_{n=0}^{\infty} 4c_n t^n = 0,$$

that is,

$$\sum_{n=0}^{\infty} \big((n+2)(n+1)c_{n+2} + 4c_n\big)\, t^n = 0,$$

for the coefficients $c_0, c_1, c_2, \ldots$. This implies the **recursion relation**

$$(n+2)(n+1)c_{n+2} + 4c_n = 0,$$

that is,

$$c_{n+2} = -\frac{4}{(n+2)(n+1)}\, c_n,$$

which is actually a second-order difference equation. Unfortunately, we cannot use a characteristic equation to help solve it. Fortunately, it is simple enough that we can guess the solution: Because

$$c_2 = -\frac{4}{2\cdot 1}\, c_0, \quad c_4 = -\frac{4}{4\cdot 3}\, c_2 = \frac{4^2}{4\cdot 3\cdot 2\cdot 1}\, c_0, \quad \text{and} \quad c_6 = -\frac{4}{6\cdot 5}\, c_4 = \cdots,$$

we guess that

$$c_{2k} = (-1)^k\, \frac{2^{2k}}{(2k)!}\, c_0,\ k = 1, 2, \ldots.$$

We can easily check that it satisfies the recursion relation. Thus, among the solutions of the ODE are

$$y(t) = \sum_{k=0}^{\infty} c_{2k}t^{2k} = c_0t^0 + c_2t^2 + c_4t^4 + c_6t^6 + \cdots$$

$$= c_0 \cdot \left(1 - \frac{2^2}{2!}\, t^1 + \frac{2^4}{4!}\, t^4 - \frac{2^6}{6!}\, t^6 + \cdots\right) = c_0 y_1(t),$$

where c_0 is an arbitrary constant and $y_1(t) \triangleq \cos 2t$. Of course, we knew this already, but it's good to see that our new method gives a result that agrees with our old method for solving $\ddot{y} + 4y = 0$.

But, wait, there are other solutions: Because

$$c_3 = -\frac{4}{3\cdot 2}\, c_1, \quad c_5 = -\frac{4}{5\cdot 4}\, c_3 = \frac{4^2}{5\cdot 4\cdot 3\cdot 2}\, c_1, \quad \text{and} \quad c_7 = -\frac{4}{7\cdot 6}\, c_5 = \cdots,$$

we guess that

$$c_{2k+1} = (-1)^k\, \frac{2^{2k}}{(2k+1)!}\, c_1,\ k = 1, 2, \ldots.$$

We can easily check that it satisfies the recursion relation. Thus, among the solutions of the ODE are

$$y(t) = \sum_{k=0}^{\infty} c_{2k+1} t^{2k+1} = c_1 t^1 + c_3 t^3 + c_3 t^3 + c_5 t^5 + \cdots$$

$$= c_1 \cdot \left(t^1 - \frac{2^2}{3!} t^3 + \frac{2^4}{5!} t^5 - \frac{2^6}{7!} t^7 + \cdots \right)$$

$$= \frac{1}{2} c_1 \cdot \left(2t - \frac{2^3}{3!} t^3 + \frac{2^5}{5!} t^5 - \frac{2^7}{7!} t^7 + \cdots \right) = c_1 y_2(t),$$

where c_1 is an arbitrary constant and $y_2(t) \triangleq \frac{1}{2} \sin 2t$. Of course, we knew this already, but it's good to see that our new method gives a result that agrees with our old method for solving $\ddot{y} + 4y = 0$.

Putting all of the solutions together, we see that the general solution of the ODE $\ddot{y} + 4y = 0$ is

$$y(t) = \sum_{k=0}^{\infty} c_{2k}\, t^{2k} + \sum_{k=0}^{\infty} c_{2k+1}\, t^{2k+1} = c_0 y_1(t) + c_1 y_2(t) = c_0 \cdot \cos 2t + c_1 \cdot \frac{1}{2} \sin 2t,$$

where the functions $y_1(t)$ and $y_2(t)$ were defined earlier. ○

Interestingly, the two solutions $y_1(t)$ and $y_2(t)$ are the basic solutions predicted theoretically by Theorem 3.9 in Section 3.3.

C.1.1 Convergence

First, we recall some definitions and results concerning

$$\sum_{n=0}^{\infty} a_n, \tag{C.5}$$

a series of real numbers.

Definition C.1

(a) The **sequence of partial sums** is $\{s_N\}_{N=0}^{\infty}$, defined by

$$s_N \triangleq \sum_{n=0}^{N} a_n. \tag{C.6}$$

(b) Series (C.5) **converges** if the sequence of partial sums converges.

(c) Series (C.5) **converges absolutely** if the series

$$\sum_{n=0}^{\infty} |a_n|$$

converges.

(d) Series (C.5) **diverges** if the sequence of partial sums does not converge.

Theorem C.1

Suppose $\sum_{n=0}^{\infty} a_n$ and $\sum_{n=0}^{\infty} b_n$ are absolutely convergent series and $\sum_{n=0}^{\infty} h_n$ is not absolutely convergent. Then for all complex constants β and γ,

(a) $\sum_{n=0}^{\infty}(\beta a_n + \gamma b_n)$ is absolutely convergent.
(b) $\sum_{n=0}^{\infty}(\beta a_n + \gamma h_n)$ is not absolutely convergent unless $\gamma = 0$.

Given a series $\sum_{k=1}^{\infty} a_k$, *if possible*, we calculate

$$L \triangleq \lim_{k\to\infty} \left| \frac{a_{k+1}}{a_k} \right| . \tag{C.7}$$

Theorem C.2

(a) If $L < 1$ exists, then series (C.5) converges absolutely.
(b) If $L > 1$ exists, then series (C.5) diverges.

Note that Theorem C.2 does not discuss what happens if L does not exist or if $L = 1$.

Definition C.2

(a) A finite number R is called the **radius of convergence** for power series

$$\sum_{n=0}^{\infty} c_n t^n, \tag{C.8}$$

if it both (a) converges absolutely for $|t - t_0| < R$ and (b) diverges for $|t - t_0| > R$.

(b) The radius of convergence of power series (C.8) is **defined to be** $R = \infty$, and we say the radius of convergence is infinite, if the power series (C.8) converges at all t.

Given a power series (C.8), for any fixed value of t, we can define the corresponding series of real numbers

$$\sum_{n=0}^{\infty} a_n t^n,$$

where $a_n \triangleq c_n t^n$.

Theorem C.3

Suppose there exists

$$\alpha \triangleq \lim_{k\to\infty} \left| \frac{c_{k+1}}{c_k} \right| . \tag{C.9}$$

(a) If $0 < \alpha < \infty$, then the power series (C.8) converges absolutely for $|z - z_0| < R \triangleq \frac{1}{\alpha}$ and diverges for $|z - z_0| > R$, hence has radius of convergence equal to R.

(b) If $\alpha = 0$, then the power series (C.8) converges at all z, hence has infinite radius of convergence.

Example C.3

Find the radii of convergence of the two solutions of $\ddot{y} + 4y = 0$ we found in C.2.

Method: The first solution is

$$y_1(t) = c_0 \sum_{k=0}^{\infty} C_k t^{2k} = c_0 \cdot \cos 2t,$$

where we define $C_k \triangleq c_{2k}/c_1 = (-1)^k \frac{2^{2k}}{(2k)!}$. We calculate

$$\alpha \triangleq \lim_{k\to\infty} |C_{k+1} \div C_k| = \lim_{k\to\infty} \left| (-1)^{k+1} \frac{2^{2(k+1)}}{(2(k+1))!} \div \left((-1)^k \frac{2^{2k}}{(2k)!} \right) \right|$$

$$= \lim_{k\to\infty} \left| (-1) \frac{2^2}{(2k+2)(2k+1)} \right| = 0,$$

so $y_1(t)$ has radius of convergence $R = \infty$.

The second solution is

$$y_2(t) = c_1 \sum_{k=0}^{\infty} (-1)^k \frac{2^{2k}}{(2k+1)!} t^{2k+1} = \frac{1}{2} c_1 \cdot \sin 2t,$$

where we define $C_k \triangleq = c_{2k+1}/c_1 = (-1)^k \frac{2^{2k}}{(2k+1)!}$. We could use similar work to explain why $y_2(t)$ has radius of convergence $R = \infty$. Instead, let us use the recursion relation to find α:

$$\alpha \triangleq \lim_{k\to\infty} \left|C_{k+1} \div C_k\right| = \lim_{k\to\infty} \left|c_{2(k+1)+1} \div c_{2k+1}\right| = \lim_{k\to\infty} \left|c_{(2k+1)+2} \div c_{2k+1}\right|$$

$$= \lim_{k\to\infty} \left| -\frac{4}{(2k+1+2)(2k+1+1)}\, c_{2k+1} \div c_{2k+1} \right| = \lim_{k\to\infty} \left| \frac{4}{(2k+3)(2k+2)} \right| = 0.$$

So, the radius of convergence of $y_2(t)$ is $R = \infty$. ◯

C.1.2 Nonconstant Coefficients ODEs

We know from Chapter 3 how to solve all constant coefficients scalar ODEs. The power series method gives us a way to systematically look for solutions of nonconstant coefficients ODEs.

To do that, it helps to become skilled at some manipulations of power series. The first manipulation involves multiplying a power series by a polynomial.

Example C.4

If $y(t) = \sum_{n=0}^{\infty} c_n t^n$, rewrite $(t+2)\dot{y}$ in terms of power series that uses t^n in the summation(s).

Method: Without rigorous analysis of convergence, we have

$$(t+2)\dot{y} = (t+2)\frac{d}{dt}\left[\sum_{n=0}^{\infty} c_n t^n\right] = (t+2)\sum_{n=0}^{\infty} \frac{d}{dt}\left[c_n t^n\right] = (t+2)\sum_{n=1}^{\infty} n c_n t^{n-1}.$$

Note that in the sum, the term for $n = 0$ drops out because $nc_n t^n = 0$ when $n = 0$. Another way of seeing this is to note that the $n = 0$ term in $y(t) = \sum_{n=0}^{\infty} c_n t^n$ is the constant term $c_0 t^0 = c_0$, whose derivative with respect to t is zero.

Continuing, we calculate

$$(t+2)\dot{y} = t\sum_{n=1}^{\infty} n c_n t^{n-1} + 2\sum_{n=1}^{\infty} n c_n t^{n-1} = \sum_{n=1}^{\infty} n c_n t^n + \sum_{n=1}^{\infty} 2 n c_n t^{n-1}.$$

The first term is a power series that uses t^n in the summation, but the second is not. To fix this, we use a shift of index by substituting $m = n - 1$, that is, $n = m + 1$:

$$\sum_{n=1}^{\infty} 2 n c_n t^{n-1} = \sum_{m=0}^{\infty} 2(m+1) c_{m+1} t^m,$$

after noting that $m+1 = n = 1, 2, \ldots$ corresponds to $m = 0, 1, \ldots$. To get back to n-powers rather m-powers, substitute $n = m$ to get

$$\sum_{m=0}^{\infty} 2(m+1) c_{m+1} t^m = \sum_{n=0}^{\infty} 2(n+1) c_{n+1} t^n.$$

Putting the two terms back together, we have

$$(t+2)\dot{y} = \sum_{n=1}^{\infty} n c_n t^n + \sum_{n=0}^{\infty} 2(n+1) c_{n+1} t^n \,. \bigcirc \tag{C.10}$$

When we solve an ODE in Example C.5 as follows, we will want to rewrite (C.10) in terms of a single power series that starts at $n = 1$. To do that, we "peel off" the first term from the second series in (C.10):

$$\sum_{n=0}^{\infty} 2(n+1)c_{n+1}t^n = 2(0+1)c_{0+1}t^0 + \sum_{n=1}^{\infty} 2(n+1)c_{n+1}t^n.$$

So, (C.10) can be rewritten as

$$(t+2)\dot{y} = \sum_{n=1}^{\infty} nc_n t^n + 2c_1 + \sum_{n=1}^{\infty} 2(n+1)c_{n+1}t^n = 2c_1 + \sum_{n=1}^{\infty} \left(nc_n + 2(n+1)c_{n+1}\right) t^n .$$

Example C.5

Find a power series solution of the ODE $(t+2)\dot{y} + y = 0$.

Method: Substitute the usual power series solution in the form (C.4), that is, $y(t) = \sum_{n=0}^{\infty} c_n t^n$, into the ODE and do the manipulations as in Example C.3:

$$0 = (t+2)\dot{y} + y = (t+2)\frac{d}{dt}\left[\sum_{n=0}^{\infty} c_n t^n\right] + \sum_{n=0}^{\infty} c_n t^n$$

$$= t\sum_{n=1}^{\infty} nc_n t^{n-1} + 2\sum_{n=1}^{\infty} nc_n t^{n-1} + \sum_{n=0}^{\infty} c_n t^n$$

$$= \sum_{n=1}^{\infty} nc_n t^n + \sum_{n=1}^{\infty} 2nc_n t^{n-1} + \sum_{n=0}^{\infty} c_n t^n$$

$$= \sum_{n=1}^{\infty} nc_n t^n + \sum_{n=0}^{\infty} 2(n+1)c_{n+1}t^n + \sum_{n=0}^{\infty} c_n t^n .$$

Recall that the second term was written in terms of powers t^n rather than powers t^{n-1} by doing a change of index from n to $m = n-1$ and then back to n.

In order to get a recursion relation, we need to combine all of the series into one series. In order to do that, we peel off all of the t^0 terms to get

$$0 = (t+2)\dot{y} + y = \sum_{n=1}^{\infty} nc_n t^n + 2c_1 + \sum_{n=1}^{\infty} 2(n+1)c_{n+1}t^n + c_0 + \sum_{n=1}^{\infty} c_n t^n .$$

So, the ODE is

$$0 = (t+2)\dot{y} + y = 2c_1 + c_0 + \sum_{n=1}^{\infty} \left(nc_n + 2(n+1)c_{n+1} + c_n\right) t^n .$$

Sorting by powers of t gives

$$0 = 2c_1 + c_0 \tag{C.11}$$

and

$$0 = nc_n + 2(n+1)c_{n+1} + c_n, \quad \text{for } n = 1, 2, \ldots .$$

The latter can be rewritten as

$$0 = (n+1)c_n + 2(n+1)c_{n+1}, \quad \text{for } n = 1, 2, \ldots;$$

hence,

$$c_{n+1} = -\frac{(n+1)}{2(n+1)}\, c_n = -\frac{1}{2}\, c_n, \quad \text{for } n = 1, 2, \ldots . \tag{C.12}$$

Combining (C.11) and (C.12), this situation is simple enough that we can guess the solution: Because

$$c_1 = -\frac{1}{2}\, c_0, \quad c_2 = -\frac{2}{2}\, c_1 = \frac{(-1)^2}{2^2}\, c_0, \quad \text{and} \quad c_3 = -\frac{1}{2}\, c_2 = \frac{(-1)^3}{2^3}\, c_0, \ldots,$$

we guess that

$$c_n = \frac{(-1)^n}{2^n}\, c_0, \; n = 1, 2, \ldots .$$

We can easily check that it satisfies the recursion relation. Thus, the solutions of the ODE are

$$y(t) = \sum_{n=0}^{\infty} c_n t^n = c_0 t^0 + c_1 t^1 + c_2 t^2 + c_3 t^3 + \cdots = c_0 \cdot \left(1 + \frac{(-1)^1}{2^1}\, t^1 + \frac{(-1)^2}{2^2}\, t^2 + \cdots\right)$$

$$= c_0 \sum_{n=0}^{\infty} \left(\frac{-t}{2}\right)^n = c_0 \cdot \frac{1}{1 - (-t/2)} = c_0 \cdot \frac{2}{2+t}\,.$$

where c_0 is an arbitrary constant. The solutions of $(t+2)\dot{y} + y = 0$ are

$$y(t) = c_0 y_1(t),$$

where c_0 is an arbitrary constant and $y_1(t) \triangleq \frac{2}{2+t}$. ◯

In fact, this first-order ODE can be solved by the method of separation of variables and gives the same conclusion. Note that using the recursion relation, we can calculate that the radius of convergence of $y_1(t)$ is

$$R = \left(\lim_{k\to\infty} \left|c_{k+1} \div c_k\right|\right)^{-1} = \left(\lim_{k\to\infty} \left|-\frac{1}{2}\, c_k \div c_k\right|\right)^{-1} = \left(\lim_{k\to\infty} \left|-\frac{1}{2}\right|\right)^{-1} = 2.$$

This makes sense because $y_1(t)$ blows up at $t = -2$.

Example C.6

Find a power series solution of the ODE $(1+x^2)y'' + 10xy' + 20y = 0$.

This problem is more complicated than the previous examples because the ODE is second-order *and* it has nonconstant coefficients.

Method: Substitute the usual power series solution in the form (C.3), that is, $y(x) = \sum_{n=0}^{\infty} c_n x^n$, into the ODE and do manipulations similar to those in Example C.3 and afterward:

$$0 = (1+x^2)y'' + 10xy' + 20y = \left((1+x^2) \sum_{n=2}^{\infty} c_n n(n-1)x^{n-2} \right)$$

$$+ 10x \sum_{n=1}^{\infty} c_n n x^{n-1} + 20 \sum_{n=0}^{\infty} c_n x^n$$

$$= \left(\sum_{n=2}^{\infty} n(n-1)c_n x^{n-2} + \sum_{n=2}^{\infty} n(n-1)c_n x^n \right) + \sum_{n=1}^{\infty} 10 n c_n x^n + \sum_{n=0}^{\infty} 20 c_n x^n$$

$$= \sum_{n=0}^{\infty} (n+2)(n+1)c_{n+2}x^n + \sum_{n=2}^{\infty} n(n-1)c_n x^n + \sum_{n=1}^{\infty} 10 n c_n x^n + \sum_{n=0}^{\infty} 20 c_n x^n$$

$$= 2c_2 + 6c_3 x + \sum_{n=2}^{\infty} (n+2)(n+1)c_{n+2}x^n + \sum_{n=2}^{\infty} n(n-1)c_n x^n + 10c_1 x + \sum_{n=2}^{\infty} 10 n c_n x^n$$

$$+ 20c_0 + 20c_1 x + \sum_{n=0}^{\infty} 20 c_n x^n$$

$$= 2c_2 + 20c_0 + (6c_3 + 10c_1 + 20c_1)\, x$$

$$+ \sum_{n=2}^{\infty} \left((n+2)(n+1)c_{n+2} + n(n-1)c_n + 10 n c_n + 20 c_n \right) x^n .$$

Sorting by powers of x gives

$$0 = 2c_2 + 20c_0, \quad 0 = 6c_3 + 10c_1 + 20c_1 = 6c_3 + 30c_1, \tag{C.13}$$

and

$$0 = (n+2)(n+1)c_{n+2} + n(n-1)c_n + 10 n c_n + 20 c_n, \text{ for } n = 2, 3, \ldots .$$

The latter can be rewritten as

$$c_{n+2} = -\frac{n(n-1) + 10n + 20}{(n+2)(n+1)}\, c_n = -\frac{(n+5)(n+4)}{(n+2)(n+1)}\, c_n, \text{ for } n = 2, 3, \ldots . \tag{C.14}$$

Combining (C.13) and (C.14), this situation is simple enough that we can guess a solution: Because

$$c_2 = -\frac{20}{2}\,c_0 = -\frac{5\cdot 4}{2\cdot 1}\,c_0 \quad \text{and} \quad c_4 = -\frac{7\cdot 6}{4\cdot 3}\,c_2 = \left(-\frac{7\cdot 6}{4\cdot 3}\right)\left(-\frac{5\cdot 4}{2\cdot 1}\,c_0\right)$$

$$= (-1)^2\,\frac{7\cdot 6\cdot 5\cdot 4}{4\cdot 3\cdot 2\cdot 1}\,c_0,$$

$$c_6 = -\frac{9\cdot 8}{6\cdot 5}\,c_4 = \left(-\frac{9\cdot 8}{6\cdot 5}\right)\left(\frac{7\cdot 6\cdot 5\cdot 4}{4\cdot 3\cdot 2\cdot 1}\,c_0\right) = (-1)^3\,\left(\frac{9\cdot 8\cdot 7\cdot 6\cdot 5\cdot 4}{6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1}\right)c_0\,,$$

we guess that

$$c_{2k} = (-1)^k\,\frac{(2k+3)!}{3!(2k)!}\,c_0 = (-1)^k\,\frac{(2k+3)\cdot(2k+2)\cdot(2k+1)}{3\cdot 2\cdot 1}\,c_0,\ k = 1, 2, \ldots.$$

So, the ODE has solutions $y(x) = c_0 y_1(x)$, where c_0 is an arbitrary constant and

$$y_1(x) = \frac{1}{c_0}\sum_{k=0}^{\infty} c_{2k}\,x^{2k} = \frac{1}{6}\sum_{k=0}^{\infty}(-1)^k\,(2k+3)(2k+2)(2k+1)x^{2k}\,.$$

Similarly, the odd powers naturally group together to give

$$c_3 = -\frac{30}{6}\,c_1 = -\frac{6\cdot 5}{3\cdot 2}\,c_1, \quad c_5 = -\frac{8\cdot 7}{5\cdot 4}\,c_3 = \left(-\frac{8\cdot 7}{5\cdot 4}\right)f\left(\frac{6\cdot 5}{3\cdot 2}\,c_1\right) = (-1)^2\,\frac{8\cdot 7\cdot 6\cdot 5}{5\cdot 4\cdot 3\cdot 2}\,c_1,$$

$$\text{and} \quad c_7 = -\frac{10\cdot 9}{7\cdot 6}\,c_5 = \left(-\frac{10\cdot 9}{7\cdot 6}\right)\left(\frac{8\cdot 7\cdot 6\cdot 5}{5\cdot 4\cdot 3\cdot 2}\,c_0\right) = (-1)^3\,\left(\frac{10\cdot 9\cdot 8\cdot 7\cdot 6\cdot 5}{7\cdot 6\cdot 5\cdot 4\cdot 3\cdot 2}\right)c_0.$$

We guess that

$$c_{2k+1} = (-1)^k\,\frac{(2k+4)!}{4!(2k+1)!}\,c_1 = (-1)^k\,\frac{(2k+4)\cdot(2k+3)\cdot(2k+2)}{4\cdot 3\cdot 2\cdot 1}\,c_1, k = 1, 2, \ldots\,.$$

So, the ODE has solutions $y(x) = c_1 y_2(x)$, where c_1 is an arbitrary constant and

$$y_2(x) = \frac{1}{c_1}\sum_{k=0}^{\infty} c_{2k+1}\,x^{2k+1} = \frac{1}{24}\sum_{k=0}^{\infty}(-1)^k\,(2k+4)(2k+3)(2k+2)x^{2k}.$$

The general solution of the ODE is

$$y(x) = \sum_{k=0}^{\infty} c_{2k}\,x^{2k} + \sum_{k=0}^{\infty} c_{2k+1}\,x^{2k+1} = c_0 y_1(x) + c_1 y_2(x),$$

where the functions $y_1(x)$ and $y_2(x)$ were defined earlier in terms of power series.

C.1.3 Power Series in $(x - a)$

It may make sense to take the solution form to be a power series in powers of $(x - a)$ rather than powers of $(x - 0)$, that is, to assume

$$y(x) = \sum_{n=0}^{\infty} c_n (x-a)^n.$$

C.2 Method of Frobenius

The method of Frobenius is a modification of the power series solution method that turns out to be very useful for classical special functions, such as Bessel functions, used in mathematical physics.

In a sense, the method is a generalization of the method we used in Section 3.5 to solve Cauchy–Euler ODEs

$$x^2y'' + p_0xy' + q_0y = 0, \text{ where } ' = \frac{d}{dx} \text{ and } p_0, q_0 \text{ are constants,}$$

in Section 3.5. Recall that we substituted in a desired solution in the form $y(x) = x^n$, a single power of x, and we got a characteristic equation $n(n-1) + p_0n + q_0 = 0$, assuming p_0 and q_0 are constants.

Definition C.3

A function $p(x)$ is **real analytic** on an open interval $-\delta < x < \delta$ if there is a power series whose radius of convergence is greater than or equal to δ, for which $p(x)$ equals that power series for all x in the interval $-\delta < x < \delta$.

Definition C.4

An ODE of the form

$$x^2y'' + xp(x)y' + q(x)y = 0 \tag{C.15}$$

has $x = 0$ as a **regular singular point** if $p(x)$ and $q(x)$ are real analytic on an open interval $-\delta < x < \delta$.

The **method of Frobenius** consists of substituting a solution in the form

$$y(x) = x^n \sum_{m=0}^{\infty} c_m x^m \tag{C.16}$$

into ODE (C.15).

Expand $p(x)$ and $q(x)$ in power series:

$$p(x) = \sum_{m=0}^{\infty} p_m x^m, \quad \text{and} \quad q(x) = \sum_{m=0}^{\infty} q_m x^m.$$

Substitute (C.16), that is, $y(x) = \sum_{m=0}^{\infty} c_m\, x^{m+n}$, into ODE (C.15) to get

$$\begin{aligned}
0 &= x^2 \cdot \sum_{m=0}^{\infty} (m+n)(m+n-1)c_m\, x^{m+n-2} \\
&\quad + \Big(p_0 + p_1 x + p_2 x^2 + \cdots\Big) \cdot x \cdot \sum_{m=0}^{\infty} (m+n)c_m\, x^{m+n-1} \\
&\quad + \Big(q_0 + q_1 x + q_2 x^2 + \cdots\Big) \cdot \sum_{m=0}^{\infty} c_m\, x^{m+n} \\
&= \sum_{m=0}^{\infty} (m+n)(m+n-1)c_m\, x^{m+n} + p_0 \sum_{m=0}^{\infty} (m+n)c_m\, x^{m+n} \\
&\quad + p_1 x \cdot \sum_{m=0}^{\infty} (m+n)c_m\, x^{m+n} + \cdots + q_0 \sum_{m=0}^{\infty} c_m\, x^{m+n} + q_1 x \cdot \sum_{m=0}^{\infty} c_m\, x^{m+n} + \cdots .
\end{aligned}$$

Sort by powers of x and isolate the x^{0+n} terms, which correspond to $m = 0$, to get

$$0 = x^n \cdot \big((0+n)(0+n-1)c_0 + p_0(0+n)c_0 + q_0 c_0\big) = c_0 x^n \cdot \big(n(n-1) + np_0 + q_0\big).$$

This implies that n, the power of x that multiplies the power series, must satisfy the **indicial equation**

$$n(n-1) + np_0 + q_0 = 0. \tag{C.17}$$

This is a generalization of the characteristic equation for Cauchy–Euler equations. Why? Because if $p(x) \equiv p_0$ and $q(x) \equiv q_0$ are constants, then ODE (C.15) is a Cauchy–Euler equation whose characteristic equation is (C.17).

Note that

$$p_0 = \lim_{x \to 0} p(x) \quad \text{and} \quad q_0 = \lim_{x \to 0} q(x)\,, \tag{C.18}$$

because we assumed $p(x)$ and $q(x)$ are real analytic on $-\delta < x < \delta$.

Theorem C.4

Suppose $x = 0$ is a regular singular point for ODE (C.15), where $p(x)$ and $q(x)$ are real analytic on interval $-\delta < x < \delta$ and p_0, q_0 are as in (C.18). Suppose n_1 and n_2 are the two solutions of indicial equation (C.17), with $\mathcal{R}e(n_1) \geq \mathcal{R}e(n_2)$. Then ODE (C.15) has a basic solution in the form

$$y_1(x) = |x|^{n_1} \sum_{m=0}^{\infty} c_m x^m$$

that is convergent for $0 < |x| < \delta$ and has $c_0 = 1$.

In addition, there is a second basic solution $y_2(x)$ that is convergent for $0 < |x| < \delta$:

(a) If $n_1 - n_2$ is not an integer, then $y_2(x)$ is in the form

$$y_2(x) = |x|^{n_2} \sum_{m=0}^{\infty} \widetilde{c}_m x^m,$$

where $\widetilde{c}_0 = 1$.

(b) If $n_1 = n_2$, which we call the **first exceptional case**, then $y_2(x)$ is in the form

$$y_2(x) = |x|^{n_1+1} \sum_{m=0}^{\infty} \hat{c}_m x^m + (\ln|x|) y_1(x),$$

where $\hat{c}_0 \neq 0$.

(c) If $n_1 - n_2$ is a nonzero integer, which we call the **second exceptional case**, then $y_2(x)$ is in the form

$$y_2(x) = |x|^{n_2} \sum_{m=0}^{\infty} \check{c}_m x^m + c(\ln|x|) y_1(x),$$

where $\check{c}_0 \neq 0$ and c is a constant, possibly equal to zero.

C.2.1 Bessel's Equation of Order α

The ODE

$$x^2 y'' + xy' + (x^2 - \alpha^2) y = 0 \tag{C.19}$$

is called **Bessel's equation of order** α, if α is a nonnegative real constant.

Example C.7

For Bessel's equation of order one, find a power series solution involving only positive powers.

Method: The ODE is

$$x^2 y'' + xy' + (x^2 - 1) y = 0, \tag{C.20}$$

and it has a regular singular point at $x = 0$. The indicial equation is

$$0 = n(n-1) - n - 1 = n^2 - 1,$$

so the method of Frobenius says there is a solution of the form

$$y(x) = x^1 \cdot \sum_{m=0}^{\infty} c_m x^m = \sum_{m=0}^{\infty} c_m x^{m+1}, \text{ for } 0 < x < \delta, \tag{C.21}$$

for some $\delta > 0$. Substitute this into ODE (C.20) to get

$$0 = x^2 \cdot \sum_{m=1}^{\infty} (m+1)mc_m\, x^{m-1} + x \cdot \sum_{m=0}^{\infty} (m+1)c_m\, x^m + (x^2-1) \cdot \sum_{m=0}^{\infty} c_m\, x^{m+1}$$

$$= \sum_{m=1}^{\infty} (m+1)mc_m\, x^{m+1} + \sum_{m=0}^{\infty} (m+1)c_m\, x^{m+1} + \sum_{m=0}^{\infty} c_m\, x^{m+3} - \sum_{m=0}^{\infty} c_m\, x^{m+1}.$$

Because so many of the series are in terms of powers x^{m+1}, it makes sense to get all of the series in terms of powers x^{m+1} instead of powers x^m. This gives us

$$0 = \sum_{m=1}^{\infty} (m+1)mc_m\, x^{m+1} + \sum_{m=0}^{\infty} (m+1)c_m\, x^{m+1} + \sum_{m=2}^{\infty} c_{m-2}\, x^{m+1} - \sum_{m=0}^{\infty} c_m\, x^{m+1}.$$

Among the four series, the highest *first* power is x^{2+1}, that is, x^{m+1} for $m = 2$. So, we peel off x^{0+1} and x^{1+1} powers so that all of the series will start with the index $m = 2$:

$$0 = 2c_1x^2 + \sum_{m=2}^{\infty} (m+1)mc_m\, x^{m+1} + c_0x + 2c_1x^2 + \sum_{m=2}^{\infty} (m+1)c_m\, x^{m+1} + \sum_{m=2}^{\infty} c_{m-2}\, x^{m+1}$$

$$- c_0x - c_1x^2 - \sum_{m=2}^{\infty} c_m\, x^{m+1}$$

$$= (c_0 - c_0)x + (2c_1 + 2c_1 - c_1)x^2 + \sum_{m=2}^{\infty} \left((m+1)mc_m + (m+1)c_m + c_{m-2} - c_m\right) x^{m+1}$$

$$= 3c_1x^2 + \sum_{m=2}^{\infty} \left((m+2)mc_m + c_{m-2}\right) x^{m+1}.$$

Sorting by powers of x gives

$$c_1 = 0 \tag{C.22}$$

and

$$(m+2)mc_m + c_{m-2} = 0 \text{ for } m = 2, 3, \ldots.$$

The latter can be rewritten as

$$c_{m+2} = -\frac{1}{(m+4)(m+2)}\, c_m, \text{ for } m = 2, 3, \ldots. \tag{C.23}$$

Combining (C.22) and (C.23), we can guess the solution: Note first that

$$c_2 = -\frac{1}{4 \cdot 2}\, c_0 = -\frac{1}{2 \cdot 2 \cdot (2 \cdot 1)}\, c_0 \quad \text{and} \quad c_3 = -\frac{1}{5 \cdot 3}\, c_1 = 0.$$

Using (C.23), we see that all of the odd-indexed coefficients are zero, and the even-numbered coefficients are

$$c_4 = -\frac{1}{6 \cdot 4}\, c_2 = (-1)^2\, \frac{1}{6 \cdot 4} \cdot \frac{1}{2 \cdot 2 \cdot (2 \cdot 1)}\, c_0 = (-1)^2\, \frac{1}{2 \cdot 2 \cdot (3 \cdot 2) \cdot 2 \cdot 2 \cdot (2 \cdot 1)}\, c_0$$

$$= (-1)^2\, \frac{1}{2^4} \cdot \frac{1}{3!2!}\, c_0.$$

Next, we have

$$c_6 = -\frac{1}{8 \cdot 6} c_4 = -\frac{1}{2^2 \cdot 4 \cdot 3} (-1)^2 \frac{1}{2^4} \cdot \frac{1}{3!2!} c_0 = (-1)^3 \frac{1}{2^6} \cdot \frac{1}{4!3!} c_0 .$$

We guess that

$$c_{2k} = \frac{(-1)^k}{2^{2k}(k+1)!k!}.$$

So, the ODE has solution $y(x) = c_0 y_1(x)$, where c_0 is an arbitrary constant and

$$y_1(x) = x \cdot \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(k+1)!} \left(\frac{x}{2}\right)^{2k} . \bigcirc$$

The **Bessel function of the first kind and order one** is defined by

$$J_1(x) \triangleq \frac{x}{2} \cdot \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(k+1)!} \left(\frac{x}{2}\right)^{2k} = \frac{1}{2} y_1(x).$$

More generally, for any nonnegative real number α, the **Bessel function of the first kind and order** α is defined by

$$J_\alpha(x) \triangleq \left(\frac{x}{2}\right)^{\alpha} \cdot \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\Gamma(k+\alpha+1)} \left(\frac{x}{2}\right)^{2k} .$$

Using the method of Frobenius, $J_\alpha(x)$ is derived to be the first solution, $y_1(x)$, of Bessel's equation of order α, that is,

$$x^2 y'' + xy' + (x^2 - \alpha^2) y = 0, \tag{C.24}$$

corresponding to the greater root $n = \alpha$ of the indicial equation $n(n+1) - n - \alpha^2 = 0$. Note that the **gamma function** is defined by

$$\Gamma(z) \triangleq \int_0^{\infty} e^{-x} x^{z-1} \, dx \tag{C.25}$$

and satisfies $\Gamma(\ell + 1) = \ell!$ for nonnegative integers ℓ.

The second solution of Bessel's equation of order α is notated $Y_\alpha(x)$. If α is a nonnegative integer, we have the first exceptional case (when $\alpha = 0$) or the second exceptional case of Theorem C.4. It turns out that

$$\lim_{x \to 0^+} |Y_\alpha(x)| = \infty$$

when α is a nonnegative integer.

The **modified Bessel equation of order** α is

$$x^2y'' + xy' - (x^2 + \alpha^2)y = 0 \tag{C.26}$$

and has solutions $I_\alpha(x)$ and $K_\alpha(x)$, called **modified Bessel functions of order** α of the first and second kind, respectively. They have the useful properties that $I_\alpha(x) > 0$ for all $x > 0$ and $\lim_{x\to 0^+} |K_\alpha(x)| = \infty$.

C.3 Polynomial Solutions of ODEs

Sometimes when we substitute a power series solution into an ODE, the recursion relation gives us a solution where all of the coefficients of the power series are zero past a certain index. Such a solution is a polynomial. It turns out that such polynomials are very useful classical special functions, such as Legendre polynomials, used in mathematical physics.

Example C.8

For Legendre's equation

$$(1 - x^2)y'' - 2xy' + \nu(\nu + 1)y = 0, \tag{C.27}$$

where ν is a constant parameter, along with boundary conditions

$$|y(-1^+)| < \infty, |y(1^-)| < \infty, \tag{C.28}$$

find the five polynomial solutions of lowest degree.

Method: Substitute the usual power series (C.3), that is,

$$y(x) = \sum_{n=0}^{\infty} c_n x^n,$$

into the ODE to get

$$\begin{aligned}
0 &= (1 - x^2) \sum_{n=2}^{\infty} n(n-1)c_n x^{n-2} - 2x \sum_{n=1}^{\infty} nc_n x^{n-1} + \nu(\nu+1) \sum_{n=0}^{\infty} c_n x^n \\
&= \sum_{n=2}^{\infty} n(n-1)c_n x^{n-2} - \sum_{n=2}^{\infty} n(n-1)c_n x^n - \sum_{n=1}^{\infty} 2nc_n x^n + \sum_{n=0}^{\infty} \nu(\nu+1)c_n x^n \\
&= \sum_{n=0}^{\infty} (n+2)(n+1)c_{n+2} x^n - \sum_{n=2}^{\infty} n(n-1)c_n x^n - \sum_{n=1}^{\infty} 2nc_n x^n + \sum_{n=0}^{\infty} \nu(\nu+1)c_n x^n \\
&= 2c_2 + 6c_3 x + \sum_{n=2}^{\infty} (n+2)(n+1)c_{n+2} x^n - \sum_{n=2}^{\infty} n(n-1)c_n x^n - 2c_1 x - \sum_{n=2}^{\infty} 2nc_n x^n \\
&\quad + \nu(\nu+1)c_0 + \nu(\nu+1)c_1 x + \sum_{n=2}^{\infty} \nu(\nu+1)c_n x^n .
\end{aligned}$$

Sorting by powers of x gives

$$0 = 2c_2 + \nu(\nu+1)c_0, \quad 0 = 6c_3 - 2c_1 + \nu(\nu+1)c_1 \tag{C.29}$$

and

$$0 = (n+2)(n+1)c_{n+2} - n(n-1)c_n - 2nc_n + \nu(\nu+1)c_n \text{ for } n = 2, 3, \ldots.$$

The latter can be rewritten as the recursion relation

$$c_{n+2} = \frac{n(n+1) - \nu(\nu+1)}{(n+2)(n+1)}\, c_n, \text{ for } n = 0, 1, \ldots. \tag{C.30}$$

For any positive integer n, when $\nu = n$, the recursion relation shows that $c_{2k+2} = 0$ for all $k \geq n$. We have even polynomials

$$\begin{aligned}
&\textit{For } \nu = 0, \quad y(x) = c_0.\\
&\textit{For } \nu = 2, \quad y(x) = c_0 + c_2x^2 = c_0\left(1 - 3x^2\right).\\
&\textit{For } \nu = 4, \quad y(x) = c_0 + c_2x^2 = c_0\left(1 - 10x^2 + \frac{70}{6}x^4\right).
\end{aligned}$$

and odd polynomials

$$\begin{aligned}
&\textit{For } \nu = 1, \quad y(x) = c_1.\\
&\textit{For } \nu = 3, \quad y(x) = c_1x + c_3x^3 = c_1\left(x - \frac{5}{3}x^3\right). \bigcirc
\end{aligned}$$

The corresponding **Legendre polynomials** $P_n(x)$ are constant multiples of these polynomials "normalized" to have $P_n(1) = 1$:

$$\begin{aligned}
&\textit{For } \nu = 0, \quad P_0(x) \equiv 1.\\
&\textit{For } \nu = 1, \quad P_1(x) = x.\\
&\textit{For } \nu = 2, \quad P_2(x) = -\frac{1}{2}\left(1 - 3x^2\right).\\
&\textit{For } \nu = 3, \quad P_3(x) = -\frac{3}{2}\left(x - \frac{5}{3}x^3\right).\\
&\textit{For } \nu = 4, \quad P_4(x) = \frac{3}{8}\left(1 - 10x^2 + \frac{35}{3}x^4\right).
\end{aligned}$$

Learn More about It

The *Handbook of Mathematical Functions*, edited by Milton Abramowitz and Irene Stegun, Dover Publications © 1964, is a good reference for all special functions, such as Bessel functions of all kinds and Legendre polynomials. *An Introduction to Ordinary Differential Equations*, by Earl A. Coddington, Prentice-Hall © 1961, and *Intermediate Differential Equations*, 2nd edition, by Earl D. Rainville, The Macmillan Co. © 1964, are very good textbooks for learning about special functions, including the explanation for Theorem C.4 on the method of Frobenius. *The Theory of Bessel Functions*, by G. N. Watson, Cambridge University Press © 1922 (reprinted 1996), is a standard reference for all things concerning those functions.

C.3.1 Problems

In problems 1–4, find the recursion relation when a power series about $x = 0$ is used to solve

1. $(2x^2 + 3)y'' - 5xy' + 7y = 0$
2. $(x^2 + 1)y'' + 5xy' - xy = 0$
3. $(x^2 + 3)y' + y = 0$
4. Bessel's equation of order α, where α is an unspecified nonnegative number

In problems 5–9, solve the IVP.

5. $(x^2 + 1)y'' + 6y' + 6y = 0,\ y(0) = -1,\ y'(0) = 0$
6. $(1 - x^2)y'' - 2xy' + 20y = 0,\ y(0) = -2,\ y'(0) = 0$ [Hint: Use the initial conditions to find c_0, c_1 *early* in your work.]
7. $(x^2 + 1)y'' + 6y' + 6y = 0,\ y(0) = 0,\ y'(0) = 1$
8. $(x^2 + 1)y'' - 3y + 3y = 0,\ y(0) = 0,\ y'(0) = 7$
9. $(x^2 + 1)y'' - 26y = 0,\ y(0) = 0,\ y'(0) = 1$
10. Express the general solution of $y'' + xy' + 2y = 0$ in the form $y(x) = c_0 z_0(x) + c_1 z_1(x)$. Write down formulas for the functions $z_0(x)$ and $z_1(x)$.

In problems 11 and 12, find polynomial approximations of two linearly independent solutions $z_0(x)$ and $z_1(x)$. Use all powers of x up to and including x^8.

11. $(1 + 2x^2)y'' + 6xy' + 2y = 0$
12. $(x^2 + 1)y'' + 6xy' + 4y = 0$
13. Solve $(x^2 + 5)y'' + xy' - y = 0$. State clearly which constants are arbitrary. State the interval of validity of the solution.
14. Solve $(x^2 + 4)y'' + 6xy' + 4y = 0$. State clearly which constants are arbitrary. State the interval of validity of the solution.

In problems 15–17, find a complete set of basic solutions of the ODE and find the radius of convergence of your choice of one of those solutions.

15. $(x^2 + 1)y'' + 6xy' - 6y = 0$
16. $x^2 y'' - 2y = 0$
17. Bessel's equation of order two
18. Suppose that $\psi(z) \triangleq J_n(z)$ satisfies ODE $z\frac{d}{dz}(z\frac{d\psi}{dz}) + (z^2 - n^2)\psi(z) = 0,\ z = e^x$, and $y(x) \triangleq \psi(e^x)$. What ODE does $y(x)$ satisfy?

Index

A

Abel's theorem, 238, 332–333, 371, 426
Adams–Bashforth method, 705
Adams–Moulton formula, 706
Adaptive step size, 708
Adjugate matrix, 53–54, 79–82
Aitken's δ^2 method, 656–657
Algebraic inverse, 183
Algebraic multiplicity, 81, 132, 138–139, 387
Ampère—Maxwell's law, 840
Ampère's circuital law, 566
Amplification factor, 1003
Amplitude, 232–235, 268–269, 275, 277–278, 284
Analog-to-digital converter (A/D or ADC), 340
Angle of attack, 1257
Annihilator method, 93
Anti-derivative, 551–552
Anti-partial derivative, 572, 601
Areal coordinates, 731–732
Augmented matrix, 5–6, 10, 40–41, 63

B

Banach contraction mapping theorem, 666
Banach spaces, 666
Barycentric coordinates, 732
Beats phenomenon, 274, 283–287
Bendixson–Dulac theorem, 1352
Bessel function of the first kind and order one, 1402
Bessel's equation of order α, 1400–1403
Bessel's equation of order zero, 714
Bi-harmonic operator, 1299
Block matrices, 43, 153–154
Boole's rule, 671–673
Bounded as $t \to \infty$, 202–203
Bounded function, 202
Bounded linear functionals and operator adjoint, 186–187
Brachistochrone problem, 1083–1085
Bromwich contour, 1285, 1287
B-spline curve, 729–731
Butterfly effect, 1364
BVP, *see* Two-point boundary value problem (BVP)

C

Calculus of variations
  constraints
    differential equation, 1103–1105
    integral, 1100
    isoperimetric problem, 1100
    Lagrange multiplier (λ), 1101–1103
  eigenvalue problems
    Laplacian, 1108–1109
    ODE-BVP, 1106–1108
    Rayleigh–Ritz method, 1106
    Sturm–Liouville problem, 1109–1110
  Euler–Lagrange equations
    Dirichlet boundary condition, 1093
    divergence theorem, 1093
    Fourier series, 1090
    linear approximation, 1090
    variation, 1089–1090
    variational method, 1091
  finite element methods (*see* Finite element methods)
  Hamilton's principle
    continuous media, 1097
    Hamilton's equations, 1095
    Lagrangian, 1095
    pendulum, 1095
    total action during time interval, 1096
  MacLaurin series, 1088
  minimization
    admissible function, 1081
    brachistochrone problem, 1083–1085
    functional, 1081
    natural boundary conditions, 1082
    nonlinear ODE-BVP, 1083
    potential energy, 1082–1083
    Rayleigh–Ritz method, 1085–1087
  natural boundary condition, 1082, 1094–1095
  stationarity, 1088
Camber line, 1255
Camber ratio, 1255, 1257–1258
Casorati determinant, 425–427
  forward difference operator, 330
  higher order difference operators, 331
  n-th order linear homogeneous difference equation, 332
  Wronskian determinant, 331
Cauchy complete, 183–185
Cauchy equation, 845

Cauchy–Euler ODEs
boundary conditions, 259
characteristic polynomial, 255–258
standard form, 255
Cauchy–Goursat integral theorem, 1200–1204, 1278
Cauchy principal value, 1218–1219
Cauchy–Riemann equations, 1140–1143, 1256
Cauchy–Schwarz inequality, 103–105, 111–112
Cauchy's residue theorem, 1212
Cayley–Hamilton theorem, 415–417
CDF, *see* Cumulative distribution function (CDF)
Central difference operator, 696
Characteristic equation, 76, 78, 227, 245, 255, 258
Characteristic multipliers, 442, 444
Characteristic polynomials, 76, 227, 326–327, 397
Cholesky factorization, 168–169
Chord length, 1257
Clairaut's theorem, 212, 485, 583
Closed form solution, 204, 297
Closed subspace, 185
Collinear, 1234
Collocation method, 726
Column space, 115
Compartment models, 358–360
Complementary error function (erfc), 1281
Complex eigenvalues, 83–84, 381–385, 427–428
Complex inner product space, 176
Complex integration methods
Cauchy principal value, 1218–1219
Hilbert transform, 1219–1220
improper integrals, 1215–1217
integration of periodic functions, 1214–1215
Complex numbers, 82–83, 1171
Complex variable, 1130–1132
analytic function, 1148
Cauchy–Goursat integral theorem, 1200–1204
Cauchy–Riemann equations, 1140–1143
Cauchy's integral formulas
Dumb-bell contour, 1207–1209
Laurent series integration, 1210–1212
partial fractions, 1206
Cauchy's residue theorem, 1212
closed contour integration, 1198–1200
complex conjugate, 1123
complex integration methods
Cauchy principal value, 1218–1219
Hilbert transform, 1219–1220
improper integrals, 1215–1217
integration of periodic functions, 1214–1215
complex plane, 1194
continuous at z_0, 1137
contour, 1194–1195
converges absolutely, 1175
converges uniformly, 1177
derivatives, 1138–1139
diverges, 1175
elementary functions
$arg(z)$, 1157–1159
branches of logarithms, 1164
$exp(z)$, 1159–1160
$log(z)$, 1160–1164
power functions, 1164–1169
entire function, 1149
geometric series, 1175
harmonic conjugate, 1152–1155
harmonic functions, 1150–1152
holomorphic, 1148
imaginary part, 1123
initial point and terminal point, 1194
Laurent series
domains, 1183
geometric series, 1180
Maclaurin series, 1184–1185
open annulus, 1180–1182
power series, 1179
removable singularity, 1184
L'Hôpital's rule, 1150
modulus/absolute value, 1123
open disk, 1135
orthogonal families, 1143–1144
parametrized curve, 1194, 1198
polar form, complex number
DeMoivre's theorem, 1125
Euler's formula, 1124
exponential form, 1125
polar coordinates, 1124
polygonal curve approximation, 1195
polynomial in z of degree, 1138
positively oriented, 1194–1195
power series about z_0, 1175
punctured disk, 1135
radius of convergence, 1177
Riemann sum, 1196
roots, 1127–1130
streamlined version, 1197
Taylor series
Cauchy's integral formula, 1178
open disk $\mathcal{D}$, 1179
product of, 1185
real analytic, 1179
trigonometric functions
complex numbers, 1171

Euler's formula, 1171
set union, 1174
zeros and poles
singularities, 1188–1193
zero of order, 1186–1188
Conformal mappings
elementary functions, 1227
harmonic functions, 1226–1227
Laplace's equation
boundary conditions, 1247
boundary values, circles, 1248–1251
Dirichlet boundary condition, 1246
electric field lines, 1248
Joukowsky map, 1251–1253
Neumann boundary condition, 1246
open mapping theorem, 1246
Zhukovskii airfoils (*see* Zhukovskii airfoils)
linear mapping
boundary point, 1224, 1226
definition, 1223
domain $\mathcal{D}$, 1224, 1226
magnification, 1223, 1225
punctured line, 1232
rotation, 1223–1225
tangent vectors, 1223, 1226
translation, 1223, 1225
Möbius transformations
circles, 1230–1232
concentric circles, 1233–1240
cross-ratio method, 1240–1242
definition, 1228
distinct points, 1240
inversion mapping, 1230
lines, 1232–1233, 1241–1243
orientations, 1242
test value, 1243
Conjugate gradient method, 1046
Constant coefficient matrix, 355
Contact resistance, 854
Continuum mechanics
body forces, 844
Cauchy equation, 845
conservation of mass, 842–843
contact force, 843
displacement/deformation, 843
divergence theorem, 844–845
Jacobian determinant, 842
Lagrangian description, 840–841
reference state, 840
rigid rotation, 841–842
stress tensor, 843
stress/traction vector, 843
total linear momentum, 843
Contour plot, 502, 940–945
Convergence acceleration technique, 656
Converges, 1390
Convolution, 348, 1268
definition, 312
Laplace transform properties, 312
particular solution, 314–318
properties, 312, 314
reality check, 313
Convolution on the real line, 643
Coriolis acceleration, 503, 542–543
Correlation coefficient, 105
Countably infinite, 1284
Courant–Friedrichs–Lewy (CFL) criterion, 1017
Cox–de Boor recursive, 728
Cramer's rule, 54–56
Crank–Nicholson method, 1005–1006
Critically damped case
homogeneous ordinary differential equations, 225–228, 237, 256
second-order, linear, constant coefficients homogeneous ordinary differential equations (LCCHODEs), 237
Cross ratio, 1240–1242
Cubic basis polynomials, 732
Cubic B-splines
blending function, 723
chain rule, 724
collocation method, 726
cubic spline uniform basis functions, 723
horizontal translation, 723
linear combination, 724
ODE-BVP, 725–726
solid curve, 726–727
Cubic spline uniform basis functions, 723, 1115
Cumulative distribution function (CDF), 636–637
Cyclotron motion, 463

D

D'Alembert wave solutions, 1014
real line, 882–883
step function, 888
support $\Phi(x)$, 888–889
zero initial displacement
intersection symbol, 891
interval cases, 891–893
IVP, 889–890

piecewise continuous function, 890–891
transitional cases, 893–895
zero initial velocity
chain rule, 883–884
decomposition, 884–885
graphical method, 884
moving interval, 885–886
overlap, 886–887
Dantzig's simplex procedure
pivoting process, 1055–1056
unit cost reduction
basic feasible solution, 1057–1058
Mathematica command, 1060
matrix-tableau form, 1057
maximum unit reduced cost, 1058
minimum reduction, 1056–1058
row operations, 1058
swapping variable, 1059
wiggle room, 1059
Decay constant, 219
Definite integral, 400–401, 549–551, 639, 672
Deflation of polynomials/deflation, 690
Delta functions, 318–320, 787, 1385
DeMoivre's theorem, 1125
Density function, 559, 595, 597, 636
Determinant, 48–52, 77, 141–143
De-tuning parameter, 280
Diagonal matrix, 21–22, 85–86, 95–98, 132
Differential equation constraints, 1103–1105
Differential operator, higher-order linear ordinary differential equations (ODE), 251–253
Directional derivative, 488–489
Direct methods, 682
Dirichlet boundary condition, 1093, 1108, 1246
Dirichlet problem, 864–865, 1249–1251
Discrete convolution, 798, 1185
Discrete Fourier transform (DFT)
aliasing property, 794–796
analysis equation, 791
digital filter, 792
discrete convolution, 798–799
Euler's formula, 794–795
fast Fourier transform
discrete convolution, 802
divide and conquer strategy, 801–802
even integer, 800–801
indirect method, 802
odd integer, 800–801
Hermitian conjugate, 792
Mathematica™ command and MATLAB®, 796–798
N-th root of unity, 791
power spectrum, 796
Riemann sum, 790
symmetric matrix, 791–792
time series, 791
unitary matrix, 792–793
Discretization matrix, 993, 1002
Discretized eigenvalue problem, 715
Dispersion equation, 1015
Divergence theorem, 621–624
Diverges, 1391
Dot product, 101, 457–458
Double Fourier series, 951
Double integral
anti-partial derivative, 572
double Riemann sum, 571
element of area, 572
Fubini's theorem, 574
iterated integral, 572
mass density, 571, 573
polar coordinates, 580–582
volume
average value of a function, 580
double Riemann sum, 575
horizontal line segment, 579
mass density, 577
solid tetrahedron, 577–578
triple integrals, 575
type II region, 579
vertical line segment, 576–577, 579
Double Riemann sum, 571, 575
2D stokes potential fluid, 1155
Dual basis, 1022
Dumb-bell contour, 1207–1209

E

Eccentric circles, 1233
Eigenspaces, 91–92
Eigenvalues and eigenvectors
adjugate matrix method, 79–82
annihilator method, 93
characteristic equation, 76, 78
characteristic polynomial, 76
column operation, 77
complex conjugate pairs, 83–84, 94
complex numbers, 82–83
deficient eigenvalues, 98
determinant, 77
eigenspaces, 90–92
elementary row operations, 77
geometry, 75–76

linear homogenous systems, 362
MATLAB® and *Mathematica*™, 86–87
nonresonant nonhomogenous systems, 362
triangular and diagonal matrices, 85–86
trivial solution, 79
Electric field lines, 1248
Electrostatic potential difference, 566
Elementary column operations, 52
Elementary functions, complex variable
arg(*z*), 1157–1159
branches of logarithms, 1164
exp(*z*), 1159–1160
log(*z*), 1160–1164
power functions, 1164–1169
Elementary matrices, 24–25
Elementary row operations, 3, 77
Empty set, 892, 894
Energy density function, 881
Envelope, solution graph, 235
Equivalent system, 1
Error tolerance, 183–184
Euclidean norm, 664
Euler–Lagrange equations
Dirichlet boundary condition, 1093
divergence theorem, 1093
Fourier series, 1090
linear approximation, 1090
variation, 1089–1090
variational method, 1090
Euler's conservation equation, 624–627
Euler's formula, 230–231, 1171
Even function, 744–747, 756
Exactness criterion, 212–214, 492
Exact ordinary differential equation (ODE)
Clairaut's theorem, 212
definition, 211
exactness criterion, 212–214
explicit solution, 214
partial derivatives, 211–212
verification, 211
Explicit solution, exact ordinary differential equation (ODE), 214
Exponential order, 1286, 1379

F

False position method, 654
Fast Fourier transform
discrete convolution, 802
divide and conquer strategy, 801–802
even integer, 800–801
indirect method, 802
odd integer, 800–801
Fick's law of diffusion, 523
Finite difference methods
heat equations
boundary conditions, 996
central difference, 991
Dirichlet boundary conditions, 997
finite difference stencil, 992
forward difference, 991
Neumann/Robin boundary conditions, 997
nonlinearity, 998–1000
PDE-BVP-IVP, 992–996
stencil/replacement equations, 992
time-dependent boundary conditions, 997
potential equations
boundary conditions, 1010–1011
central difference approximations, 1007
finite differences approximate solution, 1008–1010
five-point Laplacian stencil, 1008
replacement equations, 1007
wave equations
D'Alembert wave solutions, 1014
Lax scheme, 1016–1017
scalar hyperbolic problem, 1014–1016
wave speeds, 1014
Finite element methods
Galerkin method, 1116–1117
least squares, 1117–1118
linear differential operator, 1112
Mathematica commands, 1112–1115
PDEs, 1118
Rayleigh–Ritz method, 1111–1112, 1117
trial functions, 1111
uniform tent basis spline functions, 1111
Finite jump discontinuity, 740
First exceptional case, 1400
Five-point Laplacian stencil, 1008
Fixed end boundary condition, 1094
Floquet representation, 442–445
Forced oscillations
beats phenomenon, 283–287
examples, 273–274
forced spring-mass-damper system, 273–274
maximum frequency response, 281–283
resonance case, 274–276
steady-state oscillation, 276–281
Forced spring-mass-damper system, 273–274
Forward difference operator, 331, 695
Forward Euler scheme, 1015–1016

Fourier cosine series
arbitrary function $f(t)$, 762–763
coefficients, 756
convergence, 757–758
even function, 756–757
finite sum $y(t)$, 764–765
IVP, 765–767
ODE, 763–764
steady-state solution, 765
Fourier cosine transform, 1269, 1272–1278
Fourier series, 1090
complex Fourier series, 778–780, 788–789
convolution
definition, 785–786
delta function, 787–788
nonzero constants, 787
ramp function, 786
cosine series
arbitrary function $f(t)$, 762–763
coefficients, 756
convergence, 757–758
even function, 756–757
finite sum $y(t)$, 764–765
IVP, 765–767
ODE, 763–764
steady-state solution, 765
DFT (*see* Discrete Fourier transform (DFT))
Fourier transform
absolutely integrable, 780–781
finite pulse, 783–784
finite wave train, 784–785
inversion theorem, 781–782, 789
Laplace transform operator, 783
non-rigorous way, 782–783
principal value, 782
Riemann sum, 782
generalized series expansion
characteristic equation, 769
coefficients b_n, 770
eigenfunction, 768, 771
eigenvalue, 768–769, 771
equilibrium temperature distribution, 772–773
graphical method, 774–775
ODE-BVP, 765, 773–774
periodic boundary condition, 775–776
square integrability condition, 769
orthogonality and coefficients (*see* Fourier series coefficients)
Parseval's theorem (*see* Parseval's theorem)
Rayleigh quotient
first integral, 818
linear function $X(x)$, 822
quadratic functionals, 820
regular Sturm–Liouville problem, 819–820
second integral, 819
single trial function $X(x)$, 820–822
separation of variables, 755
sine series
arbitrary function $f(t)$, 762–763
coefficients, 758
convergence, 759–760
finite sum $y(t)$, 764–765
IVP, 763–767
least period, 760–761
nonzero terms, 761–762
odd function, 758–759
ODE, 763–764
steady-state solution, 765
Sturm–Liouville problems (*see* Sturm–Liouville problems)
Fourier series coefficients
average value, 743
complex Fourier series, 751–753
convergence of, 739–741
even and odd function, 744–747
extend periodically, 748
Fourier expansion, 742
Fourier representation of, 752–753
function graph, 749–751
harmonic components, 737
inner product function, 741–743
integral calculation, 748
interval of length, 754–755
Mathematica™ command, 738–739
nice function, 743–744
orthogonality relations, 741–742
partial sums, 738–739
periodic extension, 748–749
span vector, 741
trigonometric function identity, 753–754
Fourier sine series
arbitrary function $f(t)$, 762–763
coefficients, 758
convergence, 759–760
finite sum $y(t)$, 764–765
IVP, 765–767
least period, 760–761
nonzero terms, 761–762
odd function, 758–759
ODE, 763–764
steady-state solution, 765
Fourier sine transform, 1272–1278
Fourier's law of heat conduction, 523, 846
Fourier transform
absolutely integrable, 780–781

convolution, 1268
definition, 1263
$F(\omega)$, 1267
finite pulse, 783–784
finite wave train, 784–785
inversion theorem, 781–782, 789
Laplace transform operator, 783
non-rigorous way, 782–783
operator symbols, 1263–1264
PDEs
first-order ODE, exponential decay, 1271
Fourier cosine and sine transforms, 1272–1278
heat kernel, 1272
Parseval's theorem, 1279
separation of variables, 1270
spatial variable, 1270
principal value, 782
rectangular contour, 1266
residues and real integral method, 1264
Riemann sum, 782
semicircular contours, 1264–1265
Fréchet derivative, 666
Free end boundary condition, 1094
Frequency response, 232, 279, 281–283
Frobenius method
Bessel's equation of order α, 1400–1402
definition, 1398–1400
Frobenius norm, 181, 679
$F(s) \to 0$ uniformly on that family, 1284
Fubini's theorem, 574
Fully plastic moment, 1054
Fundamental theorem of calculus, 552
Fundamental theorem of line integrals, 562–564

G

Galerkin method, 1116–1117
coercive, 1019–1020
differential operator, 1020–1021
Euclidean inner product, 1022
N-dimensional vector, 1019
nonlinear problems, 1023–1024
ODE-BVP, 1018–1019
PDE, 1022–1023
space of functions, 1018
Gaussian distribution, 641
Gaussian elimination methods, 6, 9, 679
Gaussian error function, 642
Gauss–Jordan method, 7, 11–12, 124
Gauss–Seidel iteration, 683–684
Gauss's theorem, 621
Generalized eigenvalue problem, 141
Generalized eigenvector, 389
Generalized normal equations, 147, 1047
Geometrical calculus
arctan, 464
area of the triangle, 505
change of variables, area, and volume, 516–519
contour plot, 502
Coriolis acceleration, 503
cross product, 457–458
curves
ellipse in the plane, 472–473
figure eight curve, 474–475
line segment, 471
ODE system, 476–477
parametrized, 470
Pythagorean, 472
tangent vector, 474–475
vector field, 476
Zhukovskii airfoil, 477–478
cyclotron motion, 463
cylindrical coordinates, 465–466
differential operators and curvilinear coordinates
Clairaut's theorem, 528
curl operator, 523
cylindrical coordinates, 533
div, *curl*, and Laplacian operators, 532
divergence, 523
Fick's law of diffusion, 523
Fourier's law of heat conduction, 523
gradient operator, 522
incompressible, 523
intersect orthogonally, 525
irrotational, 523
Laplacian operator, 524, 527–528
matrix-vector equation:, 530
multivariable chain rule, 526–527
properties of, 524
scale/length factors, 529
spherical coordinates, 532
three curves, 525–526
vorticity, 523
dot product, 457
element of surface area, 519–520
equation of the plane, 462
geostrophic wind, 503
isobars, 502–503
level curves, 502, 505
level sets, 502
linear transformations

area and volume, 514–516
mapping, 511–512
stretching and/or compressing transformations, 513
transformation or map, 511
unit square, 512
lines, 459
orthogonal matrix, 509–510
orthonormal basis, spherical coordinates, 468–470
parallelogram, 505–507
parametric equations, 462
partial derivatives
Clairaut's theorem, 485
definition, 485
gradient vector in $\mathbb{R}^3$, 492
linear approximation, 486–489
multivariable chain rules, 489–491
scalar potential functions, 493–497
planes, 459–461
polar coordinates, 463
range/image, 463
relationships to the standard o.n. basis, 470
right-handed orthogonal bases for $\mathbb{R}^3$, 467–468
rotating coordinate frames (*see* Rotating coordinate frames)
spherical coordinates, 465–467
surfaces
cylindrical coordinates, 479
definition, 482
double cusped horn, 481–482
paraboloidal, 478
parametrized surface, 478
sphere and cone intersection, 480
sphere of radius, 481
spherical coordinates, 481
tangent plane and normal vector, 500–502
triangle with vertices, 506
trigonometric functions, 463
volume of parallelepiped, 507–509
Geometric LU factorization, 165–167
Geometric multiplicity, 81–82
Geostrophic wind, 503
Gill's method, 711
Global optimization
constrained optimization problem, 1038–1039
global express guaranteed shipping, 1040–1042
Lagrange multiplier, 1038–1039
unit basis vector, 1042–1043
Gradient vector, 487
Gramian/Gram matrix, 125
Gram–Schmidt process, 106–109, 118–119, 148–149
Green's function, 948; *see also* Heat kernel
Green's theorem
annulus, 585–586
Clairaut's theorem, 583
planar vector field, 582–583
polar coordinates, 586–587
rectangle, 587–588

H

Half-life, 219
Half-range expansion, *see* Fourier cosine series; Fourier sine series
Hamilton's principle
continuous media, 1097
Hamilton's equations, 1095
Lagrangian, 1095
pendulum, 1095
total action during time interval, 1096
Hankel transforms
Bessel's differential equation, 1295–1296
bi-harmonic operator, 1299
boundary conditions, 1295–1296
definition, 1295
integral transforms, 1297
inversion, 1295
Laplacian operator, 1296
PDE solution, 1298
polar/cylindrical coordinates, 1294
Weber's second exponential integral, 1298
Heat equations
arbitrary constant, 909
boundary conditions, 996
boundary value problem, 853–854
central difference, 991
composite rod, 854, 913–914
Dirichlet boundary conditions, 997
equilibrium solution, 907–908
finite difference stencil, 992
first-order constant coefficient, 904
forward difference, 991
Fourier series problem, 911
Fourier sine series, 905–907
Fourier's law of heat conduction, 846
heat flux, 903
inhomogeneous boundary condition, 907–908
initial conditions, 911–913

Neumann boundary conditions, 848, 997
nonlinearity, 998–1000
ODE, 909–910
PDE-BVP-IVP, 992–996
polar coordinates, 967–969
product solution, 904–905
rate of decay, 907
spatial dimension, 850–851
specific heat temperature, 846–847
steady-state temperature, 849
stencil/replacement equations, 992
temperature distribution, 851–853
temperature evolution, 908–909
thermal conductivity, 846–847
thermal diffusivity, 847
time constant, 907
time-dependent boundary conditions, 914–918, 997
xt-plane, 903–904
Heat kernel, 1272
Hessian matrix, 1063
Higher-order linear ordinary differential equations (ODE)
Abel's theorem, 248–249
basic solutions, 245
complete set of basic solutions, 245, 247–249
differential operator notation *D*, 251–253
general solution, 245
LCCHODEs, 250–251, 253
shift theorem, 253
solution for DC circuit, 246–247
standard form, 244
Wronskian determinant, 248
Hilbert space, 185, 188, 1077–1078, 1116
Hilbert transform, 1219–1220
Hill's equation, 447–448
Homogeneous Dirichlet boundary conditions, 997
Householder matrix, 120, 691–693

I

Identity matrix, 21
Implicit partial difference equation, 1005
Implicit partial pivoting, 682
Implicit solution, 196, 216
Improper integrals, 553–554, 1215–1217
Improved tangent method (ITM), 702
Incompressible fluid, 1144
Indefinite integral, 551
Inertia matrix, 140
Initial condition, 196, 200–201, 214, 216–217, 224–225, 237, 248
Initial value problem (IVP)
Peano's existence theorem, 215–217
Picard's existence and uniqueness theorem, 217
Picard's theorem with interval of existence, 217–218
Integrals
anti-derivative, 551–552
area calculation, 549–550
average value of a function, 553
constraints, 1100
divergence theorem, 621–624
double (*see* Double integral)
Euler's conservation equation, 624–626
fundamental theorem of calculus, 552, 554–555
Gauss's theorem, 621
Green's theorem (*see* Green's theorem)
improper integrals, 553–554
indefinite integral, 551
integration by parts, 552
line integrals
Ampère's circuital law, 566
arclength function, 556
density function, 559
electrostatic potential difference, 566
element of arclength, 557
fundamental theorem, 563–564
helix, 559–560
kinetic energy (KE), 567
linear approximation of a curve, 556
magnetic flux density, 567
magnetic permeability, 566
Newton's second law of motion, 567
path direction, 565
tangent vector, 565–566
vector-valued functions, 560–562
piecewise-defined function, 551
regular partition, 549
Riemann sum, 549
Stokes' theorem (*see* Stokes' theorem)
substitution method, 552, 554–555
surface integrals and applications (*see* Surface integrals)
triple integrals and applications (*see* Triple integrals and applications)
union of intervals, 550
Integral transform methods
Fourier transform (*see* Fourier transform)
Hankel transforms (*see* Hankel transforms)

inverse Laplace transform (*see* Inverse Laplace transform)
Integrating factor, 197–198, 200–201, 204, 208
Integration by parts, 552
Inverse Fourier transform, *see* Fourier transform
Inverse Laplace transform, 301–303
Bromwich contour, 1285, 1287
class L, 1284
contour integration, 1283
countably infinite, 1284
exponential order, 1286
$F(s) \to 0$ uniformly on that family, 1284
Heaviside's method, 1288
poles, 1284
rational function, 1288
table lookup method, 1283
unbounded family of left half circles, 1284
wave equation, 1288–1291
Inverse matrix
partitioned matrix, 43–44
row reduction algorithm, 40–43
square matrices, 37–40
Inverse power method, 670
Inversion mapping, 1131–1132, 1230
Irreducible quadratic, 1373–1374
Isolated singularity, 1188
Isoperimetric problem, 1100
Iterated integral, 572

J

Jacobian matrix, 516, 664, 666–667
Jacobi iteration converges, 685
Joint probability density (function)., 643
Joukowsky map, 1251–1253

K

Karush–Kuhn–Tucker (KKT) condition, 1064–1065
Kronecker delta, 106
Kutta condition, 1257

L

Lagrange multipliers (λ), 1101; *see also* Global optimization
Lagrangian (L), 1095
Lamé constants, 878
Laplace equation, 849
arbitrary constants, 1247
boundary values, circles, 1248–1251
contour/3D plot
homogeneous boundary condition, 943–945
isotherms, 941
Mathematica command, 940–941
rotated graph solution, 942
Dirichlet boundary condition, 1246
electric field lines, 1248
electrostatic potential,935–937
equilibrium heat problem, 932–933
Joukowsky map, 1251–1253
Neumann boundary condition, 1246
open mapping theorem, 1246
parallel sides, 933–935
polar coordinates
Cauchy–Euler equation, 960
clairvoyance, 964–966
continuity boundary condition, 962
double root, 963–964
eigenvalue problem, 959–960
equilibrium heat problem, 958–959
flux vector, 961
Fourier sine series, 961
linearity solution, 960–961
nontrivial product solution, $R_n(r)$, 960
periodic boundary condition,962–963
physical boundary, disk, 964
temperature distribution, 961
using clairvoyance, 937–940
Zhukovskii airfoils (*see* Zhukovskii airfoils)
Laplace transforms, 305–306
definitions, 1379
inverse Laplace transform (*see* Inverse Laplace transform)
linear systems, 390–391
operator, 783
of periodic function, 321
properties and derivations, 299, 301–302, 312
$\mathcal{L}[cf(t)](s) = c\mathcal{L}[f(t)](s)$, any constant c, 1380
$\mathcal{L}[\cos \omega t](s) = \frac{s}{s^2+\omega^2}$, 1381
$\mathcal{L}[\delta(t-c)](s) = e^{-cs}$, 1385
$\mathcal{L}[e^{at}f(t)](s) = F(s-a)$, where $F(s) = \mathcal{L}[f(t)](s)$, 1382
$\mathcal{L}[e^{at}](s) = \frac{1}{s-a}$, 1380
$\mathcal{L}[f(t)](s) = \frac{1}{1-e^{-sT}} \int_0^T f(t)\, e^{-st}\, dt$, *for* $s > 0$, 1385–1386
$\mathcal{L}[g(t)\ \text{step}(t-c)](s) = e^{-cs}.\ \mathcal{L}[g(t+c)](s)$, 1382–1383
$\mathcal{L}[(f*g)(t)](s) = \mathcal{L}[f(t)](s).\ \mathcal{L}[g(t)](s)$, 1383

$\mathcal{L}[\sin \omega t](s) = \frac{\omega}{s^2+\omega^2}$, 1381
$\mathcal{L}[\text{step}(t-c)](s) = \frac{e^{-cs}}{s}$, 1382
$\mathcal{L}[t \cos \omega t](s) = (s^2 - \omega^2)/(s^2 + \omega^2)^2$, 1384–1385
$\mathcal{L}[t](s) = n!/s^n$, $n = 0, 1, 2, \ldots$, 1380–1381
$\mathcal{L}[t \sin \omega t](s) = 2\omega s/(s^2 + \omega^2)^2$, 1384
$\mathcal{L}[\ddot{y}](s) = s^2\mathcal{L}[y](s) - sy(0) - \dot{y}(0)$, 1382
$\mathcal{L}[\dot{y}](s) = s\mathcal{L}[y(t)](s) - y(0)$, 1380
solution in frequency domain, 300
steady-state solution, 304–305
unit step function (*see* Unit step function)
Laplacian operator, 524
LaSalle invariance principle, 1341–1343
Laurent series
domains, 1183
geometric series, 1180
integration of, 1210
Maclaurin series, 1183–1185
open annulus, 1180–1182
power series, 1179
removable singularity, 1184
Lax scheme, 1016–1017
LCCHODEs, *see* Second-order, linear, constant coefficients homogeneous ordinary differential equations
Leading entry, 2
Least squares, 1117–1118
Least squares solutions (L.S.S.), 158–159
best approximation problem, 122
normal equations
basic theoretical method, 123–124
Gramian/Gram matrix, 125
orthogonal matrices, 127–129
regression line, 125–127
P_A projection, 121–122
residual, 121
vector subspace, 122
Legendre polynomials, 1404
Leibniz's rule, 497
Levinson–Smith theorem, 1347–1349
L'Hôpital's rule, 639, 1150
Liapunov functions
autonomous system, 1332
coefficients homogeneous systems, 1329
differential equations, 1329
Hamiltonian mechanics, 1329
instability, 1336–1338
Liapunov theory, 1333–1334
negative definite, 1331
negative semi-definite, 1331
nonlinear ODEs, 1328–1329
positive definite, 1331
positive semi-definite, 1331
quadratic forms, 1331, 1334–1336
real-valued function, 1331
stability equilibrium point, 1338
Liapunov stability, 1313
Liapunov theory, 1333–1334
Linear constant coefficients homogeneous difference equation (LCCHΔE), 325–327, 333–334, 340, 343–347
Linear, constant coefficients homogeneous ordinary differential equations (LCCHODEs), 264–265, 372–375
Linear constant coefficients homogeneous system (LCCHS), 372–375, 1309, 1313–1314
Linear, constant coefficients, homogeneous system of difference equations (LCCHSΔ), 422, 427
Linear control theory
Cayley–Hamilton theorem, 415–417
controllability condition, 418–419
multi variable control system, 417, 420
observability criterion, 421
optimal control, 420
piecewise continuous, 414
Riemann sums, 415
scalar-valued function, 412–414, 419
single input control system, 412, 420
vector-valued function, 412
Linear convergence, 655
Linear first-order ordinary differential equations (ODE)
basic solutions, 199
complete set of basic solutions, 199
general solution, 199, 201–202
integrating factor, 197–199
particular solution, 199–200
standard form, 196
Linear homogenous systems
arbitrary constants, 365–366
complete set of basic solutions, 364–365
eigenvalues and eigenvectors, 362
fundamental matrix, 368–372
general solution, 364–366
initial conditions, 363
initial value problem, 366–367
Maclaurin series, 375–376
nonconstant coefficients, 376–377
second-order LCCHODE and LCCHS, 372–375
unique solution, 363
Linearity principle, 27

Linear mapping
boundary point, 1224, 1226
definition, 1223
domain $\mathcal{D}$, 1224, 1226
magnification, 1223, 1225
punctured line, 1232
rotation, 1223–1225
tangent vectors, 1223, 1226
translation, 1223, 1225
Linear programming (LP) problem
Acme company, 1049–1050
convex set and lines, 1048–1049
linear inequalities, 1048
slack variable and standard form
feasible solution, 1051–1052
nonnegative number, 1050–1051
total cost, 1053–1054
wheat bran, oat flour, and rice flour proportions, 1052–1053
structural optimization, 1054
Linear superposition principle, 31
Linear system, 353
compartment models, 358–360
complex eigenvalues, 381–385
constant coefficient matrix, 355
deficient eigenvalues, 387–390
of difference equations
Casorati determinant, 425–427
color blindness, 423–425
complex eigenvalues, 427–428
ladder network electrical circuits, 429–434
LCCHSΔ, 422
second-order scalar difference equation, 428–429
stability, 434–436
homogeneous systems of second-order equations, 385–387
Laplace transforms, 390–391
linear control theory
Cayley–Hamilton theorem, 415–417
controllability condition, 418–419
multivariable control system, 417, 420
observability criterion, 421
optimal control, 420
piecewise continuous, 414
Riemann sums, 415
scalar-valued function, 412–414, 419
single input control system, 412, 420
vector-valued function, 412
linear homogenous systems
arbitrary constants, 365–366
complete set of basic solutions, 364–365
eigenvalues and eigenvectors, 362
fundamental matrix, 368–372
general solution, 364–366
initial conditions, 363
initial value problem, 366–367
Maclaurin series, 375–376
nonconstant coefficients, 376–377
second-order LCCHODE and LCCHS, 372–375
unique solution, 363
matrix–vector form, 354–355
nonhomogeneous linear systems
characteristic polynomial, 397
definite and indefinite integration, 396
fundamental matrix, 395–396, 398–400
law of exponents, 397
solution, 400
variation of constants, 397
variation of parameters, 397
nonresonant nonhomogeneous systems
computer hardware obsolescence model, 407–408
constant forcing function, 406
eigenvalues and eigenvectors, 404
general solution, 404–405, 407
Mathematica^{TM}, 407
nonresonance assumption, 405
particular solution, 405
sinusoidal forcing, 408–411
undetermined coefficients method, 403
periodic linear differential equations
Floquet representation, 442–445
Hill's equation, 447–448
period doubling, 440
principal fundamental matrix, 440
stability, 445–447
stroboscopic/return map, 441–442
T-periodic solution, 440, 448–451
RLC series circuit, 353–354
RLC two-loop circuit, 355–356
second-order equations, 357–358
spring–mass–damper system, 356
stability, 391–392
Linear transformations, 511–513
Line integrals
Ampère's circuital law, 566
arclength function, 556
density function, 559
electrostatic potential difference, 566
element of arclength, 557
fundamental theorem, 563–564
helix, 559–560
kinetic energy (KE), 567
linear approximation of a curve, 556

magnetic flux density, 567
magnetic permeability,566
Newton's second law of motion, 567
path direction, 565
tangent vector, 565–566
vector-valued functions, 560–563
Line search, 1046
Liouville's theorem, 371
Logarithmic decrement, 235–236
Lorentz force, 458
Lower triangular matrix, 22–23

M

Maclaurin series, 375–376, 400, 1088, 1184–1185
Magnetic permeability, 566
Magnification, 1223, 1225
Mass matrix, 140
*Mathematica*TM, 64, 86–87, 214, 407, 477–478, 557, 559–560, 601, 603, 605, 623, 673, 716, 738–739, 796–797, 822, 924, 1040, 1086, 1305
MATLAB®, 86–87, 597, 662, 796–798, 1010
Matrix
addition, 16–17
adjugate matrix, 53–54
augmented matrix, 5–6
basis, 64–65
bounded linear functionals and operator adjoint, 186–187
Cauchy criterion, 185
Cauchy–Schwarz inequality, 103–105, 111–112
Cholesky factorization, 168–169
convergence of sequences, 183–184
Cramer's rule, 54–56
determinant, 48–52
diagonal matrix, 21–22, 95–98
direction vector, 103
dot product, 101
eigenvalues and eigenvectors
adjugate matrix method, 79–82
annihilator method, 93
characteristic equation, 76, 78
characteristic polynomial, 76
column operation, 77
complex conjugate pairs, 83–84, 94
complex numbers, 82–83
deficient eigenvalues, 98
determinant, 77
eigenspaces, 91–92
elementary row operations, 77
geometry, 75–76
*Mathematica*TM, 86–87
MATLAB®, 86–87
triangular and diagonal matrices, 85–86
trivial solution, 79
elementary matrices, 24–25
elementary row operations, 3–4
Gram–Schmidt process, 106–109, 118–119
Hilbert space, 185, 188
homogeneous system
$Ax = 0$ form, 26
basic solutions, 30–31
general linear combination of vectors, 27
general solution, 28
linearity principle, 27
nullity, 31–32
spanning set, 28–30
trivial solution, 27
identity matrix, 21
inner product, 102, 175–177
inverse matrix
partitioned matrix, 43–44
row reduction algorithm, 40–43
square matrices, 37–40
leading entry, 2
least squares solutions (*see* Least squares solutions)
linear functionals and operators, 177–179
linear independence, 62–64
lower inequality, 145
lower triangular matrix, 22–23
LU factorization, 165–168
Moore–Penrose generalized inverse, 160–162
multiplication, 18–20
nonhomogeneous system, 33–36
nonzero row, 2
norm and bounded linear operators, 179–183
orthogonal matrix, 116–118
orthogonal projections, 109–111
orthogonal set of vectors, 105–106
orthogonal sets and bases, 114–116
pivot column, 3
pivot positions, 2
positive definite matrix (*see* Positive definite matrix)
positive semi-definite matrix, 136
powers, 23–24
QR factorization, 148–151
rank of C, 3
Rayleigh quotient
continuous systems, 170
property, 172

real matrix, 171
theorem, 172–174
row echelon form, 2
RREF
basic variable, 8
free variable, 8
Gaussian elimination methods, 9
Gauss–Jordan method, 6–7
lemma, 10–11
linear system, 11–13
matrix of coefficients, 10
signal restoration, 187–188
SVD
block matrices, 153–154
construction, 152
factorization, 154–158
least squares solution (L.S.S.), 158–159
positive eigenvalues, 152
pseudo-diagonal matrix, 151, 153
uniqueness, 155
zero eigenvalues, 152
symmetric matrices
A^2, A^{-1}, $\sqrt{A}$ formulas, 137
diagonal matrix, 132
orthogonality for eigenvectors, 132–134
orthogonally similar, 131
spectral formula, 135
unique least squares solution, 138–139
transpose, 24
triangle inequality, 103–104, 145
upper inequality, 145
upper triangular matrix, 22–23
vector spaces and subspaces, 65–69
weak convergence and compactness, 189–191
zero matrix, 21
zero row, 2
Matrix addition, 16–17
Matrix multiplication, 18–20
Matrix norm induced, 664
Matrix–vector form, 354–355
Maxwell's integral equations
Ampère—Maxwell's law, 840
electromotive force, 838
magnetic flux density, 838–840
magnetomotive force, 840
physical law, 839
Method of successive approximations, 656
Midpoint method, 702
Möbius transformations, 1228–1229
circles, 1230–1232
concentric circles
centered, 1234
collinear, 1234
eccentric circles, 1233
interior *vs.* exterior relationship, 1236–1239
positive and negative orientation, 1238
quadratic equation, 1235
cross-ratio method, 1240–1242
distinct points, 1240
inversion mapping, 1230
lines, 1232–1233
orientations, 1242
test value, 1243
Modified Bessel equation of order α, 1403
Modified Newton's method, 653
Modified (or corrected) Newton's method, 653
Momentum density function, 881
Monodromy matrix, 442
Moore–Penrose generalized inverse, 139, 160–163
Multi-index, 732
Multivariable chain rules, 489–491

N

Natural frequency, 232
Neumann boundary condition, 1246
Neumann–Goldstine criterion, 1003–1005
Newton–Cotes formula, 670
Newton–Raphson method
in $\mathcal{R}^n$, 661–664
scalar equation
bisection method, 652–653
cycling in, 652–653
iterate, 650–651
point-slope formula, 649
quadratic convergence, 651
software implementation, 651
tangent line, 649
Newton's law of cooling, 357
Nonhomogeneous Dirichlet boundary conditions, 997
Nonhomogeneous linear systems
characteristic polynomial, 397
definite and indefinite integration, 396
fundamental matrix, 395–396, 398–400
law of exponents, 397
solution, 400
variation of parameters/constants, 397
Nonlinear differential equations
continuous dependence, 1359–1361
delay equations

characteristic equation, 1369–1370
Euler's method, 1370
horseshoe map and chaos
Brownian motion, 1365
butterfly effect, 1364
entropy, 1365
homoclinic orbit, 1367
periodic solution, 1365
Poincaré map, 1366
Smale horseshoe map, 1365–1366
initial value problem (IVP), 1303
LaSalle invariance principle, 1341–1343
Liapunov functions (*see* Liapunov functions)
limit cycles
Hopf bifurcation, 1349–1350
Levinson–Smith theorem, 1347–1349
orbitally stable, 1344
periodic linearization, 1346
periodic solution, 1346–1347
stable and unstable, 1344
time translation, 1344
local Lipschitz condition, 1355–1356
Peano's existence theorem, 1355
phase plane
autonomous systems, 1304–1306
damped oscillator system, 1305
equilibria, 1306
initial value problem (IVP), 1303
LCCHS, 1309
nonlinear system, 1305–1306
phase line, 1307–1309
saddle point case, 1309–1311
stable spiral point, 1305
steady-state oscillation, 1303–1304
undamped harmonic oscillator system, 1304
unstable spiral point, 1305
Picard's theorem, 1356–1358
stability of equilibrium point
asymptotically stable, 1314
attractor, 1314
LCCHS, 1313–1314
Liapunov stability, 1313
linearization stability, 1315–1317
linear system, 1314
stable and unstable, 1313
undamped harmonic oscillation, 1314
using $r(t)$, 1318–1319
variation of parameters, linearization
almost linear system, 1321
periodic solutions, 1326–1327
saddle point theorem (*see* Saddle point theorem)
Nonlinear programming problem
complementarity condition, 1066–1067
convex programming (CP) problem, 1064
convex set, 1063
dual LP problem, 1068–1071
feasible region, 1064
geometric dimensioning/tolerancing, 1071–1072
Hessian matrix, 1063
KKT condition, 1064–1065
Lagrangian, 1064
line segment, 1062
real, constant, positive semi-definite matrix, 1063
Slater's condition, 1064
stationarity condition, 1066–1068
Nonlinear Volterra integral equations, 1328
Nonresonant nonhomogeneous systems
computer hardware obsolescence model, 407–408
constant forcing function, 406
eigenvalues and eigenvectors, 404
general solution, 404–405, 407
Mathematica™, 407
nonresonance assumption, 405
particular solution, 405
sinusoidal forcing, 408–411
undetermined coefficients method, 403
Nontrivial solution, 715
Nonuniform tent–spline functions, 727–729
Normal equations
basic theoretical method, 123–124
Gramian/Gram matrix, 125
orthogonal matrices, 127–129
regression line, 125–127
Normal random variable, 640–642
Norm and bounded linear operators, 179–183
Numerical methods
$Ax = b$
absolute error, 678
column diagonally dominant, 685
convergence and stability, 677
Frobenius norm, 684, 686
Gaussian elimination algorithm, 677
Gauss–Jordan algorithm, 679–682
Hilbert matrix, 686
induced matrix norm, 684–685
iterative methods, 682–684
Jacobi iteration converges, 685
perfectly conditioned matrix, 679
relative error, 678
row diagonally dominant, 685
standard condition number, 677

BVP (*see* Two-point boundary value problem (BVP))
derivative approximations
central difference approximation, 699
central difference operator, 696
forward difference operator, 695
linear approximation, 695
Taylor series, 696–698
finite difference methods (*see* Finite difference methods)
Galerkin method (*see* Galerkin method)
integrals approximation
Boole's rule, 671–672
iteration of quadrature, 674–675
midpoint rule, 670, 672
Newton–Cotes formula of closed type, 670
numerical drift, 675
quadrature rules, 670, 672–673
Riemann sums, 669–670
Simpson's rule, 670, 672
step size, 670
trapezoidal rule, 670, 672
two-edged sword, 675
linear algebraic eigenvalue problems
almost triangular matrix, 692
deflation and similarity, 690
elementary method, 687–688
extended sign function, 691
Householder matrix, 691–693
lower Hessenberg form, 692
power methods, 688–690
QR algorithm, 693
similarity transformations, 690–691
upper Hessenberg form, 692
numerical stability
amplification factor, 1003
consistent, 1001
convergent method, 1001
Crank–Nicholson method, 1005–1006
discretization matrix, 1002
Neumann–Goldstine criterion, 1003–1005
stable, 1002
sum of sines formula, 1003
ODE-IVPs
Euler's method (EM), 700, 702
global error, 702
ITM, 702
linear approximate solution graph, 700
local error, 701
multistep methods, 705
multivariable chain rule, 701, 703
numerical instability and stiffness, 707–708
Predictor–Corrector methods, 706
Runge–Kutta methods, 704–705
Taylor's formula method, 706–707
Taylor's theorem, 703
scalar equation (*see* Scalar equation)
splines
cubic B-splines (*see* Cubic B-splines)
curve in $\mathbb{R}^n$, 729–730
horizontal translation, 721
nonuniform splines, 727–729
piecewise linear approximation, 720
piecewise polynomial, 720
support interval $\mathcal{I}$, 722
surface, 731
tent–spline function, 721
triangular surface patches, 731–732
uniform tent basis functions, 721
uniform tent–spline functions, 720
system of equations (*see* System of equations)

O

Odd function, 744–747
ODE, *see* Ordinary differential equation (ODE)
One compartment model, 207
One-sided z-transforms, 340–341
Optimization
continuous function, 1028–1029
convex function, 1029–1030
critical number, 1028
global minimum (value), 1027
insulation, critical radius of, 1029
local minimum (value), 1027
LP problem (*see* Linear programming (LP) problem)
nonlinear programming
complementarity condition, 1066–1067
convex programming (CP) problem, 1064
convex set, 1063
dual LP problem, 1068–1071
feasible region, 1064
geometric dimensioning/tolerancing, 1071–1072
Hessian matrix, 1063
KKT condition, 1064–1065
Lagrangian, 1064
line segment, 1062
real, constant, positive semi-definite matrix, 1063

Slater's condition, 1064
stationarity condition, 1066–1068
Rayleigh–Ritz method
dot product, 1076
global minimizer, 1076–1077
Hilbert space eigenvalue problem, 1077–1078
maximization problem, 1075–1076
real symmetric matrix, 1073–1075
simplex procedure (*see* Dantzig's simplex procedure)
variables function
critical/stationary point, 1033–1034
feasible direction, 1032–1033
Fermat's theorem, 1034–1035
global optimization (*see* Global optimization)
Hessian matrix, 1035
interior, 1033
local minimizer, 1033
minimization problem, 1032
numerical minimization, 1045–1046
objective function, 1032
saddle point, 1035
second derivative test, 1036–1038
steepest descent method, 1045–1046
Taylor's theorem, 1036
Order, definition, 195
Ordinary differential equation (ODE)
Fourier sine series, 763–764
linear system (*see* Linear system)
polynomial solutions, 1403–1404
power series solutions (*see* Power series solutions, ODEs)
Orthogonal matrix, 116–118, 127–129
Orthogonal projections, 109–111
Orthogonal sets and bases, 114–116
Overdamped, 225–228, 237
Over-relaxation methods, 683

P

Parametric equations, 459
Parseval identity, 116, 160, 830
Parseval's theorem
Bessel's inequality, 829
best approximation, 829
complex Fourier series, 829–830
Fourier cosine series, 826–827
Fourier series, 827
Fourier transforms, 831–832
mass density, 824–825
square integrable, 827–828
Partial differential equations (PDEs)
cylindrical coordinates
Bessel's equation of order zero, 974–975
double integrals, 980–981
eigenfunctions, 979
Fourier–Bessel coefficients, 977–978
modified Bessel's equation of order n, 980–981
nonhomogeneous boundary conditions, 976
orthogonality relation, 977
periodicity, 973–974
θ-dependence, 973
using clairvoyance, 975–976
zero component, 978–979
D'Alembert wave solution (*see* D'Alembert wave solution)
energy conservation, 897–900
first-order ODE, exponential decay, 1271
Fourier cosine and sine transforms, 1272–1279
heat equations
arbitrary constant, 909
boundary value problem, 854–855
composite rod, 854, 913–914
equilibrium solution, 907–908
first-order constant coefficient, 904
Fourier series problem, 911
Fourier sine series, 905–907
Fourier's law of heat conduction, 846
heat flux, 903
inhomogeneous boundary condition, 907–908
initial conditions, 912–913
Neumann boundary condition, 848
ODE, 909–910
product solution, 904–905
rate of decay, 907
spatial dimension, 850–851
steady-state temperature, 849
temperature distribution, 853–855
thermal conductivity, 846–847
thermal diffusivity, 847
time constant, 907
time-dependent boundary condition, 914–918
time evolution, 908–909
xt-plane, 903–904
heat kernel, 1272
integral equation
arbitrary control volume, 835, 837–838

continuum mechanics (*see* Continuum mechanics)
divergence theorem, 835–837
fluid mass, 835
heat energy density, 836
heat flux vector, 836
Maxwell's equation (*see* Maxwell's integral equations)
Stokes' theorem, 838
Laplace equation (*see* Laplace equation)
Laplacian and application
eigenfunction, 949–950
steady-state temperature, 953–954
surface waves, 954–956
temperature distribution function, 952
time-dependent heat flow, slab, 951–952
transverse vibrations, rectangular membrane, 952–953
Parseval's theorem, 1279
polar coordinates
heat equation, 967–969
Laplace equation (*see* Laplace equation, polar coordinates)
potential equation
boundary condition, 861
closed ball, 865
Dirichlet problem, 865
irrotational, 860
Laplace's equation, 861
magnetostatics, 860–861
Poisson's formula, disk, 864–865
potential flow, 860
solvability condition, 862–864
vorticity, 860
separation of variables, 1270
spatial variable, 1270
spherical coordinates
electric potential, u, 982
Legendre polynomials, 983–984
Legendre's equation, 982–983
Neumann problem, 985
orthogonality relation, 984
ρ- and ϕ-dependence, 982
solvability condition, 984–985
wave equation (*see* Wave equation)
Partial fractions
complex form, 1377
proper rational function, 1373–1377
real form, 1373–1377
Partial pivoting, 681
Partial sums, 738–739
Partitioned matrix, 17, 43–44
PDEs, *see* Partial differential equations (PDEs)
Peano's existence theorem, 215, 1355–1359
Pendulum, 1095
Period doubling, 440, 1367
Periodic function, 286
Periodic linear differential equations
Floquet representation, 442–446
Hill's equation, 447–449
period doubling, 440
principal fundamental matrix, 440
stability, 445–447
stroboscopic/return map, 441–442
T-periodic solution, 440, 448–451
Phase form, 232–234
Picard's theorem, 1358–1361
Piecewise continuous, 414
Piecewise smooth, 740
Plane wave, 897
Plasticity, 1054
Poincaré–Andronov–Hopf bifurcation, 1351–1352
Poincaré map, 1368
Poisson's equation, 849, 860
Poisson's ratio, 879
Position vector, 459
Positive definite matrix, 136
generalized eigenvalue problem, 140
positive definiteness and determinants, 141–143
scalars, 139
vibrations, 140
Positive semi-definite matrix, 136
Potential equations
boundary conditions, 1010–1011
central difference, 1007
finite differences approximate solution, 1008–1010
five-point Laplacian stencil, 1008
replacement equations, 1007
Potential flow, 627, 860, 1144
Power functions, 1164–1169
Power series solutions, ODEs
basic methods, 1387–1390
convergence
converges, 1390
converges absolutely, 1391
diverges, 1391
radius of convergence, 1391–1393
sequence of partial sums, 1390
Frobenius method
Bessel's equation of order α, 1400–1403
definition, 1398–1400
nonconstant coefficients, 1393–1397
$x - a$, 1397

Practical resonance, 280
Predictor–Corrector methods, 706
Principal minors, 142
Principal nth root of unity, 1129
Probability density function (PDF), 636
Probability distributions
  bell curve, 641
  CDF, 636–637
  expectation, 638
  exponential distribution, 638
  frequency interpretation, 635
  Gaussian distribution, 641
  Gaussian error function, 642
  joint distribution, 643–644
  L'Hôpital's rule, 639
  mean of random variable, 640
  mode of continuous random variable, 640
  normal random variable, 640
  PDF, 636
  problems, 645
  random variable, 635
  standard deviation, 640
Propagator, *see* Heat kernel
Pseudo-diagonal matrix, 151, 153
Punctured line, 1232
Pythagorean theorem, 463

Q

QR algorithm, 693
Quarter-chord point, 1255
Quasi-frequency, 232
Quasiperiodic function, 236, 286

R

Radius of convergence, 1177, 1391–1393
Rayleigh quotient
  continuous systems, 170
  first integral, 818
  linear function $X(x)$, 822
  property, 172
  quadratic functionals, 820
  real matrix, 171
  regular Sturm–Liouville problem, 919–820
  second integral, 819
  single trial function $X(x)$, 820–822
  theorem, 172–174
Rayleigh–Ritz method
  dot product, 1076
  eigenvalue problems, 1106
  finite element methods, 1111–1113, 1117
  global minimizer, 1076–1077
  Hilbert space eigenvalue problem, 1077–1078
  maximization problem, 1075–1076
  minimization, 1085–1087
  real symmetric matrix, 1073–1075
Real analytic function, 1179, 1398
Real inner product space, 176
Real linear factor, 1373
Reciprocal/dual lattice, 522
Recursion relation, 1387, 1389
Reference configuration, 869–870
Reflection transformations, 513
Regular singular point, 1398
Relaxation method, 683
Replacement equations, 992
Riemann sums, 341, 415, 549, 574, 670, 1196
Root mean square (RMS), 830
Rotating coordinate frames
  ODEs, 537–539
  velocity and acceleration
    acceleration of the particle relative, 542
    angular acceleration vector, 543
    angular velocity vector, 543
    constant angular speed, 540
    Coriolis acceleration, 541
    Earth's equator, 542–543
    intermediate frame, 543
    rotating coordinate frame, 541
Rotation transformations, 513
Round-off-error, 654–655
Row reduced echelon form (RREF)
  basic variable, 8
  free variable, 8
  Gaussian elimination methods, 9
  Gauss–Jordan method, 6–7
  lemma, 10–11
  linear system, 11–13
  matrix of coefficients, 10
Row reduction algorithm, 40–43
Runge–Kutta method of order four, 711
Runge–Kutta method of order three, 711
Runge–Kutta method of order two, 711
Runge–Kutta methods, 704–705

S

Saddle point theorem
  dashed solution curves, 1323
  differentiable surface, 1324

invariant, 1324
k-dimensional surface, 1324
ODE, 1322
phase plane, 1323
planar linear system, 1322
stable and unstable manifold, 1325
twice continuously differentiable surface, 1324
Sawtooth function, 323
Scalar equation
Aitken's δ^2 method, 656–657
fixed point problem iteration, 656
modified Newton's method, 653
Newton–Raphson method
bisection method, 652–653
cycling in, 652–653
iterate, 650–651
point-slope formula, 649
quadratic convergence, 651
software implementation, 651
tangent line, 649
secant method, 654–656
Scalar hyperbolic problem, 1015–1016
Scalar ordinary differential equation (ODE)
Casorati determinant
forward difference operator, 330
higher order difference operators, 331
n-th order linear homogeneous difference equation, 332
Wronskian determinant, 331
Cauchy–Euler equations, 325
characteristic polynomials, 326–327
complex number, polar form, 327
complex roots, 328
Fibonacci numbers, 324
forced oscillations
beats phenomenon, 283–287
examples, 273–274
forced spring-mass-damper system, 273–274
maximum frequency response, 281–283
resonance case, 274–276
steady-state oscillation, 276–281
homogeneous problems, 195–196
Cauchy–Euler ODEs, 255–259
exact ODEs (*see* Exact ordinary differential equation (ODE))
higher-order linear ODEs (*see* Higher-order linear ordinary differential equations (ODE))
initial value problem (IVP), 215–217
linear first-order ODE (*see* Linear first-order ordinary differential equations (ODE))
second-order LCCHODEs (*see* Second-order, linear, constant coefficients homogeneous ordinary differential equations (LCCHODEs))
separable ordinary differential equation (ODE), 209–211
steady-state solution, 203
transient solutions, 202, 204
Laplace transforms, 305–306
inverse Laplace transform, 301–303
of periodic function, 321
properties, 299, 301–302
solution in frequency domain, 300
steady-state solution, 304–305
unit step function (*see* Unit step function)
LCCHΔE, 325–327
nonhomogeneous
general solution, 263–264, 269
LCCHODEs, 264–265
method of undetermined coefficients, 265–269
particular solution, 263–264, 269
RHS function, 264
shift theorem, 270–271
nonhomogeneous linear difference equation, 334
n-th order difference equation, 324
polar coordinates/trigonometry, 327
second-order difference equation, 329
undetermined coefficients method, 334–337
variation of parameters
general solution, 293
methods, 294–297
particular solution, 294
second-order nonhomogeneous ODE, 292
z-transforms (*see* z-transforms)
Scalar potential functions, 493–497
Scalar triple product, 508
Secant method
scalar equation, 654–656
system of equations, 666–668
Second-order equations, 357–358
Second-order, linear, constant coefficients homogeneous ordinary differential equations (LCCHODEs)
Abel's theorem, 238–329
amplitude and phase form, 232–234
basic solutions, 239
critically damped case, 237
existence and uniqueness, 229

existence of complete set of basic solutions, 229–230
general solution, 228–229
series *RLC* circuit, 225–228
solution graph
envelope, 235
logarithmic decrement, 235
quasi-period, 236
spring–mass–damper systems, 222–225
standard form, 222
underdamped case
basic solution, 230–232
Euler's formula, 230–231
Wronskian determinant, 238
Separable ordinary differential equations (ODE), 209–211
Separation of variables method, *see* Partial differential equations (PDEs)
Sequence of partial sums, 1390
Series *RLC* circuit, 225–228
Set union, 1174
Shear modulus, 878
Shear transformations, 24, 513, 515
Shift operator, 341
Shift theorem, 253, 270–271
Shooting methods, 717
Signal restoration, 187–188
Similarity transformation, 690–691
Simpson's rule, 670, 672–673, 1001
Singular value decomposition (SVD)
block matrices, 153–154
construction, 152
factorization, 154–158
least squares solution (L.S.S.), 158–159
positive eigenvalues, 152
pseudo-diagonal matrix, 151, 153
uniqueness, 155
zero eigenvalues, 152
Slater's condition, 1064
Smale horseshoe map, 1365–1366
Sobolev space, 1109
Solution of ODE, definition, 195
Spanning set, 28–30
Spectral formula, 135
Splines
cubic B-splines (*see* Cubic B-splines)
curve in $\mathbb{R}^n$, 729–730
horizontal translation, 721
nonuniform splines, 727–729
piecewise linear approximation, 720
piecewise polynomial, 720
support interval $\mathcal{I}$, 722
surface, 730
tent–spline function, 721
triangular surface patches, 731–732
uniform tent basis functions, 721
uniform tent–spline functions, 720
Spring–mass–damper systems, 222–225
Square matrices, 37–40
Square wave function, 316, 450, 747
Standard form
Cauchy–Euler ODEs, 255
higher-order linear ordinary differential equations (ODE), 244
linear first-order ordinary differential equations (ODE), 196
second-order, linear, constant coefficients homogeneous ordinary differential equations (LCCHODEs), 222
Standard orthonormal, 467
Steady-state oscillation, 276–280, 1306
Steady-state solution, 203, 269, 276–278, 304, 345–348, 849
Steffensen's method, 657
Stencil equations, 992
Stiffness matrix, 140
Stokes' theorem
chain of three curves, 630
Green's theorem, 627, 634
inverse square law force field, 632
Poisson's equation, 632
positively oriented boundary curve, 627–629
total flux of *curl* (**F**), 631–632
Streamlines, 1144
Stress/traction, vector, 843
Stretching and/or compressing transformations, 513
Stroboscopic/return map, 441–442
Sturm–Liouville problems, 1109–1110
adjugate matrix method, 812
chain rule, 811–812
eigenvalues, 806
fourth-order ODE-BVP, 813–815
interface condition, 810
match, 810–811
ODE, 805
orthogonality relation, 806–807
regular, 805
singular Sturm–Liouville problem, 808–810
Super linear convergence, 655
Surface integrals
annulus geometry, 609–610
computer algebra system, 619
element of surface area, 609–611
parametric surface, 607–608
scalar-valued function, 611–613

surfaces of revolution, 620–621
twisted cylinder, 617–618
vector field
closed surface, 616
flux of **F** across, 614
iterated double integral, 614
piecewise smooth, 615
positive orientation, 616–617
tangent plane, 613
total flux of electric field, 615
upward orientation, 616
vector-valued function, 613–614
z-coordinate centroid, 618–619
SVD, *see* Singular value decomposition (SVD)
Symmetric matrices
A^2, A^{-1}, $\sqrt{A}$ formulas, 137
diagonal matrix, 132
orthogonality for eigenvectors, 132–134
orthogonally similar, 131
positive definite matrix, 136
positive semi-definite matrix, 136
spectral formula, 135
unique least squares solution, 138–139
Symmetric space, 1116
System of equations
fixed point problem iteration, 666
Newton–Kantorovich theorem, 664–666
Newton's method in $\mathbb{R}^n$
iterates converge, 662–663
Jacobian matrix, 661
MATLAB®, 662
quadratic convergence, 662–664
software implementation, 662
secant method, 666–668

T

Taylor series, 706–707, 874
Cauchy's integral formula, 1178
open disk $\mathcal{D}$, 1179
product of, 1185
real analytic, 1179
Taylor's theorem, 697–698, 702–703, 1035–1036
Tent basis functions, 727
Thermal conductivity, 846–847
Thermal diffusivity, 847, 908, 915
Thickness distribution, 1255
Thickness ratio, 1255, 1258
Time constant, 227–228, 251, 447, 907
Torque, 458
Total linear momentum, 843, 871
Transient solutions, 202, 204
Translation, 1223, 1225
Transverse shear forces, 879–880
Triangle inequality, 103–104, 145
Triangular and diagonal matrices, 85–86
Trigonometric functions
complex numbers, 1171
Euler's formula, 1171
set union, 1174
Triple integrals and applications
cylindrical coordinates
average value of a function, 601
boneshaft, 597–598
circular symmetry, 597
cortical bone material, 597
solid, 598–600
total mass of the bone, 598
trabecular bone material, 597
polar moment of inertia, 595–594
solid tetrahedron's centroid, 596
spherical coordinates, 602–605
triple Riemann sum, 595
Triple Riemann sum, 595
Trivial solution, 27
Two-point boundary value problem (BVP)
Bessel's equation of order zero, 714
finite difference method, 712
IVP numerical methods, 716–717
ODE-BVP eigenvalue problem, 715–716
periodic solutions of linear problems, 717–718
replacement equations, 712–713

U

Undamped natural frequency, 232
Underdamped, 225–226, 230–232
Undetermined coefficients method, 265–269, 403
Uniaxial stretching, 879
Uniform tent basis functions, 721, 1112
Unilateral z-transforms, 340–341, 350
Unit square, 512
Unit step function
convolution
definition, 312
Laplace transform properties, 312
particular solution, 314–318
properties, 312, 314
reality check, 313
delta functions, 318–320
graphical method, 309–312

periodic function, 321
properties, 307–308
Upper triangular matrix, 22–23

V

Variation of parameters
general solution, 293
methods, 294–297
particular solution, 294
second-order nonhomogeneous ODE, 292
Vector field conservative, 564
Virtual work principle, 881–882

W

Wave equations
D'Alembert wave solutions, 1014
Fourier sine series problem, 925–926
guitar string, 868–869, 926–927
infinite series, 925
Lax scheme, 1016–1017
linear elasticity
linearized strain tensor, 876–877
plane strain, 877–878
principal strains, 877
shear modulus, 878
linear elastostatics, 879–880
product solution, 925
scalar hyperbolic problem (*see* Scalar hyperbolic problem)
speed of sound, 924
adiabatic conditions, 875–876
chain rule and Taylor series, 874
continuity equation, 87
linearization system, 874
mass density, 873
momentum equation, 872
perturbation method, 875
small disturbances, 873
taut string, 927–928
undamped harmonic oscillator, 924
vibrating string
contact/tension force, 870
elastic, 870, 872
local elongation, 870
Newton's second law, 871
reference configuration, 869–870
wave speeds, 1014
Wave number, 897
Wronskian determinant, 238–239, 248, 295, 331, 368

Y

Yield stress, 1054
Young's modulus, 879

Z

Zero matrix, 21
Zhukovskii airfoils, 477–478
camber line, 1255
camber ratio, 1255
centered at the origin, 1253–1254
law of cosines, 1253–1254
lift force
angle of attack, 1257
Cauchy–Riemann equations, 1256
chord length, 1257
complex potential function, 1256
complex velocity, 1256
2D flow, 1255–1257
Kutta condition, 1257
potential flow, 1255
stagnation point, 1257
nondimensional parameter, 1253
quarter-chord point, 1255
small parameter, 1253
thickness distribution, 1255
thickness ratio, 1255
Zoo of Solutions of LCCHODEs, 250–251, 253
z-transforms
convolution, 348
geometric series, 340
LCCHΔE, 342–344
linearity properties, 342
Riemann sum, 341
shift operator, 341
sinusoidal signals, 344–345
steady-state solution, 345–347
transfer function, 348–349
unilateral/one-sided, 340–341

ADVANCED ENGINEERING MATHEMATICS

Beginning with linear algebra and later expanding into calculus of variations, **Advanced Engineering Mathematics** provides accessible and comprehensive mathematical preparation for advanced undergraduate and beginning graduate students taking engineering courses. This book offers a review of standard mathematics coursework while effectively integrating science and engineering throughout the text. It explores the use of engineering applications, carefully explains links to engineering practice, and introduces the mathematical tools required for understanding and utilizing software packages.

- Provides comprehensive coverage of mathematics used by engineering students
- Combines stimulating examples with formal exposition and provides context for the mathematics presented
- Contains a wide variety of applications and homework problems
- Includes over 300 figures, more than 40 tables, and over 1500 equations
- Introduces useful *Mathematica*™ and MATLAB® procedures
- Presents faculty and student ancillaries, including an online student solutions manual, full solutions manual for instructors, and full-color figure slides for classroom presentations

Advanced Engineering Mathematics covers ordinary and partial differential equations, matrix/linear algebra, Fourier series and transforms, and numerical methods. Examples include the singular value decomposition for matrices, least squares solutions, difference equations, the *z*-transform, Rayleigh methods for matrices and boundary value problems, the Galerkin method, numerical stability, splines, numerical linear algebra, curvilinear coordinates, calculus of variations, Liapunov functions, controllability, and conformal mapping.

This text also serves as a good reference book for students seeking additional information. It incorporates *Short Takes* sections, describing more advanced topics to readers, and *Learn More about It* sections with direct references for readers wanting more in-depth information.

K11552

CRC Press
Taylor & Francis Group
an **informa** business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

ISBN: 978-1-4398-3447-3
90000
9 781439 834473